SÍNTESIS DE INFORMACIÓN: DESARROLLO Y EVALUACIÓN DE UN MODELO INTERACTIVO

Enrique Amigó Cabrera

A mis padres, Enrique y Ana,
y a mis hermanas, Elena y Patricia

Agradecimientos

Quisiera agradecer a Anselmo Peñas y Julio Gonzalo su dedicación, paciencia y rigor al guiar este trabajo y poner sobre la mesa multitud de buenas ideas que han podido materializarse en este libro, pero sobre todo por ayudarme a encontrar mi profesión. Quería expresar también mi agradecimiento a Felisa Verdejo, por su confianza a lo largo de estos años y por construir y dirigir tan eficazmente este espacio, a pesar de todas las dificultades, en el que muchos hemos podido evolucionar como investigadores.

A Jesús Giménez, por su empuje, su capacidad, y por dar continuidad e impacto a este trabajo. A Víctor Peinado, por colaborar en diferentes trabajos y ayudarme con sus conocimientos lingüísticos y técnicos. Pero sobre todo, por su paciencia con un compañero que nunca se ha llevado bien con las máquinas. Por supuesto, a todos aquellos que han dedicado parte de su tiempo como sujetos de prueba en nuestros experimentos. Sin ellos hubiera sido inviable este trabajo.

A Horacio Rodríguez, por su interés y meticulosidad en revisiones de artículos y trabajos de investigación, por conversaciones en momentos puntuales que han supuesto un elemento clave en mi trabajo. A mi bisabuelo Blas Cabrera, científico, por darme la ilusión de que podría haber un poco de nuestros antepasados escondido en nosotros. A mi padre, por enseñarme a ver las cosas siempre desde otro punto de vista, a mi madre por enseñarme a ver las cosas siempre desde el punto de vista más importante, y a mi hermana Patricia por crecer conmigo. Y muy especialmente a mi mujer, Ana y a mis hijos Nicolás, Carla y Guillermo.

Índice general

Índice de figuras

Capítulo 1

Introducción

Supongamos que un periodista llega a su puesto de trabajo para redactar un artículo sobre el primer hombre que pisó la Luna, y necesita consultar fuentes, como artículos periodísticos o libros que hablen del astronauta. Es decir, el periodista necesita realizar un trabajo de *acceso a la información textual*. Para ello, emplea una herramienta que automáticamente encuentra documentos relativos al astronauta y que extrae y analiza los méritos obtenidos por éste a lo largo de su carrera, además de datos relevantes sobre su vida y algunas anécdotas. La herramienta estructura automáticamente toda esta información y presenta los resultados en un informe. Finalmente, el periodista retoca ligeramente el informe e introduce algunas conclusiones. En una hora ha terminado su trabajo y vuelve a su casa.

Lamentablemente esta herramienta no existe. En su lugar, un periodista probablemente introduciría varias consultas en un buscador de Internet o en bases de datos documentales hasta obtener un conjunto de, por ejemplo, cien documentos con información potencialmente relevante. Posteriormente tendría que leer los documentos para extraer y organizar información susceptible de aparecer en el informe, y finalmente elaborar la versión final de su artículo. Posiblemente tardaría uno o dos días en realizar esta tarea. En esta memoria llamaremos "Síntesis de Información" a este proceso de recopilación, análisis y elaboración de información.

Ahora bien, supongamos que dicho sistema existe. Otro periodista redacta un artículo sobre el mismo tema para otra revista. Éste emplea exactamente la misma herramienta, que le devuelve un informe inicial. Sin embargo, el periodista no está satisfecho con este boceto, dado que éste no contempla aspectos de la vida personal que el periodista considera relevantes: qué le motivó a seguir esa carrera profesional, cómo afectó en su vida el hecho de estar por encima de la media en cuanto a sus capacidades, dónde conoció a su esposa, etc. Además, el periodista es incapaz de precisar al sistema exactamente el tipo de artículo que desea redactar. Finalmente se ve obligado a repasar todos los documentos potencialmente relevantes y realizar el trabajo manualmente, lo que le lleva dos días.

El problema de este sistema ideal es que no considera la subjetividad inherente a la Síntesis de Información. Los dos periodistas escriben artículos distintos, y el proceso de acceso a la información realizado es diferente en ambos casos. Este problema podría haberse solventado con la incorporación de mecanismos apropiados

de interacción con el usuario. De esta forma, el sistema no realiza la tarea completa de forma automática, y su función es la de asistir al periodista, no la de sustituirle. Es por tanto el usuario el que en última instancia determina qué información es relevante, o qué relación existe entre las distintas piezas de información.

En este libro, analizaremos la tarea de Síntesis de Información desde la perspectiva del acceso a la información textual. Desarrollaremos y evaluaremos un modelo interactivo. Es decir, estudiaremos cómo ayudar al usuario a analizar y recopilar la información necesaria para la elaboración de un informe sobre un conjunto voluminoso de documentos relevantes.

1.1. Necesidad de Síntesis de Información

La Síntesis de Información es una tarea costosa que se plantea en diversos contextos reales. Prueba de ello es la existencia de diferentes servicios de información en los que se elabora para el cliente una síntesis de acuerdo con sus necesidades de información. En este apartado proponemos dos ejemplos: servicios de vigilancia tecnológica y el servicio Google Answers, para finalmente centrarnos en el dominio periodístico, sobre el que trabajaremos a lo largo del libro.

Conocer la tecnología relacionada con un tipo de producción es clave para mantener el rendimiento de una empresa. Actualmente existen consultoras que ofrecen servicios de "vigilancia tecnológica". La figura 1.1 muestra una batería de consultas proporcionada por la Red de Servicios Avanzados de Vigilancia Tecnológica [1]. El objetivo general de esta red de organismos es contribuir a mejorar la competitividad de las pequeñas y medianas empresas españolas, facilitando servicios de alerta de información avanzados que ofrezcan un valor añadido a las empresas usuarias.

Algunas de las respuestas a estas preguntas se podrían encontrar en un único documento, como por ejemplo: *"¿Cuál es la composición de un material del que sólo se conoce su denominación?"*. Sin embargo, la gran mayoría requiere un proceso de recopilación de información dispersa, como es el caso de *"Bibliografía sobre citoquinas y su aplicación en cosméticos"*, es decir, extraer y organizar datos distribuidos a lo largo de varios documentos. Otras de las consultas de la figura podrían requerir, además de búsqueda de documentos y recopilación de datos, un proceso de elaboración. Por ejemplo, considerando la pregunta *"¿Cuál es el mejor proceso para obtener un material con una micro-estructura determinada?"*, la respuesta consistirá en un informe o resumen en el que se debe comparar varios procesos en base a sus características (requisitos, costes, etc.), para finalmente seleccionar el proceso más adecuado. Es necesario para ello relacionar entre sí datos procedentes de distintas fuentes. Este proceso de recopilación, organización, y elaboración se ajusta a lo que denominamos Síntesis de Información.

En el entorno de la WEB, son conocidos los buscadores que ayudan a los usuarios a localizar documentos relevantes a partir de una necesidad de información expresada en forma de consulta (Google, Altavista, etc.). En muchos casos, sin embargo, este sólo es el primer paso en el proceso mediante el cual se satisfacen

[1] `http://www.fedit.es`

- Mecanización de segmentos de pistón de forma oval por control numérico
- Viabilidad de montaje de cables de transmisión de datos a través de tubos de conducción de aguas
- Reducir el nivel de ruido producido por los engranajes hasta los límites autorizados por la UE
- ¿Qué material es sustitutivo de este otro que empleo actualmente?
- ¿Cuál es la composición de un material del que se conoce únicamente su denominación?
- Envejecimiento de envases de polietileno de alta densidad que contienen peróxidos orgánicos
- Bibliografía sobre citoquinas y su aplicación en cosméticos.
- ¿Qué tipos de materiales son capaces de soportar unas determinadas condiciones de corrosión y temperatura?
- Adhesivos que permitan una perfecta estanqueidad entre u rango de temperatura (-40º y −125º)
- Características técnicas de los diferentes tipos de teflon
- ¿Para qué puedo utilizar este material que es un residuo de mi proceso productivo?
- Procesado de conservas de túnidos
- ¿Cuál es el mejor proceso para obtener un material con una microestructura determinada?
- Listado de especies comerciales de pescado en la C.A.P.V.
- Envasado de platos preparados en atmósferas controladas
- Equipos de nanofiltración para el sector alimentario.
- Tipos de tejidos para tapizado de muebles
- Posibilidad de reciclaje de residuos de prefabricados de hormigón
- Espesor del vidrio laminado
- Proceso de elaboración de puré de fresa y frambuesa
- Propiedades nutricionales del verdel
- Búsqueda de patentes españolas de aplicación de metales sobre bizcocho cerámico
- Bibliografía sobre la gestión de la cadena de suministro
- Bibliografía sobre la satisfacción del cliente y la evaluación de la fidelidad
- Búsqueda de artículos de revista que traten sobre la aplicación de la norma UNE_EN ISO 9001: 2000 en centros de formación

Figura 1.1: Batería de consultas en la Red de Servicios Avanzados de Vigilancia Tecnológica

las necesidades del usuario. El siguiente paso consiste en extraer, organizar y relacionar piezas de información, con el fin de redactar un informe.

Un ejemplo que ilustra este problema es el de Google Answers Service, en donde los usuarios envían una consulta compleja que no puede ser resuelta simplemente analizando los primeros documentos de una lista devuelta por un buscador. La respuesta a dichas consultas complejas es elaborada por un experto quien, por lo general, combina sus propios conocimientos con búsquedas en Internet, devolviendo al cliente un informe con las piezas de información más relevantes y enlaces a las fuentes a partir de las cuales se ha obtenido la información. Un ejemplo de consulta real atendida por este servicio es: *La relación entre los ejercicios de desarrollo de la atención y su impacto en el rendimiento académico*. Para atender a esta consulta, el experto accede a un gran número de páginas WEB y redacta un extenso documento que incluye referencias que permiten al cliente acceder a las fuentes originales.

Síntesis de Información en el dominio periodístico

En este libro, prestaremos especial atención al dominio periodístico. Elaborar un artículo periodístico requiere en muchos casos recopilación, análisis y elaboración de información textual, es decir, lo que entendemos por Síntesis de Información.

A modo de ejemplo, la figura 1.2 muestra una noticia proporcionada por la agencia de información EFE. En este documento se trata información que justifica el nombramiento de Jermaine Dye como jugador de la semana en la competición de Béisbol de la Liga Americana. Para la realización de este informe, el periodista debe recuperar datos que justifiquen el nombramiento (victorias, carreras realizadas, competidores por el premio, etc.), es decir, debe recopilar información a partir de informes publicados con anterioridad e integrar toda esta información en un nuevo artículo. Posiblemente la mayoría de los datos han sido recopilados a partir de otros informes.

Una primera cuestión que hemos de plantearnos es qué tipo de necesidades de información tiene un periodista. En [NM97] se identifican dos tipos de necesidades de información en este contexto: comprobación de datos de tipo factual (fechas, nombres, lugares, etc.) y adquisición de conocimiento general en relación a un tema. En este segundo caso es necesario recopilar y relacionar entre sí piezas distribuidas a lo largo de varias fuentes para elaborar información nueva, es decir, un proceso de síntesis.

Una segunda cuestión que podríamos plantearnos es de qué herramientas dispone un periodista para resolver dichas necesidades. Una de las herramientas de acceso a la información que más ha influido en el trabajo de los periodistas son los sistemas de acceso a Internet. En [WN97] se realiza un análisis sobre el impacto ejercido por la aparición de Internet en el dominio periodístico. Una de las conclusiones a las que llega el autor es que los periodistas consideran que disponen de sobrada información, pero poco tiempo y muy poco espacio para expresarse. Esto implica que, aun teniendo disponible la información, el usuario sigue sin ver

Nueva York, 1 may (EFE).- El jardinero Jermaine Dye, de los Reales de Kansas City, fue nombreado hoy como ganador del premio Jugador de la Semana dentro de la competición del béisbol profesional de la Liga Americana para el periodo que finalizó el pasado domingo.

La Producción ofensiva de Dye permitió a los Reales conseguir cinco triunfos en los seis partidos que disputaros para volver a recuperar posiciones en la División Central de la Americana. Dye llegó a los 10 partidos consecutivos que pegó de hit y tenía 11 indiscutibles en 21 viajes al plató con tres cuadrangulares de cuatro carreras impulsadas y 11 remolcadas al finalizar la semana. El jardinero de los Reales lidera la Americana con 12 dobletes y las 28 carreras impulsadas que es una nueva marca en la historia del equipo para el mes de abril.

El mayor rival de Dye para el premio fue el puertoriqueño José Valentín, de los Medias Blancas de Chicago, y su compañero de equipo Greg Norton. Valentín logró el pasado jueves ciclo completo en el tunro al bate y Norton terminó con 500 de promedio. EFE

Figura 1.2: Informe de la agencia EFE

satisfechas sus necesidades, siendo necesario por tanto ayudar al usuario a realizar un trabajo de síntesis de manera eficiente.

Esta necesidad de analizar y elaborar información junto con la sofisticación de las herramientas de acceso a la información ha generado lo que se denomina periodismo asistido por ordenador –CAR: Computer-Assisted Reporting–. El siguiente párrafo es un texto de Fran Casal (periodista) en el que se describe el concepto de CAR:

"...El Periodismo Asistido por Ordenador ha sido utilizado para realizar los reportajes ganadores del Premio Pulitzer de los últimos 10 años. El CAR implica que los periodistas comienzan a usar los ordenadores no sólo como máquinas de escribir, sino también para realizar investigaciones complejas a través de bases de datos, para trabajar con ingentes cantidades de números y estadísticas públicas, para analizar esos datos y utilizar ese análisis para conseguir historias de alto nivel con un contexto más profundo. El CAR no reemplaza a las técnicas periodísticas tradicionales, sólo las mejora y las complementa. No es algo ajeno al periodismo tradicional, sino la esencia de la supervivencia del periodismo en el siglo 21, y lo que es más importante: el CAR está en el corazón del buen periodismo de servicio público."

El CAR plantea sin embargo nuevos problemas, entre los que está la necesidad de evaluar la calidad de la información recuperada, dado que algunas fuentes como Internet, no aseguran la fiabilidad del autor. Otro nuevo reto que incorpora el CAR es el uso y manejo de toda esta información disponible, sobre todo si se encuentra en formato textual, que se aborda mediante un proceso de síntesis.

1.2. Características de la Síntesis de Información

Los ejemplos anteriores describen la Síntesis de Información en varios contextos: vigilancia tecnológica, servicios de información en internet y el dominio periodístico. A partir de estos ejemplos podemos identificar algunas características del acceso a la información en la SI:

1. **La información requerida no se encuentra en una única pieza de texto como un fragmento o documento**. Considerando los dominios descritos a modo de ejemplo en el apartado anterior, es muy posible que la información necesaria para responder a una pregunta del tipo *"¿Cuál es el mejor proceso para obtener un material con una micro-estructura determinada?"* se encuentre distribuida a lo largo de varios documentos, y dentro de estos, a lo largo de varios fragmentos. Por otro lado, el experto de Google Answers, genera el informe a partir de múltiples búsquedas en Internet. Por último, la redacción del artículo periodístico requiere acceso a datos publicados en diversos informes anteriores.

2. **La información requerida no se encuentra de forma explícita en las fuentes**. Esto implica que es necesario no sólo recopilar, sino analizar e interpretar la información recopilada. Por ejemplo, dada la consulta *¿Cuál es el mejor proceso para obtener un material con una micro-estructura determinada?*, el usuario debe determinar entre otras cosas el coste de cada proceso. Es posible que esta información no aparezca de forma explícita, sino implícita en enunciados que han de ser interpretados. En el segundo ejemplo (Google Answers), saber si las actividades relacionadas con la atención afectan al rendimiento académico implica interpretar estudios y opiniones publicadas por especialistas. En el caso del periodista que redacta un artículo sobre la elección del jugador de la semana, es necesario inferir, por ejemplo, que las victorias de su equipo son en realidad también suyas, o que los dobletes o carreras impulsadas por el jugador son méritos que justifican la obtención del título. El artículo del ejemplo es bastante básico; por supuesto, la necesidad de identificar información implícita es más patente en periodismo de investigación.

3. **Se requiere una respuesta elaborada.** Para que la información obtenida durante el proceso de síntesis sea útil, ésta debe presentarse, cubriendo de forma ordenada los aspectos que satisfacen las necesidades de información. Por ejemplo, en el caso de la consulta *"¿Cuál es el mejor proceso para obtener un material con una micro estructura determinada?"* el informe final debería exponer y justificar de forma organizada los costes de las distintas alternativas. En el caso del servicio de Google Answers el resultado del trabajo del experto es un informe amplio y estructurado. Análogamente, en el caso del periodista, el artículo mostrado sigue una determinada estructura. En primer lugar muestra en el artículo los méritos del jugador y, en segundo lugar, los méritos de sus competidores.

En definitiva, entendemos por *Síntesis de Información* (SI) como **el proceso mediante el cuál, dada una necesidad de información compleja, se extrae, organiza e interrelaciona las piezas de información contenidas en un conjunto de documentos relevantes, con el fin de obtener un informe sin redundancias que satisfaga dicha necesidad de información.** Esta tarea implica recopilación, análisis y elaboración de información.

Problemas de acceso a la información de estas características están presentes también en multitud de dominios, como son las revisiones del estado del arte en contextos científicos, estudios de mercado, revisión de casos jurídicos, etc.

1.3. Cómo abordar la Síntesis de Información

Tanto la subjetividad como la complejidad de la tarea sugieren a la elaboración de un modelo interactivo de SI que asista al usuario en la resolución del problema. Desde la perspectiva del acceso a la información, la SI es una tarea aun no estudiada en profundidad. Por tanto, la elaboración de este modelo requiere un trabajo de investigación extenso. Requiere definir y acotar el problema, elaborar un corpus de referencia en condiciones controladas sobre el que poder evaluar, definir una metodología de evaluación apropiada, y finalmente comparar distintas aproximaciones. A continuación se resumen brevemente estos requisitos que estructuran el contenido de este libro.

Acotación del problema

Para acotar el problema del acceso a la información en la SI desde un punto de vista computacional es necesario estudiar en qué medida las técnicas actuales ayudan a abordar el problema. Concretamente, se han desarrollado, por ejemplo, sistemas para la localización documentos relevantes en una colección dada una consulta introducida por el usuario. Existen también sistemas orientados a extraer de una colección de documentos la respuesta a una pregunta concreta del tipo *¿Cuál es la capital de Suiza?* , y también sistemas que resumen automáticamente un documento o un conjunto de documentos. La pregunta es hasta qué punto las técnicas empleadas en estos sistemas se adecuan a la naturaleza del problema de la SI. En este libro abordaremos esta cuestión.

En segundo lugar, elaborar un modelo interactivo de SI implica definir mecanismos de interacción entre usuario y sistema. En el campo de los sistemas de acceso a la información se han aplicado distintos tipos de interacción. Por ejemplo, un posible esquema de interacción es la exploración de contenidos a través de elementos intermedios, como puede ser un índice de materias. Otros esquemas, por ejemplo, se basa en diálogo, es decir, ciclos de pregunta/respuesta entre sistema y usuario. La cuestión qué esquema de interacción es más adecuado en el caso de la SI, cuestión que también abordaremos en este libro.

Un marco de evaluación para la Síntesis de Información

Para estudiar el problema del acceso a la información en la SI analizaremos muestras en las que sujetos reales realizan, en condiciones controladas, informes que responden a necesidades de información concretas. Es decir, partiremos de un corpus de Síntesis de Información. Este corpus nos permitirá, por un lado, estudiar el comportamiento de los sujetos durante el proceso de SI y, por otro lado, disponer de un conjunto de informes modelo sobre los que contrastar modelos computacionales de SI. Para este segundo objetivo será necesario disponer de una metodología de evaluación apropiada. Todo ello constituye nuestro marco de evaluación.

Elaborar un corpus de informes en condiciones controladas no es una tarea sencilla. Realizar un informe manualmente sobre papel es muy costoso, y no permite monitorizar el comportamiento de los sujetos de prueba. Es necesario ofrecer a los sujetos una herramienta sencilla y robusta que disminuya el esfuerzo necesario. La cuestión que surge entonces es cómo especificar un sistema que no sesgue la naturaleza de los informes obtenidos. Además, también es necesario establecer consultas y fuentes de información que se ajusten al problema de la SI. Considerando estos aspectos, se presenta en este libro ISCORPUS, un corpus de 75 informes generados manualmente.

Una vez disponible el corpus de informes, necesitamos una metodología de evaluación que permita comparar entre sí distintas aproximaciones al problema en base a dicho corpus. Es decir, necesitamos una metodología de evaluación automática sobre informes de referencia. Como hemos apuntado, la tarea de SI es subjetiva. Ante una necesidad de información, no existe un único informe apropiado. Definir una metodología de evaluación para la SI debe permitirnos caracterizar los informes generados por sujetos reales, es decir, identificar los rasgos comunes de los informes generados manualmente considerados como modelos. Para ello, es necesario estudiar las similitudes y diferencias entre informes, aplicando y combinando distintos rasgos. Veremos que, en este sentido, las metodologías de evaluación existentes sufren algunas limitaciones. Definimos en este libro la metodología de evaluación QARLA, que permite combinar, aplicar y evaluar métricas de similitud entre informes generados por sistemas e informes modelo generados manualmente. Veremos que esta metodología de evaluación es también generalizable a sistemas interactivos.

Desarrollo de un modelo interactivo de Síntesis de Información basado en conceptos clave

Una vez acotado el problema y ya definido un marco de evaluación apropiado, la siguiente cuestión es cómo abordar el desarrollo de un modelo de SI. En este libro se describe el papel que juegan en la SI los *conceptos clave* extraídos a partir de los documentos por sintetizar. En esta memoria, interpretamos concepto clave en el dominio de noticias como el conjunto de personas, organizaciones o factores que adquieren protagonismo en el asunto tratado. Por ejemplo, si consideráramos el tema “conflicto en Palestina”, algunos conceptos clave podrían ser “los atentados”, “Arafat”, o “Israel”.

Nuestra primera hipótesis consiste en que la distribución de estos conceptos en un informe es un rasgo a tener en cuenta en el proceso de elaboración y evaluación de informes. Para validar de esta primera hipótesis estudiaremos la distribución de conceptos clave en los informes. Veremos que éste es un rasgo común compartido por informes generados manualmente frente a informes generados de forma automática mediante técnicas básicas. Además, veremos que este rasgo, en combinación con otros, hace más fiable el proceso de evaluación de informes.

Por otro lado, para que se pued an explotar los conceptos clave en el desarrollo de un modelo de SI, es necesario previamente extraer automáticamente dichos conceptos clave a partir de los documentos originales por sintetizar. En este libro se comparan distintas aproximaciones a la extracción automática de los conceptos. Estudiaremos también la relación existente entre la distribución de estos conceptos en los documentos originales y en un informe manual.

La segunda hipótesis planteada en este libro consiste en que se pueden estructurar los contenidos mediante estos conceptos, facilitando el acceso a la información de cara a la elaboración del informe en el contexto de un sistema interactivo de SI. Para validar dicha hipótesis compararemos diferentes esquemas de interacción con el usuario sobre los que definir un modelo de acceso a la información. Concretamente, estudiaremos distintas estrategias basadas en la organización de los contenidos por títulos y por listas terminológicas. Veremos que la exploración de la información mediante una lista de términos representativos de los conceptos clave, facilita el acceso a la información con vistas a la elaboración de un informe.

Finalmente, a partir del estudio del rol de los conceptos clave en la SI y de la comparación de diferentes estrategias de interacción con el usuario, definimos el modelo interactivo PRISMA, que ha sido implementado en un prototipo y probado sobre una amplia colección de documentos periodísticos.

1.4. Objetivos del trabajo

El objetivo principal del trabajo descrito en este libro es el desarrollo y la implementación de un modelo interactivo, al que denominamos PRISMA, sobre la base de las dos hipótesis descritas en el apartado anterior. Estas dos hipótesis definen dos subobjetivos:

1. Estudiar el papel de los conceptos clave en la SI. Esto incluye:
    - Validar la necesidad de considerar conceptos clave en el desarrollo de sistemas de SI.
    - Estudiar diferentes estrategias de extracción automática de conceptos clave a partir de los documentos originales
    - Estudiar estrategias de predicción de la distribución de conceptos clave en un informe modelo generado manualmente.
    - Validar la utilidad de los conceptos clave como rasgo por considerar en la evaluación automática de informes.

2. Comparar distintas estrategias interactivas de exploración de contenidos en el contexto de la SI. En especial estrategias basadas en la exploración de títulos de documentos frente a estrategias basadas en la exploración de términos y fragmentos asociados a dichos términos.

Para ello, necesitamos desarrollar un marco de trabajo apropiado. Este marco estará constituido por un corpus de informes y una metodología de evaluación, lo que define los siguientes objetivos:

- Elaborar un corpus de informes y conceptos clave que soporte los experimentos realizados en este trabajo.
- Elaborar una metodología de evaluación adaptada al problema de la SI que permita combinar, aplicar y validar métricas de similitud sobre informes de referencia. Esta metodología debería permitir:
  - la aplicación y combinación de métricas de similitud.
  - la meta-evaluación de conjuntos de métricas de similitud.
  - la validación del corpus de informes automáticos e informes modelo sobre los que se aplican y validan métricas de similitud.
  - una generalización para dominios interactivos.

Por último, la definición del marco de trabajo y el desarrollo del modelo PRISMA requiere definir con precisión el concepto de Síntesis de Información, y hacer una revisión del estado del arte en cuanto a los aspectos planteados en este libro. Esto determina los siguientes objetivos:

- Definición del problema de la Síntesis de Información desde una perspectiva computacional, considerando las tareas abordadas actualmente por sistemas de acceso a la información textual.
- Realizar un estudio sobre técnicas de automatización del acceso a la información extrapolables a la SI. Concretamente nos centraremos en técnicas automáticas de resumen.
- Realizar un estudio sobre estrategias interactivas en sistemas de acceso a la información.
- Realizar un estudio sobre técnicas de evaluación extrapolables a la SI.

1.5. Estructura del libro

Este libro está dividido en tres partes: preliminares, marco de evaluación y desarrollo de un modelo interactivo.

Parte I: Preliminares

En la primera parte se acota el problema del acceso a la información en la SI. Ésta incluye los capítulos 2, 3 y 4. En el capítulo 2, se acota el problema desde un punto de vista computacional. Concretamente, se analiza la relación existente entre sistemas actuales de acceso a la información textual y la SI (sección 2.1). En este mismo capítulo se analiza en qué medida las técnicas empleadas en sistemas de resumen son aplicables al problema de la SI: técnicas generativas (sección 2.2), técnicas extractivas (sección 2.3), tratamiento de múltiples documentos (sección 2.4) y tratamiento de las necesidades de información expresadas en forma de consulta (sección 2.5). Finalmente, en la sección 2.6 se recapitula el análisis de las relaciones existentes entre la SI y las técnicas de resumen.

En el capítulo 3 se analizan las estrategias de interacción con el usuario en sistemas de acceso a la información. Este capítulo incluye un análisis de distintos modelos cognitivos del acceso a la información que muestran la necesidad de incorporar interacción entre usuario y sistema en tareas relacionadas con la SI (sección 3.1). En la sección 3.2 se propone una categorización de los distintos esquemas de interacción aplicados en sistemas de acceso a la información. En la sección 3.3 se analiza los esquemas de interacción empleados en sistemas de resumen. Por último, en las secciones 3.5 y 3.4 analizamos los requisitos de un esquema de interacción en la SI y estudiamos la aplicación de distintos esquemas de interacción para el caso concreto de la SI.

En el capítulo 4 se realiza una revisión del estado del arte en cuanto a modelos de evaluación, incluyendo la evaluación de sistemas de resumen (sección 4.1), la evaluación de sistemas componentes interactivos 4.2, y por último se analizan los requisitos de un marco de evaluación en el contexto de la SI (sección 4.3).

Parte II: Marco de evaluación

La segunda parte de este libro engloba todo lo relacionado con la metodología de evaluación empleada en este trabajo. Esto incluye en primer lugar, el desarrollo de un corpus adaptado al problema de la SI (capítulo 5). En este capítulo se describen en primer lugar los temas y documentos de partida sobre los que se han generado los informes en el corpus (sección 5.1). Se describe también el modo en que se han generado dichos informes (sección 5.2), y un conjunto de informes automáticos generados a partir de estrategias automáticas básicas (sección 5.5). Este capítulo incluye un estudio del comportamiento de los sujetos de prueba en la elaboración de informes (sección 5.6). Finalmente, se extraen algunas conclusiones sobre la información obtenida a lo largo del desarrollo de ISCORPUS (sección 5.7).

En el capítulo 6 se define una metodología de evaluación adaptada a las necesidades de la SI. El punto de partida en la elaboración de esta metodología es la definición de una serie de requisitos formales que las medidas aportadas por este marco deberían cumplir (sección 6.2). El marco QARLA incluye una medida QUEEN para la aplicación y combinación de métricas de similitud sobre informes modelo (sección 6.3), una medida KING para la meta-evaluación de conjuntos de

métricas de similitud (sección 6.4), y una medida JACK para la validación del conjunto de informes sobre los que se testean las métricas (sección 6.5). En este capítulo se generaliza también el marco QARLA para dominios interactivos (sección 6.6), y se aplica y evalúa el marco sobre un corpus de resúmenes empleado habitualmente en el área (sección 6.7). Finalmente, en la sección 6.8, se recapitula el trabajo realizado a lo largo de este capítulo.

Parte III: Desarrollo de un modelo interactivo de SI

La tercera y última parte del libro describe la elaboración del modelo PRISMA, un modelo interactivo de SI basado en la extracción automática de conceptos clave. El primer paso para la elaboración de PRISMA es el estudio empírico del papel de los conceptos clave en el problema de la SI (capítulo 7). Esto incluye un estudio sobre la necesidad de considerar conceptos clave en la elaboración de sistemas (sección 7.1), la comparación de diferentes estrategias de extracción automática de conceptos clave (sección 7.2), y la estimación de la distribución de los conceptos en un informe modelo (sección 7.3). Se cuantifica también en este capítulo cómo las métricas de similitud basadas en conceptos clave pueden aumentar la fiabilidad de la evaluación dentro del marco QARLA (sección 7.4). Finalmente, en el apartado 7.5 se recapitulan los resultados obtenidos en los experimentos descritos.

Otro paso previo a la implementación del sistema es la evaluación empírica sin usuarios de prueba de diferentes estrategias de exploración de contenidos en el contexto de la SI (capítulo 8). En este experimento se definen en primer lugar diferentes estrategias de exploración de los contenidos textuales (sección 8.1). En la sección 8.2 se describe en detalle la metodología seguida en este experimento y en la sección 8.3 se exponen los resultados obtenidos. Finalmente en la sección 8.4 se enumeran algunas conclusiones obtenidas en relación a este capítulo.

En el capítulo 9, se describe el modelo PRISMA como resultado de los estudios realizados a lo largo del libro, y un prototipo implementado que atiende a dicho modelo. Para describir este modelo, se detalla el conjunto de niveles de acceso a la información que conforman el modelo (sección 9.1), el conjunto de fases que componen el proceso de SI reflejado en el modelo (sección 9.2), y una descripción detallada de las distintas tareas que en PRISMA realiza el sistema y el usuario en cada caso (secciones 9.2.1, 9.2.2 y 9.2.3). En este capítulo se detallan, además, aspectos de implementación del sistema (sección 9.3), y se compara el modelo PRISMA con otros modelos similares presentes en el estado del arte (sección 9.4). Se analiza en el último apartado de este capítulo la evaluación de PRISMA (sección 9.5).

Finalmente, en el capítulo 10 se recogen las conclusiones de este libro.

Parte I

Preliminares

Capítulo 2

Síntesis de Información y generación de resúmenes

La Síntesis de Información (SI), tal y como la hemos definido, es un problema presente en varios dominios como es el periodístico, la toma de decisiones en negocios, la investigación médica, etc. (capítulo 1). En todos estos casos, los procesos de acceso a la información comparten una serie de características: información no necesariamente explícita en las fuentes originales, información distribuída a lo largo de varias fuentes, y necesidad de una respuesta elaborada. Estas son por tanto, las condiciones sobre las que surge la SI. De estas condiciones se deriva la necesidad de recopilación, análisis y elaboración de información, procesos que caracterizan la Síntesis de Información.

Desde una perspectiva computacional, se han definido diferentes tareas de acceso a la información textual que guardan relación con la SI, sobre las que se han desarrollado y evaluado diferentes aproximaciones. Algunas de estas tareas son la elaboración de resúmenes, la búsqueda de documentos, la extracción de información (transformación de datos en formato original a un formato estructurado), o la búsqueda de respuestas a preguntas concretas (fechas, nombres etc.). En este capítulo se estudia la relación existente entre estas tareas y la SI. Como veremos, la SI posee características en común con cada una de ellas, siendo la tarea de resumen automático la que en mayor medida se asemeja a la SI (sección 2.1).

En el desarrollo de sistemas de Resumen Automático se abordan aspectos del acceso a la información mediante técnicas que pueden resultar útiles en un modelo de SI, como son la identificación de conceptos clave, descomposición de los contenidos en piezas de texto más pequeñas, identificación de fragmentos relevantes o la detección de información redundante (secciones 2.2 y 2.3). Además, en estos sistemas se ha abordado el problema del tratamiento de información procedente de varios documentos, cuestión que afecta al proceso de Síntesis de Información (section 2.4). Se han desarrollado también modelos de Resumen Automático orientados a una necesidad de información expresada en forma de consulta (sección 2.5). La consideración de las necesidades del usuario a la hora de recopilar y organizar la información contenida en las fuentes es también un aspecto implicado en el problema de la SI.

Podremos constatar que algunas técnicas de resumen son aplicables especialmente en las fases de recopilación y análisis de la información para el caso de la SI, aunque siempre a un nivel superficial.

2.1. Síntesis de Información desde un punto de vista computacional

Para abordar el problema del acceso a la información en la SI desde un punto de vista computacional, es importante entender dónde se ubica el problema en relación a otras tareas de acceso a la información sobre las que se han desarrollado modelos computacionales. Algunos de estas tareas son: Recuperación de Documentos, Extracción de Información, Búsqueda de Respuestas y Resumen Automático. La relación existente entre las distintas tareas se ilustra en la figura 2.1 En los siguientes párrafos analizaremos estas relaciones.

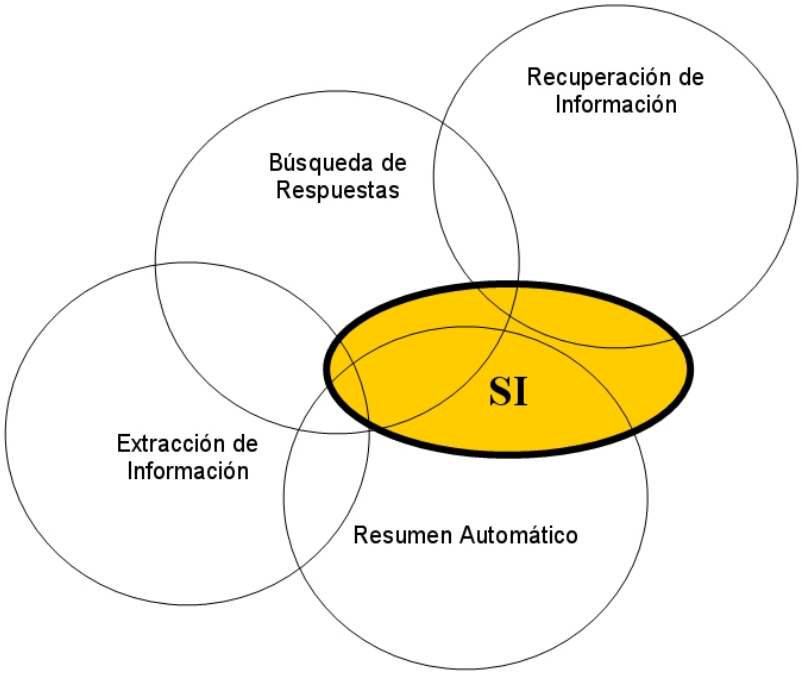


Figura 2.1: Relación entre tareas de acceso a la información textual

2.1.1. SI y Recuperación de Información (RI).

La Recuperación de Información consiste en la recuperación y selección de aquellos documentos que sean relevantes ante necesidades de información formuladas por los usuarios en forma de consulta. Por lo general, un sistema que res-

ponda a este problema devuelve al usuario una lista de documentos ordenada por relevancia de acuerdo a la consulta.

La Recuperación de Información se relaciona con la SI, en la medida que aborda el problema de encontrar un conjunto de documentos susceptibles de contener información relevante. Sin embargo, tal y como hemos planteado la SI, la recuperación de documentos es una tarea distinta, dado que en la SI partimos de una colección de documentos relevantes. Este presupuesto se ajusta a muchos casos reales. Por ejemplo, en el contexto periodístico, en general se dispone suficientes fuentes de información, pero poco tiempo y espacio para desarrollar el trabajo de síntesis [WN97].

2.1.2. SI y la Extracción de Información(EI)

Se entiende la Extracción de Información como la tarea consistente en localizar información de un tipo predefinido a partir de fuentes en formato textual. Por ejemplo, un sistema de extracción de información centrado en el dominio del terrorismo, extrae datos como nombres de terroristas, número de víctimas, armas, fechas, localidades etc. Un sistema centrado en el dominio de los negocios extrae datos como nombres de compañías, productos, etc. En esencia, se trata de una transformación de formato de información no estructurada a información estructurada. Es decir, transformación de información expresada en lenguaje natural, en información que un sistema pueda tratar.

Sin embargo, el resultado de las técnicas de extracción de información pueden ser insuficientes en el proceso de síntesis. Por ejemplo, como apunta Blake [BP02], en el dominio médico, no solo son importantes los datos en sí, sino el contexto en el que esos datos han sido obtenidos por los autores originales, por lo que las piezas de información extraídas carecen de utilidad fuera de su contexto. Otra de las limitaciones de las técnicas desarrolladas en sistemas de EI es que se centran por lo general en dominios muy específicos.

2.1.3. SI y Búsqueda de Respuestas (BR).

El problema de la Búsqueda de Respuestas hace referencia a la resolución de preguntas correctamente formuladas en lenguaje natural atendiendo a una necesidad de información precisa. El sistema extrae una respuesta a una pregunta a partir de una colección de documentos. Actualmente, estos sistemas son capaces de atender a cuestiones relativas a hechos concretos, es decir, información de tipo factual. Un ejemplo de este tipo de preguntas es, *"¿Qué país gano el mundial de fútbol de 1978?"*.

Generalmente, el sistema debe encontrar la respuesta en una colección de documentos. Para preseleccionar los documentos susceptibles de contener la respuesta, en estos sistemas se emplean técnicas de Recuperación de Documentos, existiendo por tanto una relación entre ambas tareas.

Por otro lado, tanto el análisis de la pregunta como la extracción de respuestas candidatas requiere un proceso de extracción de información, en cuanto que se

ha de transformar la información expresada en lenguaje natural, en información estructurada. Si se atiende a cuestiones relativas a hechos concretos, las técnicas de extracción de información pueden ser muy útiles en este campo, como demuestran la mayoría de sistemas de Búsqueda de Respuestas desarrollados. Estas dos áreas están por tanto estrechamente relacionadas.

Dado que actualmente los sistemas de Búsqueda de Respuestas atienden a cuestiones de tipo factual (personas, organizaciones, fechas, etc.), las técnicas empleadas en estos sistemas no son fácilmente extrapolables al problema de la síntesis. Estas técnicas asumen que el dato solicitado se encuentra de forma explícita en algún punto de las fuentes, mientras que en el caso de la SI la información a la que se necesita acceder puede aparecer de forma implícita y distribuída.

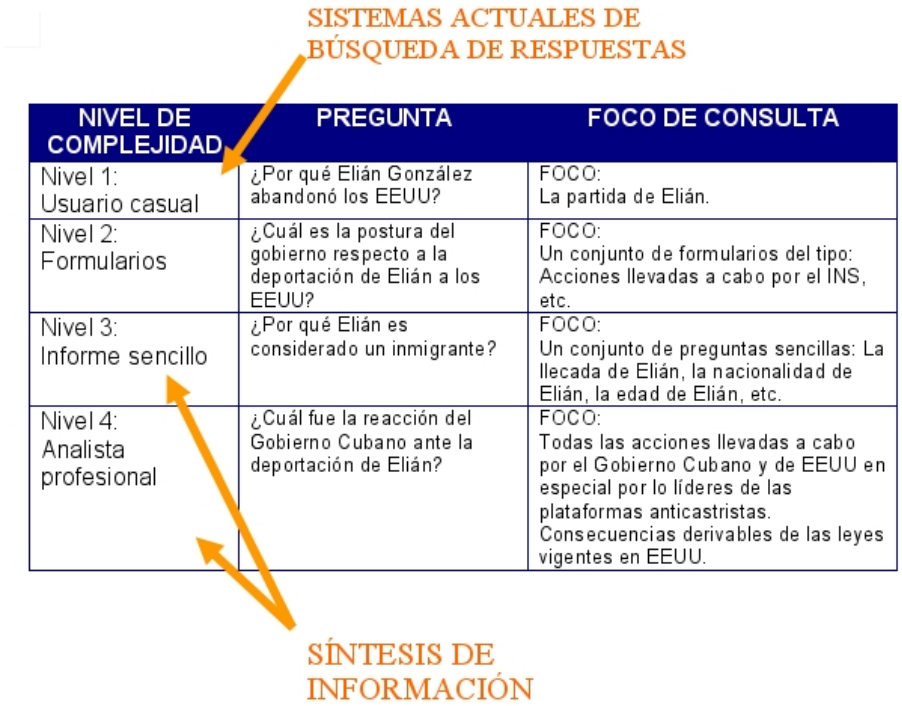
Sin embargo, se aprecia una convergencia en las líneas futuras planteadas en el área. El Q&A Roadmap Committee [BCV+01] elaboró un documento describiendo una serie de hitos para la tarea de Búsqueda de Respuestas dirigido a orientar la metodología de evaluación de estos sistemas en el foro de investigación TREC (Text Retrieval Conference). La figura 2.2 muestra una clasificación de los niveles de complejidad de una consulta propuesta en dicho documento. Actualmente los sistemas de Búsqueda de Respuestas atienden a preguntas del primer nivel o de usuario casual, mientras que no son capaces de descomponer la necesidad de información en preguntas más concretas, fusionar los resultados y mucho menos realizar las inferencias necesarias para elaborar información. Por otro lado, las consultas consideradas de tercer y cuarto nivel (informe sencillo y analista profesional) requieren, al igual que en el problema de la síntesis, recopilación, análisis y elaboración de información.

2.1.4. SI y Resumen Automático (RA).

Mani y Maybury [Man01a] proponen la siguiente definición para el concepto de Resumen Automático: *"El proceso de extraer la información más importante a partir de una fuente (o fuentes) para producir una versión abreviada orientada a un determinado usuario y tarea"*.

No existe una frontera clara entre la tarea de resumen guiada por una consulta, y la Búsqueda de Respuestas en el caso de solicitar una respuesta elaborada. Por ejemplo, la pregunta de tipo definición "¿Quién es Bill Clinton?" puede responderse en tres palabras o mediante un resumen en tres páginas. La diferencia fundamental entre ambas tareas es que en el caso de la Búsqueda de Respuestas el conjunto de documentos de partida no está predeterminado. Es decir, el Resumen Automático no requiere Recuperación de Documentos, y la Búsqueda de Respuestas sí. Por tanto, la Búsqueda de Respuestas y el Resumen Automático son dos problemas que en principio podrían estar estrechamente relacionados. Sin embargo, dado que los sistemas actuales de Búsqueda de Respuestas no devuelven respuestas tan elaboradas, existen hoy por hoy grandes diferencias entre ambas tareas a efectos prácticos.

En general, podemos constatar que, al igual que ocurre con las futuras líneas de desarrollo de los sistemas de Búsqueda de Respuestas, la evolución de los sistemas



NIVEL DE COMPLEJIDAD	PREGUNTA	FOCO DE CONSULTA
Nivel 1: Usuario casual	¿Por qué Elián González abandonó los EEUU?	FOCO: La partida de Elián.
Nivel 2: Formularios	¿Cuál es la postura del gobierno respecto a la deportación de Elián a los EEUU?	FOCO: Un conjunto de formularios del tipo: Acciones llevadas a cabo por el INS, etc.
Nivel 3: Informe sencillo	¿Por qué Elián es considerado un inmigrante?	FOCO: Un conjunto de preguntas sencillas: La llecada de Elián, la nacionalidad de Elián, la edad de Elián, etc.
Nivel 4: Analista profesional	¿Cuál fue la reacción del Gobierno Cubano ante la deportación de Elián?	FOCO: Todas las acciones llevadas a cabo por el Gobierno Cubano y de EEUU en especial por lo líderes de las plataformas anticastristas. Consecuencias derivables de las leyes vigentes en EEUU.

Figura 2.2: Categorización de consultas en búsqueda de respuestas según Q&A Roadmap Committee

de resumen se dirige hacia el problema de la Síntesis de Información. Las primeras aproximaciones al problema del Resumen Automático se centran en la generación de resúmenes sobre un único documento [KPC95, TM97], mientras que a partir de 1997 los esfuerzos comienzan a centrarse en la generación de resúmenes a partir de varios documentos [RHB00, MB97, LH02, KSH02, SOC+02]. Posteriormente, en 2003, con motivo de la serie anual de conferencias DUC (Document Understanding Conferences), principal foro de evaluación de sistemas de resumen en la actualidad, se han desarrollado sistemas guiados por necesidades de información expresadas en forma de consulta [GHW03, ROQT03, CKSZ03]. Finalmente, en 2005 y 2006, se establece como tarea en el DUC la Síntesis de Información, tomando como referencia publicaciones derivadas de este trabajo[1]. En estos foros, se establece una serie de consultas, un conjunto amplio de documentos por cada consulta y se establece como tarea la elaboración de un resumen extenso.

La tarea de Resumen automático está estrechamente relacionada con la Síntesis de Información, en cuanto que en muchos casos requiere la identificación, descomposición, organización y extracción de relaciones sobre el conjunto de piezas de información contenidas en las fuentes originales. Sin embargo, el acceso a la información en la SI posee una serie de características propias frente a la tarea de Resúmen en general:

- Debe existir una necesidad de información concreta. En muchos casos, la tarea de resumen automático se ha planteado como un proceso de reducción de los contenidos sin definir previamente una necesidad de información.

- La información debe estar distribuida a lo largo de varias fuentes. Entendemos que hablamos de SI cuando este conjunto de documentos es amplio.

- Debe de ser necesario el análisis y la elaboración de información. Por ejemplo, es necesario un trabajo de análisis si el grado de relevancia de las piezas de información contenidas en las fuentes depende de otras piezas de información, o si la información requerida no se encuentra de forma explícita en las fuentes.

El hecho de considerar como entrada al sistema de resumen un conjunto de documentos y una consulta, es decir, el resumen multidocumento orientado a consulta, implica que el sistema debe operar sobre una necesidad de información integrando datos procedentes de distintas fuentes, es decir, cumple las dos primeras características. Sin embargo, no necesariamente requiere un proceso de análisis y elaboración de información, elementos clave en la SI.

En definitiva, podemos decir que la tarea de resumen, y en especial el resumen multidocumento orientado a consulta, subsume a la Síntesis de Información. Es por ello por lo que dedicaremos parte de los preliminares de este libro a las técnicas empleadas en sistemas de Resumen Automático.

[1] http://www-nlpir.nist.gov/projects/duc/duc2005/tasks.html

2.2. Técnicas generativas de resumen automático

Desde el punto de vista de las técnicas empleadas, Mani y Maybury [Man01a] diferencian entre técnicas de resumen automático basadas en la recopilación y organización de fragmentos de texto (resúmenes de tipo extractivo), de otros que aplican técnicas de generación de lenguaje, (resumen generativo). En el primer caso se presupone que una recopilación y ordenación de frases o párrafos sueltos es suficiente, mientras que en el otro caso la información recopilada se reescribe para configurar el resumen. La mayoría de los sistemas de Resumen Automático son de tipo extractivo mientras que el conjunto de sistemas que aplican generación de lenguaje es más reducido, debido, entre otras cosas, a que el proceso de reescritura implica un procesamiento lingüístico muy sofisticado.

En lo que respecta a sistemas de resumen generativo, puede distinguirse entre dos casos. El primero de ellos cubre aproximaciones en las que se refinan aspectos formales como la fluidez o legibilidad de un resumen generado de forma automática. Algunas de las técnicas empleadas para refinar el resumen son: resolución de anáfora [HL02a], o la eliminación de sintagmas no relevantes como clausulas de relativo o adverbiales [BCD02].

El segundo tipo de sistemas que emplean generación de lenguaje incluye aquellas aproximaciones en las que todo el contenido del resumen se redacta, de forma automática, a partir de la salida de un módulo de Extracción de Información. Es decir, la información relevante contenida en los documentos originales se representa en un formato estructurado y, posteriormente, se transforma de nuevo en formato textual. La principal referencia en este tipo de aproximación es el trabajo de McKeown y Radev [MR95]. En este modelo, el sistema genera, a partir de los formularios cumplimentados por sistemas de Extracción de Información, un resumen que describe el contenido de un conjunto de documentos. Los formularios están orientados al dominio de los documentos. Por ejemplo, formularios para documentos biográficos contienen campos como *"fecha de nacimiento, lugar de estudios etc."* El sistema organiza los formularios detectando relaciones entre sus respectivos campos. Un planificador de discurso organiza la información en párrafos y, finalmente, un generador de lenguaje transforma el contenido de los formularios en textos. Este trabajo muestra que es posible generar automáticamente un resumen a partir de una representación simbólica. Sin embargo, esta aproximación está inevitablemente condicionada por las restricciones derivadas del proceso de Extracción de Información, que se desarrolla en general dentro de un dominio restringido.

En cualquier caso, generación de lenguaje no implica generación de información. Es decir, los objetivos de la generación de lenguaje aplicada en modelos de resumen automático no se corresponden con la fase de elaboración de información dentro del proceso de SI. La diferencia es que en el caso de los modelos de resumen no se genera nueva información, sino que se expresa de manera legible y resumida la información recopilada. Sin embargo, estas técnicas podrían dar legibilidad al informe generado mediante SI.

2.3. Técnicas extractivas de resumen automático

En la clasificación propuesta por Teufel (figura 2.3) el conjunto de aproximaciones de tipo extractivo se descompone distinguiendo entre casos en los que se trata de manera aislada cada frase o párrafo en el proceso de selección, respecto a casos en los que se analiza las relaciones existentes entre éstos.

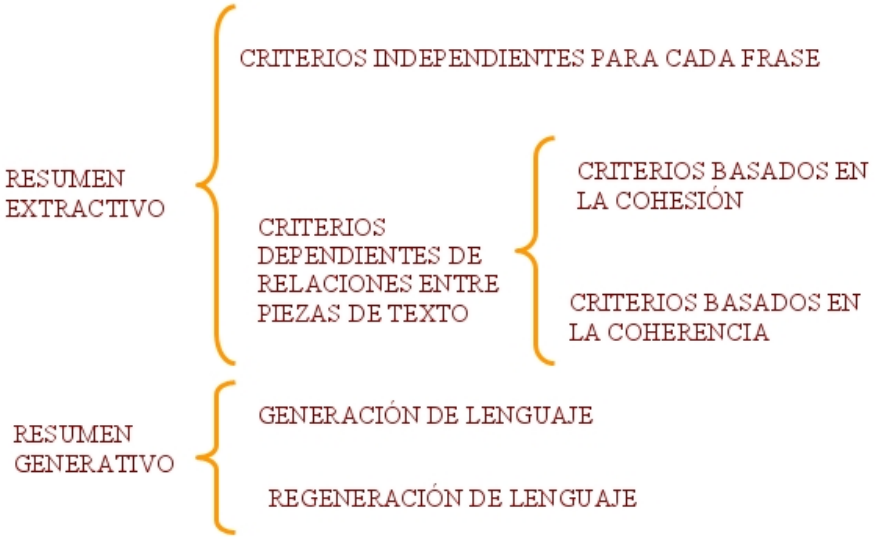


Figura 2.3: Clasificación propuesta por Simone Teufel para los sistemas de resumen automático

Aspectos que determinan de forma independiente la idoneidad de un fragmento para formar parte del resumen son: la presencia de términos característicos del tema del que trata el documento, localización de la frase, longitud de la frase, presencia de entidades o nombres propios, etc. Por otro lado, aspectos que determinan las relaciones entre fragmentos de texto son la cohesión y la coherencia.

Las técnicas basadas en la cohesión establecen relaciones entre fragmentos aislados en un proceso de agrupación, mientras que las técnicas basadas en la coherencia realizan un proceso de partición del documento. En ambos casos el resultado consiste en grupos de fragmentos de texto relacionados. Una vez que se han identificado dichos grupos, se seleccionan los fragmentos siguiendo dos objetivos:

1. Cubrir todos los aspectos del tema tratado en el texto original, que se traduce en que el conjunto seleccionado cubra una buena cantidad de estos grupos.

2. Evitar información no relevante en el resumen final, que se traduce en seleccionar solo los fragmentos más representativos de cada grupo.

A continuación describimos con más detalle cada una de estas técnicas.

2.3.1. Localización del fragmento como criterio de selección

La localización es un aspecto determinante sobre todo en el contexto de artículos periodísticos o científicos, en donde los primeros fragmentos de un documento, párrafo o sección son representativos del resto. Por ejemplo, es corriente que en la primera frase de un párrafo se plantee una idea que posteriormente se desarrolle a lo largo de éste. Lo mismo ocurre con las cabeceras de los artículos periodísticos respecto al documento completo. Un ejemplo de ello son los resultados obtenidos en el foro de evaluación DUC [BDH+00] en donde técnicas básicas basadas en la selección de primeras frases de documentos no obtienen resultados mucho peores que técnicas sofisticadas.

En diversos estudios se ha comparado la localización del fragmento en su documento frente a otros criterios de selección [Bax58, KPC95, TM97]. Se ha podido comprobar que la localización es uno de los criterios más significativos de los que determinan la idoneidad para ser incluido dentro de un resumen. En concreto, Kupiec emplea un método de aprendizaje basado en fórmulas bayesianas para analizar el peso de este criterio junto con otros como expresiones indicativas de relevancia como *"en resumen"*, o la presencia de términos del título, términos frecuentes del documento original, etc., siendo la localización el criterio más determinante.

2.3.2. Términos y expresiones indicativas de relevancia

Se trata de términos o expresiones predefinidas, es decir, independientes del contenido de las fuentes que se pretenden sintetizar, cuya presencia en fragmentos representa un criterio en el proceso de selección. La ocurrencia de estos términos denota idoneidad del fragmento para ser seleccionado, o lo contrario.

Elementos que denotan idoneidad son expresiones del tipo *"En resumen.."*,*"En definitiva"*, etc. [KPC95, HSIM02] o determinados tipos de términos como nombres propios [GMCC00, Sch02]. Elementos que actúan en detrimento de la relevancia del fragmento son expresiones temporales, pronombres al comienzo de frase, frases que empiezan por expresiones del tipo *"pero"* o *"a pesar de"*, paréntesis, formas del verbo *"decir"* etc. [LH02, KSH02, Sch02]. En el caso de Teufel y Moens [TM97] se utiliza 1670 sintagmas indicativos de relevancia agrupados en 5 categorías de calidad, tanto positivas como negativas (-1..+3).

En [Sch02] los términos o expresiones indicativas se identifican automáticamente a partir de córpora independientes. En este trabajo se extrae mediante métodos estadísticos una serie de expresiones típicas ocurrentes en las líneas de cabecera de los artículos periodísticos. Además, se aplica técnicas de procesamiento lingüístico a nivel sintáctico con el objetivo de detectar verbos asociados a tipos concretos de entidades (ladrón/detener, presidente/elegir etc..). La co-ocurrencia de estos tipos de entidad y verbos asociados denota idoneidad del fragmento para ser seleccionado.

2.3.3. Longitud de los fragmentos como criterio de selección

La longitud de las frases candidatas a pertenecer al resumen se tiene en cuenta como criterio de selección en varios trabajos [TM97, Sch02, ORL02, HSIM02, KSH02, SOC$^+$02], dándose preferencia a frases más cortas. Se comprueba que las ideas principales contenidas en un texto son enunciados que vienen expresados mediante frases cortas, mientras que explicaciones y detalles vienen expresadas mediante frases más largas.

Contrariamente a los casos anteriores, si los fragmentos seleccionados no son frases sino segmentos en la estructura de discurso [MG01], los resultados mejoran cuando se da preferencia a fragmentos más largos. Aquellas unidades de discurso más largas serán a su vez piezas de información más relevantes en sus respectivos documentos.

2.3.4. Identificación del tema

Esta técnica consiste en identificar y representar previamente el tema tratado en el documento o el conjunto de documentos que se pretende sintetizar, para posteriormente seleccionar aquellos fragmentos que guarden relación directa con éste.

Existen distintas técnicas de representación del tema. Una de estas aproximaciones consiste en extraer los términos del título [Edm69, PW94, Mah99, KPC95, TM97] o de la cabecera del documento [HSIM02]. Los fragmentos en los que aparezcan dichos términos tendrán más peso en el proceso de selección. La representación del tema mediante términos del título se aplica únicamente a resumen mono-documento, en donde existe un único título.

Otras aproximaciones representan el tema como un conjunto de términos extraídos a partir del contenido completo de los textos mediante técnicas estadísticas [KSH02, SOC$^+$02, LH02]. Estas técnicas se apoyan en la distribución de los términos a lo largo de los documentos para identificar aquellos que, por frecuencia y distribución, sean más representativos en las fuentes que se desea resumir. En algunas propuestas se considera de forma especial para la representación del tema a las entidades como nombres de persona, organización , lugar, etc. [Sch02, GHW03].

En la aproximación propuesta por Schiffman y McKeown [Sch02] se genera una representación del tema a nivel léxico-conceptual. Es decir, no se consideran los términos originales sino una representación semántica de los mismos. De esta forma, se relacionan entre sí términos sinónimos o hiperónimos, abordando el problema de la sinonimia. Por ejemplo, la aparición del término “aeroplano” y del término “avión” en un mismo documento introduce un único concepto en la representación del tema. Para ello, se emplea el diccionario de sentidos WordNet de donde se puede extraer todos los posibles sentidos de un término y otros conceptos asociados mediante relaciones de hiperonimia (“avión-vehículo”).

En otras aproximaciones se representa el tema como un punto en un espacio vectorial [ORL02, RHB00]. Los términos contenidos en un documento o fragmento determinan un punto en el espacio. El tema se representa como el punto centroide respecto al conjunto de documentos a resumir. La idoneidad de un fragmento en

el proceso de selección se determina en función de su posición relativa respecto al tema en el espacio vectorial.

2.3.5. Técnicas basadas en cohesión entre fragmentos

El concepto de cohesión fue introducido por Halliday y Hasan [HH76] y representa las relaciones existentes entre distintas partes de un texto. Las técnicas de selección de fragmentos basadas en la cohesión parten de la idea de buscar enlaces entre piezas de información para emplearlos como criterios de relevancia a la hora de extraer los fragmentos de un documento en la elaboración del resumen. Se presupone que los fragmentos más susceptibles de ser incluidos en el resumen son aquellos que se encuentran más relacionados con otros fragmentos del texto original.

La cuestión es qué criterios determinan las relaciones entre fragmentos, o lo que es lo mismo, como determinar la distancia entre dos fragmentos de texto en relación a sus contenidos. Algunos de estos criterios son:

Distribución y co-ocurrencia de términos. En la mayoría de los casos se emplea distancias basadas en distribución y co-ocurrencia de términos para establecer distancias entre fragmentos. En algunos sistemas se identifican conjuntos de fragmentos contiguos y estrechamente relacionados dentro de un documento [MSB97, SSMB97]. En otras aproximaciones se identifican grupos de fragmentos no contiguos y similares entre sí [ORL02, RHB00, MG01, LH02, SOC$^+$02, SBW00]. Estas técnicas tienen la ventaja de que son sencillas de implementar e independientes del lenguaje de los textos.

Contigüidad y co-ocurrencia de elementos léxico-conceptuales (cadenas léxicas) Las aproximaciones anteriores no tienen en cuenta aspectos como la aparición de términos sinónimos o semánticamente relacionados. Una alternativa que sí contempla estos aspectos es la identificación de cadenas léxicas. Estas cadenas determinan las relaciones de cohesión entre fragmentos [BME99]. Una cadena léxica enlaza una serie de términos cercanos en el texto que mantienen cierta relación a nivel semántico.

Según el tipo de relación, Barzilay clasifica en tres grupos estas cadenas en función de las relaciones existentes entre los términos que las componen: Extrafuertes (repetición del mismo término), fuertes (términos sinónimos), medio fuertes (otros tipos de relación como hiperimia). La relaciones más fuertes se pueden establecer entre términos más distantes entre sí, mientras que relaciones débiles son tenidas en cuenta únicamente entre términos cercanos.

Los conjuntos de frases que comparten una cadena léxica representan una unidad temática en el texto. La aproximación consiste en escoger una de las frases como elemento representativo del conjunto.

La aproximación basada en cadenas léxicas presenta algunas limitaciones. Una de ellas es que los conjuntos de fragmentos que forman una unidad de

cohesión no son divisibles en unidades de cohesión más pequeñas. Es decir, no se puede "jugar" con la granularidad o nivel de detalle del resumen. Otra limitación viene dada por conceptos que son referenciados anafóricamente. Por ejemplo, pronombres ("él se fue..") o términos que adquieren sentido solo en su contexto ("el presidente..."). Una técnica que solventa este problema es el uso de cadenas de correferencia.

Cadenas de correferencia. Son similares a las cadenas léxicas, solo que no se componen de términos semánticamente relacionados, sino de referencias a una misma entidad, resolviéndose así el problema de la anáfora. Los criterios de selección son similares al caso de las cadenas léxicas [BM98, BCD02]

Representación semántica del contenido. Algunos autores analizan las cohesión del resumen mediante una representación semántica de la información contenida en los documentos. En el trabajo de Mani y Bloedorn [MB97] se calcula, para cada documento, un grafo cuyos nodos representan conceptos, que se asocian mediante relaciones expresadas en enunciados del documento. Posteriormente se calcula la conjunción de varios grafos, escogiendo para el resumen final aquellos fragmentos que han generado los nodos del grafo inclusión.

Otro caso en el que se emplean técnicas de este tipo es el de la aproximación de Barzilay [Bar03]. En este caso se extrae oraciones simples de los textos originales transformándolas en un formato común. Este paso requiere complejas técnicas de procesamiento lingüístico dado que una misma oración puede expresarse de múltiples formas (pasiva, activa, como proposición de relativo, etc.). El criterio de selección de información se basa en la similitud entre las oraciones normalizadas. A diferencia de los otros sistemas basados en Extracción de Información, éste no extrae para la generación del resumen final las frases completas a partir de las cuales ha extraído la información, sino que reescribe las proposiciones empleando reglas de generación de lenguaje natural.

Sea cual sea el criterio de cohesión entre fragmentos, todos los autores coinciden en varios aspectos:

1. En general, los fragmentos más relevantes mantienen relaciones de cohesión con otros fragmentos en los documentos originales.

2. Dentro de un conjunto de fragmentos cohesionados en los documentos originales, la localización de estos fragmentos es un criterio de relevancia más determinante que el grado de cohesión con otros fragmentos del mismo conjunto.

3. No necesariamente los fragmentos que componen el resumen final mantienen relaciones de cohesión.

En definitiva, las relaciones de cohesión permiten identificar aspectos relevantes del tema tratado en los documentos, pero no son en sí mismas un criterio suficiente para la selección de fragmentos en el proceso de resumen. Este criterio se complementa con otros criterios como la localización del fragmento.

2.3.6. Técnicas basadas en la coherencia del texto

La coherencia del texto viene dada por las estructuras retóricas que contiene, es decir, por el modelo de discurso del documento. Las técnicas de resumen automático basadas en coherencia asumen que se puede fragmentar un texto en sus diferentes partes de forma jerárquica siguiendo la línea de discurso. Por ejemplo, un documento que trate el problema de la fiebre, si está correctamente redactado, se podrá dividir en segmentos contiguos que traten asuntos como *"causas de la fiebre, diagnóstico de la fiebre, tratamiento de la fiebre etc."*. A su vez el proceso de segmentación será aplicable a cada una de sus partes.

Algunos sistemas de generación automática de resumen emplean diferentes tipos de evidencias para identificar el modelo de discurso y así fragmentar el texto y extraer los fragmentos más relevantes. Dependiendo de la teoría de discurso empleada se analiza el texto identificando segmentos y organizándolos en una jerarquía que incluye frases, párrafos u oraciones simples. A los nodos de la jerarquía, según su localización en ésta, se le asignan distintos pesos para la extracción de fragmentos relevantes.

Algoritmo de Marcu

Un algoritmo sencillo, usado en varias aproximaciones para la extracción de estructuras retóricas, fue introducido por Marcu [Mar95]. Este método asume varios aspectos en relación a la estructura de un texto:

1. Las unidades elementales de la estructura de un texto no están solapadas.

2. La estructura del discurso es jerárquica. Esto quiere decir que dos segmentos del texto pueden o bien representar aspectos distintos o bien tener una relación de núcleo-satélite, en donde uno de ellos aporta información extra en relación a la información contenida en el otro fragmento.

3. Estas estructuras pueden representarse mediante árboles binarios.

4. Las relaciones entre fragmentos son extrapolables a las subunidades.

Marcadores de discurso

En la aproximación propuesta por Marcu, la detección de estas estructuras en los textos se apoya en marcadores de discurso del tipo *"mientras que"', "dado que", "es decir"*, que son empleados como elementos de cohesión entre proposiciones o como macroconectores entre frases o párrafos. Según Redeker [Red90]

el conjunto de marcadores de discurso es suficiente para la derivación de las estructuras retóricas de un texto, aunque el comportamiento de éstos está sujeto a ambigüedades.

No resulta demasiado complicado, mediante reglas ortográficas, distinguir hasta un 80 % de los casos de ambigüedad entre conectores de discurso y conectores de frases [Hir94]. Sin embargo, resulta complicado identificar el tamaño de los fragmentos que el macro-conector asocia. Marcu propone un método para la identificación del tramo de texto cubierto por cada conector. El autor reconoce que resulta imposible determinar con exactitud estas regiones, aunque lo que si se puede hacer es reducir las variables posibles teniendo en cuenta restricciones generales del árbol de discurso. Marcu se basa en las hipótesis asumidas antes descritas para restringir el conjunto de posibilidades mediante un sistema basado en restricciones.

Finalmente, tras haber identificado las estructuras de discurso se asume que aquellos que se encuentran en un punto más alto de la jerarquía son más relevantes. Hay que decir que, con anterioridad ya se había planteado este modelo de extracción de estructuras retóricas, aunque para el caso concreto del japonés [OSM94], distinguiendo nexos entre oraciones de nexos entre párrafos.

Coherencia basada en cadenas léxicas

Otro caso interesante es el de Ken [KCC+98], en el que, tras una segmentación del documento completo basada en distribución de términos, se extrae una estructura jerárquica de discurso dentro de cada segmento. Esta estructura de discurso se identifica a partir del solapamiento de cadenas léxicas. Una unidad de discurso está asociada a una cadena léxica, y se descompone en otras subunidades en función de las cadenas léxicas contenidas. Esta aproximación tiene la ventaja de ser independiente del lenguaje, ya que no requiere la identificación de nexos de discurso.

Roles de discurso predefinidos

Teufel y Moens [TM98] plantean también la estructura retórica del texto como recurso para la generación de un resumen, pero desde otra perspectiva. Proponen a priori un conjunto de 5 roles retóricos del tipo *"introducción", "conclusiones"* etc. Cada frase del texto se alinea con un determinado rol. Este alineamiento se realiza en base a criterios independientes para cada frase. Los resultados mostraron que las expresiones indicativas *"en resumen...", "este trabajo..."* superan a otros criterios como localización o longitud de las frases tanto en el proceso de extracción de oraciones relevantes como en la alineación con roles retóricos.

Saggion y Lapalme [SL00] aplican una heurística semejante solo que esta vez tienen en cuenta hasta 52 roles retóricos. Esto es posible dado que la generación automática del resumen está en concreto orientado a artículos científicos, aunque dentro de cualquier dominio. Los roles son del tipo *"identificación del problema", "identificación de la solución"*, etc. Al igual que en el caso anterior se emplean expresiones o términos indicativos para la alineación entre fragmentos y roles. Esta

distinción tan fina entre roles retóricos permite que el sistema pueda generar tanto resúmenes de tipo indicativo como resúmenes de tipo informativo.

Una ventaja del estudio de la coherencia en base a roles predefinidos es que no está sujeta a suposiciones relativas a la contigüidad y organización de los elementos del discurso. Sin embargo, los roles predefinidos son más dependendientes de dominio. Por ejemplo, roles como *"identificación del problema"* o *"identificación de la solución"* derivan de la naturaleza de los artículos científicos.

Criterios de coherencia en resumen multi-documento

Estas técnicas basadas en la coherencia no son aplicables a textos distribuidos en varios documentos, dado que, por definición, la estructura retórica está ligada a un único texto. Sin embargo, algunas aproximaciones al problema del resumen multi-documento extraen unidades de discurso de cada una de las fuentes [BCD02, MG01]. En cualquier caso, en resumen multi-documento no basta con analizar el árbol que representa la estructura de discurso para extraer los fragmentos más relevantes, sino que es necesario estudiar las relaciones de cohesión entre las piezas de información de los distintos textos.

2.4. Tratamiento de varios documentos

El tratamiento de varios documentos es un aspecto clave del problema de la Síntesis de Información. En este apartado se analiza la forma en que se aborda esta cuestión en sistemas de Resumen Automático.

Muchas de las técnicas empleadas en resumen mono-documento se pueden aplicar igualmente al caso multi-documento. Por ejemplo, en [KSH02] se comprueba que, entre otros criterios orientados al tratamiento de varios documentos, la localización del fragmento en su documento sigue siendo un aspecto clave en la elaboración de los resúmenes.

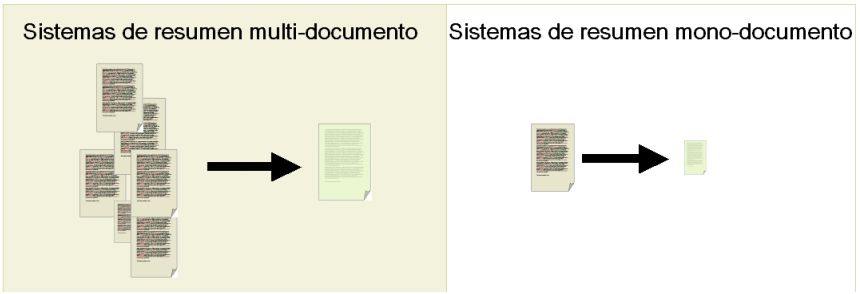


Figura 2.4: Resumen multidocumento frente a resumen mono-documento

2.4.1. Características de un resumen multi-documento

En [SOC+02] se realizó un estudio acerca de las características de los resúmenes manuales generados por sujetos de prueba sobre el corpus empleado en el DUC

(Document Understanding Conference), en donde se establece un marco de evaluación sobre el que comparar distintos sistemas de resumen. En el corpus sobre el que se evalúan estos sistemas, los conjuntos de documentos por resumir constan de menos de 20 documentos y los resúmenes se componen de unas 100 palabras. Algunas conclusiones obtenidas fueron:

- Los sujetos prestan mucha más atención a algunos documentos que a otros.
- No parece haber ningún patrón de comportamiento en la extracción de fragmentos en función de la estructura de los textos.
- Un 60 % de la información contenida en un resumen manual no aparece en otro de los resúmenes manuales.

2.4.2. Tipos de documentos de partida

En varios trabajos se ha comprobado que existen diferencias en la tarea de resumen dependiendo del tipo de conjuntos de documentos a resumir [MS02, KSH02]. Básicamente, se consideran tres tipos:

- Conjuntos de documentos que tratan un mismo asunto que evoluciona a lo largo del tiempo, por ejemplo, *"la elección Clinton como presidente"* (resumen mono-evento).
- Conjuntos de documentos que tratan distintas instancias de un mismo asunto, por ejemplo *"Incendios en EEUU"* (resumen multi-evento).
- Biografías o conjuntos de documentos que tratan diferentes aspectos de una entidad.

Por ejemplo, en el sistema orientado a resumen de noticias on-line Columbia Newsblaster, se distinguen estos tres tipos temas a resumir. Para el caso de los documentos centrados en un mismo asunto, se emplea el sistema MultiGen [BME99] que aplica técnicas estadísticas para la identificación de oraciones de contenido similar, con el fin de extraer la intersección de los temas tratados en los documentos. Sin embargo, en el caso de los grupos de documentos que tratan un tema general o biográfico, se emplea el sistema DEMS [SNM02], dando especial importancia a piezas de texto que ofrecen información nueva, o características del tipo de entidad en la que se centran los documentos.

2.4.3. Dificultades en resumen automático multi-documento

Desde un punto de vista computacional, el tratamiento de varios documentos implica nuevas dificultades añadidas [GMCC00]:

1. **Presencia de información redundante.** La redacción correcta de un documento lleva consigo la eliminación de información redundante, por lo que el problema de la redundancia en las fuentes en resumen mono-documento

supone menor dificultad. Sin embargo, en un conjunto de documentos relacionados aparece mucha información repetida, necesaria para mantener la coherencia de cada documento individual. Es necesario en este tipo de sistemas un mayor control sobre la información redundante.

2. **El aumento del ratio de compresión**. La disminución del tamaño del resumen respecto al de las fuentes originales, exige un mayor cuidado en la selección de los fragmentos.

3. **Necesidad de tratar una dimensión temporal**, típica en documentos de tipo periodístico. Informes posteriores completan la información mostrada en informes anteriores en relación a una mismo evento.

4. **Mismos conceptos referenciados de distinta forma**. La terminología empleada en cada documento puede diferir aun refiriéndose al mismo concepto.

5. **Necesidad de contextualizar los fragmentos recopilados**. Una misma información, por ejemplo, *"la bolsa subió un 2 %"*, puede tener distinto sentido dependiendo del documento al que pertenece, por ejemplo, ¿cuándo subió?, ¿qué bolsa?.

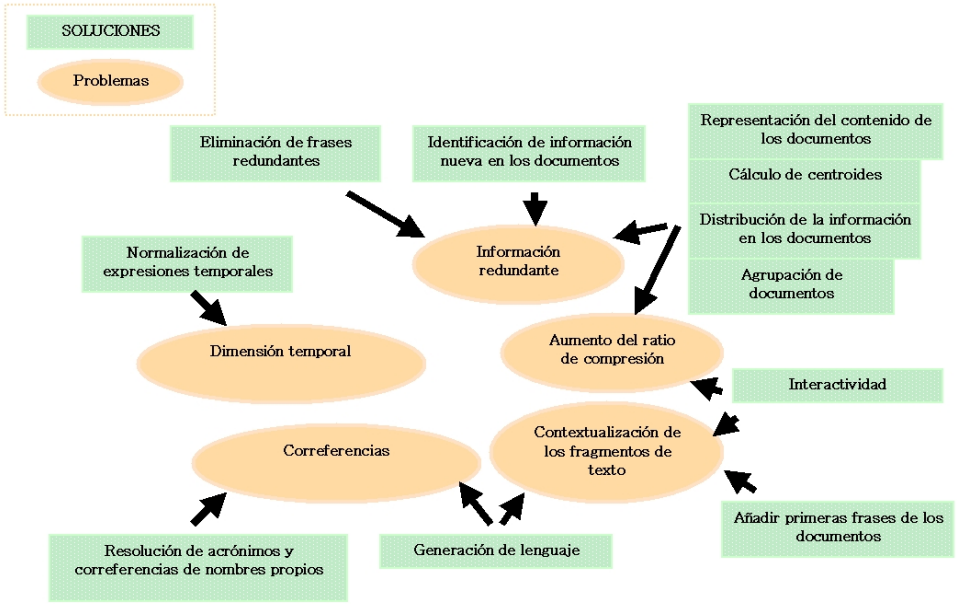


Figura 2.5: Problemas y aproximaciones en sistemas de resumen multi-documento

2.4.4. Técnicas empleadas en resumen multi-documento

Ante estos problemas, se han desarrollado técnicas de selección de fragmentos complementarias a las aplicadas en resumen mono-documento. Algunos de los aspectos en los que se apoyan estas técnicas son (figura 2.5):

Distribución de la información en distintos documentos. La repetición de la información a lo largo de varios documentos denota representatividad dentro del conjunto. Por otro lado, la recopilación de información relevante y exclusiva de cada documento de la colección asegura una buena cobertura del resumen final. [Sch02, LH02, ORL02, BME99].

Caracterización de la información nueva en los documentos. Para evitar incluir información redundante en el resumen, es necesario reconocer las piezas de texto que introducen nueva información en el conjunto de documentos. Por ejemplo, para el desarrollo del sistema DEMS [SNM02] se identificaron las características típicas de estas piezas de texto, analizando el tipo de términos que aparecen en las primeras frases de cada documento periodístico de un extenso corpus.

Cálculo de un centroide. (Apartado 2.3.4) Cada documento o fragmento de la colección, según los términos que contiene, se sitúa en un punto de un espacio vectorial. Se calcula el centroide de todos estos puntos y se escoge fragmentos de texto dependiendo de su posición relativa a dicho centroide. [ORL02, RHB00].

Eliminación de frases redundantes. Para abordar el problema de la redundancia de la información en distintos documentos, en [RHB00], se identifican relaciones de subsunción entre pares de frases, en las que el contenido de una de ellas aparece en la otra. Se identifica además relaciones de equivalencia en la que ambas frases se subsumen mutuamente.

En [GMCC00], se emplea el algoritmo MMR (Maximal Marginal Relevance), más sofisticado. Este algoritmo, entre otros criterios, penaliza aquellos fragmentos que:

- se asemejen a fragmentos ya seleccionados anteriormente
- pertenezcan a grupos de fragmentos similares de entre los que ya se ha seleccionado anteriormente alguno de ellos
- pertenezcan a documentos de los que ya se ha seleccionado algún otro fragmento.

A pesar de que los autores pudieron observar que los resúmenes generados siguiendo estos criterios contenían menos información redundante, tanto en estos mismos experimentos como en los realizados en [KSH02], no se obtuvo mejoras cuantitativas en los resultados en relación a otros algoritmos de eliminación de redundancias.

Normalización de expresiones temporales. La fecha de edición de los documentos se emplea como criterio de selección, dando preferencia a las ocurrencias de un contenido en las últimas ediciones o a las primeras [BME99, GMCC00, SNM02]. Para considerar la dimensión temporal de los hechos descritos en los documentos, se han empleado técnicas de extracción de información con el fin de reconocer y formatear fechas [BCD02, BME99].

Añadir a los fragmentos recopilados la primera frase de su documento. Esta aproximación permite contextualizar, especialmente en el dominio periodístico, los fragmentos recopilados a partir de distintos documentos [LH02]. En [BCD02] se aplica una variante introduciendo oraciones previas en donde se resuelve anáforas de la oración seleccionada.

Generación de lenguaje Algunas aproximaciones, con el fin dar cohesión a los fragmentos extraídos a partir de distintos documentos, aplican técnicas de generación de lenguaje [BME99, MR95].

Resolución de acrónimos y correferencias de nombres propios . Mediante el reconocimiento y tratamiento de las entidades, es posible resolver parcialmente el problema de las referencias a un mismo concepto de distinta forma en distintos documentos [Sch02].

Ajuste entre representaciones de los documentos Esta aproximación consiste en generar una representación de cada uno de los documentos, de forma que permita identificar piezas de información comunes a todos ellos. Por ejemplo, el sistema MultiGen [BME99] obtiene la intersección de frases procedentes de distintos documentos mediante el alineamiento de árboles sintácticos. En [MB97] se generan representaciones en formato de grafo, en donde los nodos representan entidades y las relaciones entre nodos vienen dadas, tanto por criterios de localización, como de relaciones semánticas extraídas del texto. Para la generación de resúmenes, se identifican los nodos comunes y únicos de los distintos grafos. Análogamente en [MR95, HL02a] se contrastan formularios, resueltos de forma automática para cada documento mediante técnicas de extracción de información.

Agrupación de documentos La agrupación de documentos permite obtener información útil para el proceso de resumen, por ejemplo, asegurando la cobertura mediante la selección de piezas de texto pertenecientes a cada uno de los grupos [SBW00, RHB00].

Interactividad Atendiendo al problema de la contextualización, Goldstein sugiere la incorporación de paradigmas interactivos en los sistemas de resumen, de tal forma que sea el propio usuario el que contextualice los fragmentos recopilados por el sistema [GMCC00].

2.5. Tratamiento de la consulta

Dentro del foro de evaluación DUC (Document Understanding Conference), en el año 2003 aparecen aproximaciones centradas en generar un resumen a partir de una necesidad de información expresada mediante una pregunta o punto de vista [GHW03, ROQT03, CKSZ03]. Las técnicas empleadas por estos sistemas no difieren demasiado de las empleadas en generación de resumen no orientado a consulta. Por lo general, la pregunta o punto de vista se integra en el proceso de selección de

fragmentos, otorgando un mayor peso a aquellos que mantienen relación con los términos de consulta.

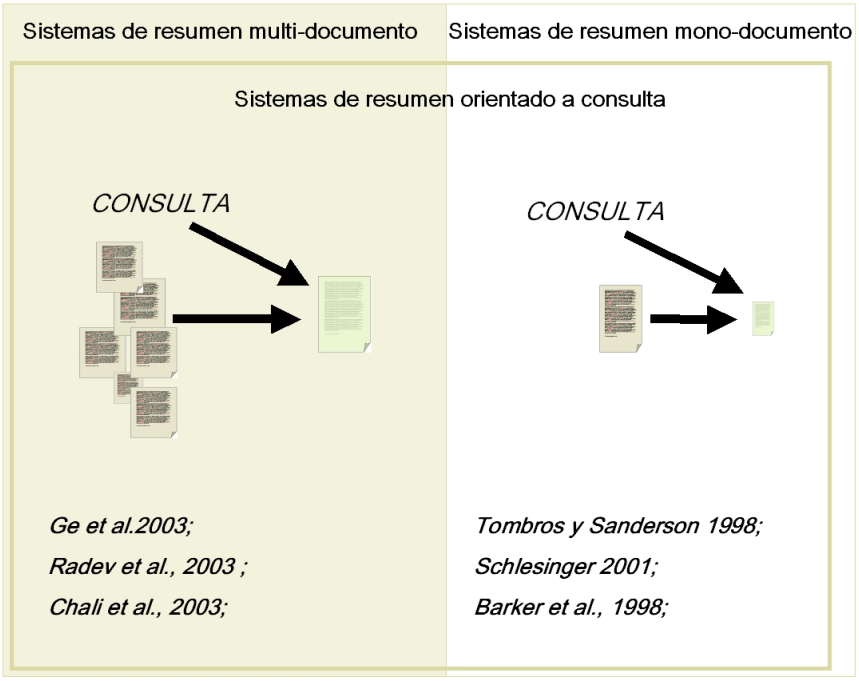


Figura 2.6: Resumen orientado a consulta

Por ejemplo, Jiayin Ge [GHW03] basa su estrategia en la identificación de entidades (nombres propios) representativas en el asunto y en la recopilación de fragmentos que las contienen. Los términos de consulta son considerados como entidades y cobran más peso que las entidades que aparecen únicamente en los documentos.

El sistema propuesto por Dragomir Radev [ROQT03] genera resúmenes a partir de la relación existente entre fragmentos contenidos en los documentos y el tema central, representado como un centroide en un espacio vectorial. Integra la pregunta o punto de vista que guía el resumen incorporando los términos de consulta en la representación del tema central. Esta aproximación obtuvo los mejores resultados para la tarea concreta de generación de un resumen corto en respuesta a una pregunta en el DUC 2003.

Yllias Chali [CKSZ03] emplea técnicas basadas en cadenas de términos semánticamente relacionados que ocurren en puntos próximos del texto (cadenas léxicas, apartado 2.3.5). Para orientar la generación del resumen hacia la consulta se otorga más peso a aquellos fragmentos que, además de contener términos pertenecientes a cadenas léxicas, contienen también términos de consulta.

En [MB97] se genera grafos a partir de los documentos, que representan relaciones entre entidades. La consulta se introduce en el proceso de resumen como un conjunto de nodos a tener en cuenta en mayor medida que el resto.

Como se puede ver en estos cuatro ejemplos, las técnicas empleadas relacionan

piezas de información con la consulta a un nivel muy superficial. Básicamente, se otorga más peso a los términos que aparecen en la consulta. Esto hace que los resultados dependan en gran medida de la idoneidad de los documentos en relación a la necesidad de información.

2.6. Conclusiones: aplicación de técnicas de resumen a Síntesis de Información

El acceso a la información en la SI es un proceso que requiere extracción y análisis de información, en el que se parte de un conjunto amplio de documentos relevantes y de una necesidad de información que puede expresarse en forma de consulta. Analizaremos la relación entre las técnicas actuales de resumen automático y la SI en función de cómo podrían abordarse estos aspectos.

Recopilación de información

En primer lugar, la recopilación de información se ha abordado en el contexto de los sistemas de resumen automático mediante distintas técnicas de selección de fragmentos. La localización del fragmento en el documento es el criterio más empleado para la elaboración de resúmenes. En distintos estudios se ha podido comprobar que la localización es un rasgo más determinante que otros como la longitud del fragmento, la densidad de términos relevantes o la aparición de expresiones indicativas de relevancia (sección 2.3.1). Además, la localización ha sido aplicable tanto a resumen mono-documento como multi-documento, lo que lo hace extrapolable al caso de la SI, y es complementaria a otros criterios de recopilación más sofisticados como son las estructuras de cohesión (sección 2.3.5). Incluso las estructuras retóricas, extraídas en técnicas basadas en la coherencia (sección 2.3.6), se apoyan en la localización de los fragmentos para determinar la posición que el fragmento ocupa en el árbol de discurso. Se considera que las primeras posiciones las ocupan fragmentos más susceptibles de ser seleccionados en el proceso de resumen. La importancia de la localización del fragmento en el proceso de recopilación de información es innegable. Sin ir más lejos, una recopilación de los primeros fragmentos de cada uno de los capítulos de este libro, podría constituir un resumen aceptable.

Otros rasgos que también podrían ser empleados en SI son la longitud del fragmento, la aparición de expresiones indicativas de relevancia, o la densidad de términos relevantes (sección 2.3). Estas técnicas son, en principio, independientes del dominio y aplicables tanto a resumen mono-documento, como a resumen multi-documento. Sin embargo, criterios de selección como la longitud del fragmento o la presencia de expresiones indicativas pueden perder efectividad con el aumento del ratio de compresión al considerar múltiples documentos. De hecho, en resumen multi-documento no se emplea por lo general estas técnicas, sino sobre todo, técnicas basadas en la cohesión (sección 2.4), como eliminación de frases redundantes, resolución de correferencias, ajustes entre representaciones de los documentos o agrupación de documentos.

Análisis de la información

Las técnicas de resumen cubren además, aunque superficialmente, procesos de análisis de la información. Uno de los primeros trabajos realizados en este sentido es la identificación del tema tratado en los documentos (sección 2.3). Algunas técnicas de resumen asumen que los fragmentos más relevantes contienen elementos del tema, como términos y conceptos representativos o una distribución de vocabulario semejante a una distribución centroide extraída a partir de los documentos originales. Sin embargo, este tipo de técnicas introducen precisión en el resumen final, pero no cobertura sobre los contenidos. Es posible que se pierdan detalles relevantes no relacionados directamente con el tema principal. Surge entonces el análisis de contenidos basado en la cohesión.

Este segundo nivel de análisis basado en la cohesión permite agrupar y relacionar entre sí los fragmentos de texto contenidos en los documentos originales, identificando los distintos temas tratados (sección 2.3.5). Existen múltiples criterios para la asociación de fragmentos, como la co-ocurrencia de términos, cadenas de correferencia, ajuste entre representaciones semánticas, etc. Estas técnicas son extrapolables a resumen multi-documento, y además, es posible asociar fragmentos con una necesidad de información expresada textualmente. Por tanto, el análisis de la cohesión de los documentos originales puede ser muy útil en la Síntesis de Información.

Los contenidos de los documentos originales también pueden ser analizados mediante la identificación de estructuras retóricas (sección 2.3.6). Con vistas a la Síntesis de Información, este tipo de análisis tiene la limitación de no ser, por definición, extrapolable a conjuntos de documentos, sino que es sólo aplicable sobre documentos individuales. Este análisis no parece en principio especialmente útil para la SI.

Por último, en algunos casos se ha realizado un análisis más sofisticado de los documentos originales mediante técnicas de extracción de información (sección 2.2). Estas técnicas se basan en la identificación de patrones que son altamente dependientes de dominio (por ejemplo, ”lugar y fecha de atentados”, y no son, por tanto aplicables en un sistema robusto de SI de propósito general.

Multiplicidad de documentos

La SI toma como entradas un conjunto amplio de documentos y una necesidad de información. En el desarrollo de sistemas de resumen, se ha podido constatar que la consideración de múltiples documentos acarrea algunas dificultades como son: el aumento del ratio de compresión, la pérdida de contexto de fragmentos aislados, la resolución de correferencias, etc. (sección 2.4). En este capítulo hemos apuntado algunas técnicas que permiten abordar este tipo de problemas. Además, se ha podido comprobar que el proceso de resumen se ve influido por la forma en que se distribuyen los contenidos en los distintos documentos, distinguiéndose entre documentos fuente centrados en un mismo tema y documentos fuente que describen temas distintos relacionados de alguna forma entre sí.

Tratamiento de consultas

Por último, también en tareas de resumen se ha considerado en algunos casos, una necesidad de información concreta expresada en términos de consulta (Sección 2.5). Las técnicas empleadas en estos casos se apoyan en un tipo de análisis muy superficial. Concretamente, se otorga más peso a fragmentos que contienen elementos presentes en la consulta, como términos, entidades o conceptos. Un acercamiento más detallado a la necesidad de información requiere una especialización del sistema, como es el caso de los sistemas de resumen basados en técnicas de extracción de información que se se centran en dominios concretos (sección 2.2).

En definitiva, las técnicas de resumen automático pueden ser útiles en el desarrollo de un sistema de SI, por lo menos en las fases de recopilación y análisis de la información. Sin embargo, hay que tener en cuenta que estas técnicas automáticas no son capaces de realizar un análisis del discurso comparable con el que podría realizar un humano. Aunque sea posible estructurar o relacionar automáticamente fragmentos de textos en base a distribuciones de términos o entidades, es obvio que un sistema no posee el conocimiento del mundo necesario para realizar un análisis en profundidad de los contenidos.

Estas limitaciones son más patentes en el tratamiento que estas técnicas hacen de la consulta cuando la tarea está guiada por una necesidad de información, dado que las consultas son consideradas únicamente a un nivel muy superficial.

Capítulo 3

Interactividad en sistemas de acceso a la información

Se han desarrollado distintos modelos cognitivos de tareas de acceso a la información relacionadas con la SI. Estos modelos sugieren que las tareas de acceso a la información son altamente subjetivas (sección 3.1). Por ejemplo, en el caso de la SI, distintos sujetos podrían realizar informes distintos en las mismas condiciones, siendo estos informes igualmente válidos. Esta característica de la Síntesis de Información introduce el problema de cómo elaborar un modelo de acceso a la información para la SI que se adapte a las necesidades de distintos usuarios. Una solución a este problema consiste en introducir en el modelo interacción entre usuario y sistema.

Un sistema interactivo de SI no realiza la tarea por sí solo, sino que asiste al usuario en el proceso de síntesis, ayudándole a precisar y resolver sus necesidades de información. Existen distintos esquemas de interacción aplicados en sistemas de acceso a la información textual. En este capítulo propondremos una categorización de estos esquemas, analizando en detalle cada una de ellos (sección 3.2). Algunos de estos esquemas de interacción han sido aplicados también en sistemas interactivos de resumen (sección 3.3). Como veremos, no todos los esquemas son aplicables a la SI, aun cuando han sido aplicados en la tarea de resumen. Esto se debe a que un esquema de interacción en SI debería cumplir ciertas características, como ofrecer al usuario una vista global de los contenidos, o permitir la generación de informes extensos.

3.1. Acceso a la información textual desde una perspectiva cognitiva

Existen distintos modelos de comportamiento de un sujeto durante el proceso de acceso a la información, los cuales presentan grandes diferencias entre sí. Sin embargo, en conjunto, caracterizan una serie de rasgos que podemos particularizar para la Síntesis de Información. Estas características sugieren la necesidad de incorporar componentes interactivos en un sistema de acceso a la información

orientado a SI. Estas características son:

Subjetividad y conocimiento del mundo

El acceso a la información es una tarea subjetiva, en cuanto que depende de la asimilación de información por parte del sujeto y de cómo interprete sus necesidades. Además, este proceso depende del conocimiento previo del sujeto, es decir, su conocimiento del mundo.

Por ejemplo, en el modelo de Dervin, [Der77] se interpreta el concepto de "necesidad de información" como un vacío existente entre el conocimiento que posee y el conocimiento que requiere en una situación dada. Esta necesidad de información se representa en este modelo como una pregunta, y la respuesta a dicha pregunta da paso a una nueva situación. Este modelo interpreta el acceso a la información como un proceso estrechamente relacionado con el conocimiento del mundo por parte del sujeto.

La interacción entre usuario y sistema permite adaptar el proceso de acceso a la información al usuario, abordando así el problema de la subjetividad.

Integración de la clarificación de las necesidades de información en el proceso

En muchos casos, en el proceso de acceso a la información el sujeto no puede precisar sus necesidades al comenzar el proceso, sino a lo largo de éste. Por ejemplo, en [Bel80], se propone un modelo (ASK) en el que la necesidad de información no es un objeto que el usuario pueda entender y representar con precisión. Es decir, la clarificación de la necesidad de información es un proceso integrado dentro del acceso a la información.

En esta misma línea, el modelo de Bates, al que denominó modelo *Berry Picking* [Bat90] interpreta el acceso a la información como un proceso de exploración en el que el usuario asimila información a medida que recorre distintas fuentes hasta haber asimilado la información que realmente necesita. El proceso en sí de exploración no es solo un paso intermedio, sino que forma parte del acceso a la información, posibilitando la clarificación de sus necesidades.

Esta idea implica que el usuario de un sistema de acceso a la información, para clarificar sus necesidades, necesita interactuar con las fuentes. Esto sugiere la incorporación de interacción entre usuario y sistema. Es decir, la interactividad es un rasgo que parece natural en un sistema de acceso a la información.

Fases en el proceso de acceso a la información

Existen varios modelos en los que se ha tratado de identificar las fases que componen el proceso de acceso a la información. Uno de ellos es el de Kuhlthau [Kuh91]. Este autor estructura el proceso en cinco fases: iniciación, exploración, formulación, recopilación y presentación. En un principio realiza experimentos con un gran número de estudiantes de secundaria, y en trabajos posteriores valida el modelo analizando el comportamiento de abogados [Kuh01]. En este segundo tra-

bajo, se confirma que es necesario un proceso de exploración previo a la formulación de la necesidad de información.

El modelo de Kuhlthau, impone la restricción de linealidad en las distintas fases que lo componen y, además, no contempla la manipulación de la información como parte del proceso. Sin embargo, puede existir retro-propagación o iteración en las distintas fases que componen un proceso de acceso a la información. Por ejemplo, el modelo de McKenzie [J.98] propone un ciclo iterativo de planificación, exploración, síntesis, evaluación y revisión.

Otro ejemplo de esto es el modelo propuesto por Catherin Blake [BP02] para el concepto de SI, aplicado en su caso al dominio del conocimiento médico. Este modelo se compone de tres procesos interrelacionados: recuperación de información, en el que los científicos identifican un conjunto de citas a partir de una base de datos documental, extracción de datos a partir de los documentos recopilados y análisis de la información recopilada. La necesidad de información es un elemento dinámico y central, en el sentido de que los tres procesos interactúan con ésta, rompiéndose la estructura lineal del proceso.

Un modelo más detallado en cuanto a las tareas que componen el acceso a la información es BIG6[1], desarrollado por Mike Eisenberg y Bob Berkowitz en 1988. Este modelo ha sido aplicado especialmente en entornos docentes. Se trata de una metodología de enseñanza aplicada en cientos de instituciones educativas, corporaciones y programas de educación de adultos. El modelo Big6 es representa un proceso sistemático de búsqueda, uso, aplicación y evaluación de información sobre necesidades y tareas específicas. Las seis fases que componen el modelo son:

1. **Definición del problema:** Esta fase incluye definir el problema a resolver e identificar las necesidades de información.

2. **Estrategias de búsqueda de información:** En esta fase se determinan todas las posibles fuentes y se seleccionan las mejores.

3. **Localización y acceso:** Se localizan intelectualmente y físicamente las fuentes, y se encuentra la información dentro de las fuentes.

4. **Uso de la información**: La información es asimilada (lecura, escucha, visión, etc.), extrayéndose la más relevante.

5. **Síntesis:**Se organiza la información procedente de distintas fuentes, y se presenta.

6. **Evaluación:**Se juzga el producto resultante (efectividad) y el proceso realidad (eficiencia)

Al igual que en los modelos de Blake o McKenzie, los autores recalcan que no necesariamente las fases son ejecutadas de forma secuencial. Sin embargo, sus estudios muestran que en los casos en los que el problema de acceso a la información es resuelto satisfactoriamente, aparecen todas estas fases.

[1] `http:www.big6.com`

Las dos primeras fases del modelo (definición del problema y estrategias de búsqueda de la información) no son directamente extrapolables a lo que entendemos como Síntesis de Información textual, dado que el problema a resolver en el caso de la SI es la elaboración del informe, y suponemos que se dispone de un conjunto de documentos relevantes.

Considerando los distintos modelos cognitivos hemos visto que el proceso de acceso a la información es subjetivo y complejo. Estos aspecos se pueden particularizar para la SI. Además, diversos estudios muestran que el proceso de acceso a la información no es lineal, sino que consiste en ciclos de iteración en los que se recopila, analiza y elabora información nueva, definiéndose progresivamente las necesidades del sujeto. Estas conclusiones sugieren el desarrollo de sistemas interactivos, en los que el sistema asista al usuario en la realización de la SI. Los sistemas interactivos permiten que el usuario no tenga que precisar sus necesidades previamente y pueden adaptarse a diferentes usuarios. Queda abierta la cuestión de qué tipo de interacción entre usuario y sistema es más apropiada. En las próximas secciones analizaremos diferentes tipos de interacción.

3.2. Esquemas de interacción

Denominamos esquema de interacción al conjunto de herramientas mediante las cuales el usuario interactua con el sistema a lo largo de la realización de la tarea. En el desarrollo de sistemas de acceso a la información se han incorporado esquemas de interacción, como por ejemplo, la agrupación de documentos para la identificación de documentos relevantes (tarea de Recuperación de Información) o esquemas de diálogo entre usuario y sistema para satisfacer las necesidades de información.

Para la elaboración de un modelo interactivo de SI es necesario definir el modo en que el sistema ayuda al usuario en la realización de la tarea. Es decir, es necesario fijar un modelo de interacción que permita explotar los procesos automáticos de forma que ahorren tiempo y esfuerzo al usuario.

Algunos esquemas de interacción en modelos de acceso a la información son:

1. **Diálogo.** Se trata de modelos en los que se establece una interacción por medio de un diálogo entre usuario y sistema [Chu97]. Este diálogo, emula una conversación entre el usuario y un asistente humano, de forma que el sistema ayuda a clarificar y expresar las necesidades de información.

   Este esquema requiere costosos mecanismos de tratamiento del lenguaje, dado que no solo es necesario interpretar las consultas del usuario sino que además el sistema debe elaborar nuevas consultas e interpretar las respuestas de éste. Este tipo de interacción se aplica en dominios restringidos y datos concretos, como consultas de horarios de transporte u operaciones bancarias.

2. **Control sobre parámetros del sistema.** En este esquema de interacción el usuario tiene acceso al modo de operar del sistema. Por ejemplo, selección de la estrategia de búsqueda en un sistema de Recuperación de Información,

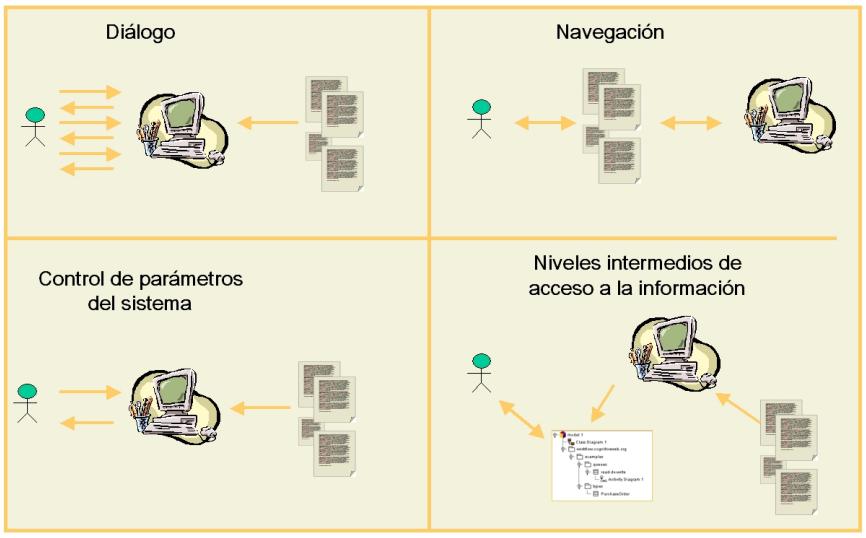


Figura 3.1: Esquemas de interacción en sistemas de acceso a la información

longitud del resumen en un sistema de Resumen Automático, tipo de fuentes a las que se quiere acceder, etc.

El inconveniente de estos esquemas de interacción es que el proceso en sí de acceso a la información no se realiza de forma interactiva sino que es ejecutado por el sistema. El usuario, en función de criterios iniciales o en función de los resultados generados en interacciones anteriores, modifica ciertos parámetros del sistema. Es decir, aunque existe interacción entre usuario y sistema, no existe colaboración durante el proceso de acceso a la información. El proceso de acceso a la información se repite sobre nuevos parámetros hasta que el usuario queda satisfecho con los resultados.

3. **Navegación.** Este esquema de interacción se ajusta a un modelo poco estructurado de acceso al información. Consiste en proporcionar al usuario enlaces entre puntos del espacio de exploración. El usuario interacciona directamente con las fuentes de información textual.

   Un caso representativo es el de los navegadores sobre internet que, a partir de los enlaces preestablecidos en el código fuente de las páginas WEB, dan la posibilidad de saltar de unos documentos a otros a lo largo de la red. Además, algunos sistemas son capaces de generar estos enlaces dinámicamente. Por ejemplo, el buscador Google permite, a partir de cualquier elemento de la lista de títulos de documentos recuperados, acceder a *"páginas similares"*. Esta funcionalidad da acceso a otros documentos que guardan relación con el primero.

4. **Niveles intermedios de acceso a la información.** En este esquema de interacción, el sistema establece pasos intermedios que ayudan al usuario a acceder a la información que busca. Un ejemplo ilustrativo es la clasifica-

ción de libros en una biblioteca. En este caso el nivel intermedio de acceso a la información está compuesto por conceptos como *"Narrativa"* o *"Poesía"*.

Análogamente, en el caso de documentos en soporte digital, se han generado jerarquías que estructuran el contenido de una colección de documentos. Por ejemplo, portales de búsqueda en la red como Altavista o Google, ofrecen una jerarquía de niveles en los que se organiza el contenido de la colección por temas como *.Economía"*, *"Arte"*, *"música"*, etc. Otros ejemplos de niveles intermedios de acceso a la información aplicados al problema de la búsqueda de documentos son la generación automática de un resumen de cada documento de la lista recuperada o la visualización de contextos de los términos de consulta en el documento (Google). En otras aproximaciones el sistema, por medio de mecanismos de extracción de terminología, sugiere al usuario términos (sintagmas nominales) relacionados con una consulta inicial, ayudándole a refinar dicha consulta [PnGV01].

Como se expondrá más adelante, también se han propuesto esquemas de interacción de este tipo aplicados a sistemas para la generación de resúmenes. Por ejemplo, sistemas que permiten profundizar en el grado de detalle sobre distintos puntos de un resumen [LLS03].

5. **Presentación de la información.** En muchos casos, una interacción básica, como la exploración de una lista de títulos en sistemas de Recuperación de Información, o la visualización del contenido completo de un documento para el acceso a fragmentos relevantes, puede optimizarse mostrando la información de forma que el usuario pueda identificar con rapidez lo que busca.

   Un ejemplo son aproximaciones que agrupan los títulos de una lista de documentos en función de las semejanzas en sus contenidos [CKPT92]. Otro ejemplo, para el caso de la visualización de documentos, el subrayado automático de términos (Google) o de fragmentos relevantes [LLS03]. Estas aproximaciones ayudan al usuario a encontrar lo que busca mostrando la misma información solo que presentada de forma estructurada.

   La visualización de la información está presente en cualquier modelo interactivo de acceso a la información. Por ejemplo, siguiendo un esquema de interacción basado en niveles intermedios, es necesario también estudiar cuál es el mejor modo de presentar dichas estructuras intermedias.

   El elemento común en todos los modelos teóricos de acceso a la información es la interacción entre el usuario y el texto original [Bel93]. En un modelo interactivo de acceso a la información textual, no se puede negar al usuario la oportunidad de acceder a un documento completo. Es necesario en este caso estudiar cuál es el mejor modo de presentar al usuario dicho documento.

3.3. Interactividad en modelos de Resumen Automático

En los proceso de evaluación de sistemas realizados en el foro DUC, se ha podido comprobar que los criterios que guían la realización de un resumen son en gran medida subjetivos. Es decir, dependen de cómo el usuario interprete del contenido de los documentos o sus necesidades de información. Es por ello por lo que se plantea la necesidad de incorporar componentes interactivos en los sistemas, [BKB+98, GMCC00] dotándoles de flexibilidad en relación a las preferencias del usuario.

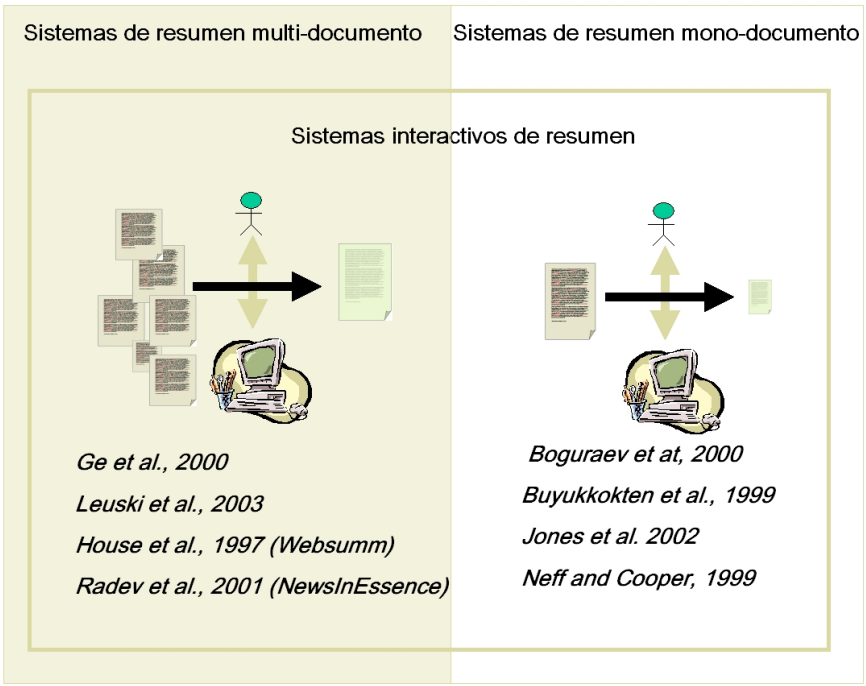


Figura 3.2: Sistemas interactivos de resumen

Existen aún pocos modelos interactivos para la generación de resúmenes y pocos experimentos y comparaciones entre distintas aproximaciones. En este apartado se describe algunas propuestas en las que el usuario interactúa con el sistema durante el proceso de elaboración del resumen.

Siguiendo la clasificación de esquemas interactivos propuesta (apartado 3.2) analizamos en este apartado algunas aproximaciones interactivas al problema del resumen:

Control sobre parámetros del sistema

Estos esquemas de interactividad consisten en el control por parte del usuario de variables que determinan el modo de trabajar del sistema. Por ejemplo, varias aproximaciones ofrecen la posibilidad de controlar:

1. **Rangos de tiempo** en las fechas de publicación de los documentos a resumir [RBGZR91, Hou97]. El usuario determina el intervalo de tiempo dentro del cuál se deben enmarcar los acontecimientos sobre los que desea realizar un resumen. El parámetro temporal es especialmente significativo en el contexto de documentos de tipo periodístico.

2. **Fuente de la que proceden los documentos** [RBGZR91]. Por ejemplo, generación de resúmenes en función de la información aportada por un conjunto de periódicos seleccionados por el usuario.

3. **Longitud del resumen** [Hou97, LLS03, JLP02]. El usuario determina el tamaño del resumen, y por lo tanto el grado de compresión que desea.

4. **Términos clave** [NC99, JLP02, LLS03]. En estas aproximaciones el resumen es generado tomando como entrada los conceptos clave seleccionados por el usuario a partir de una lista sugerida por el sistema.

Estos esquemas de interacción tienen el inconveniente de que no permiten refinar de forma parcial el resumen resultante, sino que el sistema debe realizar de nuevo el proceso completo de resumen tras cada variación de los parámetros por parte del usuario, obteniéndose un resumen nuevo en cada paso de la interacción.

Niveles intermedios de acceso a la información

Algunas aproximaciones interactivas al problema del resumen presentan una primera versión del resumen generado como nivel intermedio de acceso a la información. El usuario puede entonces profundizar en aquellos aspectos del resumen sobre los que desee obtener más información. Esto se consigue dando la oportunidad al usuario de acceder al contexto de donde ha sido extraído cualquiera de los fragmentos que componen el resumen.

En alguno de estos casos [LLS03, GMCC00], el sistema permite acceder a los documentos originales desde enlaces incluidos en un resumen multi-documento generado de forma automática. Otros sistemas que emplean un primer resumen como nivel intermedio, establecen más niveles de acceso a la información. El usuario puede profundizar en diferentes grados sobre distintos puntos del resumen inicial. Esto se puede realizar de dos maneras distintas. Una posibilidad para ofrecer varios niveles e detalle es mostrar al usuario el contexto original del fragmento incluido en el resumen mediante ventanas de texto progresivamente más amplias [BKB$^+$98]. Se muestra en un primer nivel la frase completa de donde se ha extraído un enunciado, en un segundo nivel las frases contiguas, en un tercer nivel el párrafo completo y así sucesivamente.

Otra opción es la que se ha denominado resumen ”fractal”[YW03, BGMP01]. Cualquier fragmento del resumen inicial sintetiza un bloque de texto en las fuentes originales. El usuario profundiza en un fragmento del resumen sustituyéndose éste por un nuevo resumen elaborado a partir de dicho bloque. Este esquema se

puede aplicar de forma recursiva, dado que el nuevo resumen incrustado en el resumen original contiene a su vez fragmentos que sintetizan bloques de texto más pequeños.

La extracción de listas de términos clave como nivel intermedio de acceso a la información es también una aproximación bastante habitual en sistemas interactivos de resumen [BKB+98, BGMP01, RPH+95]. En este tipo de aproximaciones el usuario accede a piezas de información organizadas por conceptos clave, refinando así progresivamente el resumen.

Todos estos modelos interactivos de resumen tienen como rasgo común la existencia de dos fases fundamentales en el proceso de acceso a la información. Es decir, la interacción entre el usuario y el sistema consiste en una serie de ciclos sobre las fases de:

Vista global de los contenidos: En este paso se obtiene una vista general de los contenidos. Esta visión global vienen dadas por los niveles intermedios de acceso a la información, como pueden ser una lista de conceptos clave o un primer resumen generado por el sistema.

Contextualización: En esta fase el usuario profundiza en cualquiera de las piezas de información que constituyen los niveles intermedios accediendo al contexto original de dichas piezas, ya sea el fragmento de texto al que pertenece o el documento del que ha sido extraído.

Presentación de la información.

En algunos casos, la información sintetizada por el sistema no se muestra con el formato de un resumen tradicional, sino que se incorporan elementos gráficos.

Por ejemplo, el sistema i-Neast [LLS03] identifica nombres de lugar en documentos de tipo periodístico y sitúa las distintas noticias sobre un mapa. Esa aproximación permite al usuario visualizar los distintos acontecimientos sobre una dimensión espacial. Por otro lado, el sistema descrito en [ABBN00] identifica y resume noticias asociadas a grupos de documentos, mostrando junto con el resumen de cada noticia, un gráfico con los documentos asociados ordenados por relevancia. En general, existen pocas aproximaciones que combinen elementos gráficos con mecanismos de Resumen Automático.

En este mismo sistema se subraya dentro de cada documento y de forma automática fragmentos relevantes. Los criterios empleados para el subrayado automático corresponden con algunas técnicas aplicadas en Resumen Automático para la selección de fragmentos relevantes. Por ejemplo, fragmentos que contengan términos relevantes, localización de los fragmentos, etc.

En general, podemos afirmar que los sistemas interactivos resumen actuales son en realidad adaptaciones de sistemas automáticos a un dominio interactivo. Los elementos de interacción con el usuario, como parámetros del sistema, resumen fractal o listas de términos clave, son elementos ya tratados en la elaboración de sistemas automáticos de resumen.

3.4. Requisitos de un esquema de interacción en Síntesis de Información

Existen varios motivos para introducir un componente interactivo en un modelo computacional de SI. Uno de ellos es, como ya hemos descrito en otros apartados, la naturaleza subjetiva de la SI. Además, como hemos podido constatar en la revisión del estado del arte en cuanto a sistemas de resumen, las técnicas actuales de tratamiento de contenidos textuales no permiten realizar un análisis en profundidad de los contenidos (sección 2.6). Por tanto, el uso de estas técnicas puede tener limitaciones al aplicarse a la SI. Estas limitaciones pueden compensarse con la implicación del usuario en el proceso.

La primera cuestión que ha de plantearse es qué esquemas de interacción entre sistema y usuario son más apropiados en el caso del acceso a la información en la SI. Un esquema de interacción en SI debería satisfacer algunas condiciones:

- **Mostrar al usuario una visión global de los contenidos.** La recopilación y el análisis de un conjunto amplio de documentos puede facilitarse en gran medida si se dispone de una vista global. Mediante estas vistas, el usuario puede centrar su atención en los aspectos que considere más relevantes. Diversos estudios con sujetos (sección 3.1) muestran que la exploración, o el análisis no exhaustivo, inicial de los contenidos es un proceso habitual en el acceso a la información textual.

- **Contextualización de los fragmentos de información mostrados al usuario.** La recopilación de piezas de información a partir de distintas fuentes requiere una contextualización de las mismas. Es decir, el sentido de un fragmento de texto puede depender del documento del que se ha extraído. Este hecho ha podido constatarse tanto en estudios con sujetos de prueba [BP02] (sección 3.1), como en el desarrollo de sistemas de resumen multidocumento (sección 2.4). Por ejemplo, en algunos sistemas de resumen se adjunta el primer fragmento del documento a los fragmentos recopilados.

- **Generación de informes extensos.** Un informe generado mediante SI puede ser extenso, conteniendo varias páginas. Por tanto, el usuario debe tener la posibilidad de construir progresivamente su informe, abordando sucesivamente diferentes aspectos del tema tratado en los documentos.

- **Independencia de dominio de los documentos originales.** La adaptabilidad a diferentes dominios es una característica deseable en un sistema de SI, o por lo menos la adaptabilidad a diferentes tipos de contenido dentro de un mismo dominio. Por ejemplo, un sistema interactivo de SI podría adaptarse a todo tipo de noticias dentro del dominio periodístico o, en el otro extremo, ser capaz de analizar, por ejemplo, únicamente noticias relativas a atentados terroristas.

- **Adaptabilidad a distintos usuarios y necesidades de información.** Diferentes usuarios pueden tener diferentes necesidades de información o incluso

diferentes formas de interpretar cómo dichas necesidades deben de ser satisfechas. El usuario debe por tanto tener la posibilidad de guiar el proceso en relación a sus necesidades de información. Es decir, la relevancia de las distintas piezas de información debe de ser un elemento flexible, no prefijado por el sistema.

3.5. Conclusiones: esquemas de interacción en Síntesis de Información

En los apartados anteriores se ha descrito distintos esquemas de interacción en sistemas de acceso a la información. No todos los esquemas satisfacen las condiciones descritas en este apartado. Por ejemplo, un esquema basado en diálogo permite en muchos casos el acceso interactivo a datos concretos. Aunque el acceso a datos concretos puede ayudar en la elaboración de un resumen, puede no ser eficiente para la elaboración de un informe extenso. Además, un esquema basado en diálogo, no aporta una vista global de los contenidos tratados en los documentos.

Otra posibilidad es el uso de esquemas de interacción basados en el control sobre parámetros del sistema. Se han elaborado sistemas interactivos de resumen en los que el usuario puede controlar parámetros como la longitud del resumen, fuentes, intervalos de tiempo, etc. En estos sistemas, el usuario ajusta progresivamente los parámetros del sistema, hasta obtener un resumen que se ajuste a sus necesidades. En cada paso, el usuario debe supervisar los resultados. Este esquema de interacción puede ser útil en la elaboración de resúmenes de poca extensión. Sin embargo, supervisar en cada ciclo de interacción un informe completo puede ser costoso para el usuario. Este esquema por tanto, no se ajusta al problema de la SI.

Los niveles intermedios de acceso a la información, al igual que la

presentación estructurada de información al usuario, son esquemas de interacción que satisfacen las condiciones descritas anteriormente. En primer lugar, pueden ofrecer al usuario una vista global de los contenidos que se pretende sintetizar. Por ejemplo, algunos sistemas interactivos de resumen proponen como nivel intermedio un primer resumen, o listas de términos representativos de los contenidos. Estos recursos pueden ser muy útiles también el contexto de la Síntesis de Información. Por otro lado, el subrayado automático como presentación de la información facilita una visión global del documento que se está visualizando.

En segundo lugar, ambos esquemas de interacción ofrecen la posibilidad al usuario de contextualizar aquellas piezas de información que considere más relevantes. Por ejemplo, los resúmenes fractales o las ventanas de texto de amplitud variable permiten situar en su contexto a las piezas de información escogidas por el usuario.

En tercer lugar, estos esquemas permiten explorar sucesivamente distintos aspectos del tema tratado en los documentos. Esta cualidad facilita la elaboración progresiva de un informe extenso. Además, estas estrategias pueden emplearse en distintos dominios y atender a distintas necesidades de información. Estas necesidades son definidas por el usuario mediante el proceso de exploración de niveles

intermedios o visualización de la información estructurada por el sistema.

En resumen, este análisis sugiere que los esquemas de interacción entre sistema y usuario más apropiados en un sistema interactivo de SI son los niveles intermedios de acceso y la presentación estructurada de la información.

Capítulo 4

Metodologías de evaluación

Para el desarrollo de un modelo de acceso a la información orientado a la SI es necesario disponer de una metodología de evaluación sobre la que comparar distintas aproximaciones. Como hemos apuntado en capítulos anteriores, la SI está estrechamente relacionada con el problema del Resumen Automático. Así mismo, las metodologías de evaluación más relacionadas con el problema de la SI, son las metodologías empleadas en sistemas de resumen. En este capítulo analizaremos las metodologías desarrolladas para la evaluación de resúmenes y estudiaremos su adecuación al problema de la SI. Como veremos a lo largo de este capítulo, las metodologías de evaluación automática aplicadas en sistemas de resúmenes poseen deficiencias que han de ser resueltas para abordar el problema de la SI (sección 4.1).

El problema de la evaluación se complica aún más al introducir mecanismos de interacción entre usuario y sistema, dado que incorporar usuarios de prueba conlleva costes. Como veremos, existen metodologías de evaluación con diferentes grados de implicación por parte de sujetos de prueba. En este capítulo analizaremos estas metodologías y, concretamente, cómo se han aplicado en la evaluación de sistemas interactivos de resumen (sección 4.2).

4.1. Evaluación de sistemas de resumen automático

En la tarea de resumen, el proceso de evaluación acarrea dificultades que adicionales respecto a otras tareas como la Recuperación de Documentos o la Búsqueda de Respuestas. En [Man01b] se describen varios aspectos que convierten el problema de la evaluación de resúmenes en un asunto difícil de tratar. Algunos de estos son:

1. No hay un único resumen ideal. Existe la posibilidad de que se pueda generar un buen resumen diferente de los resúmenes considerados como modelos de referencia.

2. Emplear juicios emitidos por humanos para evaluar la calidad de un resumen incrementa el coste de evaluación.

3. Dependiendo del grado de compresión de la información en el resumen, unos modelos de resumen se pueden comportar mejor que otros [JBME98].

4. La tarea a la que está destinada un resumen influye en gran medida en los criterios de calidad de éste.

En este apartado describiremos algunas estrategias de evaluación mediante las cuales se ha abordado estos problemas. La figura 4.1 muestra una clasificación de los distintos métodos para la evaluación de resúmenes.

Figura 4.1: Clasificación de los métodos de evaluación de resúmenes

En un primer nivel, se puede clasificar las estrategias de evaluación de resúmenes en dos grupos: métodos intrínsecos y métodos extrínsecos. Los métodos intrínsecos se basan en cualidades del resumen independientes de la función que éste realiza. Por ejemplo, algunas de estas cualidades son la coherencia o cantidad de información que aporta *"informativeness"*. Los métodos extrínsecos se apoyan en la utilidad del resumen para la realización de otras tareas evaluables de forma objetiva. En los próximos apartados analizaremos diferentes métodos de evaluación. A su vez, los criterios intrínsecos se pueden clasificar en criterios de coherencia y criterios basados en la información preservada en el resumen. En el primer caso se evalúa características de forma o rasgos superficiales del resumen, mientras que en el segundo caso la evaluación se centra en la cobertura de contenidos. Por último, dentro de los métodos intrínsecos basados en cobertura de contenidos puede distinguirse aquellos casos en los que se toma como referencia los contenidos de las fuentes originales de los casos en los que se toma como referencia una serie de resúmenes modelo generados manualmente.

En los próximos apartados describiremos distintos tipos de evaluación: intrínseca basada en la coherencia, intrínseca basada en cobertura de contenidos sobre las fuentes, sobre informes modelo, y evaluación intrínseca. Algunos de ellos requie-

ren el uso de jueces humanos, mientras que otros se basan en métricas de evaluación automática.

4.1.1. Métodos basados en la coherencia del resumen

La calidad de un texto desde el punto de vista de la coherencia depende de aspectos como la aparición de anáforas sin resolver, aparición de vacíos en la estructura retórica, calidad de la gramática, legibilidad etc.

Resulta complicado captar de forma automática estos rasgos en un resumen, por lo que el papel de los métodos de evaluación automática está muy limitado y es necesario contar con juicios humanos. La calidad de la coherencia de un resumen tiene un alto grado de subjetividad, dependiendo de las consideraciones de los evaluadores. Sin embargo, en foros de evaluación, como es el caso del DUC (Document Understanding Conferences) [BDH+00], se ha comprobado que existe por lo general acuerdo entre los evaluadores en lo que respecta a este tipo de criterios.

4.1.2. Métodos basados en cobertura de contenidos respecto a las fuentes

Existen varios métodos de evaluación de resúmenes basados en la cantidad de información contenida en las fuentes originales que queda preservada en el resumen. En este caso el propio proceso de evaluación requiere un trabajo de síntesis, por lo que es necesaria la intervención de jueces humanos. La cuestión que se plantea es cómo establecer un protocolo que guíe el proceso de evaluación, de forma que los resultados generados por distintos evaluadores sean comparables.

La estrategia llevada a cabo en el TIPSTER SUMMAC para la evaluación de resúmenes orientados a consulta [MHK+98] consistió en ofrecer a los evaluadores los resúmenes generados por los sistemas junto con las fuentes originales. Los documentos originales fueron previamente etiquetados marcándose aquellas frases relacionadas directamente con la pregunta a la que está dirigida el resumen. Para la realización de este etiquetado de fragmentos relevantes, se descompuso el tema central en una serie de cuestiones que debían de ser cubiertas por el resumen. Por ejemplo, para el tema *"Saturación en las cárceles"*, los resúmenes deben cubrir aspectos como *"¿Cuáles son los nombres de las cárceles con exceso de presos?"*. Se marcó las frases que daban respuesta a cada una de estas preguntas. De esta forma, los evaluadores podían visualizar en los documentos originales qué aspectos deben estar incluidos en los resúmenes. Disponiendo de los documentos originales marcados, los evaluadores debían de asignar una puntuación a cada resumen. Esta puntuación cubría los valores *"correcta", "parcialmente correcta" e "incorrecta"*.

Una aproximación más sencilla, aplicada en [SL00], consistió en suministrar a los evaluadores una lista de conceptos clave que debían de ser tratados por los resúmenes. Los evaluadores debían asociar las listas de conceptos con los fragmentos contenidos en los resúmenes.

En algún caso se ha empleado métodos semi-automáticos para evaluar en qué medida el resumen preserva la información contenida en las fuentes sin introducir

redundancias. Es posible, aunque costoso, etiquetar cada uno de los fragmentos contenidos en los textos originales a resumir [RHB00]. En este marco de evaluación, se etiqueta cada uno de ellos en virtud de su susceptibilidad para pertenecer al resumen. Se puede entonces, de forma automática, calibrar la calidad de un resumen en función de los fragmentos escogidos por éste. Es necesario también anotar las relaciones de subsumción entre fragmentos de los textos originales para evaluar la aparición de redundancias en los resúmenes. Además de costoso, la anotación de frases relevantes solo es aplicable a resúmenes de tipo extractivo. Por otro lado, la asignación de pesos a fragmentos es un modo poco natural de elaborar un resumen, por lo que el etiquetado manual no es sencillo de realizar.

4.1.3. Métodos basados en cobertura de contenidos respecto a resúmenes de referencia

En estos métodos se evalúa el resumen en función de su semejanza a resúmenes modelo de referencia. Estos resúmenes modelo son generados manualmente en condiciones controladas. La cuestión es medir la cantidad de información presente en los resúmenes modelo que queda preservada en los resúmenes evaluados.

Este tipo de evaluación puede llevarse a cabo por medio de juicios humanos. Por ejemplo, en el caso del foro de evaluación DUC [BDH$^+$00], se presenta a una serie de jueces humanos un conjunto de resúmenes modelo frente a resúmenes a evaluar. Mediante la herramienta SEE [HL02b], los jueces supervisan la proporción de fragmentos de texto presentes en los modelos que están representados en los resúmenes evaluados.

Por otro lado, es posible aplicar métricas de evaluación automáticas basadas en la semejanza entre el resumen evaluado y el conjunto de resúmenes modelo. La principal ventaja que ofrece este tipo de métricas es que no requieren juicios humanos, sino únicamente disponer de resúmenes modelo. Por tanto, el proceso de evaluación de nuevas aproximaciones no implica un coste adicional. Estas metodologías permiten comparar múltiples variaciones de un sistema en la fase de desarrollo.

Criterios de similitud

Frente a estas ventajas, para emplear esta estrategia es necesario definir un criterio de similitud automatizable entre resúmenes. Algunos de los criterios de similitud empleados en distintos estudios son:

1. **Co-selección de frases en los resúmenes**. Dos resúmenes serán más similares cuanto más coincidan en las frases seleccionadas para la elaboración de los mismos. Este criterio se puede aplicar con facilidad en resúmenes de tipo extractivo. Sin embargo, cuando se trata de resúmenes en los que se abstrae información, es decir, en los que se genera nuevos fragmentos que sintetizan la información contenida en los textos originales, es necesario aplicar mecanismos de alineamiento entre fragmentos del resumen y fragmentos del texto original [KPC95, JBME98, MG01]. La aplicación de estas estrategias

de alineación entre fragmentos puede ser útil también en la comparación de resúmenes extractivos, dado que, dos fragmentos de texto extraídos de las fuentes originales, aun distintos, pueden contener la misma información.

La co-ocurrencia entre frases contenidas en dos textos se puede medir en términos de precisión (P)y cobertura (C) en donde:

$$P = \frac{A \cap B}{A} \quad C = \frac{A \cap B}{B}$$

siendo A el conjunto de frases que aparecen en el resumen evaluado y B el conjunto de frases que aparecen en el modelo.

Sin embargo, estas medidas no tienen en cuenta el grado de acuerdo entre dos textos causado por fenómenos aleatorios. Para ello, se ha desarrollado una métrica denominada Kappa [Sei88]. Esta métrica se establece como:

$$K = \frac{Pr - P_a}{1 - P_a}$$

En donde P_r representa la precisión de un resumen en relación al resumen de referencia y Pa representa la precisión de un resumen generado mediante la selección aleatoria de fragmentos respecto al resumen de referencia. Se ha podido comprobar [ROQT03] que los resultados ofrecidos por la métrica kappa se corresponden en mayor medida con los criterios de semejanza establecidos por evaluadores.

2. **Similitud del vocabulario empleado**. Esta estrategia se puede aplicar tanto a resúmenes de tipo extractivo como a resúmenes en los que se abstrae información. Otra de las ventajas de esta métrica es que la similitud entre dos textos no depende necesariamente del número de frases contenidas en ambas. Una limitación de este tipo de métricas es que no tiene en cuenta diferencias en los contenidos expresadas mediante negaciones, u orden entre las palabras. Sin embargo, en los trabajos realizados en SUMMAC [MHK$^+$98], se pudo comprobar que existía una correlación muy fuerte entre los juicios de los evaluadores y la alineación basada en vocabulario entre los resúmenes y los modelos.

   Existen varias métricas para determinar la similitud de textos en función de los términos que contienen. Algunas de ellas se basan en la co-ocurrencia de n-gramas (ROUGE) [LH03a] o la función coseno [Sal89]. Existen múltiples variantes de ROUGE. Algunas de ellas se basan en cobertura o precisión de n-gramas de 1,2,3 o 4 elementos. Otra variante, por ejemplo, se basa en secuencias no contiguas de términos (ROUGE-W).

3. **Representación semántica de los fragmentos contenidos en los resúmenes.** En la aproximación propuesta en [HT03] las frases contenidas en los resúmenes se descomponen en enunciados a los que denominan ”Factoids”. Por ejemplo, la frase:

"La policía arresto a un hombre blanco de origen holandés"

se descompone en:

- F1: *"La policía realizó un arresto."*
- F2: *"Un sospechoso fue arrestado."*
- F3: *"El sospechoso era hombre."*
- F4: *"El sospechoso era blanco."*
- F5: *"El sospechoso era de origen holandés."*

Se pudo comprobar que la co-ocurrencia de "factoids" es más consistente que métricas basadas en vocabulario, concretamente, la co-ocurrencia de unigramas empleadas en el DUC (métrica ROUGE-1).

En los trabajos realizados por Nenkova y Passonneau [NP04], se define una métrica de evaluación basada en SCUs (Summarization Contents Units). Una SCU representa el significado de una pieza de texto no mayor que una proposición. Estas unidades de información son anotadas manualmente, junto con relaciones de subsumción. Además, se asigna un peso a cada SCU dependiendo del número de resúmenes modelo que contienen dicha SCU. Los resultados del trabajo muestran que existe una alta correlación entre esta métrica de evaluación y ROUGE basado en unigramas.

La representación semántica de los contenidos de un resumen no es completamente automatizable, dado que requiere la anotación manual de estas unidades de representación. Si bien también es cierto que este proceso de anotación es una tarea de extracción de información que sí podría ser en un futuro automatizable. Además, estos criterios de evaluación son objetivos, es decir, no están sujetos a desacuerdos entre jueces humanos.

Acuerdo entre resumidores humanos

Como hemos apuntado, las métricas basadas en semejanza a un conjunto de resúmenes modelo poseen la cualidad de ser automatizables, por ejemplo, mediante medidas de similitud basadas en el vocabulario empleado o en la co-selección de frases. Sin embargo, es necesario validar la fiabilidad de dichas métricas. Estas medidas de similitud se aplican por lo general sobre más de un resumen modelo. Se asume por tanto que existe un cierto grado de acuerdo entre resumidores humanos. El grado de acuerdo ha sido estudiado en relación a diferentes criterios de similitud:

- **Co-selección de frases:** En [JBME98] se estudia la co-selección de frases como un rasgo común en los resúmenes generados por resumidores humanos. El solapamiento obtenido es alto (96 %) en relación a resultados obtenidos en otros estudios: 71 % en [Mar97], 46 % en [SSMB97] o 25 % en [RRS61]. Estas diferencias se deben a la estructura de los documentos originales. Esto es, el solapamiento de frases es dependiente de las características de los documentos a partir de los cuales los resúmenes son generados.

- **Solapamiento en el vocabulario empleado:** El estudio realizado por Copeck [CS04] incluye un análisis del solapamiento de vocabulario entre resúmenes generados por humanos. El solapamiento es medido en términos de palabras, y en términos de sintagmas nominales seleccionados manualmente. Se obtuvo un solapamiento del 9 % en palabras y 22 % en sintagmas nominales.
- **Solapamiento de unidades semánticas:** En [HT03] se realizó un estudio análogo esta vez sobre ”factoids”. El estudio muestra que no existe un grado de acuerdo total entre resumidores humanos. Sin embargo, ciertos “factoids” son especialmente frecuentes en los resúmenes. Además, pudieron comprobar que son necesarios entre 30 y 40 resúmenes mono-documento de un mismo texto para obtener un corpus de referencia estable.Esto supone un alto coste asociado a metodologías de evaluación basadas en juicios humanos.

Fiabilidad de las métricas

Una vez que se asume un grado de acuerdo entre resúmenes modelo, es necesario comprobar la fiabilidad de las métricas de evaluación. Es decir, si son mejores aquellos sistemas que obtienen una mejor puntuación según la métrica de evaluación. Este problema puede abordarse desde distintas perspectivas:

- **Midiendo la correlación existente entre los valores dados por distintas métricas de evaluación** [Cou03]. Esta perspectiva ayuda a entender qué métricas de evaluación aportan nueva información respecto a otras.
- **Midiendo la correlación entre los juicios generados por métricas de evaluación y los juicios generados por evaluadores humanos** [TM03, LH03a]. Esta es la estrategia más común para verificar la fiabilidad de una métrica de evaluación automática. El cálculo de la correlación se establece mediante el coeficiente Pearson, que mide la correlación lineal, o bien mediante el coeficiente Spearman, que se basa en el ordenamiento por calidad de los sistemas.

  Sin embargo esta estrategia asume que los juicios generados por evaluadores humanos son objetivos, mientras que en realidad, dependen del protocolo y las directrices dadas a los evaluadores. Además, si los sistemas son mejorados a lo largo del tiempo, no podemos asegurar que las métricas de evaluación obtengan la misma correlación que la que obtuvieron con juicios humanos sobre los sistemas antecesores.
- **Comparando la calidad de los sistemas con la calidad de los modelos de referencia** [CR03, LH03b]. Esta estrategia parte del supuesto de que los modelos son de una calidad superior a la de los sistemas que queremos evaluar. Es decir, este tipo de procesos de meta-evaluación nos indican si las métricas de evaluación son capaces de caracterizar los rasgos propios de los resúmenes modelo, distinguiéndolos de resúmenes generados de forma automática.

  Esta perspectiva solventa los problemas planteados en el punto anterior, dado que no requiere la elaboración de juicios humanos, y los criterios de validación de las métricas son objetivos. Por otro lado, al no considerar juicios

humanos, es imposible tratar aspectos concretos de carácter subjetivo, como la cobertura o la fluidez de un resumen.

El modelo más representativo de este tipo de estrategias es ORANGE, definido originalmente sobre la tarea de traducción automática [Lin04a]. ORANGE establece la calidad de una métrica de evaluación como la posición promedio de una traducción modelo en un ranking que incluya tanto traducciones modelo como traducciones generadas por sistemas. Existen sin embargo algunos aspectos aún no cubiertos por el modelo ORANGE. Por ejemplo, en ORANGE la calidad de una métrica de evaluación es sensible a elementos repetidos en el conjunto de sistemas considerados. Es decir, si introducimos dos veces el mismo sistema en el ranking, la posición que ocupa es resumen modelo se ve afectada. Además ORANGE no ofrece ningún criterio para medir la representatividad del corpus de resúmenes automáticos. Dicho de otro modo, no considera la heterogeneidad de los sistemas de resumen. Por último, tampoco ofrece la posibilidad de combinar distintas métricas de evaluación.

En definitiva, la dificultad de generar buenas métricas intrínsecas de evaluación de resúmenes parte del hecho de que un resumen es un objeto elaborado y complejo, por lo que la evaluación mediante juicios humanos es costosa. Por otro lado, la evaluación mediante métricas automáticas es compleja, dado que no existe un único resumen modelo, ni un único criterio de semejanza entre un resumen y el modelo. Aunque hemos visto que existen algunos criterios de meta-evaluación de métricas automáticas, como ORANGE, no aparece en el estado del arte un marco de evaluación que permita aplicar, evaluar y combinar diferentes métricas automáticas de evaluación.

4.1.4. Métodos extrínsecos

Los métodos extrínsecos para la evaluación de resúmenes se apoyan en la eficiencia con que se realiza una determinada tarea en la que están implicados dichos resúmenes. Por ejemplo, en qué medida ayuda un resumen a un usuario a identificar documentos relevantes en relación a una necesidad de información o a contestar una pregunta determinada. En muchos casos el problema de la evaluación se simplifica, dado que resulta más sencillo evaluar tareas como recuperación de documentos o elaboración de respuestas que evaluar directamente la calidad de un resumen.

Una de estas estrategias extrínsecas consiste en que sujetos de prueba respondan a un conjunto de cuestiones relacionadas con el documento original, teniendo en mano el resumen que se pretende evaluar. Por ejemplo, en la aproximación propuesta en [MKA92] se escogieron una serie de ejercicios de lectura y comprensión. Una serie de sujetos respondían a estas preguntas disponiendo únicamente de los resúmenes evaluados. Estas respuestas se compararon con otras elaboradas a partir de los textos originales. En la aproximación propuesta para la evaluación del sistema SUMGEN [May95], las preguntas a las que debían responder los sujetos de

prueba consistieron en datos específicos del dominio, como el nombre, los participantes o la duración de misiones en documentos que describen batallas simuladas. El problema de estas estrategias de evaluación es que los criterios de calidad dependen del tipo de preguntas planteadas para el proceso de evaluación.

Otra aproximación consiste en medir el esfuerzo empleado en reeditar un resumen generado automáticamente al transformarlo en un informe aceptable en el contexto de una determinada tarea, como por ejemplo, la reconstrucción de la información esencial de un documento [HM98]. Aquí aparece un problema análogo al anterior, dado que los criterios de calidad de los resúmenes dependerán del tipo de tarea para la cuál el sujeto de prueba reedita el texto contenido en el resumen.

La generación del resumen puede estar embebida dentro de otro sistema. Por ejemplo, un sistema de recuperación automática de documentos puede indexar la información a partir de resúmenes, o también, un sistema de Búsqueda de Respuestas puede resumir las fuentes para después aplicar un procesamiento más fino de extracción de la respuesta. En estos casos, se puede comparar los resultados del sistema aplicando distintas estrategias en la fase de resumen, aunque siempre dentro del contexto de un sistema determinado. Por ejemplo, dependiendo de la aproximación empleada en la recuperación de documentos será más eficiente un determinado tipo de resúmenes.

En cualquier caso, los métodos extrínsecos son muy útiles cuando se quiere evaluar la calidad de los resúmenes en función de una tarea de usuario o sistema en el que está embebido.

4.2. Evaluación de sistemas interactivos

En esta sección analizamos las estrategias de evaluación aplicadas en sistemas interactivos de acceso a la información.

4.2.1. Tipos de evaluación

Existen distintas perspectivas desde la que evaluar un sistema interactivo de acceso a la información. En [Sch03] se identifican distintos aproximaciones para la evaluación de este tipo de sistemas:

Evaluación sobre córpora: (Scientific metrics) Este tipo de métricas son aplicadas en base a corpus de referencia, sin emplear usuarios de prueba. Un ejemplo son las evaluaciones de sistemas de RI sobre una colección de documentos candidatos, consultas y documentos relevantes asociados a dichas consultas. El sistema de RI se ejecuta con la consulta inicial como único elemento representativo de la interacción con el usuario. Otro ejemplo es la tasa de respuestas correctas en un sistema de búsqueda de respuestas o de diálogo partiendo de preguntas definidas a priori. La mayoría de estas métricas de evaluación se basan en medidas de precisión y cobertura.

Evaluación mediante sujetos de prueba: (Component metrics) Estas métricas son específicas para cada funcionalidad del sistema, y se apoyan en evalua-

ciones con usuarios de prueba en interacción con el prototipo evaluado. Por ejemplo, en el caso de un sistema de diálogo algunas de estas métricas pueden ser la calidad de la información obtenida por el usuario, o el número de interacciones necesarias para obtener dicha información. Por ejemplo. en un sistema interactivo de recuperación de documentos, una métrica de este tipo podría ser el tiempo empleado o la precisión y cobertura de la lista de documentos relevantes identificados por el usuario.

Evaluación en un entorno real: (Impact metrics) En estos casos se evalúa en qué medida el sistema facilita el trabajo al usuario en un entorno real. Algunos criterios de evaluación pueden ser la confianza en el sistema, el ahorro de tiempo en la realización de las tareas de usuario, la fiabilidad del sistema, etc.

La principal limitación de la evaluación en entornos reales (impact metrics) radica en el coste, dado que es necesario la implementación de un sistema acabado, y se requiere mucho tiempo para analizar el impacto del sistema en el entorno de usuario. Este inconveniente se resuelve mediante la evaluación con sujetos de prueba en condiciones controladas ("component metrics"), dado que el hecho de realizar experimentos en condiciones controladas permite reducir la cantidad necesaria de sujetos de prueba, reduciendo también el tiempo necesario para el proceso de evaluación.

Frente a esto, la evaluación en condiciones controladas ("Component metrics") no contempla aspectos como la usabilidad o la satisfacción del usuario [HT00], que sólo se podrían medir en un contexto real. Es decir, resulta complicado extrapolar los resultados obtenidos en laboratorio a casos reales. Para solventar este problema la metodología PARADISE [WLKA97] propone, para el caso de sistemas basados en diálogo, una técnica de estimación de la satisfacción del usuario a partir de resultados obtenidos mediante "Component Metrics", analizando la correlación entre ambos tipos de evaluación.

Tanto las evaluaciones en entornos reales como la evaluación mediante sujetos de prueba en alboratorio (impact metrics y component metrics) sufren una serie de limitaciones derivadas del uso de sujetos reales:

1. **Diferencias entre distintos usuarios utilizados en la evaluación influyen notoriamente en los resultados del proceso** [Hea99]. Cada usuario puede tener su propia interpretación de las necesidades de información. A modo de ejemplo, durante el proceso de generación de corpora para sistemas de recuperación de documetos se ha podido comprobar que el grado de acuerdo entre diferentes anotadores en relación a qué documentos son o no son relevantes no supera un 70 % [VH00], aun disponiendo de tiempo ilimitado y accediendo al contenido completo de cada uno de los documentos potencialmente relevantes. Por lo tanto, las consideraciones de cada individuo pueden ser distintas y afectar igualmente a cualquier proceso interactivo de acceso a la información.

2. **Los experimentos no son replicables** [Bel02]. Un mismo usuario no puede repetir bajo las mismas condiciones un proceso de acceso a la información, dado que en este proceso el usuario adquiere conocimiento. Por ejemplo, en el caso de la recuperación de documentos, un usuario que busque información sobre *"la producción de plátanos en Canarias"*, no lo hará de la misma forma por segunda vez, dado que ya sabrá qué empresas exportan el producto, o qué elementos como el turismo, reservas de agua etc., afectan a la producción. Además, el usuario habrá adquirido habilidad en el proceso de búsqueda, sea cual sea la necesidad de información. Dado que los resultados dependen del usuario, estos experimentos no son, por lo tanto, repetibles.

3. **Las diferencias entre resultados obtenidos mediante distintas aproximaciones son pequeñas**[HO02]. En general, los usuarios son capaces de adaptarse a las herramientas de las que dispone. Es decir, en muchos casos los sujetos de prueba son capaces de suplir deficiencias de los sistemas, por lo que las variaciones entre resultados generados mediante distintas herramientas no son tan visibles como en el caso de los sistemas automáticos.

Estos tres inconvenientes se resumen en un alto coste del proceso de evaluación con sujetos reales. En definitiva, en un primer estadio del desarrollo de sistemas es necesario disponer de métricas de evaluación que no requieran la implicación de usuarios reales (scientific metrics).

4.2.2. Evaluación de sistemas interactivos de resumen

En el caso de los modelos interactivos de resumen, se han realizado escasos trabajos en relación al problema de la evaluación.

En [BGMP01] se aplica una metodología extrínseca, concretándose los objetivos del usuario en forma de Búsqueda de Respuestas, al estilo de la tarea interactiva definida en el TREC 2002. En este modelo el usuario profundiza en las piezas de información mostradas por el sistema, accediendo a fragmentos de texto progresivamente más amplios. El sistema es evaluado en función del número de acciones que el usuario debe realizar hasta llegar a la respuesta de la pregunta planteada.

Por otro lado, se han llevado a cabo también evaluaciones intrínsecas, pero únicamente sobre sistemas interactivos de resumen basados en el control de parámetros del sistema. En estos casos se evalúa el sistema fijando valores en los parámetros del sistema. En el caso de [JLP02], en donde el usuario controla los parámetros de resumidor, como el tema o la longitud del resumen se evalúa la calidad de los resúmenes generados de forma automática en base a resúmenes modelo. Algo semejante ocurre en el caso de [YW03], en donde se evalúan los resúmenes generados con parámetros fijos en el sistema mediante cuestionarios a sujetos de prueba y medidas de precisión sobre oraciones relevantes.

Sin embargo, que sepamos, no se ha profundizado en la evaluación de aspectos interactivos en sistemas de resumen basados en niveles intermedios de acceso a la información, dado que el uso de usuarios reales es costoso y resulta complicado

predecir el comportamiento de un usuario real a la hora de explorar estos niveles intermedios.

4.3. Conclusiones: metodologías de evaluación y Síntesis de Información

La evaluación de sistemas interactivos en el contexto de la SI presenta una doble dificultad. En primer lugar, aparecen las dificultades heredadas de la evaluación intrínseca de resúmenes. En segundo lugar, hereda también las dificultades de la evaluación de sistemas interactivos de acceso a la información, es decir, cómo evaluar la interacción entre usuario y sistema. La cuestión que nos planteamos en este punto es qué tipo de marco de evaluación es más apropiado para el estudio de la SI considerando estos dos aspectos.

Evaluación y caracterización de informes

Como hemos visto en esta revisión del estado del arte, las metodologías de evaluación se clasifican en primer lugar en métodos intrínsecos y extrínsecos. Los métodos extrínsecos evalúan los resúmenes en función de una tarea concreta. En nuestro caso, no hemos orientado los informes elaborados en ISCORPUS a ninguna tarea en particular, y nos centraremos en métodos intrínsecos.

Dentro de los métodos intrínsecos se distinguen los métodos basados en la coherencia de los métodos basados en los contenidos. En nuestro caso, los informes generados son de tipo extractivo, por lo que la coherencia dentro de una frase queda asegurada por la propia coherencia de los documentos originales. La coherencia entre frases está relacionada con la fase de presentación, mientras que en esta monografía nos centraremos en la fase de recopilación de información. En definitiva, nos centraremos en metodologías intrínsecas de evaluación basadas en contenidos.

Por último, existen métodos que evalúan contenidos en función de los documentos originales y métodos que evalúan en función de resúmenes modelo. En el segundo caso, hemos visto que es posible automatizar el proceso de evaluación. Esta automatización es clave en nuestro trabajo, dado que necesitamos poder comparar múltiples aproximaciones sin estar sujetos al coste asociado al uso de jueces humanos. En definitiva, necesitamos una metodología de evaluación automática, intrínseca, y basada en contenidos preservados en el informe evaluado en relación a informes modelo.

Para ello, hemos de disponer de informes modelo y medidas de similitud automátizables entre el informe evaluado y los informes modelo. Esto presenta principalmente dos problemas. Por un lado, como hemos visto en este capítulo, existen múltiples criterios de similitud entre dos resúmenes, y análogamente, entre dos informes. Por otro lado, dada la subjetividad la tarea de SI, existen múltiples informes modelo sobre los que evaluar un informe.

En base a esto, consideramos dos características deseables para un marco de evaluación que no están presentes en las metodologías descritas en este capítulo. En

primer lugar, es posible que diferentes métricas de similitud posean características complementarias. El marco de evaluación debería permitir combinar entre sí diferentes criterios de similitud. De cada criterio de similitud nos interesa en qué rasgos del informe se basa. No nos interesa en cambio la topología de la medida de similitud, es decir, su escala. En definitiva, sería deseable poder combinar y validar combinaciones de métricas de similitud con independencia de las propiedades de escala de cada métrica.

El segundo aspecto es que la subjetividad de la tarea de SI provoca que existan múltiples soluciones posibles. Es decir, existe múltiples informes modelo para una misma necesidad de información. Las métricas automáticas de evaluación existentes realizan algún tipo de promediado para considerar múltiples modelos. Los resultados son por tanto, sensibles a las propiedades de escala de dichas métricas. Este problema no ha sido aún resuelto.

Habitualmente, la fiabilidad de las métricas de similitud se estima mediante la correlación con juicios humanos (sección 4.1.3). Sin embargo, como hemos apuntado, los juicios humanos son costosos, especialmente en el caso de la SI. El marco de evaluación ORANGE aborda este problema, eliminando la necesidad de juicios humanos bajo la suposición de que la calidad de los modelos es siempre superior a la calidad de los elementos evaluados. Sin embargo, la medida ORANGE de meta-evaluación se ve afectada si las muestras de elementos automáticos está sesgada.

En definitiva, necesitamos un marco de evaluación automática basado en cobertura de contenidos sobre informes modelo que permita combinar métricas de similitud y considerar diferentes modelos sin verse afectado por las propiedades de escala. Necesitamos también una medida de meta-evaluación que no requiera juicios humanos y que resuelva el problema de los sesgos en las muestras de elementos automáticos. El marco QARLA propuesto posee estas cualidades (capítulo 6).

Evaluación de la interacción en un sistema de SI

La evaluación más realista de un sistema interactivo es sin duda el análisis del comportamiento del sistema y de la satisfacción de los usuarios en un entorno real (“impact metrics”). El problema de esta evaluación es que no permite comparar entre sí distintas aproximaciones. Esto es posible en cambio mediante experimentos controlados en laboratorio con usuarios de prueba (“component metrics”). En este caso se evalúan aspectos concretos del sistema y es posible comparar distintas aproximaciones entre sí. Sin embargo, el número de aproximaciones comparadas está limitado por el coste asociado al uso de sujetos de prueba. Este hecho se acentúa en el caso de la Síntesis de Información, en donde la elaboración de un informe mediante una herramienta interactiva requiere un esfuerzo considerable por parte del usuario.

Una tercera opción es la evaluación sin usuarios de prueba (“scientific metrics”). Mediante este tipo de métricas no podemos evaluar aspectos como la amabilidad o la usabilidad del sistema. Es decir, no podemos evaluar directamente la interacción con el usuario, aunque sí otros aspectos del sistema. La principal ven-

taja de este tipo de evaluación es que podemos comparar múltiples aproximaciones sin coste adicional. Este tipo de métricas de evaluación es especialmente apropiado para una primera fase en el estudio de aproximaciones interactivas al problema de la Síntesis de Información.

Sin embargo, no existe en el estado del arte una metodología de evaluación intrínseca aplicable a sistemas interactivos de resumen basados en niveles intermedios de acceso a la información (ver sección 4.2.2). Es esta monografía propondremos una generalización de QARLA para dominios interactivos que se ajusta a estas características.

Parte II

Marco de evaluación

Capítulo 5

ISCORPUS: un corpus de Síntesis de Información

Hemos definido la tarea de Síntesis de Información como el proceso consistente en generar un informe en base a una necesidad de información expresada en forma de consulta, a partir de un conjunto de documentos relevantes y relacionados entre sí (capítulo 1). Existen múltiples muestras de este tipo de informes, dado que se trata de un problema bastante común en diversos dominios (sección 1.1). Sin embargo, no se ha elaborado hasta el momento (2004), un corpus controlado de informes generados por un conjunto de personas en las mismas condiciones, sobre las mismas fuentes y ante las mismas necesidades de información. Disponer de un corpus de estas características es esencial para la realización de estudios relativos al acceso a la información orientada a SI. Por ello, hemos desarrollado ISCORPUS, un corpus en español que contiene 74 informes generados por 9 personas a partir de 8 conjuntos de documentos periodísticos asociados a distintas necesidades de información. En este capítulo se describen los criterios sobre los que se han escogido los temas tratados en los informes, los documentos originales, y cómo se ha diseñado las instrucciones dadas a los sujetos de prueba y la herramienta empleada para la elaboración de informes.

En este libro, nos centramos en el papel que los conceptos clave juegan en el proceso de SI. Consideramos como conceptos clave en el contexto periodístico al conjunto de personas, organizaciones o factores significativos que aparecen en los documentos originales y que guardan relación con las necesidades de información propuestas. ISCORPUS incluye de conceptos clave extraídos manualmente por los mismos autores de los informes.

Finalmente, la monitorización de las acciones realizadas por los sujetos de prueba durante la elaboración de los informes y los cuestionarios cumplimentados nos han permitido extraer conclusiones sobre la naturaleza de la SI. En líneas generales, veremos que la SI puede entenderse con un proceso fundamentalmente extractivo, al menos, en un primer estadio. Veremos también que existen notables diferencias dependiendo del tipo de tema tratado, estableciéndose dos categorias: temas que tratan un mismo asunto que evoluciona a lo largo del tiempo y temas que tratan instancias de un mismo tipo de evento.

5.1. Selección de temas

Un primer paso en la elaboración del corpus es la selección de conjuntos de documentos y necesidades de información sobre las que generar los informes. A este respecto, el corpus debe poseer ciertas características propias de la Síntesis de Información. Entendiendo que la Síntesis de Información es un paso que sigue a la recuperación de documentos, parece natural partir de temas propuestos en el área de recuperación de documentos. En el contexto de la evaluación de sistemas de recuperación de documentos, un tema determina una necesidad de información que debe ser cubierta por lo documentos recuperados por el sistema. El tema viene expresado en forma de consulta. Algunos foros de evaluación de sistemas de recuperación de documentos son TREC[1], CLEF[2] o NTCIR[3]. En estos foros, el tema viene representado por un título, una descripción breve y una descripción detallada. La siguiente tabla muestra un ejemplo de tema en la campaña de evaluación CLEF.

Título	Campañas europeas contra el racismo
Descripción breve	Encontrar documentos que hablen sobre campañas contra el racismo en Europa
Descripción detallada	Los documentos relevantes describen campañas informativas o educativas contra el racismo étnico, religioso o hacia inmigrantes) en Europa. Deben hacer referencia a campañas organizadas, y no a meras opiniones contra el racismo.

Estos temas son complejos y están bien definidos, por lo que poseen las características necesarias para la elaboración de ISCORPUS. Estas descripciones pueden reorientarse hacia la elaboración de un informe sustituyendo frases como "*Encontrar documentos que...*" por frases como "*Elaborar un informe que describa...*". Para la elaboración de ISCORPUS hemos considerado una colección de artículos periodísticos escritos en español aportados por la agencia EFE, y empleados en los foros de evaluación CLEF 2001-2003.

El corpus del CLEF contiene, para cada necesidad de información, un conjunto de documentos identificados manualmente como relevantes en relación a la necesidad de información. Con el fin de disponer de suficiente cantidad de documentos sobre los que generar los informes, hemos escogido los ocho temas con mayor cantidad de documentos relevantes asociados. Todos los conjuntos seleccionados contienen más de 100 documentos relevantes. Por homogeneidad, hemos restringido a los primeros 100 documentos todos los conjuntos, ordenados cronológicamente.

La siguiente lista muestra el conjunto de temas sobre los que se elaboran los informes en ISCORPUS. Su definición se basa en la reescritura de la descripción detallada de los 8 temas seleccionados a partir del corpus CLEF.

TEMA TT1 En 1994 algunas de las naciones europeas que participaban en la misión de paz en Bosnia quisieron retirar sus tropas. Obtener información

[1] http://trec.nist.gov
[2] http://www.clef-campaign.org
[3] http://research.nii.ac.jp/ntcir/

sobre las razones para proponer esa retirada.

TEMA TT2 Generar un resumen con la información más importante en relación a la invasión de Haití por los soldados de la ONU y de los EEUU, tanto acerca de la discusión sobre la decisión de la ONU de enviar tropas americanas a Haití, como la invasión misma. Se hablará también de sus consecuencias directas.

TEMA TT3 Razones y causas subyacentes de la intervención de las tropas rusas en Chechenia. Declaraciones de políticos rusos, incluido el presidente Yeltsin que justifican el envío de las tropas rusas a Chechenia.

TEMA TT4 Información más relevante en relación al levantamiento de campesinos indígenas en Chiapas (Méjico). Motivos y desarrollo de la rebelión de la población indígena en Chiapas. Reacciones del gobierno mexicano.

TEMA TT5 Negociadores del tratado de paz en el Medio Oriente entre Israel y Jordania. Información más relevante sobre el tratado.

TEMA TT6 Durante el conflicto entre los Hutus y los Tutsis, Francia inició la operación Turquesa en el suroeste de Ruanda con el fin de proporcionar ayuda humanitaria a la población. Obtener información sobre esta operación.

TEMA IE1 Campañas informativas o educativas contra el racismo (étnico, religioso o hacia inmigrantes) en Europa. Se considerarán campañas organizadas y no meras opiniones sobre el racismo.

TEMA IE2 Información relacionada con huelgas de hambre organizadas con fin de atraer la atención hacia una causa. Identificar casos en que una huelga de hambre haya sido convocada, incluido el motivo de la huelga y el resultado, si se conoce.

Dentro de este conjunto de 8 temas podemos distinguir dos grupos. En los seis primeros, es necesario estudiar cómo evoluciona el asunto a lo largo del tiempo. Por ejemplo, "La invasión de Haití, por parte de los EEUU". Este tipo de necesidades de información ha sido definidas como resumen "mono-evento" [MBE+01]. El autor describe esta categoría como "los documentos giran en torno a un único asunto, ubicado en un lugar y espacio de tiempo, incluyendo los mismo agentes y acciones". En este trabajo las identificamos como TT, dado que el conjunto de documentos podría corresponderse con la salida de un sistema de "Topic Tracking" [YCG+99], es decir, identificación de documentos que describen la evolución de un mismo asunto.

En las otras dos necesidades de información (IE1 e IE2), el sujeto debe identificar instancias de un mismo tipo de evento, como son las campañas antiracistas o las huelgas de hambre. En [MBE+01] se define esta categoría como resumen "multi-evento". El autor la describe como "varios eventos que transcurren en diferentes lugares y espacios de tiempo y generalmente con diferentes protagonistas". En nuestro trabajo las identificamos como IE, por tratarse de un problema que en

cierta medida se aproxima a la Extracción de Información, en cuanto que el sujeto debe extraer y recopilar piezas de información que atienden a un mismo patrón.

Un informe mono-evento requiere un tratamiento más elaborado de la información contenida en las fuentes, por lo que parece más interesante desde el punto de vista del estudio de la Síntesis de Información. Sin embargo, hemos incluido en el corpus las necesidades tipo IE, dado que nos pueden ofrecer información que permite contrastar ambos tipos de problemas.

5.2. Generación de informes

Un informe requiere una mayor cantidad de espacio que un resumen de los evaluados en otros estudios (sección 2.3). Por ello, hemos establecido un máximo de 50 frases por informe, es decir, un promedio de una frase por cada dos documentos originales. Este límite satisface varias condiciones: es lo suficientemente amplio como para albergar la información esencial del tema tratado, requiere un esfuerzo de comprensión considerable por parte del sujeto, y queda descartada la estrategia de seleccionar directamente el primer párrafo de cada documento, dado que se dispone del doble de documentos que el número máximo de frases permitidas.

Se decidió que la elaboración del informe consistiría en una tarea de tipo extractivo, consistente en la selección de frases de los documentos originales. Obviamente, un proceso real de síntesis incluye re-escritura del informe generado. Sin embargo, simplificar el problema a un trabajo de recopilación implica varias ventajas. En primer lugar, facilita el análisis de los informes y la evaluación de futuros informes generados de forma automática. En segundo lugar, reduce el esfuerzo requerido por los sujetos para la elaboración de los informes.

Para la elaboración de informes se ha empleado nueve sujetos de entre 25 y 35 años de edad, todos ellos con estudios superiores y experiencia como usuarios en sistemas de recuperación de documentos. Previamente se ha descrito en detalle en qué consiste la tarea de Síntesis de Información a los sujetos, con el fin de minimizar divergencias en la interpretación de la misma. En primer lugar se les indicó que debían generar un informe con un máximo de 50 frases en relación a cada uno de los temas propuestos. Este proceso se realizaría mediante la recopilación de frases pertenecientes al conjunto de documentos originales asociados a cada tema. El tiempo límite para la elaboración de cada informe fue de 30 minutos, que es suficiente para la elaboración del informe y evita fatigas que puedan distorsionar las características del corpus.

Por otro lado, se les indicó que en un plazo de seis meses se les examinaría mediante una serie de cuestiones acerca de cada uno de los temas propuestos. En ese momento solo dispondrían del informe generado. Los sujetos serían premiados en virtud de la calidad de las respuestas. Estas instrucciones tienen como objetivo mejorar los resultados estableciendo criterios de competitividad. Además, los sujetos explotarían en la medida de lo posible el espacio disponible para la elaboración del informe. De hecho, la gran mayoría de informes generados contienen un número muy cercano a las 50 frases.

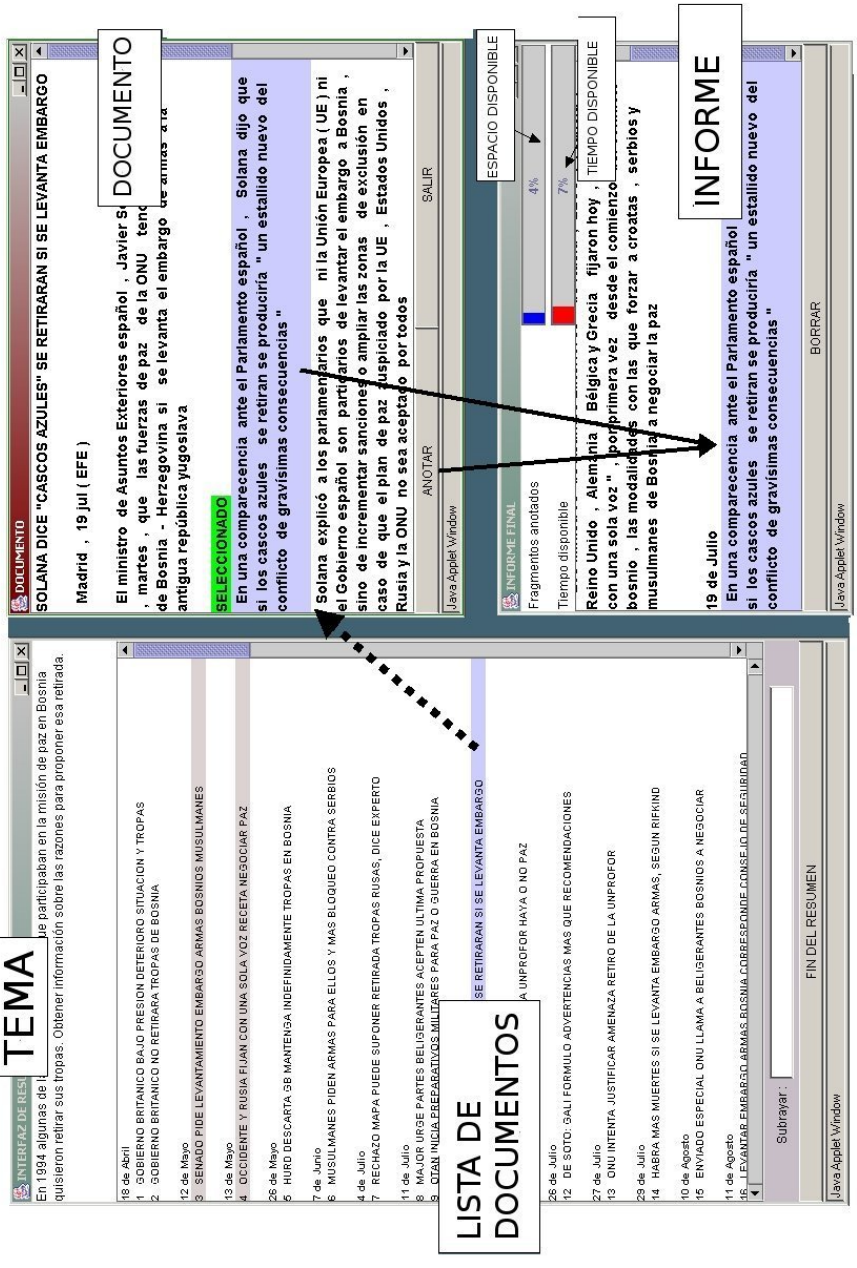


Figura 5.1: Interfaz para la realización manual de informes en ISCORPUS

Los sujetos han realizado la tarea por medio de un interfaz (figura 5.1) que permite la elaboración de informes de tipo extractivo sin demasiado esfuerzo. El sistema muestra, en primer lugar, la lista de títulos de documentos asociados a la necesidad de información (marco izquierdo en la figura). Pinchando en cualquiera de los títulos, el sistema muestra el contenido del mismo (marco derecho superior en la figura). En este marco, el usuario puede seleccionar una frase del documento original mediante un simple "clic" sobre el documento. La frase seleccionada se añade automáticamente en el informe final (marco inferior derecho en la figura), organizándose en orden cronológico según la fecha de edición del documento al que pertenecen. Durante los 30 minutos el sujeto tiene la posibilidad de superar el número máximo de frases permitidas en el informe. Una vez terminado este intervalo de tiempo, el sujeto debe eliminar frases hasta obtener un informe de 50 frases. El tiempo y espacio disponible se muestran en el interfaz mediante dos barras (marco inferior derecho en la figura).

5.3. Cuestionarios

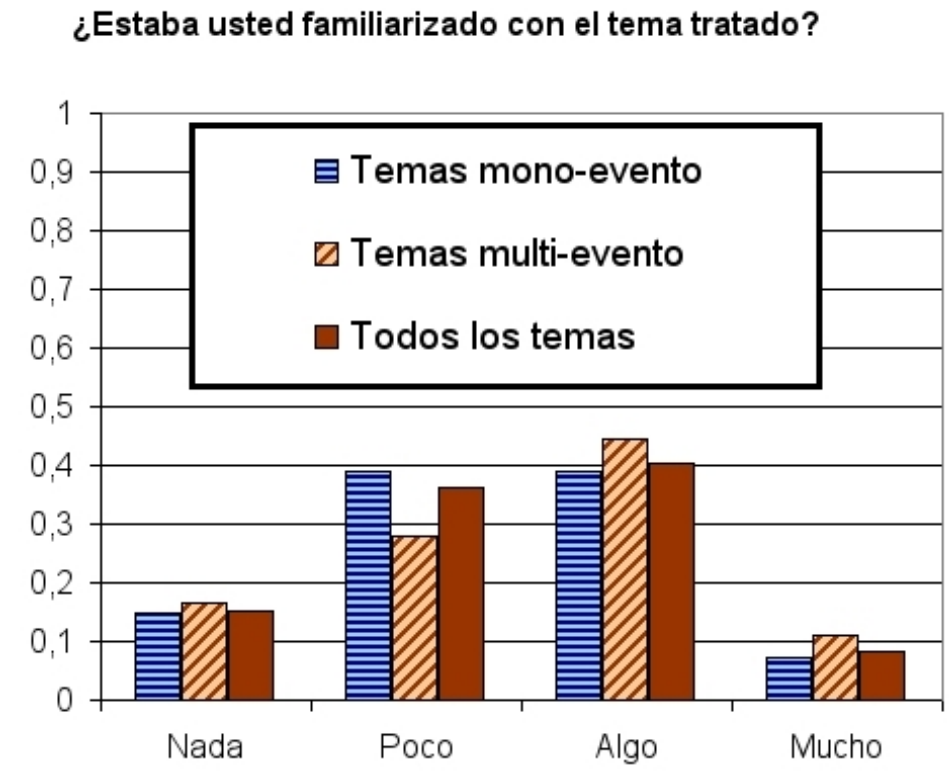
Una vez finalizado el proceso de elaboración del informe el sujeto debía de completar el siguiente formulario:

Q1 ¿Quienes son las principales personas implicadas en el asunto?
Q2 ¿Cuáles son las principales organizaciones implicadas en el asunto?
Q3 ¿Cuáles son los principales factores implicados en el asunto?
Q4 ¿Estaba usted familiarizado con el tema tratado?
Q5 ¿Considera que ha sido costosa la generación del informe?
Q6 ¿Ha sentido usted la necesidad de introducir anotaciones o reescribir el informe generado?
Q7 ¿Considera que ha elaborado un buen informe?
Q8 ¿Se encuentra usted cansado?

Las respuestas a las tres primeras cuestiones son empleadas para la anotación de conceptos clave en el corpus. En cuanto al resto de cuestiones, se ofrece al usuario la posibilidad de elegir entre cuatro posibilidades: *"nada"*, *"poco"*, *"algo"* o *"mucho"*.

Las figuras 5.2 y 5.3 muestran la proporción de cada tipo de respuesta (cuestiones Q4 a Q7) en temas mono-evento, multi-evento y en ambos. En primer lugar, en el 73 % de los casos el sujeto se sintió satisfecho o muy satisfecho con el informe realizado (Q7).

Otro aspecto a resaltar es que los sujetos consideraron más costosa (Q5) la tarea de SI sobre temas mono-evento que sobre temas multi-evento. Por ejemplo, las respuestas *"nada"* y *"poco"* son más frecuentes en temas mono-evento que en multi-evento, mientras que para las respuestas *"algo"* o *"mucho"* sucede lo contrario. Este dato sugiere que los temas mono-evento requieren un proceso más costoso de selección y análisis de información.



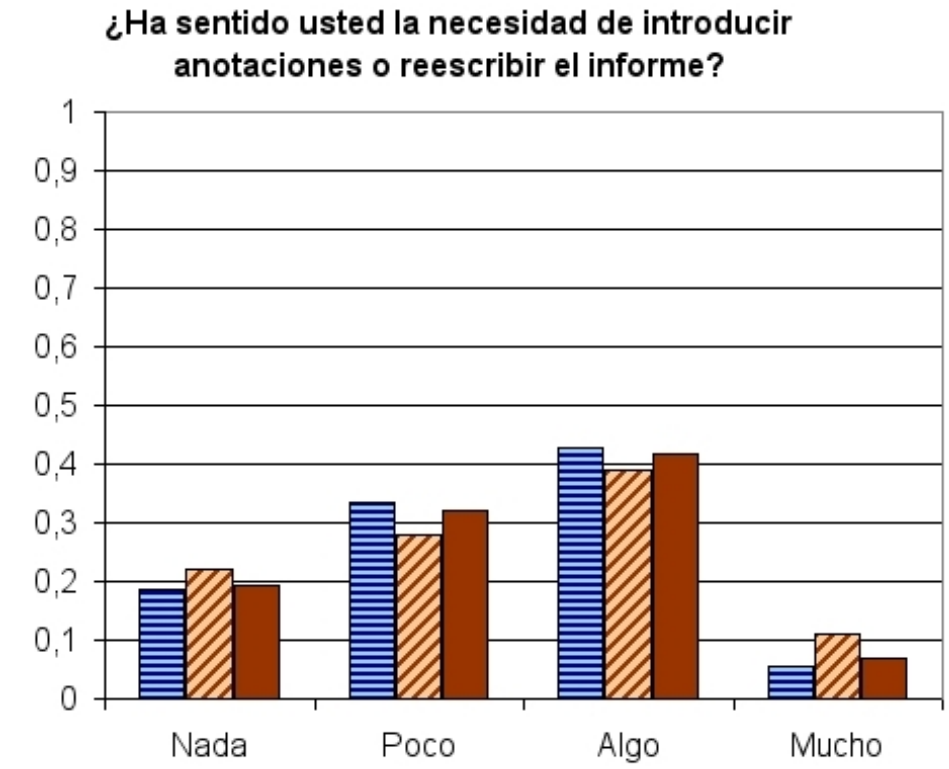


Figura 5.2: Resultados de los cuestionarios presentados a sujetos de prueba en IS-CORPUS (I)

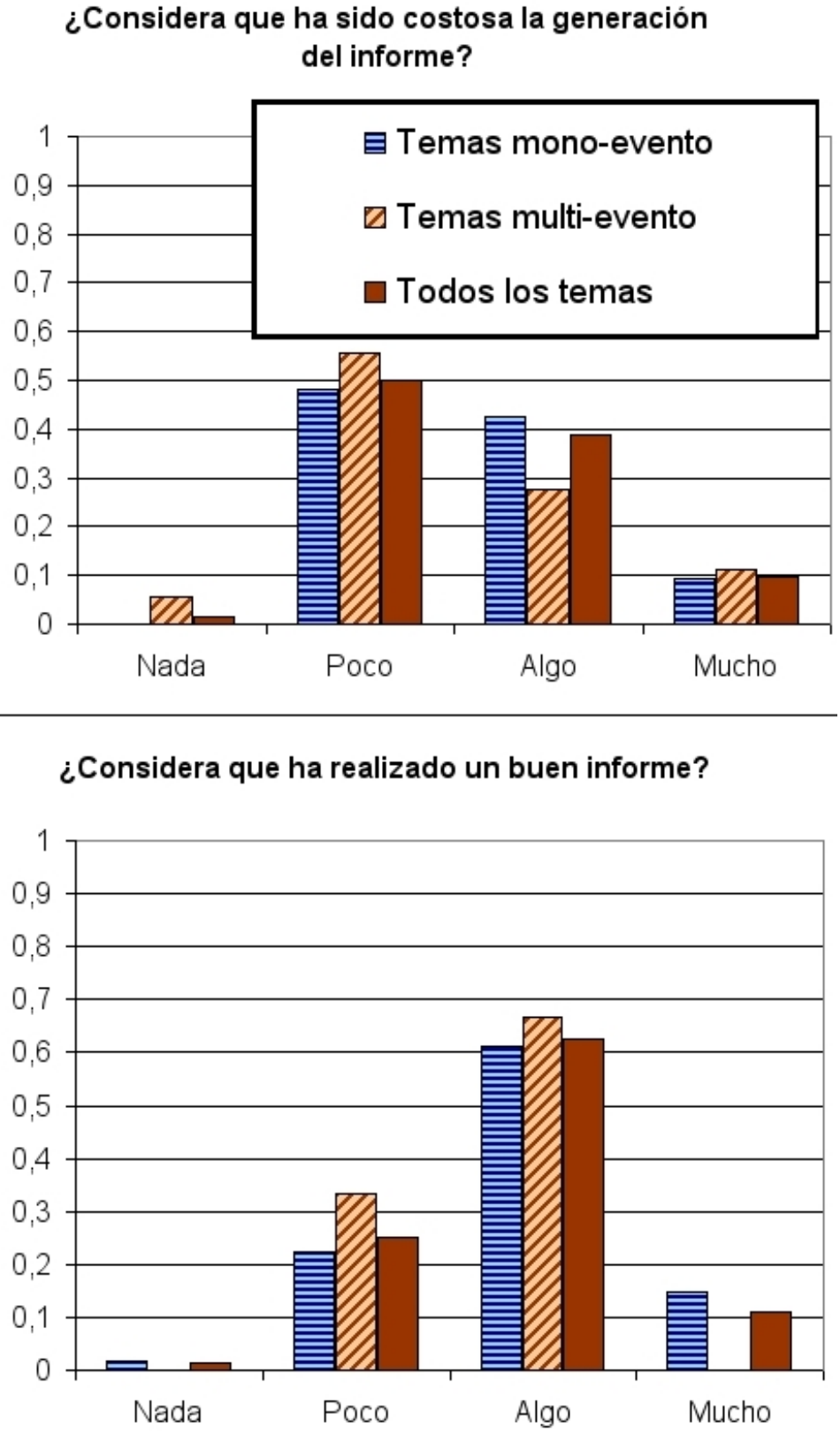


Figura 5.3: Resultados de los cuestionarios presentados a sujetos de prueba en ISCORPUS (II)

Además, parece que los sujetos se sienten más familiarizados (Q4) con temas multi-evento que con temas mono-evento. Este dato parece sorprendente si tenemos en cuenta que los temas mono-evento tratan asuntos de gran impacto como son el conflicto entre palestinos e israelíes, la invasión de Haití, el conflicto de Chiapas o la guerra en Bosnia. Sin embargo, los sujetos afirman sentirse más familiarizados con huelgas de hambre en general o campañas antiracistas. Una posible explicación a este hecho es que en el caso de los temas multi-evento, los sujetos se han visto menos obligados a profundizar en el tema tratado.

Por último, otro aspecto remarcable es que solo en un 6 % de los casos el sujeto sintió la necesidad (*"mucho"*) de reescribir o introducir anotaciones en su informe (Q6). Además, en más de la mitad de los casos los sujetos respondieron con *"nada"* o *"poco"* a esta cuestión. Dado que a los sujetos sólo se les permitió extraer frases completas de los documentos originales, este resultado sugiere que la extracción de frases puede ser una estrategia apropiada en el proceso de Síntesis de Información, o cuanto menos, un primer estadio en el proceso de elaboración del informe.

5.4. Anotación de conceptos clave

Las tres primeras cuestiones del cuestionario descrito en el punto anterior tienen como objetivo la anotación de los conceptos clave que aparecen en los documentos originales. Definimos conceptos clave como el conjunto de personas, factores y organizaciones relevantes que participan en el asunto descrito en los documentos originales.

En respuesta a las preguntas planteadas en los cuestionarios, los sujetos anotaron conceptos libremente en una serie de cajas de texto. Se estableció un máximo de 8 cajas por pregunta, es decir, un máximo de 24 (8x3) conceptos clave introducidos por cada sujeto. En el momento de introducir los conceptos clave, los sujetos podían visualizar el informe generado previamente. Ésta es, por ejemplo, la respuesta de un sujeto en relación al tema "la invasión de Haití por parte de EEUU" (tema TT2).

Gente	**Organizaciones**	**Factores**
Jean Bertrand Aristide	ONU	militares golpistas
Clinton	EEUU	golpe militar
Raoul Cedras	OEA	restaurar la democracia
Philippe Biambi		
Michel Josep Francois		

Finalmente, para cada tema hemos generado una única lista a partir de las respuestas de los sujetos. Con el fin de evitar redundancias, agrupamos aquellos conceptos clave que, en el contexto del tema tratado, se pueden entender como conceptos equivalentes. Por ejemplo, la siguiente tabla muestra la agrupación realizada para el tema TT2:

Concepto clave	Conceptos equivalentes
intervención	decisión, invasión
EEUU	Gobierno de los EEUU, Estados Unidos
Haití	Gobierno de Haití, Gobierno de facto de Haití
acuerdo	acuerdos
expulsión	exilio
Brasil	Gobierno brasileño
Argentina	Gobierno de Argentina
democracia	restauración de la democracia

La figura 5.4 muestra el número de conceptos clave (eje vertical) anotados por cierta cantidad de sujetos (eje horizontal). Como puede verse en las gráficas, existe una clara diferencia entre temas mono-evento y temas multi-evento. En temas mono-evento, especialmente en lo que respecta a personas y organizaciones, existe un conjunto considerable de conceptos anotados simultáneamente por muchos de los sujetos. Es decir, existe un grado de acuerdo entre anotadores relativamente alto. Esto es debido a que en temas multi-evento no necesariamente están presentes los mismos agentes en los distintos documentos por sintetizar. Este resultado muestra que posiblemente los conceptos clave jueguen un papel relevante en el caso de los temas mono-evento.

5.5. Generación de informes automáticos

Para analizar las características de la SI y desarrollar estrategias automáticas o semi-automáticas, es necesario disponer de un conjunto de aproximaciones automáticas básicas de referencia. Es decir, aquellas que queremos mejorar. Mediante estas estrategias de referencia obtendremos un conjunto de informes automáticos que representa las capacidades de un sistema de síntesis sencillo. Para cada uno de los temas incluidos en ISCORPUS, hemos desarrollado 30 informes. Para ello, hemos combinado diferentes criterios estándar de selección de frases:

- *Localización de la frase en el documento:* La selección de primeras frases de cada documento es un criterio empleado comúnmente en la generación automática de resúmenes multi-documento. En este trabajo, los criterios basados en localización que han sido empleados son: (1) primera frase del documento, (2) primera y segunda frase del documento, (3) primera, segunda o tercera frase del documento y (4) extracción de todas las frases del documento. En nuestro caso, la longitud límite de los informes es de 50 frases, y se dispone de 100 documentos originales por tema. Por tanto, es necesario realizar algún proceso de filtrado de documentos. Por ello, hemos considerado diferentes series en la selección de documentos ordenados cronológicamente, como los primeros documentos, los últimos documentos, documentos alternados, etc.

- *Relevancia del documento* La selección de documentos para la extracción de frases se basa en este caso en la relevancia del documento dada la necesidad

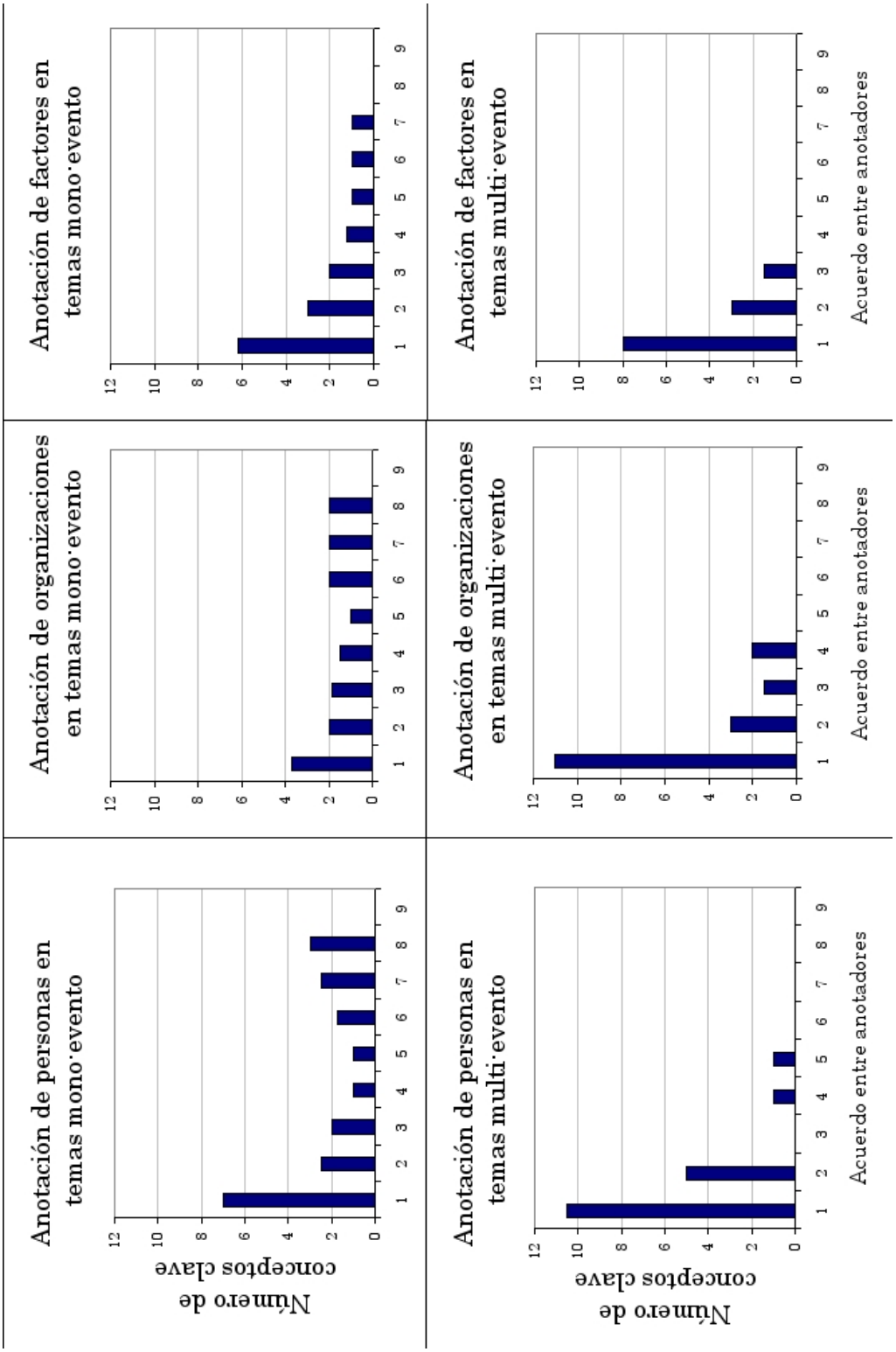


Figura 5.4: Estadísticas de conceptos clave anotados en ISCORPUS

de información planteada. La relevancia de cada documento ha sido calculada mediante el motor de búsqueda INQUERY [ACS+98], introduciendole como entradas el corpus completo de noticias y la necesidad de información tal y como le fue presentada a los sujetos.

- *Fecha del documento.* Podemos seleccionar únicamente un documento por cada fecha de edición. En el dominio periodístico es habitual que artículos escritos el mismo día contengan información redundante.

La siguiente tabla muestra el conjunto de aproximaciones automáticas de referencia, que combinan los criterios de selección de frases antes descritos:

Aproximación	**Estrategia de selección**
Aproximación 1	Primera frase de los primeros 50 documentos
Aproximación 2	Primera frase de los últimos 50 documentos
Aproximación 3	Primera frase del primer documento de cada día
Aproximación 4	Primera frase del último documento de cada día
Aproximación 5	Los primeros documentos completos
Aproximación 6	Los últimos documentos completos
Aproximación 7	Los documentos más relevantes completos
Aproximación 8	Primeras cinco frases de los documentos más relevantes
Aproximación 9	Primeras dos frases de los documentos más relevantes
Aproximaciones 10..22	Primera frase de distintas series de documentos
Aproximaciones 23..26	Primera y segunda frase de distintas series de documentos
Aproximaciones 27..32	Tres primeras frases de distintas series de documentos

Las aproximaciones 3 y 4 consideran la fecha de edición de los documentos en el proceso de selección de frases. Las aproximaciones 7, 8 y 9 apoyan en la identificación de documentos relevantes. El resto de aproximaciones son combinaciones del número de frases seleccionadas por documento y series de documentos tomando como criterio de selección la localización de la frase en su documento.

5.6. Análisis del comportamiento de los sujetos

El interfaz empleado en la realización de los informes incluye una monitorización de trazas de tiempo, es decir, instantes en los que el sujeto se introduce en un documento, sale del documento o selecciona un fragmento para ser incluido en el informe. Analizando estas trazas, hemos podido identificar algunos aspectos interesantes.

Tiempo empleado en visualizar documentos

La figura 5.5 muestra el tiempo medio empleado desde que el sujeto se introduce en un documento hasta que anota un fragmento (gráfica superior) o sale del do-

cumento (gráfica inferior). En general, tanto el tiempo de selección de una frase como de visualización del documento es menor en el caso de los temas multi-evento. Por ejemplo, el 55 % de las frases son seleccionadas en menos de 10 segundos en el caso de temas multi-evento, mientras que en el caso de los temas mono-evento, esto ocurre en un 42 % de los casos. Cerca del 18 % de los documentos descartados son examinados en menos de 5 segundos en temas multi-evento, mientras que esta cifra desciende al 5 % en el caso de los temas mono-evento. Es decir, en general en temas mono-evento los sujetos invierten más tiempo en el análisis de los documentos. Por otro lado, se confirman las observaciones de [SOC$^+$02] en el sentido de que los sujetos prestan más atención a unos documentos frente a otros.

Proporción de documentos visualizados

Otro aspecto interesante es el porcentaje de documentos que el sujeto necesita visualizar para generar su informe. La gráfica de la figura 5.6 muestra estos porcentajes. Como puede verse, en el caso de los temas tipo multi-evento (IE1 e IE2), son visualizados una mayor cantidad de documentos (58 % y 70 %), mientras que en el caso de los temas tipo mono-evento, en promedio sólo son visualizados un 45 % de los documentos. Esto implica que los sujetos han podido descartar muchos documentos simplemente por su título. Parece lógico pensar que en el caso de los temas mono-evento, existe un mayor número de documentos con información redundante.

Solapamiento de frases y documentos anotados

La figura 5.7 muestra para cada tema el solapamiento promedio de frases entre informes. Como puede verse, el solapamiento entre frases está en torno al 35 %. Este grado de solapamiento está cercano a los valores mínimos obtenidos en otros estudios sobre otras tareas de resumen (sección 4.1.3). Este hecho se debe al fuerte coeficiente de reducción de los informes, en torno a un 15 %, y la presencia de información redundante en las fuentes. Por otro lado, existe una notable diferencia entre los temas mono-evento (TT) y temas multi-evento (IE). En los temas mono-evento, en los que los sujetos invierten más tiempo en el análisis de los documentos, el solapamiento es siempre menor que en los temas multi-evento. Algo análogo ocurre si observamos el solapamiento entre documentos a partir de los cuales los distintos sujetos han extraído las frases (figura 5.8). Este solapamiento sigue siendo relativamente pequeño, y con claras diferencias entre ambos tipos de temas.

Localización de las frases seleccionadas

La figura 5.9 muestra la frecuencia con que son seleccionadas frases con distinta posición en su documento. Como puede verse en todos los temas son seleccionadas con mucha más frecuencia frases que aparecen en la primera o segunda posición del documento. Esta observación confirma el peso de la localización del fragmento como criterio de selección en la elaboración de resúmenes. La predominancia de primeros o segundos fragmentos de los documentos es más acentuada en

el caso de los temas multi-evento. Esto es coherente con el hecho de que en temas mono-evento los sujetos profundizan en los contenidos de algunos documentos en más ocasiones que en temas multi-evento.

Recopilación de información

La figura 5.10 muestra el número de fragmentos seleccionados por los sujetos a lo largo del proceso de SI. Hemos dividido en diez partes (eje horizontal) el tiempo total desde el inicio del proceso hasta el última vez que el sujeto selecciona un fragmento. Como puede verse, el número de fragmentos seleccionados se mantiene constante a lo largo del proceso. Este resultado sugiere que la fase de recopilación de información está embebida dentro de todo el proceso de SI.

Lectura de contenidos

La figura 5.11 muestra, sobre la misma partición de tiempos que en el proceso anterior, el tiempo promedio en segundos dedicado a la lectura de un documento. La figura muestra una tendencia descendente. Este resultado sugiere que los sujetos dedican más esfuerzo a la lectura de contenidos en la primera fase del proceso. Este resultado refleja la transición entre las etapas de lectura y análisis de contenidos.

5.7. Caracterización de la Síntesis de Información

Con los datos obtenidos a lo largo de la elaboración y análisis de ISCORPUS, podemos extraer algunas conclusiones en relación al problema de la Síntesis de Información:

SI como un proceso extractivo

Los resultados sugieren que es posible generar un informe autocontenido sin necesidad de redición del mismo. En muy pocos casos los sujetos consideraron la necesidad de introducir anotaciones o reescribir los informes generados, aun sabiendo que debían hacer uso del informe unos meses más tarde. Es decir, aunque evidentemente la generación de un informe real requiere reescritura, ésta podría ser realizada tomando un informe primitivo generado mediante recopilación de frases. Además, los resultados muestran que la recopilación de información se distribuye a lo largo de todo el proceso de SI. Es decir, no hay una fase de análisis o de lectura de contenidos independiente del proceso de recopilación de piezas de información. En definitiva, el proceso de SI, en el contexto de ISCORPUS, es fundamentalmente extractivo. Por tanto, parece sensato plantear un primer estadio de la resolución del problema de la Síntesis de Información como la generación de un informe extractivo.

Por otro lado, a pesar de que los sujetos seleccionan la misma cantidad de fragmentos a lo largo de todo el proceso de SI, los resultados muestran que, a medida que avanza el proceso, decrece el esfuerzo empleado en la lectura de contenidos,

es decir, el tiempo empleado en la visualización de documentos. Este resultado sugiere una relación entre la lectura de documentos completos y las fases iniciales en las que el sujeto asimila información sobre el tema en general.

Temas mono-evento frente a temas multi-evento

En cuanto a la anotación de conceptos clave, existe un alto grado de acuerdo en temas mono-evento, en donde los documentos de partida tratan la evolución de un asunto (mono-evento). Esta información es una primera evidencia sobre la utilidad de considerar conceptos clave al abordar el problema de la Síntesis de Información, especialmente en temas tipo mono-evento.

Además, considerando el comportamiento de los sujetos de prueba, la Síntesis de Información sobre temas tipo multi-evento tiene un grado de complejidad menor, es decir, requiere un análisis menos profundo del contenido de los documentos. En temas multi-evento se emplea menos tiempo en el análisis del contenido de los documentos, se seleccionan con mayor frecuencia los primeros fragmentos del documento y hay un mayor solapamiento entre informes generados por distintos sujetos. Además, dado el número de documentos visualizados en cada tipo de temas, podemos decir en en temas tipo mono-evento existe más información redundante.

El siguiente paso en este trabajo consiste en evaluar distintas aproximaciones automáticas o interactivas para la Síntesis de Información. Sin embargo, para ello es necesario un marco de evaluación fiable que permita evaluar y analizar los rasgos característicos de los informes generados en ISCORPUS. En el capítulo siguiente se describe un nuevo marco de evaluación QARLA orientado a tareas con alto grado de subjetividad, como es el caso de la Síntesis de Información.

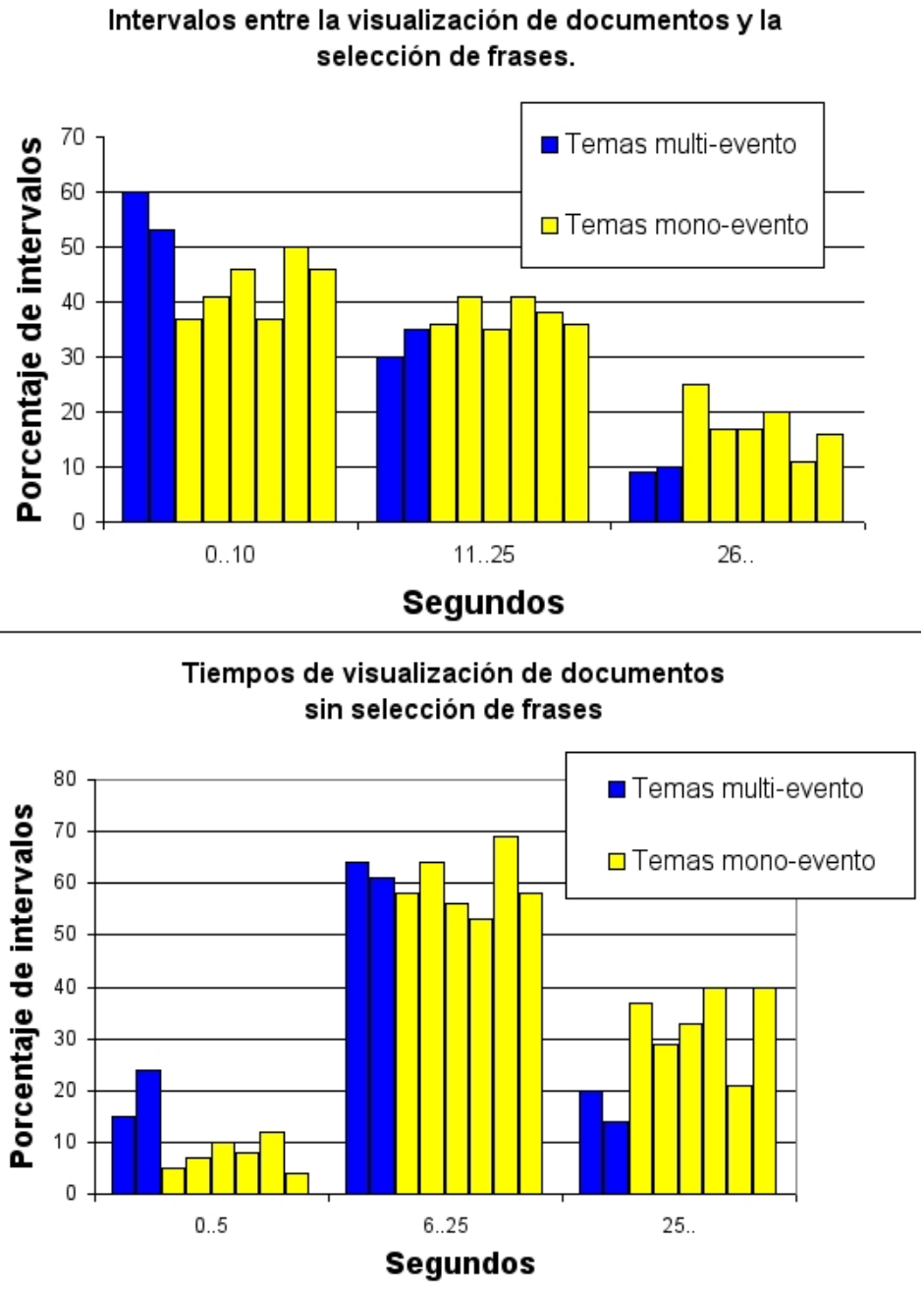


Figura 5.5: Trazas de tiempo en la generación de informes de ISCORPUS

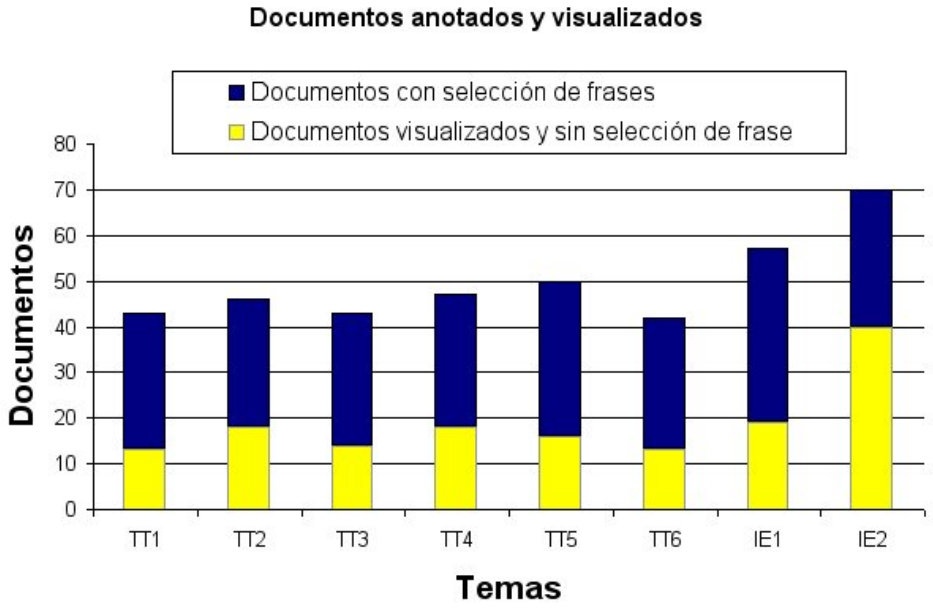


Figura 5.6: Porcentajes de documentos visualizados y anotados en ISCORPUS

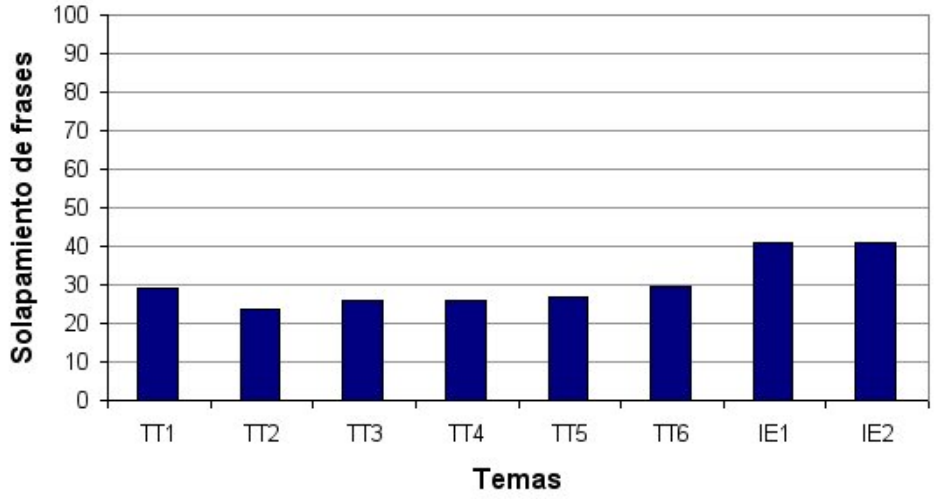


Figura 5.7: Porcentajes de solapamiento de frases en los informes generados en ISCORPUS

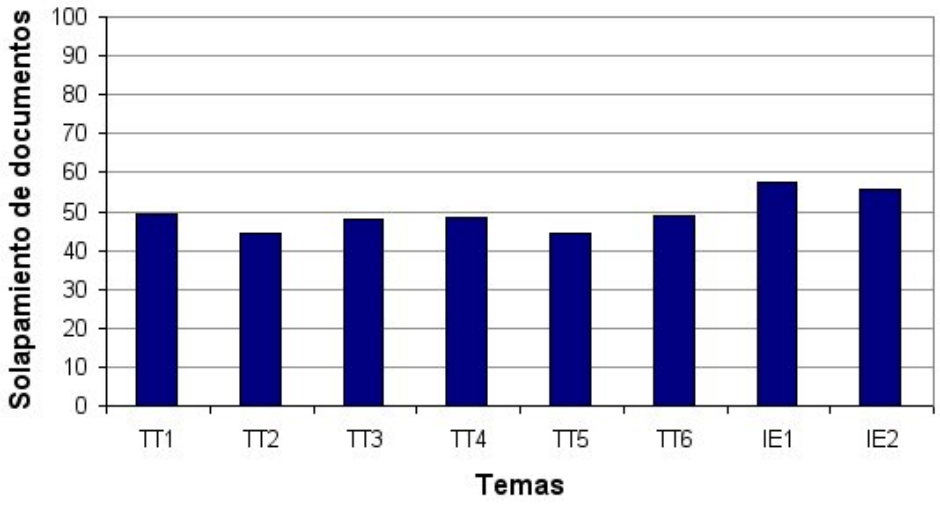


Figura 5.8: Porcentajes de solapamiento de documentos anotados en los informes generados en ISCORPUS

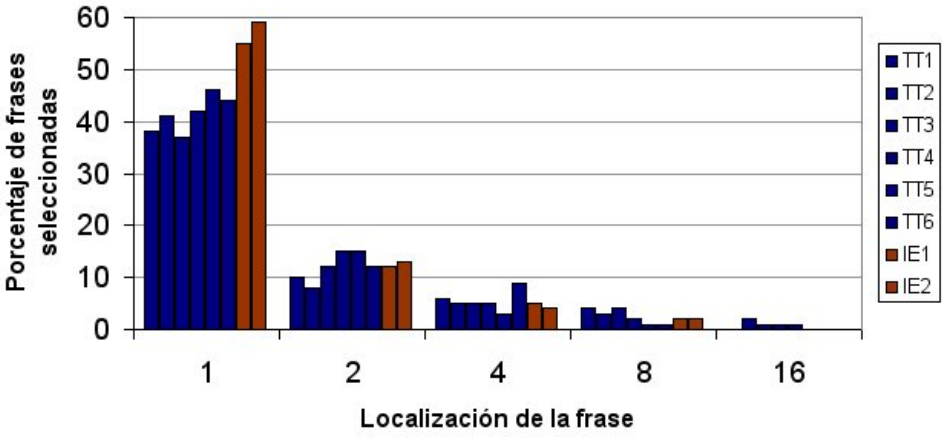


Figura 5.9: Porcentajes de frases seleccionadas desde distintas posiciones del documento original en la elaboración de informes de ISCORPUS

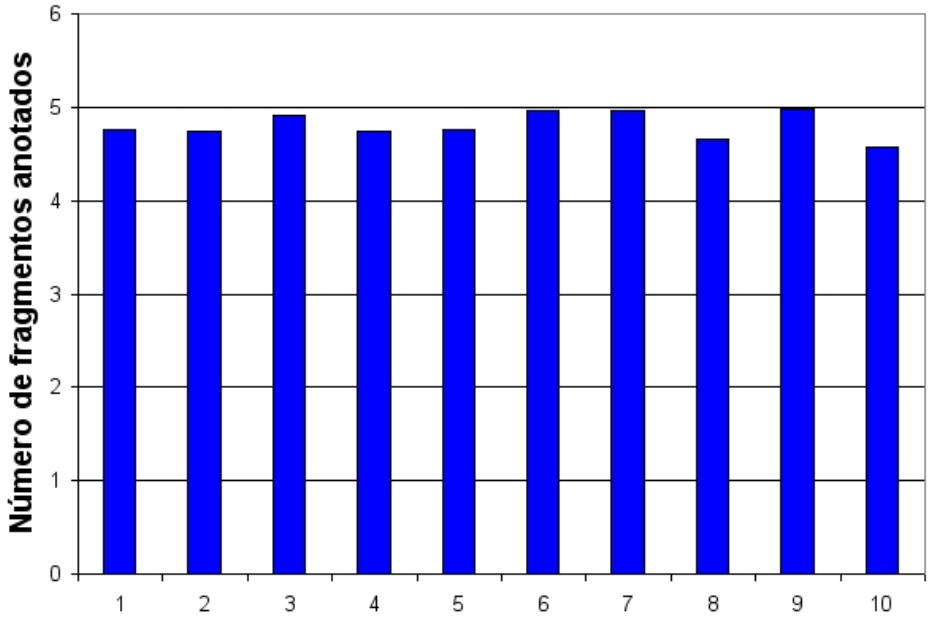


Figura 5.10: Número de fragmentos seleccionados durante el proceso de elaboración de informes de ISCORPUS

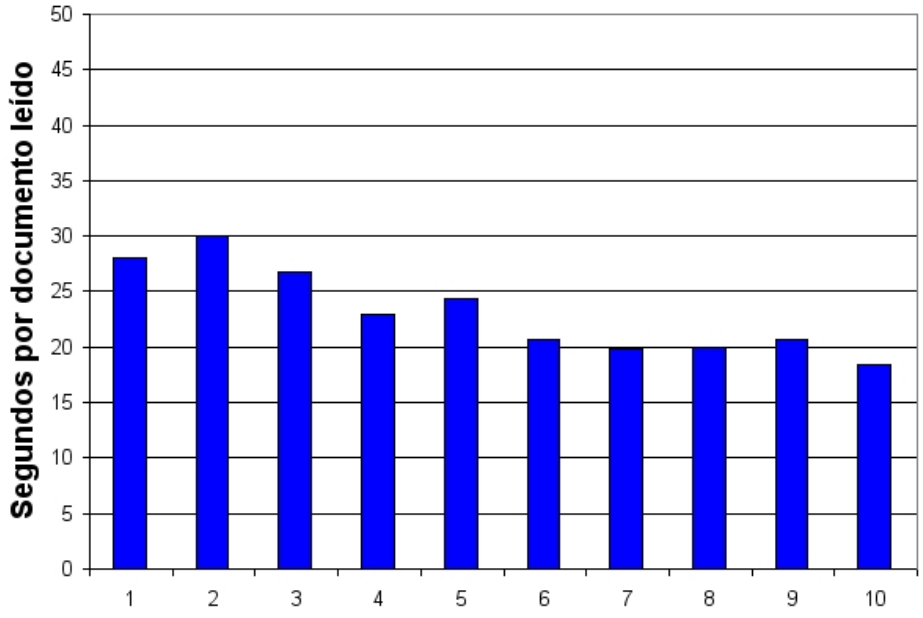


Figura 5.11: Porcentaje de tiempo empleado en la lectura de documentos en la elaboración de informes de ISCORPUS

Capítulo 6

QARLA: una metodología de evaluación automática

Para poder abordar el problema del acceso a la información en el contexto de la SI es necesario evaluar comparativamente diferentes aproximaciones. En este libro planteamos el proceso de evaluación desde una perspectiva automática, sin necesidad de incorporar jueces o usuarios de prueba, sobre los informes modelo generados en ISCORPUS. La automatización del proceso de evaluación permite la comparación de multiples aproximaciones sin coste adicional. Esto es especialmente interesante en fases iniciales del desarrollo del modelo PRISMA.

Como hemos apuntado en capítulos anteriores de este libro, la tarea de SI está íntimamente relacionada con la tarea de resumen en múltiples aspectos. Concretamente, ambas tareas presentan las mismas dificultades en cuanto a la definición de una metodología de evaluación automática. La principal dificultad reside en cómo establecer un criterio de similitud entre el resumen/informe evaluado y los modelos. Los modelos son diferentes entre sí, y existen además múltiples formas de medir la similitud entre dos textos o de combinar distintos criterios de similitud. En esta capítulo presentamos QARLA, una metodología de evaluación intrínseca basada en modelos de referencia. QARLA está orientada a sistemas de resumen y es aplicable al caso concreto de la SI, abordando las cuestiones mencionadas.

6.1. Problemas en la evaluación de resúmenes

Como apuntamos en el capítulo 4, la calidad de un resumen en relación a resúmenes modelo, puede estimarse principalmente de dos formas:

- *Juicios humanos:* Los resúmenes generados mediante diferentes sistemas son comparados manualmente con resúmenes modelo, siguiendo un protocolo determinado.

- *Proximidad a informes modelo:* Esta aproximación parte de la hipótesis de que un resumen tiene mejor calidad cuanto más se asemeje a un informe modelo de referencia.

El uso de juicios humanos implica claramente una serie de ventajas: los resultados de la evaluación son fácilmente interpretables, y permiten determinar cuáles son las virtudes y deficiencias del resumen evaluado. Por ejemplo, es posible pedir a los jueces que evalúen aspectos parciales como la fluidez del texto o la cobertura de contenidos.

Por otro lado, este tipo de evaluación presenta una serie de inconvenientes. En primer lugar, no siempre existe acuerdo entre distintos jueces humanos. En segundo lugar, contar con jueces humanos es costoso, especialmente en la fase de desarrollo de un sistema en el que se ha de comparar múltiples aproximaciones. Es decir, nuevas aproximaciones habrían de ser evaluadas nuevamente por jueces humanos.

La evaluación en base a la proximidad a resúmenes modelo permite superar estos problemas, definiendo métricas de evaluación que determinen de manera automática dicha proximidad. En primer lugar, sobre la base de una métrica de evaluación automática, tenemos una evaluación objetiva no sujeta a desacuerdos entre evaluadores. En segundo lugar, esta métrica es reusable, permitiendo la evaluación de múltiples aproximaciones sin coste adicional.

La evaluación automática sobre informes modelo de referencia requiere sin embargo solventar algunas cuestiones:

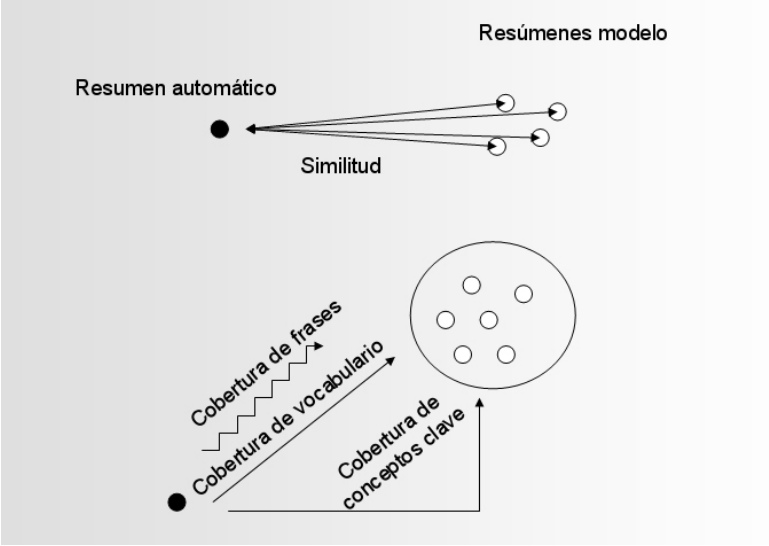


Figura 6.1: Multiplicidad de resúmenes modelo y métricas de similitud en la evaluación de resúmenes

- ¿Cómo saber si una métrica de evaluación es lo suficientemente fiable? Un resumen es un texto elaborado que posee multitud de características independientes entre si, por ejemplo, vocabulario, estructuras sintácticas, distribución de conceptos clave, longitud de las frases, etc. Por tanto, la semejanza entre dos resúmenes puede establecerse desde múltiples puntos de vista. Necesitamos por tanto, identificar el mejor criterio de similitud entre informes

evaluados y modelos.

- ¿Cómo considerar simultáneamente distintos informes modelos generados por diferentes sujetos? (parte superior de la figura 6.1) La tarea de resumen, y en concreto la de Síntesis de Información es muy subjetiva. Esto implica que, para una misma necesidad de información y un mismo conjunto de documentos originales, distintos sujetos podrían generar resúmenes diferentes.

- Si disponemos de varias métricas que consideran diferentes rasgos de los resúmenes, ¿cómo combinarlas en una sola métrica? (parte inferior de la figura 6.1) Existen infinitas maneras de asignar pesos y promediar un conjunto de métricas para combinarlas. Por tanto, la combinación resultante dependerá de la topología de las métricas individuales, es decir, de sus escalas. Sin embargo, de cada métrica nos interesa qué rasgos de los resúmenes caracteriza, no su escala. Pero, ¿es posible combinar métricas de evaluación con independencia de factores de escala?

- Meta-evaluar una métrica de evaluación implica estudiar su capacidad para estimar la calidad de, por lo menos, un conjunto de aproximaciones de muestra. Pero ¿es este conjunto de muestras representativo? Necesitamos algún modo de estimar la estabilidad, no solo de los resúmenes modelo, sino también del conjunto de aproximaciones automáticas sobre las que se ha estudiado el comportamiento de las métricas de similitud.

En este capítulo, introducimos un marco probabilístico que aborda todas estas cuestiones. Dado un conjunto de resúmenes modelo y un conjunto de resúmenes generados mediante estrategias automáticas de referencia, el marco QARLA ofrece medidas cuantitativas que permiten combinar, evaluar y aplicar métricas de similitud, y comprobar en qué medida los resúmenes de muestra son representativos.

6.2. Principios de QARLA

En este capítulo definiremos un marco de evaluación de sistemas de resumen basado en métricas de similitud y resúmenes modelo de referencia, e independiente de juicios humanos. El punto de partida sobre el que se apoya el marco estará constituido por:

- Una tarea de resumen. La tarea se caracterizará por aspectos como la longitud del resumen generado, si el resumen está orientado por una consulta, número y tipo de documentos de partida, etc.

- Un conjunto T de muestras de test. Cada muestra de test estará constituida por uno o un conjunto de documentos de partida, y en su caso, una necesidad de información concreta.

- Un conjunto M de resúmenes generados manualmente por una serie de sujetos (resúmenes modelo de referencia), y un conjunto A de resúmenes generados automáticamente mediante una serie de sistemas. Tendremos un conjunto A y M para cada muestra de test en T.
- Un conjunto X de métricas de similitud aplicables sobre un par de resúmenes cualquiera.

El marco debe aportar medidas cuantitativas para la aplicación y validación de las métricas de similitud sobre los resúmenes modelo. Es decir, el marco de evaluación lo formarán las siguientes medidas:

- Una medida $Q_{M,X}(a) \in [0,1]$ que estime la calidad de un resumen automático a empleando una métrica de similitud de X para estimar la semejanza al conjunto de modelos M. Mediante Q podremos comparar la calidad de los distintos sistemas.
- Una medida $K_{M,A}(X) \in [0,1]$ que estime la adecuación de una métrica de similitud x para la evaluación. Con K podremos seleccionar la mejor métrica de evaluación.

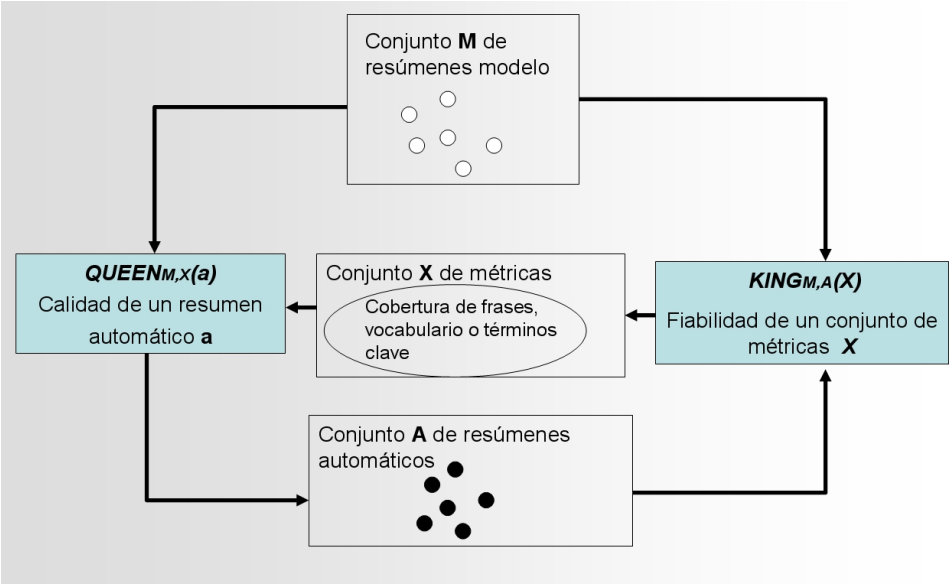


Figura 6.2: Representación de las medidas incluidas dentro del marco QARLA

La figura 6.2 muestra las medidas y elementos de partida del marco.

El marco de evaluación QARLA se apoya en la siguiente hipótesis: *todos los modelos son igualmente óptimos, aunque mantengan diferencias entre sí. Por tanto, la mejor métrica de evaluación será aquella que sea capaz de identificar y extraer los rasgos compartidos por los resúmenes modelo, discriminándolos de los resúmenes automáticos.*

En QARLA, la evaluación de resúmenes se centra en la capacidad de los sistemas de emular el resumen que generaría una persona. Desde esta perspectiva, cualquier criterio que distinga a un resumen automático de un resumen manual es igualmente relevante. La evaluación consiste en cuantificar en qué medida el resumen evaluado se asemeja a los modelos en base a estas características propias de los resúmenes manuales.

Bajo estos supuestos, podemos enfocar el trabajo definiendo un conjunto de restricciones formales que el marco K, Q debería satisfacer. Algunas de ellas están ilustradas en la figura 6.3. En este apartado nos centraremos en la aplicación y validación de métricas de similitud individuales. En apartados posteriores generalizaremos el marco QARLA para el tratamiento de conjuntos de métricas. En relación al tratamiento de métricas individuales, definimos el siguiente conjunto de restricciones sobre K y Q:

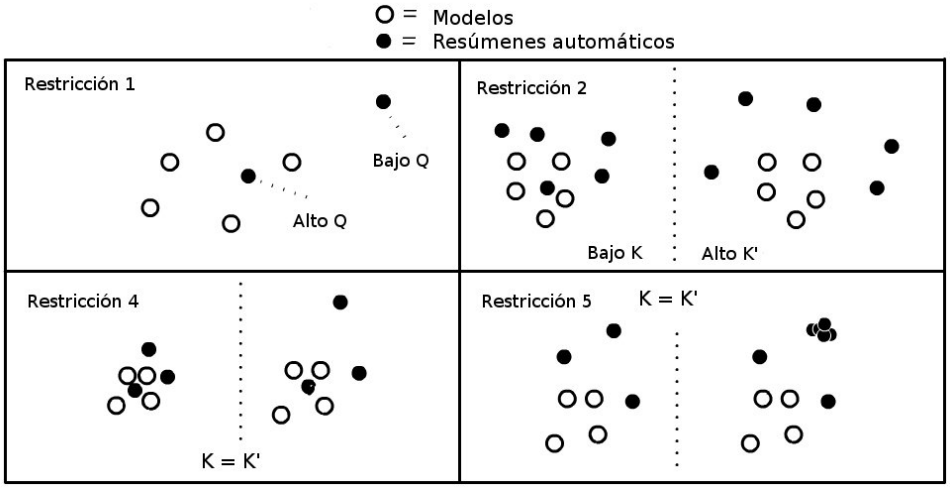


Figura 6.3: Representación de las restricciones formales en las que se apoya el marco de evaluación QARLA

Restricción 1 Dado que todos los resúmenes modelo son igualmente óptimos, el criterio Q de aplicación de métricas de similitud debería asegurar que si aumenta la distancia a cualquiera de los resúmenes modelo, entonces decrece la calidad del resumen evaluado. Formalmente, dados dos resúmenes automáticos a, a' y una métrica de similitud x, si a es igual o más distante a todos los resúmenes modelo que a', entonces a no puede tener mejor calidad que a':

$$\forall m \in M.x(a,m) \leq x(a',m) \rightarrow Q_{M,x}(a) \leq Q_{M,x}(a')$$

Restricción 2 El marco QARLA asume que la mejor métrica es aquella que caracteriza los rasgos propios de los resúmenes manuales, frente a todo aquello que no son resúmenes manuales. Es decir, una métrica de similitud x es mejor si es capaz de agrupar resúmenes manuales entre sí, distinguiéndolos de los resúmenes automáticos:

$$(\forall m, m' \in M.x(m, m') > x'(m, m')$$

$$\wedge a \in A, \forall m \in M, x(a, m) < x'(a, m)) \rightarrow K_{M,A}(x) > K_{M,A}(x')$$

Restricción 3 Desde la perspectiva de QARLA, un buen sistema es aquel que es capaz de emular un resumen generado por un ser humano. Por tanto, es evidente que los resúmenes manuales han de parecerse así mismos en mayor medida que cualquier otra cosa. Si establecemos nuestras medidas Q y K en un rango de 0 a 1, esto implica que la calidad Q de un resumen manual no puede ser 0. Formalmente, establecemos la restricción de que si x es una métrica de similitud ideal (K=1), entonces la calidad de los modelos no puede ser nula:

$$K_{M,A}(x) = 1 \rightarrow \forall m \in M.Q_{M,x}(m) > 0$$

Restricción 4 De una métrica de similitud, nos interesa los rasgos en los que se apoya. Por tanto sería deseable que la aplicación y validación de métricas de similitud no dependiera de factores de escala o rango de las métricas. Unas medidas Q y K que cumplan esta restricción facilitan la elaboración de métricas de similitud, en cuanto que no es necesario normalizar la topología de dichas métricas. Formalmente, si $x' = f(x)$ siendo f una función monótona creciente, entonces

$$K_{M,A}(x) = K_{M,A}(x') \text{ y además } Q_{M,x}(a) = Q_{M,x'}(a)$$

.

Restricción 5 El marco QARLA cuantifica la fiabilidad de una métrica de similitud en base a su capacidad de discriminación entre resúmenes manuales frente a resúmenes automáticos. El marco establecerá por tanto la fiabilidad de un métrica en base a una muestra de resúmenes automáticos A. Dado que la distribución de los resúmenes automáticos depende de las preferencias de los desarrolladores, es posible que existan elementos redundantes en este conjunto de muestra A, afectando a los resultados. Es decir, esta restricción exige que la fiabilidad K de una métrica de similitud x no debe de ser sensible a elementos repetidos en el conjunto A. Formalmente:

$$K_{M,A\cup\{a\}}(x) = K_{M,A\cup\{a,a\}}(x)$$

Restricción 6 Una métrica basada en fenómenos aleatorios de un resumen, por ejemplo, el número de ocurrencias de la letra 'a', debería de tener una calidad mínima a efectos de evaluación. Formalmente dada una métrica aleatoria x, entonces $K_{M,A}(x) = 0$.

Restricción 7 Una métrica de similitud que siempre devuelva un valor constante es también inútil a efectos de evaluación. Formalmente, dada una métrica no informativa x, por ejemplo, de valor constante, entonces $K_{M,A}(x) = 0$.

Basándonos en estas restricciones, definiremos las medidas Q y K que constituyen el marco de evaluación QARLA.

6.3. Estimación de la calidad de un resumen: medida QUEEN

Queremos definir una función $Q_{M,x}(a)$ que estime la calidad de un resumen $a \in A$, dado un conjunto de modelos M y una métrica de similitud x satisfaciendo las restricciones descritas en el apartado anterior.

Una primera aproximación al problema consiste en calcular la similitud media, o cualquier otro tipo de promediado, del resumen evaluado a los distintos modelos. Las métricas actuales de evaluación automática se basan de hecho en algún tipo de promediado sobre el conjunto de modelos. Sin embargo, esta aproximación es sensible a las propiedades de escala de la métrica de similitud empleada. Métricas de similitud con mayor escala producirán mayores valores en Q, aunque las métricas se basen los mismos rasgos de los resúmenes. Por otro lado, dependiendo de las propiedades de escala de la métrica, podría ser más apropiado un tipo de promediado distinto, por ejemplo, media geométrica o aritmética.

La siguiente definición de Q, a la que llamamos QUEEN solventa estos problemas, satisfaciendo todas las restricciones teóricas planteadas en el apartado anterior, entre ellas, la independencia de escalas:

$$\text{QUEEN}_{x,M}(a) \equiv P(x(a,m) \geq x(m',m''))$$

QUEEN representa la probabilidad estimada sobre tripletas $< m, m', m'' >$ de modelos de que el resumen evaluado a se encuentre más cercano al modelo m que los otros dos modelos m' y m'' entre sí según la métrica x. Si alteramos las propiedades de escala de x obtendremos exactamente el mismo resultado.

La figura 6.4 representa el cálculo de la medida QUEEN sobre un conjunto de tres resúmenes modelo. Con cada combinación de modelos disponemos de una muestra sobre la que calcular la probabilidad QUEEN. En el caso de la figura, el resumen evaluado a ha superado la mitad de los test a los que QUEEN somete el resumen. La calidad del resumen automático es por tanto 0.5 La figura 6.5 ilustra el comportamiento de la medida QUEEN sobre varios resúmenes automáticos evaluados. Como puede verse:

- Resúmenes automáticos que se encuentran muy alejados del conjunto de modelos obtienen siempre un valor 0 en QUEEN. En otras palabras, QUEEN no distingue entre resúmenes automáticos de baja calidad. Aunque esta propiedad reduce la granularidad del ranking generado por QUEEN, creemos

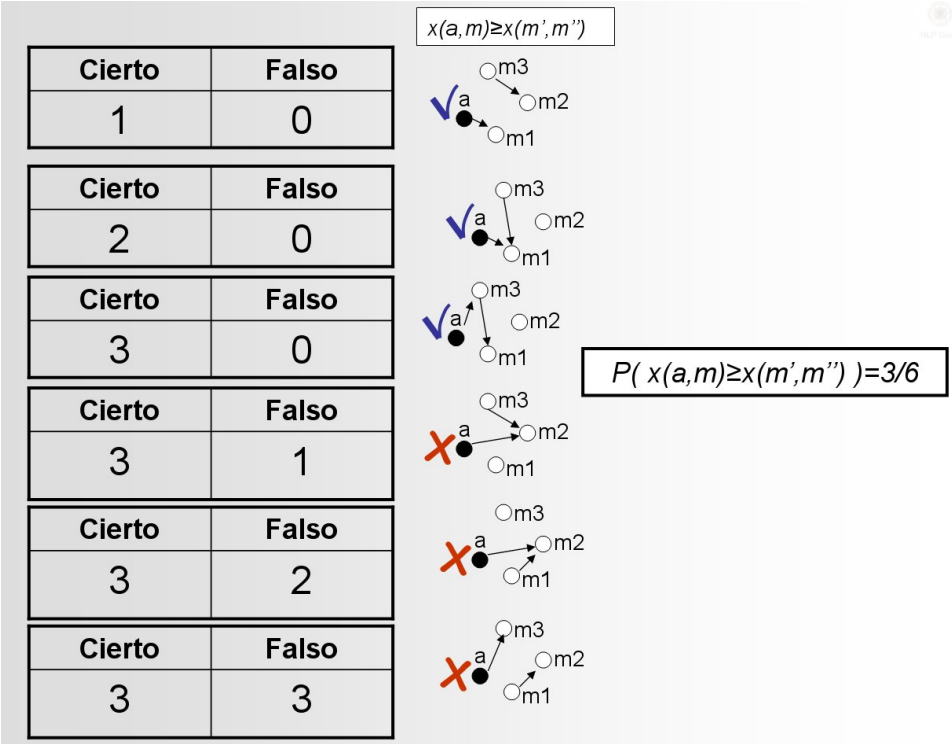


Figura 6.4: Representación del cálculo de la probabilidad QUEEN sobre una única métrica de similitud

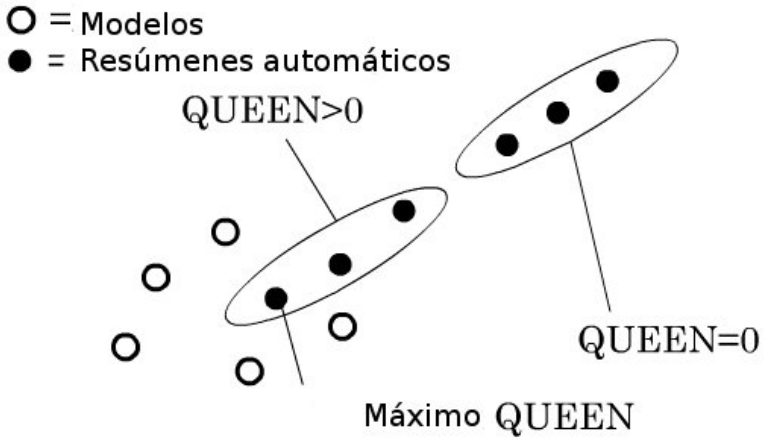


Figura 6.5: Representación de la calidad QUEEN de un conjunto de resúmenes en un espacio definido por una métrica de similitud

que esta propiedad es deseable, dado que en el caso de que todos los sistemas sean de muy baja calidad, una comparación entre éstos sería en realidad inútil.

- El valor de QUEEN se maximiza en los resúmenes automáticos que se "confunden" entre los modelos. Es decir, un resumen semejante en general a los modelos se asemeja tanto a un modelo como dos modelos entre si, obteniendo un alto valor en QUEEN.

Generalización de QUEEN para conjuntos de métricas

Diferentes métricas pueden representar diferentes rasgos de un resumen. Sería deseable por tanto estimar la calidad de un resumen considerando simultáneamente varias métricas de similitud. Supongamos por ejemplo que la mejor métrica individual sea una variante de ROUGE [LH03a] que considere únicamente de escalas unigramas. Ahora bien, supongamos que un resumen generado automáticamente contiene el mismo vocabulario que los resúmenes modelo, solo que en un orden aleatorio. En este caso concreto una medida basada en bi-gramas o tri-gramas podría aportar la información necesaria para evaluar el resumen con cierta fiabilidad.

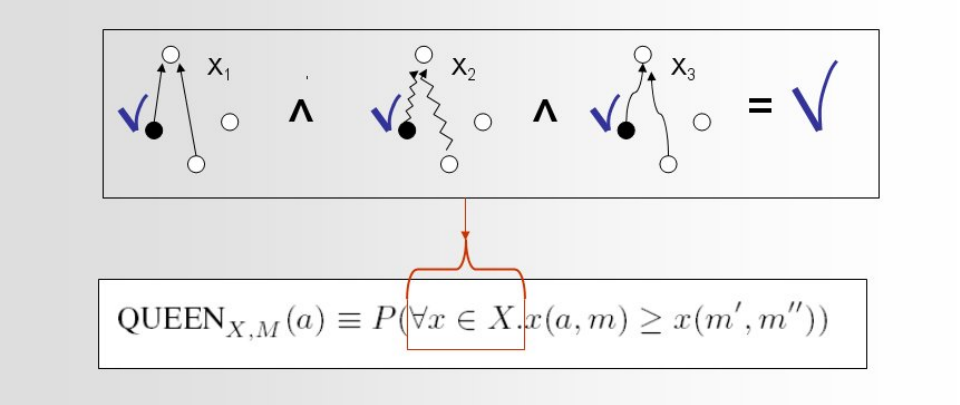


Figura 6.6: Cálculo de la probabilidad QUEEN sobre varias métricas de similitud

El problema es cómo combinar las métricas de similitud. Una aproximación sencilla consiste en aplicar algún criterio de promediado entre métricas. Sin embargo, dado que el tipo óptimo de promediado depende de las propiedades de escala de las métricas, no podemos fijar una estrategia de combinación de métricas independiente de escalas. Es decir, no se satisface la propiedad 4 descrita en el primer apartado de este capítulo.

Sin embargo, podemos asumir que un resumen es mejor si lo es para todas las métricas consideradas (figura 6.6). Desde la perspectiva de la medida QUEEN, podríamos decir que un resumen evaluado es más cercano a un modelo que dos modelos entre sí siempre que esto sea cierto para todas las métricas consideradas. Es decir, introduciendo un cuantificador universal en la definición de QUEEN:

$$\text{QUEEN}_{X,M}(a) \equiv P(\forall x \in X.x(a,m) \geq x(m',m''))$$

Podemos interpretar la medida QUEEN generalizada como un conjunto de pruebas, uno para cada métrica de similitud. Cada una de estas pruebas puede descartar la hipótesis de que el resumen evaluado tiene las características de un modelo. Si el resumen no supera alguna de las métricas, la hipótesis queda descartada para la muestra $< a, m, m', m'' >$.

Esta versión generalizada de QUEEN no se ve afectada por las propiedades de escala de las métricas de similitud, y no requiere ningún pesado de métricas en el proceso de combinación. Además, sigue cumpliendo las restricciones planteadas para Q sobre una sola métrica de similitud.

Por supuesto, los juicios generados por QUEEN son inútiles si el conjunto de métricas X empleado no captura los rasgos esenciales de los resúmenes modelo. Por tanto, necesitamos estimar la calidad del conjunto de métricas de similitud seleccionado para que QUEEN sea fiable.

6.4. Estimación de la calidad de una métrica de similitud: medida KING

Queremos definir una medida $K_{M,A}(x)$ que estime la adecuación de una métrica de similitud x para la evaluación de resúmenes automáticos en relación a un conjunto M de resúmenes modelo.

Con el fin de definir una media K fiable, partimos de la hipótesis de que la mejor métrica es aquella que mejor caracteriza los rasgos comunes de los resúmenes modelo en relación a resúmenes generados automáticamente. Es decir, será más fiable aquella métrica que agrupe en el espacio a los modelos, manteniéndolos a distancia de los resúmenes automáticos. Esta es la segunda restricción definida en el primer apartado de este capítulo (sección 6.2). Por analogía con QUEEN, podríamos definir esta medida K como:

$$K_{M,A}(x) \equiv P(x(a,m) < x(m',m'')) = 1 - \overline{(\text{QUEEN}_{x,M}(a))}$$

que representa la probabilidad de que dos modelos se encuentren más cercanos ente sí que un modelo a un resumen automático. Esta definición de K crece a medida que decrece el promedio de la calidad QUEEN de los resúmenes automáticos. La generalización de K a conjuntos de métricas es sencilla.

$$K_{M,A}(X) \equiv 1 - \overline{(\text{QUEEN}_{X,M}(a)))}$$

Sin embargo, esta definición de K no satisface las restricciones formales 3 y 5. La condición 3 es violada dado que, para un número limitado de modelos, el valor de K crece siempre que introduzcamos una cantidad suficiente de métricas de similitud, obteniendo un valor $K = 1$, es decir, la métrica ideal. Pero en esta situación, la calidad QUEEN de un modelo será también 0, dado que existe siempre

alguna métrica que pone a falso el cuantificador universal sobre el conjunto X. Por lo tanto, no se cumple la restricción 3.

Podemos proponer una formulación alternativa de K. En este caso, las métricas de mejor K deben minimizar la calidad de los resúmenes automáticos, pero tomando como referencia la calidad QUEEN de los modelos. Es decir:

$$K_{M,A}(X) = P(\text{QUEEN}(m) > \text{QUEEN}(a))$$

Según esta definición, la calidad de un conjunto de métricas X es la probabilidad de que la calidad QUEEN de un modelo sea superior a la calidad QUEEN de un resumen automático para dicho conjunto de métricas. Esta fórmula satisface todas las restricciones a excepción de la 5 ($K_{M,A\cup\{a\}}(x) = K_{M,A\cup\{a,a\}}(x)$), dado que K sería sensible a elementos repetidos en el conjunto A de resúmenes automáticos. Si introducimos un conjunto amplio de resúmenes automáticos idénticos o muy parecidos entre si, los resultados de K se verían sesgados por este conjunto.

Definimos una medida K a la que denominamos KING añadiendo un cuantificador universal sobre el conjunto A. Es decir:

$$\text{KING}_{M,A}(X) \equiv P(\forall a \in A.\text{QUEEN}_{M,X}(m) > \text{QUEEN}_{M,X}(a))$$

KING representa la probabilidad de que un modelo tenga mayor calidad QUEEN que cualquiera de los resúmenes automáticos. En términos de *rankings*, representa la probabilidad de que el resumen modelo se encuentre en una posición superior a la de todos los resúmenes automáticos.

Esta definición KING cumple todas y cada una de las restricciones descritas en el primer apartado de este capítulo. Por ejemplo, es independiente de las propiedades de escala de las métricas, dado que KING depende únicamente de QUEEN, y ésta no depende de escalas. Además, el cuantificador universal sobre A hace que sea insensible a elementos repetidos en A. Por último, y una métrica ideal tendría que considerar que los resúmenes modelo M tienen mejor calidad que los resúmenes automáticos en A, por lo que nunca otorgará calidad 0 a los resúmenes modelo.

La figura 6.7 ilustra el comportamiento de la medida KING en condiciones extremas. La gráfica de la izquierda representa un espacio definido por una métrica de similitud totalmente incapaz de discriminar modelos de resúmenes automáticos. Por tanto, podemos asegurar que

$$P(\text{QUEEN}(m) > \text{QUEEN}(a)) \approx 0{,}5$$

Dado que tenemos 5 resúmenes automáticos,

$$\text{KING} = P(\forall a \in A, \text{QUEEN}(m) > \text{QUEEN}(a)) \approx 0{,}5^5 \approx 0$$

La figura de la derecha representa una métrica que sí es capaz de agrupar modelos en relación al conjunto de resúmenes automáticos. En este caso se cumple $\text{QUEEN}(a) = 0$ para todo a, y por tanto $\text{KING}(x) = 1$.

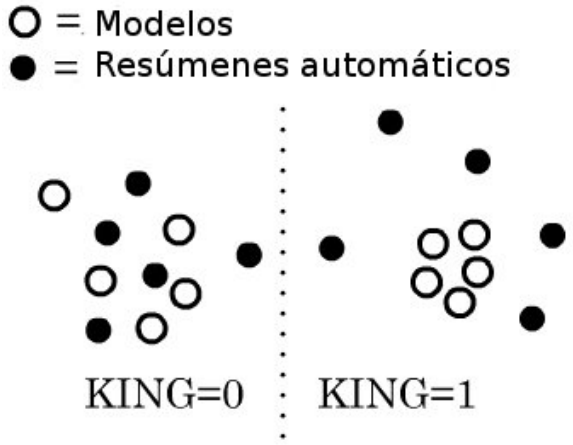


Figura 6.7: Representación del comportamiento de KING

6.5. Fiabilidad del corpus: medida JACK

Una vez que estimamos la calidad de las métricas de evaluación sobre un conjunto de muestras A de resúmenes automáticos y un conjunto M de muestras de resúmenes modelo, queda abierta la cuestión de si el corpus empleado, es decir, los conjuntos A y M son estables. Dicho de otro modo, ¿qué ocurriría si introducimos nuevos modelos en M o nuevos sistemas en A?

La aproximación más habitual a este tipo de problemas es la aplicación de un test de relevancia estadística. Sin embargo, aunque obtuviésemos resultados positivos, puede ser insuficiente. El problema es que el cálculo de probabilidades en KING y QUEEN y el test de relevancia estadística correspondiente asume que los conjuntos M y A no están sesgados. Si estos conjuntos están sesgados, el test de relevancia estadística puede dar resultados positivos y tratarse sin embargo de un corpus poco estable.

El conjunto de muestras M que representa el comportamiento de sujetos en la realización de resúmenes (modelos) puede ser de alguna forma controlado con el perfil de los usuarios escogidos y tomando un conjunto lo suficientemente amplio de personas. Pero, ¿cómo saber si el conjunto A de resúmenes automáticos es representativo? Si el conjunto A abarca sólo un tipo de aproximaciones entonces estas aproximaciones serían penalizadas en la selección de métricas mediante la medida KING. Es decir, nuevas aproximaciones sobre las que KING no ha evaluado las métricas tendrían ventaja en una evaluación posterior que aplique las mismas métricas. Necesitamos por tanto alguna medida (JACK(X, M, A)) que controle la calidad y heterogeneidad de los resúmenes automáticos A empleados para la evaluación de las métricas de similitud. Podemos establecer de nuevo un conjunto de restricciones que esta media JACK debería cumplir.

1. Dados los conjuntos M y X de modelos y métricas, si el conjunto A es más heterogéneo, éste es entonces más representativo. Es decir, el corpus dispone

de una mayor diversidad de estrategias de resumen automático. Por tanto, si aumenta las diferencias entre resúmenes automáticos de A, entonces JACK debería aumentar.

2. Dado un conjunto M y X de modelos y métricas, si los resúmenes automáticos en A se encuentran más cercanos a los modelos M, esto implica que tenemos una mayor calidad en los resúmenes automáticos, por lo que el valor de la medida JACK también debería aumentar.

3. La adición de nuevos elementos en A no puede reducir la fiabilidad del corpus, es decir, JACK debe aumentar.

4. La medida JACK debe ser insensible a elementos repetidos en el conjunto A de resúmenes automáticos, dado que no altera su heterogeneidad. Es decir, el número de aproximaciones automáticas representadas en el conjunto sigue siendo el mismo.

Una posible formulación de JACK que satisface dichos criterios es:

$$\text{JACK}(X, M, A) \equiv P(\exists a, a' \in A.\text{QUEEN}(a) > 0 \wedge$$

$$\wedge \text{QUEEN}(a') > 0 \wedge \forall x \in X.x(a, a') \leq x(a, m))$$

Esta definición representa la probabilidad sobre el conjunto de modelos M, de encontrar un par de resúmenes automáticos a, a' tal que sean más similares al modelo m que entre sí, para todas las métricas consideradas. Además a y a' deben tener un valor QUEEN mayor que 0.

La primera y segunda restricción son satisfechas, dado que podemos incrementar el valor de JACK bien decrementando la similitud entre resúmenes automáticos, o bien incrementando la similitud entre resúmenes automáticos y modelos. Además, el cuantificador $\exists$ asegura la tercera y cuarta restricción, dado que nuevos elementos en A no pueden poner a falso la condición JACK, y por otro lado, JACK no aumenta si introducimos elementos repetidos en A. La condición en la fórmula de que QUEEN(a) y QUEEN(a') no tengan valor nulo evita que únicamente dos resúmenes muy distantes de los modelos y aún más distantes entre sí, puedan obtener un valor JACK máximo.

La figura 6.8 ilustra cómo se comporta JACK. En el espacio de la izquierda, los resúmenes automáticos aparecen agrupados entre sí en relación a los modelos. Se trata por tanto de un conjunto A de resúmenes automáticos poco representativo. Considerando cualquiera de los resúmenes modelo, no existe una pareja de resúmenes automáticos más cercanos al modelo que entre si, por lo que el valor de JACK es 0. En el espacio de la derecha, los resúmenes automáticos se encuentran distribuidos en torno a los modelos. Se trata por tanto de un conjunto A heterogéneo (distantes entre si) y con elementos relativamente próximos a los modelos, obteniéndose un valor alto en JACK.

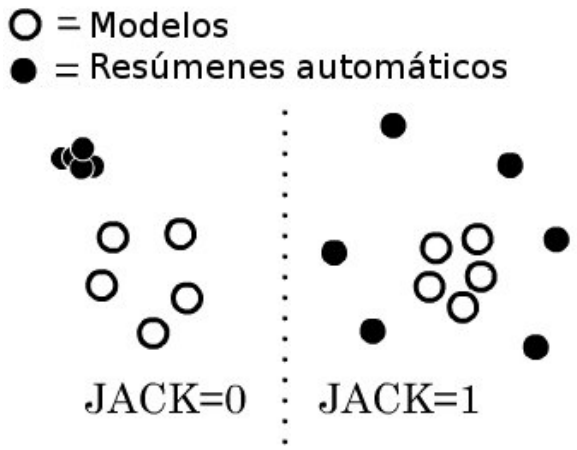


Figura 6.8: Representación del comportamiento de la medida JACK

6.6. QARLA para dominios interactivos

La calidad de un sistema interactivo de resumen se puede plantear desde la evaluación del resumen resultante o desde la perspectiva de la usabilidad, es decir, la satisfacción del usuario. En este apartado describimos una generalización de QARLA para dominios interactivos centrada en la evaluación del resumen resultante.

El marco de evaluación QARLA recibe como entrada un conjunto M de modelos, un conjunto X de métricas de similitud y un conjunto A de resúmenes generados mediante el uso de sistemas. En principio, podríamos evaluar los resúmenes resultantes de la interacción entre usuario y sistema, para cada uno de los sistemas que queremos comparar. La aplicación de QARLA en el contexto interactivo sería inmediata.

Ahora bien, la evaluación de sistemas interactivos mediante sujetos de prueba es tremendamente costosa, y sufre además una serie de limitaciones descritas en el apartado 4.2. La cuestión es si sería posible evaluar un sistema interactivo mediante QARLA sin necesidad de sujetos de prueba. En general, es dificil predecir el comportamiento de un usuario real ante un sistema. Sin embargo, existen ciertos aspectos de una aproximación interactiva que sí se pueden abordar. Este tipo de evaluación entraría dentro de la categoría de "scientific metrics" (apartado 4.2).

Un aspecto que puede ser abordado sin usuarios de prueba desde la metodología QARLA es la cobertura de contenidos de la aproximación interactiva. Es decir, la cobertura que ofrece la información más accesible desde un sistema interactivo de acceso a la información. Una forma de medir esto es mediante la cobertura sobre fragmentos de texto susceptibles de pertenecer a un resumen. Por ejemplo, aquellos fragmentos que aparezcan en más resúmenes modelo puntúan en mayor medida sobre la cobertura del sistema. Esta primera aproximación tiene los siguientes inconvenientes:

- En qué grado puntuar la cobertura sobre un fragmento de texto que aparezca en un cierto número de informes. Podría ser una relación lineal, geométrica,

etc.

- Como ponderar el número de fragmentos relevantes con la relevancia de los fragmentos. Es decir, ¿es más importante cubrir muchos fragmentos de los resúmenes de referencia, o cubrir los fragmentos más importantes?

- Puede ser que el conjunto de fragmentos más extraídos en los resúmenes modelo no baste para generar un buen resumen. Un buen resumen debe quizás incluir aspectos complementarios que no coinciden en los diferentes modelos, pero que deben de estar. Por ejemplo, todas las biografías comparten datos básicos como fecha de nacimiento o lugar donde estudió. Sin embargo, una biografía que hable sólo de esos datos frecuentes no es una buena biografía.

Evaluación de resúmenes potenciales

Para solucionar estos tres problemas, se puede plantear la evaluación desde la siguiente perspectiva: en vez de evaluar individualmente cada una de las piezas de información accesibles desde el sistema interactivo, evaluamos globalmente el conjunto de resúmenes que se podrían generar a partir de estas piezas de información. Dado que se evalúan resúmenes completos, no es necesario ponderar el peso de cada fragmento en la cobertura. La cuestión que surge entonces es cómo evaluar un conjunto de resúmenes. Esto es posible mediante una generalización de QARLA para dominios interactivos.

Sea A el conjunto de resúmenes potenciales derivados de una aproximación interactiva. Una condición que debería cumplir esta medida de evaluación es que debe existir por lo menos un resumen potencial en el conjunto A que satisfaga a cada uno de los posibles usuarios. Dado un conjunto M de resúmenes modelo generados manualmente, y un conjunto óptimo X de métricas de similitud, podemos expresar esta condición en términos de QUEEN como:

$$\text{QUEEN}_{M,X}(A) = (\exists a \in A. \forall x \in X. sim_x(a, m) \geq sim_x(m', m''))$$

Esta generalización de QUEEN representa la probabilidad sobre el conjunto de modelos de que exista por lo menos un resumen en A más próximo al modelo que dos modelos entre sí para todas las métricas consideradas. En otras palabras, representa la proporción de muestras $M \times M \times M$ que son superadas por alguno de los elementos en A.

Propiedades de QUEEN generalizada

Esta generalización de QUEEN posee ciertas propiedades interesantes, la mayoría de ellas heredadas de la definición original de QUEEN:

1. La similitud entre modelos es tomada como referencia para valorar la similitud entre un resumen potencial y un modelo. Por lo tanto, existe un criterio

absoluto de proximidad independiente de propiedades de escala de las métricas empleadas

2. La definición de QUEEN exige que por lo menos un elemento en A sea más similar al modelo que dos modelos entre si. Es decir, si obtenemos QUEEN$(A) = 1$, entonces todos los usuarios representados en el conjunto de modelos se verían satisfechos por al menos uno de los resúmenes potenciales.

3. QUEEN generalizada sigue siendo una medida probabilística que elimina la necesidad de combinar métricas mediante criterios de pesado. Es decir, permite combinar métricas con independencia de las propiedades de escala de las métricas individuales.

4. El número de resúmenes potenciales ($|A|$) derivados del conjunto de fragmentos más accesibles desde un sistema, crece exponencialmente. Esto implica un coste computacional demasiado alto en el cálculo de QUEEN. Sin embargo, dado que QUEEN es una medida definido en términos de probabilidad, podemos comparar dos sistemas si consideramos aleatoriamente el mismo número de muestras de informes potenciales en ambos sistemas. Formalmente, sean A y A' los resúmenes potenciales derivados de dos sistemas, estos pueden ser un subconjunto de todos los resúmenes potenciales siempre que $|A| = |A'|$.

En este trabajo, hemos empleado QARLA generalizada para contrastar diferentes estrategias de exploración de contenidos textuales en un sistema interactivo de Síntesis de Información. El conjunto M lo forman los informes manuales generados en ISCORPUS. Cada estrategia de exploración otorga más accesibilidad a un conjunto de frases pertenecientes a los documentos originales. Los conjuntos A asociados a cada estrategia de exploración son los informes potenciales que, mediante extracción de frases, pueden ser generados a partir de las frases más accesibles en cada modelo de exploración.

6.7. Validación de QARLA sobre el corpus de resúmenes DUC-2004

Con el fin de validar el marco QARLA, mostramos en esta sección los resultados obtenidos tras aplicar QARLA a las tareas 2 y 5 definidas en el foro de evaluación de sistemas de resumen DUC. En estos experimentos veremos que la evaluación en QARLA obtiene una buena correlación con evaluaciones realizadas mediante juicios humanos, y que QARLA aporta nuevos elementos en el análisis del problema definido en dichas tareas.

La tarea 2 consiste en generar un resumen de 100 palabras a partir de un conjunto de documentos relacionados. La tarea 5, consiste también en generar un resumen de 100 palabras a partir de un conjunto de documentos, pero partiendo de una pregunta del tipo "¿Quién es...?".

6.7.1. Análisis de métricas de similitud

Una métrica de similitud caracteriza un aspecto determinado de los resúmenes del corpus. El primer paso en estos experimentos es seleccionar un conjunto de métricas de similitud que caracterice los rasgos propios de los resúmenes modelo, en oposición a los resúmenes generados por los sistemas. En esta sección, describimos un conjunto de 59 métricas consideradas como punto de partida. Algunas de éstas aportan información redundante respecto de otras. El siguiente paso es por tanto seleccionar un conjunto de métricas que maximice la fiabilidad de la evaluación.

Métricas de similitud

En este trabajo consideramos el siguiente conjunto de métricas de similitud:

Métricas basadas en ROUGE (R): ROUGE [LH03a] es una métrica de evaluación que estima la calidad de un resumen automático basándose en la cobertura de n-gramas sobre un conjunto de resúmenes modelo. A pesar de que ROUGE es una métrica de evaluación, podemos transformarla en una métrica de similitud entre pares de resúmenes, si consideramos únicamente un resumen origen (resumen evaluado) y un resumen destino (resumen modelo). Existen múltiples variaciones de la métrica ROUGE, como por ejemplo, ROUGE-W, ROUGE-L, ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, etc. [Lin04b]. Cada una de estas métricas puede ser empleada con tres opciones de preprocesado: con eliminación de términos de parada y lematización (tipo c), únicamente con eliminación de términos de parada (tipo b), o sin ningún tipo de preprocesado (tipo c). Combinando variaciones de ROUGE con tipos de preprocesado obtenemos 24 métricas de similitud.

Métricas basadas en ROUGE e invertidas (Rpre): La métrica ROUGE está en principio orientada a cobertura. Por tanto, si invertimos la dirección en el cálculo de la similitud obtenemos necesariamente una medida orientada a precisión. Es decir, $\text{Rpre}(a, b) = \text{R}(b, a)$. Considerando las mismas variaciones que en el punto anterior, obtenemos de nuevo 24 métricas de similitud basadas en ROUGE y orientadas a precisión.

TruncatedVectModel (TVM_n): Esta familia de métricas compara la distribución en los resúmenes de los n términos más frecuentes de los documentos originales. El proceso de cálculo de esta métrica es el siguiente:

1. Se extrae los n términos más frecuentes en los documentos originales sin considerar términos de parada.
2. Se genera por cada resumen un vector con la frecuencia relativa de dichos términos en el resumen.
3. Se calcula la inversa de la distancia euclídea entre los vectores asociados a los resúmenes.

Hemos considerado 9 posibles variantes de esta métrica de similitud: $n = 1, 4, 8, 16, 32, 64, 128, 256, 512$.

AveragedSentenceLengthSim (AVLS): Consiste en comparar la longitud media de las frases en ambos resúmenes. Aunque probablemente no se trate de una métrica de evaluación fiable por sí misma, puede sin embargo aportar información sobre las características de los resúmenes generados en DUC.

GRAMSIM: Esta métrica de similitud se basa en características de formales de los resúmenes. Compara la distribución de etiquetas gramaticales en los resúmenes. Es decir, en qué proporción aparecen en los resúmenes verbos, adjetivos, nombres etc. El proceso de calculo consiste en:

1. Etiquetado de los resúmenes mediante la herramienta TreeTager [Sch94].
2. Generación de un vector con la frecuencia relativa de las distintas etiquetas en cada resumen.
3. Cálculo de la distancia euclídea entre los vectores asociados a los resúmenes.

Agrupación de métricas de similitud

Dado el conjunto de métricas de similitud descrito anteriormente, tenemos un total de 57 (24+24+9) métricas orientadas a contenidos, y 2 métricas dependientes de aspectos estilísticos. El conjunto de combinaciones posibles de métricas de similitud crece exponencialmente. Sería interesante por tanto poder agrupar conjuntos de métricas de similitud que se comporten de manera similar.

Consideramos que dos conjuntos de métricas son similares si se comportan de la misma forma respecto a la condición $QUEEN$ (predicado H de la fórmula) ante una muestra $\{a, m, m', m''\}$. Es decir, la probabilidad de que ambos conjuntos de métricas discriminen el mismo resumen automático frente a los mismos pares de modelos. Formalmente, calculamos la similitud entre dos conjuntos de métricas como:

$$sim(X, X') \equiv Prob[H_X \leftrightarrow H_{X'}]$$

$$H_X \equiv \forall x \in X x(a, m) \geq x(m', m'')$$

La tabla de la figura 6.9 muestra, sobre un número fijado en 10, los grupos de métricas de similitud obtenidos para la tarea 2 de DUC 2003. En cuanto a las métricas de tipo ROUGE, el proceso distribuye las 48 métricas en siete grupos, dependiendo de la longitud de los n-gramas y del tipo de pre-procesado de los textos. En cuanto a las 9 métricas tipo TVM, se distinguen tres grupos: tomando solo el término más frecuente, 4 u 8 términos más frecuentes y otro conjunto con el resto de configuraciones TVM.

Conjunto	DESCRIPCION	MÈTRICA DE SIMILITUD
Conjunto 1	Métricas basadas en ROUGE	R-S.b R-SU.b R-S.a R-SU.a R-1.a R-1.b R-L.b R-L.a R-W-1.2.b R-W-1.2.a R-W-1.2.c R-S.c R-SU.c R-1.c R-L.c **Rpre-W-1.2.b** Rpre-W-1.2.a Rpre-W-1.2.c Rpre-L.c Rpre-1.c Rpre-S.c Rpre-SU.c Rpre-1.a Rpre-S.a Rpre-SU.a Rpre-1.b Rpre-S.b Rpre-SU.b Rpre-L.b Rpre-L.a
Conjunto 2	ROUGE (Stemming y eliminación de stopwords, 2-gramas)	**R-2.c** Rpre-2.c
Conjunto 3	ROUGE (Stemming y eliminación de stopwords, 3-gramas)	Rpre-3.c **R-3.c**
Conjunto 4	ROUGE (Stemming y eliminación de stopwords, 4-gramas)	Rpre-4.c **R-4.c**
Conjunto 5	ROUGE (Sin stemming 2-grams)	R-2.b R-2.a **Rpre-2.b** Rpre-2.a
Conjunto 6	ROUGE (Sin stemming 3-grams)	R-3.b R-3.a **Rpre-3.b** Rpre-3.a
Conjunto 7	ROUGE (Sin stemming 4-grams)	Rpre-4.a **Rpre-4.b** R-4.b R-4.a
Conjunto 8	TVM con el término más fracuente	**TVM.1**
Conjunto 9	TVM.4 y 8 términos más frecuentes	TVM.4 **TVM.8**
Conjunto 10	TVM.>8 términos más frecuentes	TVM.16 TVM.32 TVM.64 TVM.128 TVM.256 **TVM.512**

Figura 6.9: Agrupaciones de métricas de similitud en el marco QARLA

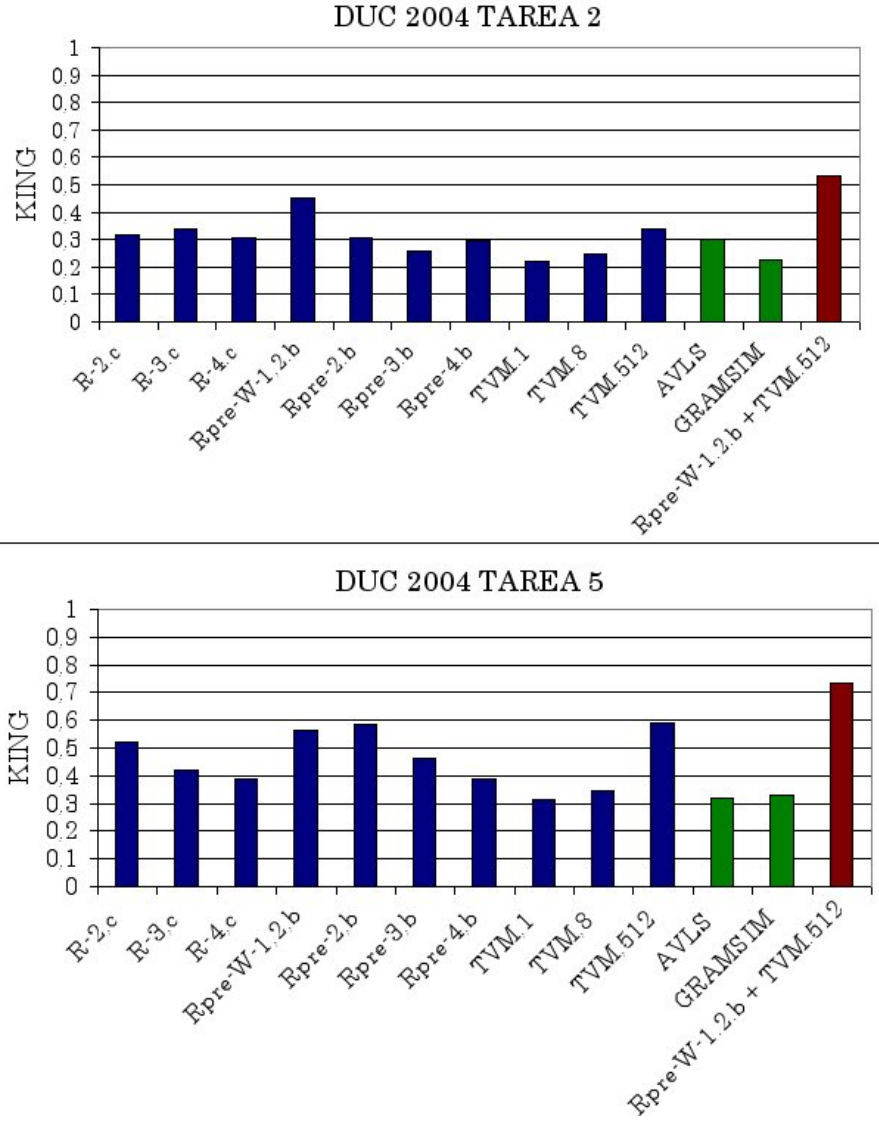


Figura 6.10: Calidad de las métricas de similitud según QARLA

Selección de métricas de similitud

Dentro de cada grupo de la figura 6.9, la métrica de similitud marcada en negrita se corresponde con la métrica de mayor KING. Como puede verse, en los conjuntos basados en n-gramas con stemming, las métricas de mayor KING son de tipo R, basadas en cobertura (grupos 2,3 y 4), mientras que en los conjuntos de métricas sobre n-gramas sin stemming (grupos 5,6 y 7) las métricas de mayor KING son de tipo Rpre, basadas en precisión. En cuanto a métricas de tipo TVM, las de mayor KING son siempre aquellas que se basan en un número mayor de términos frecuentes (grupos 8,9 y 10). Finalmente, tomamos como métricas representativas, la métrica de mayor $KING$ en cada grupo, en total 10 métricas orientadas a contenidos.

La figura 6.10 muestra el valor KING de las distintas métricas de similitud seleccionadas, es decir, la capacidad de las métricas de caracterizar a los resúmenes modelo frente a los resúmenes automáticos.

En cuanto a las métricas aisladas, Rpre-W obtiene un alto KING en ambas tareas. Esta métrica de similitud se basa en precisión de secuencias no contiguas de términos en los resúmenes (ROUGE-W-1.2). Para las métricas de tipo TVM a medida que consideramos más términos frecuentes, la métrica asciende en valores KING, es decir, caracteriza mejor a los resúmenes modelo.

La última columna representa la combinación de mayor KING de entre las posibles combinaciones de métricas de similitud. En ambas tareas esta combinación es $Rpre-W$ y TMV,512, y supera en valor KING a cualquiera de las métricas individuales. Este hecho sugiere que la capacidad de caracterización de los modelos puede aumentar por medio de la estrategia de combinación de métricas que ofrece el marco QARLA, y además que la medida KING es robusta.

6.7.2. Evaluación de resúmenes automáticos en DUC 2004

En esta sección, analizaremos los resultados de QARLA tras aplicar las métricas seleccionadas en la sección anterior en las tareas 2 y 5 del DUC.

QUEEN: Evaluación de los resúmenes

Ranking de resúmenes automáticos

La combinación de mayor KING en ambas tareas es Rpre-W y TVM.512. La figura 6.11 representa el ranking de resúmenes según dicha combinación, de acuerdo a los valores de $\text{QUEEN}_{\{\text{Rpre-W,TVM.512}\}}$ obtenidos. Los resúmenes modelo (A-H) obtienen los mayores valores de QUEEN en ambas tareas con una diferencia significativa respecto a los resúmenes automáticos, lo que refleja la capacidad de discriminación del modelo QARLA.

QUEEN frente a juicios humanos

El ranking manual generado en DUC representa un criterio subjetivo de evaluación, mientras que el modelo QARLA da más peso a aquellos aspectos que son más

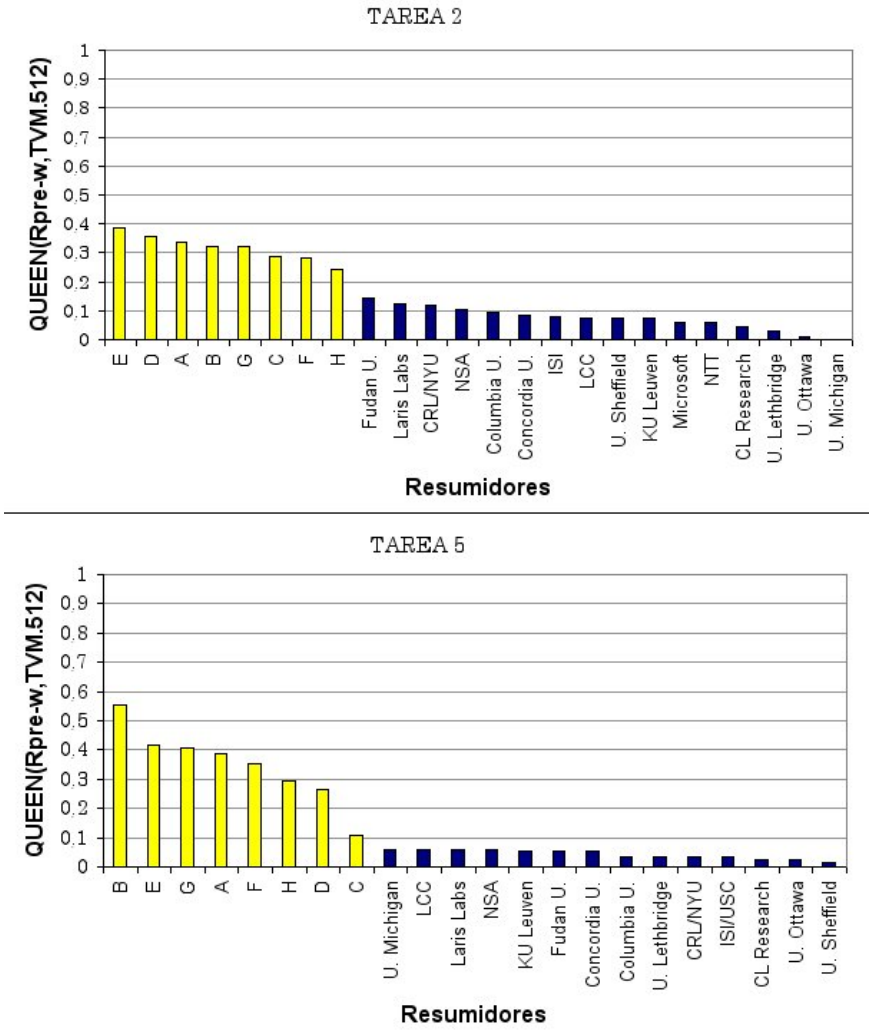


Figura 6.11: Calidad de los resúmenes automáticos generados en DUC según métricas de máximo KING en QARLA

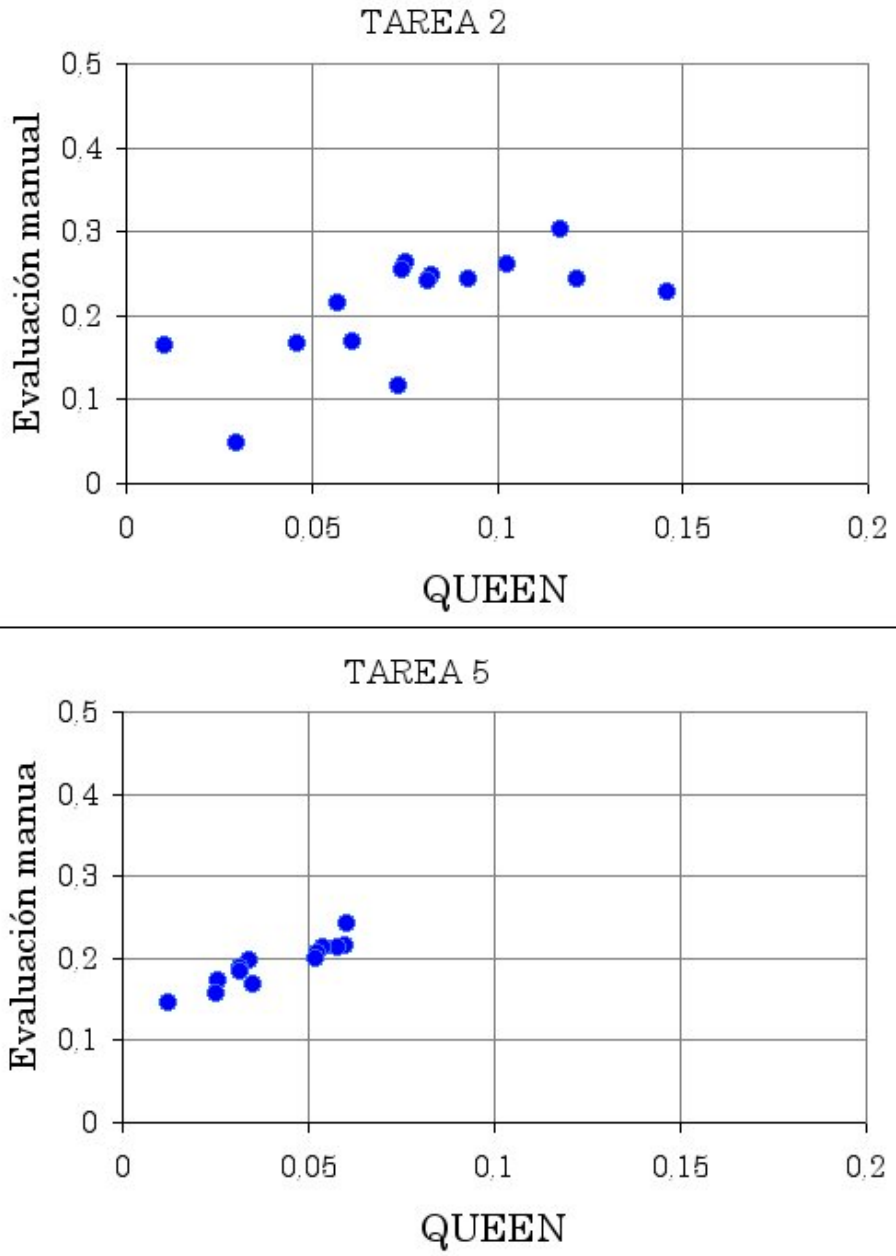


Figura 6.12: Correlación entre evaluación mediante juicios humanos en DUC y evaluación automática en QARLA

discriminativos entre modelos y resúmenes automáticos. Los resultados por tanto no tienen por qué coincidir necesariamente. Sin embargo, sí debería de existir cierta correlación. La figura 6.12 muestra la correlación existente entre ambas estrategias de evaluación, es decir valores de $\text{QUEEN}_{\{\text{Rpre-W,TVM.512}\}}$ (eje horizontal) frente a los valores de la evaluación manual (eje vertical). Esta correlación es más patente en la tarea 5 (orientada a pregunta) donde los contenidos de los resúmenes son más precisos. Por otro lado, en la tarea 2 (resúmenes genéricos) puede verse que los sistemas con mayor puntuación en QUEEN también tienen una buena valoración por parte de los jueces humanos en DUC.

KING frente a correlación con juicios humanos

Si tomamos todas las posibles combinaciones de métricas de similitud, podemos estudiar la relación existente entre los valores de KING y la semejanza a rankings producidos por juicios humanos. La figura 6.13 muestra esta relación. Cada uno de los puntos representa una combinación de métricas. El eje vertical representa la correlación Pearson entre el ranking producido por la combinación de métricas y el ranking producido por los juicios humanos en el DUC. Como puede verse, valores altos en KING garantizan una alta correlación con juicios humanos. Además, combinaciones de alto KING obtienen una correlación superior o sensiblemente superior a la del mejor marco ROUGE, y muy superior al peor ROUGE.

JACK: Heterogeneidad de los resúmenes automáticos

Mediante la medida JACK podemos analizar la heterogeneidad de los resúmenes automáticos. La figura 6.14 muestra como crece el valor JACK, tomando como métricas la combinación de máximo KING, Rpre-W y TVM.512, a medida que aumentamos el número de resúmenes automáticos. En ambas tareas los resultados tienden a estabilizarse a partir de cierto número de resúmenes, lo que indica cierta redundancia en las aproximaciones seguidas por los sistemas participantes.

Por otro lado, en la tarea 2 (resúmenes genéricos) el conjunto de resúmenes automáticos es más heterogéneo que en la tarea 5 (resumen orientado a pregunta). Es decir, se han empleado estrategias más diversas en la generación automática de resúmenes genéricos, mientras que en la tarea 5, probablemente, la pregunta haya centrado los sistemas en un mismo tipo de estrategia.

QUEEN sobre métricas aisladas

El modelo QARLA permite evaluar los resúmenes automáticos en relación a diferentes métricas de similitud. Si la métrica o el conjunto de métricas aplicado tiene un mayor KING, entonces representa un rasgo que caracteriza a los resúmenes manuales frente a resúmenes automáticos, considerándose más fiable como criterio de evaluación.

Pero, además de evaluar un sistema de resumen, QARLA permite analizar las propiedades de los resúmenes evaluados. Dado que QUEEN es independiente de la escala de las métricas, podemos comparar los valores de QUEEN obtenidos

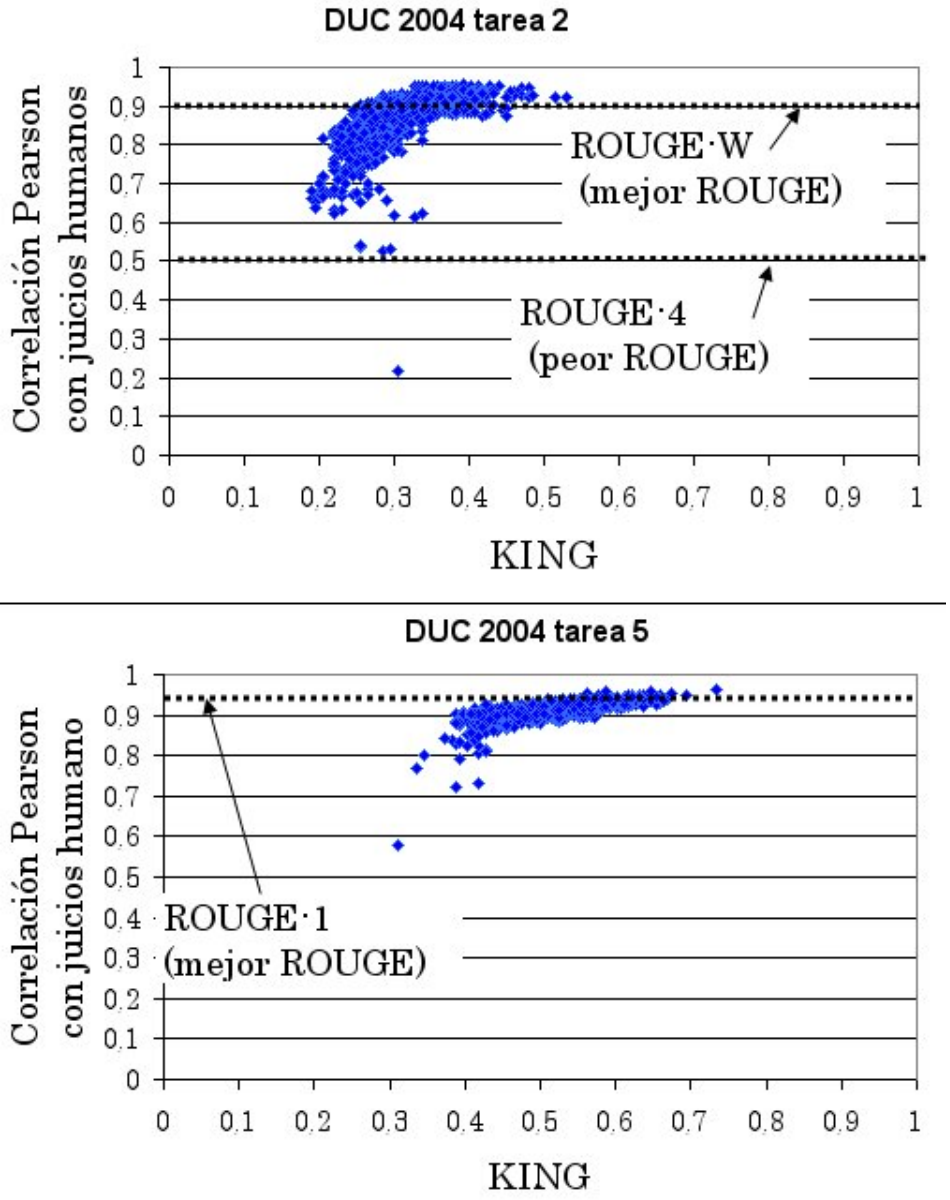


Figura 6.13: Correlación entre evaluación en DUC y QARLA sobre distintas combinaciones de métricas

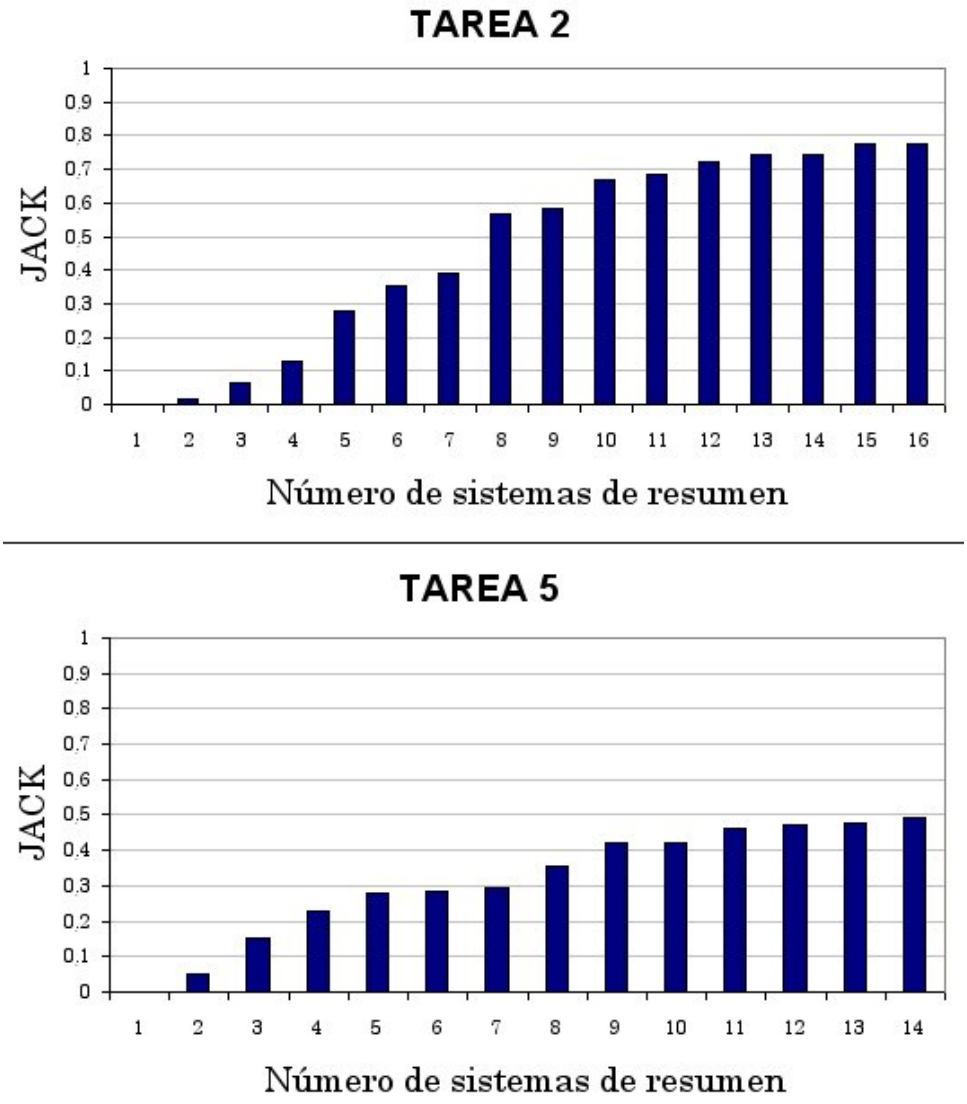


Figura 6.14: JACK frente a número de resúmenes automáticos en DUC

por los sistemas en relación a distintas métricas de similitud, siendo cada métrica representativa de algún rasgo de los resúmenes.

Las figuras 6.15 y 6.16 muestran la calidad de los resúmenes automáticos en relación a las 12 métricas de similitud seleccionadas, extraídas mediante el proceso de agrupamiento descrito. Los valores de QUEEN más elevados aparecen marcados en negrita. Los sistemas de resumen (eje vertical) están ordenados según el ranking generado manualmente en DUC. Como puede verse, en ambas tareas los primeros sistemas obtienen buena puntuación según la mayoría de los rasgos orientados a contenidos (métricas R, Rpre y TVM).

A partir de estos datos, si nos fijamos en los valores promedio (última fila) de QUEEN sobre distintas métricas de similitud, podemos extraer algunas conclusiones:

- Los resúmenes automáticos generados en el DUC no coinciden en longitud de frases con los resúmenes modelo. La puntuación promedio sobre la métrica $AVLS$ (0.26 y 0.22) es bastante reducida en relación al valor obtenido por el resto de las métricas.

- Los resúmenes automáticos son capaces de identificar los términos más relevantes (valores de QUEEN más altos en TVM.1), mientras que no identifican elementos menos frecuentes de los documentos originales pero que sí son comunes en los resúmenes modelo (valores de QUEEN más bajos en TMV.512).

- El estilo lingüístico de los resúmenes automáticos es especialmente diferente al de los modelos en la tarea 5 (resumen orientado a pregunta) frente a la tarea 2 (resumen genérico). Presumiblemente los resúmenes manuales orientados a pregunta tienen unos rasgos de estilo más definidos.

- Los resúmenes automáticos se asemejan a los modelos en secuencias de palabras largas (valores altos en R-3,R-4) en la misma medida que los modelos entre sí, especialmente en los resúmenes de tipo genérico (tarea 2).

6.8. Recapitulación

Evaluación de resúmenes en QARLA

La automatización del proceso de evaluación de resumenes consiste en calcular la similitud entre el resumen evaluado y un conjunto de resúmenes modelo generados por sujetos de prueba. Dado que se evalúa en relación a un conjunto de modelos, la similitud al conjunto debe calcularse mediante algún tipo de promediado. Esta es la estrategia llevada a cabo por las métricas estandar como ROUGE [LH03a]. Sin embargo, el promediado sobre los modelos conlleva dependencia sobre las propiedades de escala de las métricas. El modelo QARLA permite considerar conjunto de modelos sin necesidad de depender de dichas propiedades de escala.

	R-2.c	R-3.c	R-4.c	Rpre-W.b	Rpre-2.b	Rpre-3.b	Rpre-4.b	TVM.1	TVM.8	TVM.512	AVLS	GRAMSIM
NSA	**0,48**	**0,62**	**0,85**	0,19	**0,45**	**0,5**	**0,7**	**0,52**	**0,46**	0,35	0,28	**0,53**
U. Sheffield	**0,48**	**0,64**	**0,85**	0,16	0,4	**0,48**	**0,68**	**0,48**	0,33	0,26	0,28	**0,51**
Columbia U.	0,39	**0,6**	**0,85**	0,21	0,39	**0,49**	**0,69**	**0,52**	0,43	0,28	0,2	**0,55**
KU Leuven	0,39	**0,6**	**0,83**	0,2	0,4	**0,49**	**0,65**	**0,54**	0,44	0,33	0,19	0,33
ISI	0,4	**0,59**	**0,85**	0,15	0,37	**0,46**	**0,67**	**0,53**	0,4	0,29	0,27	**0,47**
CRL/NYU	**0,45**	**0,65**	**0,86**	0,41	**0,52**	**0,6**	**0,75**	**0,51**	0,39	0,19	0,2	0,38
Concordia U.	0,37	**0,58**	**0,84**	0,14	0,3	0,37	**0,62**	**0,58**	**0,49**	0,33	0,32	**0,51**
LCC	**0,48**	**0,66**	**0,85**	0,14	0,43	**0,49**	**0,67**	**0,46**	0,39	0,31	0,17	**0,48**
Laris Labs	0,43	**0,63**	**0,85**	0,22	0,38	**0,45**	**0,66**	**0,55**	**0,48**	0,35	0,32	**0,52**
Fudan U.	0,4	**0,59**	**0,84**	**0,26**	0,36	**0,44**	**0,67**	**0,59**	**0,53**	0,4	0,43	**0,55**
CL Research	0,41	**0,63**	**0,85**	0,21	0,41	**0,51**	**0,71**	0,44	0,3	0,15	0,1	0,34
NTT	0,23	**0,53**	**0,84**	0,22	0,25	0,35	**0,61**	**0,52**	0,36	0,16	0,31	0,31
U. Lethbridge	0,21	**0,52**	**0,83**	0,06	0,2	0,32	**0,61**	**0,51**	0,44	0,21	0,13	0,35
U. Michigan	0,36	**0,61**	**0,85**	0	0,3	**0,44**	**0,66**	**0,52**	0,4	0,3	0	0,3
Microsoft	0,25	**0,54**	**0,83**	0,15	0,24	0,34	**0,62**	**0,52**	0,36	0,2	0,44	0,39
U. Ottawa	0,05	0,46	**0,82**	0,02	0,03	0,21	**0,57**	0,38	0,24	0,21	**0,51**	0,13
Average	0,361	0,591	0,843	**0,171**	0,339	0,434	0,659	0,511	0,403	**0,27**	**0,26**	0,4156
	R-2.c	R-3.c	R-4.c	Rpre-W.b	Rpre-2.b	Rpre-3.b	Rpre-4.b	TVM.1	TVM.8	TVM.512	AVLS	GRAMSIM

Figura 6.15: Calidad de los resúmenes automáticos según QARLA sobre métricas individuales en la tarea 2 del DUC 2004

	R-2.c	R-3.c	R-4.c	Rpre-W.b	Rpre-2.b	Rpre-3.b	Rpre-4.b	TVM.1	TVM.8	TVM.512	AVLS	GRAMSIM
LCC	**0,25**	**0,44**	**0,67**	0,16	0,19	**0,32**	**0,45**	**0,47**	0,29	0,17	0,29	0,32
KU Leuven	0,23	0,4	0,65	0,15	0,21	0,3	0,43	**0,57**	0,4	**0,26**	0,36	0,21
U. Michigan	0,2	0,36	0,64	0,2	0,21	0,28	0,41	**0,54**	0,33	0,18	0,11	0,26
NSA	0,29	0,41	0,65	0,09	0,2	0,26	0,39	**0,54**	0,32	0,23	0,38	0,32
Laris Labs	0,24	0,42	0,66	0,13	0,17	0,27	0,41	**0,57**	0,44	0,22	0,35	**0,41**
Columbia U.	0,29	**0,44**	0,66	0,13	**0,23**	0,29	0,41	0,38	0,23	0,17	0,38	0,31
Concordia U.	**0,25**	**0,44**	0,66	0,14	0,18	0,28	0,42	**0,45**	0,29	0,18	0,34	0,31
U. Lethbridge	0,23	0,41	0,65	0,14	0,18	0,26	0,41	**0,46**	0,25	0,15	0,21	0,35
Fudan U.	0,22	0,38	0,64	0,12	0,17	0,24	0,37	0,41	0,27	0,2	0,42	0,33
U. Sheffield	0,23	0,39	0,65	0,03	0,16	0,23	0,39	0,5	0,31	0,18	0,17	0,37
CRL/NYU	0,22	0,4	0,63	**0,25**	**0,23**	0,32	0,42	**0,51**	0,29	0,08	0,26	0,18
U. Ottawa	0,13	0,3	0,62	0,13	0,09	0,16	0,3	0,24	0,11	0,08	**0,39**	0,19
CL Research	0,19	0,37	0,64	0,12	0,15	0,23	0,38	0,37	0,22	0,09	0,28	0,27
ISI/USC	0,18	0,31	0,62	0,07	0,12	0,17	0,33	**0,59**	**0,45**	**0,26**	0,37	0,31
Average	**0,13**	0,209	0,286	0,3229	**0,15**	0,227	0,3	0,336	0,37	**0,301**	**0,22**	**0,1797**
	R-2.c	R-3.c	R-4.c	Rpre-W.b	Rpre-2.b	Rpre-3.b	Rpre-4.b	TVM.1	TVM.8	TVM.512	AVLS	GRAMSIM

Figura 6.16: Calidad de los resúmenes automáticos según QARLA sobre métricas individuales en la tarea 5 del DUC 2004

Otra limitación de la aplicación directa de medidas de similitud es la combinación de métricas. El único modo de combinar métricas en una única métrica consiste en asignar cierto peso a cada una de las métricas y aplicando alguna técnica de promediado. De nuevo, esta estrategia hace depender a la evaluación de las propiedades topológicas de las métricas. En el marco QARLA, no es necesario establecer a priori un peso a cada una de las métricas de queremos combinar, y el resultado es independiente de la escala de las métricas combinadas. El único aspecto que debe ser analizado en QARLA es cuál es la mejor combinación de métricas. La medida KING permite estimar la fiabilidad de cualquier combinación de métricas de similitud.

Como limitación del modelo QARLA frente a otros marcos de evaluación automática, podríamos plantear que QARLA es incapaz de comparar entre sí resúmenes de muy baja calidad. Es decir, resúmenes que se encuentran demasiado alejados de los modelos. Este caso, todos los resúmenes tendrían calidad QUEEN nula y resulta muy sencillo encontrar combinaciones de métricas de máximo KING. Sin embargo, el bajo valor de la medida JACK obtenida nos permitiría detectar el problema. La opción más correcta en estos casos es, o bien optimizar los sistemas y evaluar de nuevo, o bien, si se desea realizar una evaluación comparativa entre sistemas, aplicar un marco de evaluación al estilo de ROUGE.

Meta-evaluación de métricas en QARLA

Dentro del el dominio de los sistemas de resumen automático, se ha abordado desde diferentes perspectivas la cuestión de cuál es la mejor métrica automática para evaluar sistemas (ver apartado 4.1.3). La metodología más cercana al marco QARLA es sin duda ORANGE [Lin04a], aplicada en en problema de la evaluación de traducciones. ORANGE considera la posición promedio de los modelos en un ranking generado por la métrica a evaluar. Al igual que en QARLA, ORANGE es capaz de evaluar automáticamente las métricas en función de su capacidad de discriminar resúmenes modelo respecto de resúmenes automáticos, sin necesidad de juicios humanos. Además, ORANGE es independiente de la escala de las métricas de evaluación evaluadas.

Sin embargo, como se ha mostrado a lo largo de este capítulo, el marco QARLA satisface algunas condiciones que no son satisfechas en ORANGE:

- Es capaz de combinar métricas de similitud, sin necesidad de establecer a priori el peso de cada una de ellas.

- QARLA no se ve afectada por elementos repetidos en el conjunto de resúmenes automáticos de muestra. En ORANGE, introducir elementos repetidos afecta a la posición que ocupan los modelos en los rankings.

- Dispone de una medida $JACK$, que permite comprobar si el conjunto de resúmenes automáticos de muestra son lo suficientemente representativos de las aproximaciones automáticas posibles.

Probablemente la ventaja más significativa respecto a ORANGE es la capacidad de QARLA de combinar métricas de similitud. Creemos que una evaluación debe capturar diferentes aspectos de los resúmenes evaluados. ORANGE, sin embargo, posee algunas ventajas frente a QARLA.

Aplicación de QARLA en DUC 2004

Sobre el corpus de resúmenes del DUC 2004, hemos podido corroborar algunas propiedades del marco de evaluación QARLA:

- El marco QARLA permite identificar y combinar criterios de similitud que caracterizan los rasgos propios de los resúmenes generados por humanos. Al aplicar dichas métricas, obtenemos una mayor calidad para los resúmenes manuales frente a resúmenes generados automáticamente (figura 6.11).
- La combinación de métricas en QARLA permite una mayor capacidad de discriminación entre resúmenes manuales y automáticos que métricas individuales (ver figura 6.10).
- QARLA permite agrupar metricas de similitud en función de los rasgos que representan. Esta funcionalidad de QARLA puede ser muy útil en la caracterización del problema del resumen automático (figura 6.9).
- Aunque en teoría no es una condición necesaria, los resultados muestran una alta correlación entre juicios humanos y juicios de QARLA sobre métricas de alto KING (ver figuras 6.13 y 6.12). Esto implica que los juicios humanos están relacionados con la capacidad de los sistemas de emular resúmenes manuales, que es el principio en el que se basa QARLA.
- El marco QARLA permite determinar la heterogeneidad de las aproximaciones automáticas en una determinada tarea de resumen (ver figura 6.14).
- QARLA permite comparar la calidad de un mismo sistema en relación a diferentes rasgos (ver figuras 6.15 y 6.16). Es decir, es posible determinar en qué aspectos un mismo sistema presenta más deficiencias en relación a los resúmenes modelo.

En general, aplicado sobre los datos del DUC-2004, los resultados de QARLA se muestran coherentes con las características reales de los sistemas, por lo general, extractivos y basados en estrategias superficiales. Los resultados son también coherentes con las diferencias entre la tarea de generación de un resumen genérico y un resumen orientado a pregunta.

En el siguiente capítulo empleamos el marco QARLA para analizar los rasgos comunes en los informes manuales de ISCORPUS, para evaluar estrategias de predicción de distribuciones de conceptos clave en un informe manual, y comprobaremos que medidas basadas en la identificación de conceptos clave pueden aportar información útil en el proceso de evaluación.

Parte III

Desarrollo de un modelo interactivo de SI

Capítulo 7

Estudio del papel de los conceptos clave en Síntesis de Información

Hemos definido la Síntesis de Información como el proceso de extraer, organizar e interrelacionar piezas de información contenidas en un conjunto de documentos relevantes con el fin de obtener un informe elaborado que satisfaga una necesidad de información. La "pieza de información" es un concepto vago. Una pieza de información puede ser un documento, un fragmento de texto o una frase, pero también un término o un "concepto" es una pieza de información. En nuestro caso, centrados en el dominio periodístico, consideramos como conceptos clave al conjunto de entidades (personas, organizaciones..) o factores (por ejemplo, "crisis" o "relaciones internacionales") con alto protagonismo en el asunto tratado en las fuentes.

Uno de los problemas a los que se enfrenta un sistema de acceso a la información es la flexibilidad del lenguaje, es decir, las múltiples formas de expresar una misma información. Existen menos formas de expresar un concepto que la información representada en una frase o párrafo. Por tanto, un concepto es una pieza de información más fácil de tratar por un sistema que el contenido de un fragmento largo de texto. Por su simplicidad, la identificación y análisis de los conceptos clave del tema tratado en las fuentes suponen un primer nivel de análisis de los contenidos en el desarrollo de un sistema de SI.

El análisis de contenidos mediante la identificación de conceptos clave ha sido parcialmente abordado en técnicas de resumen automático mediante la identificación de términos relevantes. En general estas técnicas consideran como términos relevantes a las palabras o secuencias de palabras más frecuentes (sección 2.3.4). Se asume que la presencia de estos términos hace que un fragmento sea más susceptible de pertenecer al resumen.

En principio podríamos extrapolar estas estrategias al problema de la SI. Sin embargo, en la tarea de SI, aparece mucha información redundante en las fuentes, y además, el informe resultante es un texto largo, en el que se debe abordar múltiples aspectos del asunto tratado. Es decir, es especialmente importante la eliminación de redundancias y la distribución de contenidos en el informe final. Esto nos lleva a plantearnos no solo si la aparición de conceptos clave en un fragmento denota

relevancia, sino de qué manera se distribuyen estos conceptos a lo largo de todo el informe.

En este capítulo analizaremos el papel que juegan los conceptos clave en el desarrollo y evaluación de sistemas de SI, abordando sucesivamente las siguientes cuestiones:

- **¿Es la distribución de conceptos clave un rasgo distintivo de un informe generado manualmente?** En el contexto de las tareas de resumen multi-documento en general, una aproximación sencilla y relativamente eficiente es la extracción de las primeras frases o párrafos de cada uno de los documentos. En este capítulo, veremos que mediante este tipo de aproximaciones es posible obtener una buena cobertura sobre palabras o frases contenidas en los informes modelo. Sin embargo, su distribución de conceptos clave no se asemeja a las distribuciones en informes modelo. Es decir, podemos mejorar estas aproximaciones si consideramos la distribución de conceptos clave. Dicho de otro modo, la distribución de conceptos clave es un rasgo común compartido por los modelos que les distingue de aproximaciones automáticas básicas. Por tanto, la consideración de conceptos clave es un aspecto necesario en el desarrollo de un sistema de SI.

- **¿Pueden extraerse automáticamente los conceptos clave a partir de los documentos originales?** Para que un sistema pueda hacer uso de los conceptos clave en la elaboración de informes o al asistir a un usuario, es necesario que previamente el sistema sea capaz de identificar dichos conceptos clave a partir de los documentos originales. En este capítulo veremos que la extracción automática de conceptos clave a partir de un conjunto de documentos puede optimizarse considerando la frecuencia con que aparecen en la posición inmediatamente anterior al verbo, superando en nuestro contexto a otras aproximaciones estadísticas más sofisticadas.

- **¿Existe alguna relación entre la distribución de los conceptos clave en las fuentes originales y en el informe?** Disponiendo de una herramienta para la identificación automática de conceptos clave, el siguiente paso consiste en calcular la distribución estos conceptos clave en las fuentes. Si existe alguna relación entre la distribución de los conceptos en los documentos originales y en un informe, podríamos predecir entonces la frecuencia con que debe aparecer cada uno de los conceptos en el informe final. En este capítulo, veremos que efectivamente existe esta relación.

- **¿Puede la distribución de conceptos clave aportar información útil en el proceso de evaluación de sistemas de Síntesis de Información?** En el dominio de la evaluación de resúmenes, se han empleado criterios como la cobertura de n-gramas o frases sobre resúmenes modelo. En este capítulo, veremos que considerar además los conceptos clave que aparecen en los documentos originales puede aumentar la fiabilidad del proceso de evaluación.

7.1. Necesidad de conceptos clave

En el contexto de la generación automática de resúmenes y más concretamente la Síntesis de Información, en algunos casos no es fácil mejorar los resultados obtenidos con una aproximación básica. Por ejemplo, la selección de las primeras frases de los documentos, sobre todo en el dominio de artículos periodísticos, representa en principio una técnica robusta y en algunos casos no menos eficiente que otras más sofisticadas(ver apartado 2.3.1). Sin embargo, es evidente que esta técnica no refleja el proceso cognitivo de elaboración de un informe. Es necesario por tanto identificar las deficiencias de los informes generados por este tipo de aproximaciones. Es decir, nos planteamos qué tienen en común los informes modelo que no tienen estos informes automáticos.

Con el fin de abordar esta cuestión, aplicamos distintas métricas de similitud basadas en diferentes rasgos, y analizamos qué rasgos distancian más a los modelos de los informes automáticos. Los resultados descritos en este apartado muestran que una aproximación básica puede generar un informe semejante a informes modelo en cuanto a frases o vocabulario seleccionado. Sin embargo, los resultados sugieren que la distribución de conceptos clave en los informes es un rasgo que distingue a los informes modelo de estas aproximaciones.

Para analizar el problema, estudiaremos la proximidad de los informes generados por estrategias básicas a informes modelo generados por sujetos en ISCORPUS.

7.1.1. Definición del experimento

Métricas de similitud empleadas

Para estudiar los conceptos clave como rasgo común entre informes, hemos de definir un conjunto de métricas de similitud representativas de los rasgos que queremos comparar. En este experimento estudiamos tres rasgos de los informes: frases contenidas en los informes, vocabulario y distribución de conceptos clave. Para estudiar la selección de frases como rasgo de los informes, empleamos la co-selección de frases orientada a precisión descrita en el apartado 4.1.2. Para representar la distribución de palabras como rasgo de los resúmenes empleamos la métrica de similitud R-1, derivada de la métrica de evaluación ROUGE y descrita en el apartado 6.7.1.

Para estudiar la distribución de conceptos clave como rasgo de los informes, definimos una medida NICOS que toma los conceptos clave de ISCORPUS extraídos manualmente (ver apartado 5.4). NICOS crea un vector para cada uno de los resúmenes que contiene la frecuencia relativa de cada uno de los conceptos clave identificados. El proceso de cálculo de NICOS es el que sigue:

1. Tomamos la lista de conceptos clave anotados en ISCORPUS ya agrupados (sección 5.4).

2. Para cada conjunto de conceptos clave equivalentes, anotamos el número de veces que aparece cualquiera de ellos en el resumen. Para facilitar el ajuste

entre conceptos representados mediante expresiones complejas, lematizamos y eliminamos mayúsculas tanto de los conceptos como del contenido de los informes. Obtenemos así un vector de frecuencias cuya longitud viene dada por el número de conceptos clave considerados.

3. Dividimos la frecuencia obtenida por la longitud del informe en términos de número de palabras.

4. Dados dos informes, calculamos la distancia NICOS entre ambos mediante el cálculo de la distancia euclídea entre sus respectivos vectores.

Estrategias automáticas básicas de SI

ISCORPUS contiene un conjunto de 32 informes por cada tema generados mediante distintas estrategias automáticas simples (ver sección 5.5). Estas estrategias pueden ser clasificadas en:

- Estrategias basadas en la selección de primeras frases: (aproximaciones 1..4, y 10..22).
- Estrategias basadas en la selección de documentos relevantes (aproximaciones 7..9)
- Estrategias basadas en la extracción de primeras dos frases de cada documento (aproximaciones 23..26).
- Estrategias basadas en la extracción de primeras 3 frases (aproximaciones 27..32)
- Estrategias basadas en la extracción de todas las frases de un documento (aproximaciones 5..7).

Normalización de métricas de similitud mediante QUEEN

El objetivo de este experimento es identificar las deficiencias de cada una de estas estrategias básicas de SI. Disponemos de las métricas de similitud descritas en el apartado anterior, que reflejan diferentes rasgos de los informes. Una primera forma de abordar esta cuestión consistiría en medir la similitud entre los informes generados mediante estas estrategias y los informes modelo generados en ISCORPUS en base a distintas métricas. Sin embargo, cada métrica de similitud posee sus propias propiedades de escala, luego los valores obtenidos no serían comparables entre sí.

Para solucionar este problema aplicamos dichas métricas mediante la medida QUEEN (sección 6.3). QUEEN permite aplicar métricas de similitud para calcular la distancia a un conjunto de elementos (informes modelo) con independencia de las propiedades de escala de la métrica. Por ejemplo, un valor QUEEN de 0.5 aplicado sobre una métrica implica que el informe se asemeja tanto a los modelos como

los modelos entre si. Por tanto, sabríamos que según dicha métrica de similitud, el informe automático no se distingue de un informe generado manualmente.

Aplicamos entonces la medida de evaluación QUEEN sobre cada uno de estos informes automáticos y sobre cada una de las métricas descritas en el punto anterior. De esta forma, identificaremos cuáles son los rasgos que distinguen a las estrategias básicas de los informes modelo generados manualmente.

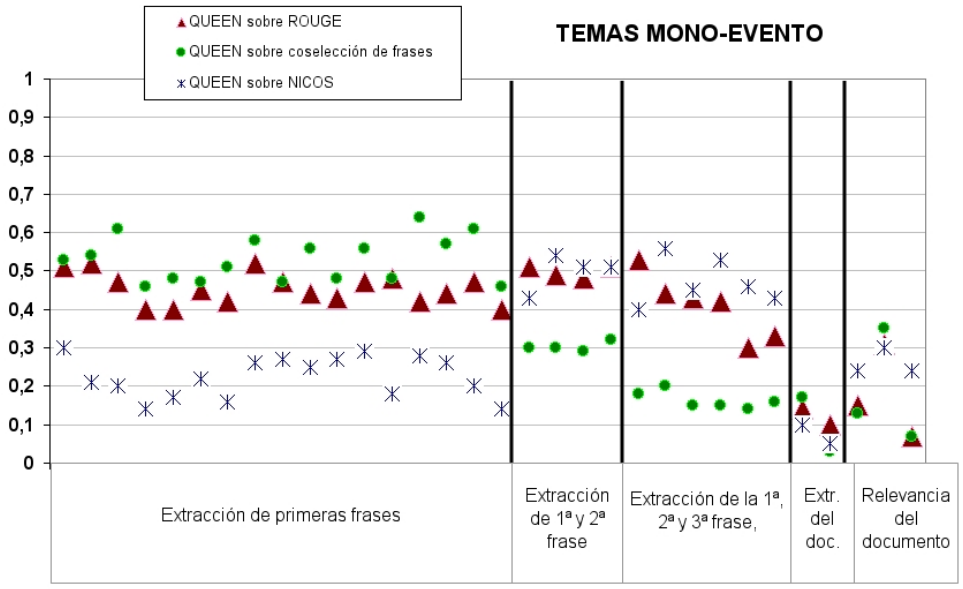


Figura 7.1: QUEEN sobre NICOS frente a Precisión de frases y R-1 en temas mono-evento

7.1.2. Resultados

Temas mono-evento

La figura 7.1 muestra los valores de QUEEN sobre NICOS, coselección de frases y R-1, aplicados entre los informes automáticos y los informes modelo para temas mono-evento. El eje horizontal representa las distintas estrategias básicas de generación de informes. El eje vertical representa el valor QUEEN promedio, sobre los seis temas, obtenido para los informes generados por cada estrategia.

Las estrategias (eje horizontal) aparecen agrupadas según las categorías descritas anteriormente. Las estrategias basadas en la selección de primeras frases (primer conjunto) generan informes similares a los informes modelo en relación a las frases seleccionadas y el vocabulario empleado (QUEEN $\approx 0{,}5$) . Sin embargo, estas estrategias no producen informes con una distribución de conceptos clave apropiada ($\text{QUEEN}_{\text{NICOS}} < 0{,}3$). Esto se debe al hecho de que, aunque las primeras frases de los documentos aparecen con frecuencia en los informes modelo, estas frases poseen una densidad de conceptos clave mayor que la de los informes modelo en general.

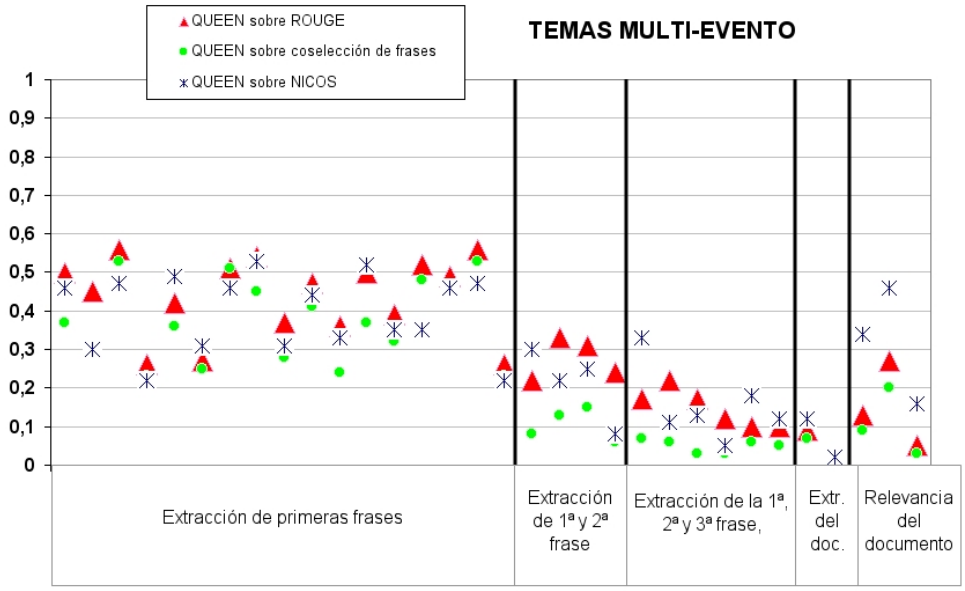


Figura 7.2: QUEEN sobre NICOS frente a precisión de frases y R-1 en temas multi-evento

Las aproximaciones basadas en selección de primeras y segundas frases del documento (segundo conjunto), generan informes similares a los modelos en cuanto a distribución de conceptos clave (NICOS) y vocabulario empleado (R-1), pero con una baja concordancia en las frases seleccionadas (coselección de frases). Como es de esperar, la extracción de documentos completos no permite generar informes semejantes a los modelos en relación a ninguna de las características.

Por último, las aproximaciones basadas en la identificación de documentos relevantes (quinto conjunto) obtienen en general valores bajos de QUEEN para cualquiera de las métricas. Posiblemente esto se deba a que el motor de búsqueda empleado para la identificación de documentos relevantes selecciona documentos similares entre sí, reduciendo por tanto la cobertura del informe final en cuanto a contenidos.

En resumen, en el caso de los temas mono-evento, ninguna de las estrategias básicas generan informes semejantes a los modelos en base a todas las características a la vez. Por ejemplo, la selección de frases y vocabulario se optimiza mediante la extracción de primeras frases de documentos, pero sería necesario considerar también la distribución de conceptos clave para generar informes semejantes a los modelos.

Temas multi-evento

Los resultados obtenidos en el caso de los temas multi-evento, difieren de los obtenidos en mono-evento. La figura 7.2 muestra que las diferentes métricas de similitud introducidas en QUEEN, coselección de frases, ROUGE y NICOS, se

comportan de manera semejante. Es decir, en el caso de los temas multi-evento, la distribución de conceptos clave no es un aspecto complementario en relación a la coselección de frases o vocabulario empleado.

En cuanto al comportamiento de las distintas estrategias automáticas de generación de informes, las estrategias basadas en la selección de primeras frases del documento iguala o supera a otras estrategias sea cual sea el criterio de similitud escogido. Ésta es una consecuencia de cómo se distribuyen los contenidos en temas muti-evento. En este tipo de temas, los documentos describen diferentes sucesos que suelen aparecer resumidos en la primera frase del artículo.

En general, las estrategias automáticas de elaboración de informes en temas multi-evento tienden a obtener valores QUEEN menores que en el caso de temas mono-evento. Como se mostró en la sección 5.6 el grado de acuerdo en cuanto a selección de frases entre sujetos es menor en temas multi-evento. Esto implica más similitud entre informes manuales y, como consecuencia, valores pequeños en QUEEN.

En definitiva, estrategias básicas de selección de frases pueden generar informes semejantes a los modelos en cuanto a frases seleccionadas o vocabulario empleado tanto en temas mono-evento como multi-evento. Sin embargo, en temas mono-evento, encontramos diferencias en cuanto a la distribución de conceptos clave en los informes. Por tanto, en este tipo de temas, los conceptos clave deben de ser considerados para superar estrategias automáticas simples.

7.2. Extracción automática de conceptos clave

El experimento descrito en la sección anterior muestra que la distribución de conceptos clave en los informes es un aspecto que debe de ser considerado al abordar el problema de la Síntesis de Información. Este aspecto solo puede ser tratado por un sistema si previamente ha sido capaz de identificar cuáles son los conceptos clave a partir del conjunto de documentos originales. Por simplicidad, entendemos el problema de identificación de conceptos clave como la extracción de una lista de términos con amplia cobertura sobre los terminos con que los sujetos representan los conceptos extraídos manualemente en ISCORPUS.

En esta sección, mostraremos que considerar información sintáctica superficial en el proceso de extracción de conceptos clave, mejora los resultados en relación a otras estrategias puramente estadísticas, en el contexto de la tarea de SI.

7.2.1. Trabajos previos

La identificación automática de términos relevantes ha sido empleada en diversos sistemas de resumen. En resumen mono-documento, los términos clave son obtenidos por lo general a partir del título del documento o de la cabecera. [Edm69, PW94, KPC95]. Sin embargo, en resumen multi-documento dentro del cual se puede englobar la Síntesis de Información, es necesario algún tipo de procesamiento para la identificación de términos clave. [LH02, KSH02, SOC$^+$02]. La mayoría de las aproximaciones se basan en criterios estadísticos.

El problema de la extracción de conceptos clave cubre dos cuestiones: ¿Qué tipo de términos deben ser considerados como candidatos? y ¿cuál es el mejor criterio de pesado para seleccionar los mejores candidatos? En cuanto a la primera cuestión, existen varias posibles respuestas. Algunos trabajos consideran sintagmas nominales [BKB+98, JLP02], palabras [BGMP01], n-gramas [LLS03, LH02] o nombres propios, nombres compuestos y abreviaturas [NC99]. En este trabajo nos centramos en términos simples como candidatos.

En cuando a los criterios de pesado, la estrategia más común en sistemas interactivos de resumen mono-documento es el uso de *tf.idf* [JLP02, BGMP01, NC99], el cual favorece términos frecuentes en el documento a resumir y poco frecuentes en el resto de la colección. En el sistema iNeast [LLS03], la identificación de términos relevantes está orientada a resumen multi-documento, y se emplea el ratio "likelihood" [Dun93], que favorece términos representativos del conjunto de documentos en oposición al resto de la colección. En otras aproximaciones se ha empleado información complementaria como por ejemplo, el número de referencias al término [BKB+98], su localización [JLP02], o la distribución del término a lo largo del documento [BGMP01, BKB+98].

7.2.2. Frecuencia de conceptos clave frente a distancia al verbo

La figura 7.3 muestra, para cada uno de los temas incluídos en ISCORPUS, la probabilidad de encontrar una palabra perteneciente a algún concepto de la lista de conceptos clave identificados en ISCORPUS, según la distancia al verbo. El eje horizontal representa la posición relativa en relación al verbo, y el eje vertical representa la probabilidad de que el término pertenezca a la lista de conceptos. La distancia ha sido calculada en términos de número de palabras, eliminando términos de parada. La línea gruesa representa el promedio sobre temas, y la línea horizontal representa la probabilidad promedio obtenida para todas las posiciones, es decir, la probabilidad de encontrar una palabra perteneciente a un concepto clave en cualquier posición de los documentos originales.

Distribución en temas mono-evento

La gráfica muestra algunas tendencias claras en los resultados. La probabilidad crece cuando nos acercamos al verbo, cayendo de forma pronunciada en posiciones posteriores a éste. Para temas mono-evento, la probabilidad de encontrar conceptos clave inmediatamente antes del verbo es un 56 % mayor que el promedio (0,39 antes del verbo y 0,25 en cualquier posición). Este hecho tiene lugar en todos los temas, por lo que no es un efecto del promediado. Este resultado es muy útil, dado que muestra que existe una correlación directa entre la posición de un término en relación al verbo y la importancia del término en el tema tratado en los documentos. Por supuesto, esta distancia al verbo debería de ser adaptada a otras lenguas con diferentes características sintácticas, y validada para distintos dominios.

Distribución en temas multi-evento

Los resultados obtenidos para temas multi-evento son sustancialmente distintos. En temas multi-evento, en general la probabilidad de encontrar un concepto clave es menor, dado que al tratarse de varios eventos no hay un conjunto bien definido de conceptos clave que adquieran protagonismo en todos los textos (ver sección 5.4). Por ejemplo, en el caso del tema "casos de huelgas de hambre", no existen demasiados puntos en común entre las diferentes huelgas que aparecen descritas en los documentos originales. La tendencia a crecer la probabilidad de ocurrencia de conceptos clave antes del verbo, y a disminuir en posiciones posteriores al verbo, aunque se mantiene en este tipo de temas, es menos pronunciada.

7.2.3. Definición del experimento

Los conceptos clave introducidos manualmente en ISCORPUS son en muchos casos expresiones complejas que pueden no estar presentes literalmente en los documentos. Por este motivo nos centramos en la extracción de términos simples que aparezcan en las expresiones introducidas como conceptos clave en ISCORPUS. En el contexto de los sistemas interactivos, asumimos que, si por lo menos uno de los términos del concepto es identificado automáticamente, es posible generar expresiones complejas empleando por ejemplo, estrategias de "exploración de sintagmas" [PnVG02]. En el contexto de sistemas automáticos de resumen, es más sencillo tratar términos simples asociados a conceptos clave que las expresiones introducidas en ISCORPUS.

Medidas de evaluación de resultados: cobertura y ruido

En este experimento, comparamos distintas estrategias de pesado para la identificación automática de términos clave, empleando dos medidas de evaluación; una medida *cobertura* que representa en qué medida los conceptos clave son cubiertos por la lista de términos identificada, y una medida *ruido* que contabiliza el número de términos que no aparecen en ningún concepto clave. Una estrategia óptima debería de tener una buena cobertura y poco ruido. Formalmente:

$$Cobertura = \frac{|\mathcal{C}_l|}{|\mathcal{C}|} \qquad Ruido = |\mathcal{L}_n|$$

donde $\mathcal{C}$ es el conjunto unión de los conceptos clave anotados manualmente; $\mathcal{L}$ representa la lista de términos ordenados por relevancia y extraídos mediante una estrategia de pesado; $\mathcal{L}_n$ es el subconjunto de términos en $\mathcal{L}$ tales que no aparecen en ninguno de los conceptos clave; y $\mathcal{C}_l$ representa el subconjunto de conceptos clave que son representados por al menos uno de los términos incluídos en la lista $\mathcal{L}$. Con el fin de dar más valor a aquellos términos que aparecen en los conceptos más anotados en ISCORPUS, no hemos eliminado elementos repetidos en la lista de conceptos.

Lo que sigue es un ejemplo ficticio de como se comporta *Cobertura* y *Ruido*:

$$\mathcal{C} = \{\text{Haití, restauración de la democracia, ONU and tropas de EEUU }\}$$

$$\mathcal{L} = \{\text{Haití, soldados, ONU, EEUU,tentativa}\}$$

$$\rightarrow \begin{array}{ll} \mathcal{C}_l = \{\textbf{Haití}, \textbf{ONU} \text{ y } \textbf{tropas EEUU}\} & Cob = 2/3 \\ \mathcal{L}_n = \{\text{soldados,tentativa}\} & Ruido = 2 \end{array}$$

Estrategias de pesado de términos

Emplando las medidas *cobertura* y *ruido*, comparamos las siguientes estrategias de pesado de términos:

TF Representa la frecuencia con que aparece el término en los documentos originales, tras eliminar términos de parada (determinantes, nexos, etc).

Likelihood ratio [LLS03] Hemos implementado el proceso descrito en [RG00] empleando únicamente uni-gramas.

OKAPImod Hemos considerado además una medida derivada de OKAPI y empleada en [RWHB$^+$92]. Esta estrategia de pesado está definida para asignar un peso a los términos de un documento. Por ello, hemos adaptado la medida de forma que se considera el conjunto de 100 documentos como un único documento.

TFSYNTAX Basándonos en los resultados descritos en el apartado anterior, TFSYNTAX calcula el peso de cada término como la frecuencia con la que aparece considerando únicamente ocurrencias inmediatamente anteriores a un verbo.

7.2.4. Resultados

La figura 7.4 muestra la relación entre cobertura (eje horizontal) y ruido (eje vertical) para cada uno de los criterios de pesado en temas mono-evento y multievento. Las curvas representan el comportamiento de la extracción de conceptos clave según diferentes criterios de pesado a medida que consideramos más elementos de la lista extraída.

Para todos los criterios de pesado se obtiene curvas semejantes menos para TFSYNTAX, que obtiene los mejores resultados para el caso de temas mono-evento. La curva se desplaza a la derecha, es decir, más cobertura para el mismo nivel de ruido. Nótese que el criterio de pesado TFSYNTAX considera únicamente un 10 % de todo el vocabulario empleado en los documentos originales, es decir, términos que aparecen en posiciones inmediatamente anteriores al verbo.

Con el fin de comprobar en qué medida estos resultados son independientes del tema, y no un efecto del promedio, comparamos la cobertura obtenida para los 50

primeros términos extraídos automáticamente. Hemos escogido la cantidad de 50 términos dado que esta cantidad produce una cobertura aceptable sobre los conceptos clave. 50 términos extraídos mediante TFSYNTAX produce una cobertura del 70 % sobre conceptos clave en temas mono-evento, aumentando solo al 80 % para 100 términos. La figura 7.5 muestra estos resultados para cada uno de los temas. Mediante el uso de TFSYNTAX se obtiene mejores resultados para todos los temas mono-evento (temas TT en la figura). En uno de los dos temas multi-evento el criterio likelihood es sensiblemente superior (temas IE en la figura).

Además de las diferencias existentes entre el criterio de pesado TFSYNTAX y el resto, cabe resaltar que estrategias estadísticas sofisticadas como OKAPI o el ratio likelihood no permiten obtener resultados significativamente mejores que el simple recuento de ocurrencias (TF). Esto se debe posiblemente al gran número de documentos (100) de los que se dispone para cada tema. Sobre cien documentos (unas 1500 frases) una medida estadística sencilla como es el pesado TF puede generar buenos resultados.

Una métrica de similitud sobre conceptos extraídos automáticamente:TFS

Análogamente a la medida NICOS, es posible definir una métrica de similitud entre informes basada en la semejanza entre distribuciones de términos relevantes extraídos automáticamente. Esta medida, dado que no requiere una anotación previa de conceptos, puede ser útil tanto en el desarrollo de sistemas de SI o en la evaluación automática de informes.

Definimos la métrica TFS_n (TFS) basándonos en la frecuencia relativa con que aparecen los n primeros términos de la lista TFSYNTAX en el informe. TFS_n es calculado por medio de los siguientes pasos:

1. Generación de una lista TFSYNTAX de términos relevantes a partir del conjunto total de documentos originales asociados al tema. Tomamos los n primeros términos de la lista

2. Cada informe es representado por un vector de n elementos que contiene la frecuencia relativa en el informe de cada uno de los términos TFSYNTAX

3. La similitud es calculada por medio de la distancia euclídea entre los vectores asociados a cada uno de los informes.

La figura 7.6 muestra la correlación existente entre los valores QUEEN de los informes automáticos calculado sobre NICOS (ver sección 7.1.2), y el QUEEN calculado sobre la métrica de similitud TFS_{64}. Es decir, muestra la semejanza QUEEN entre los informes automáticos y los informes modelo en relación a ambos métricas de similitud. Como puede verse en la figura el comportamiento QUEEN mantiene una buena correlación en el uso de ambas métricas. Esto implica que, sobre el marco QARLA, podemos sustituir NICOS por TFS_{64}, posibilitando el uso de conceptos clave en una evaluación automática de informes.

7.3. Estimación de la distribución de conceptos clave en un informe

En esta sección realizamos un experimento con el fin de estimar la distribución de los conceptos clave en un informe modelo considerando únicamente su distribución en los documentos originales.

Los resultados descritos en la sección 7.1 muestran que los conceptos clave han de ser tenidos en cuenta en la tarea de SI, especialmente en temas mono-evento. Si un sistema es capaz, además de extraer los conceptos, de estimar la distribución adecuada de estos términos en un informe, tendremos un criterio de optimización de estrategias de SI. Dado que es en los temas mono-evento donde se identifica la distribución de conceptos como una deficiencia de los sistemas, en este experimento nos centramos en este tipo de temas.

La figura 7.7 muestra la frecuencia relativa de términos extraídos mediante TFSYNTAX en los informes, en los documentos originales y en la primera frase de los documentos originales en temas mono-evento. El eje horizontal representa la localización del término en la lista generada por TFSYNTAX. El eje vertical representa la frecuencia relativa de dicho término.

En primer lugar, la figura muestra que los términos TFSYNTAX son algo más frecuentes en los informes modelo que en los documentos originales. Es decir, los sujetos tienden a incluir en los informes con más frecuencia conceptos clave que en los documentos originales. Este resultado verifica la hipótesis empleada por algunos sistemas de resumen que otorgan más peso a frases con alta densidad de términos relevantes en el tema. Sin embargo, la figura muestra también que la primera frase de los documentos contiene una mayor densidad de conceptos clave que los informes generados manualmente en ISCORPUS. Este hecho explica la baja correlación entre distribuciones de conceptos clave en los informes modelo y en los informes generados por estrategias basadas en la extracción de la primera frase de los documentos. El asunto es por tanto, cómo predecir la frecuencia relativa de un término de la lista TFSYNTAX en un informe generado manualmente. En este apartado comprobaremos que existe una relación entre la frecuencia de los conceptos clave en los documentos originales y en los informes modelo.

7.3.1. Definición del experimento

Para estudiar la relación entre la distribución de conceptos en documentos e informes, hemos modelado la frecuencia relativa de un término, extraído mediante TFSYNTAX, en un informe manual que resume un conjunto de documentos como una función lineal sobre su frecuencia en los documentos originales.

Sean:

D el conjunto de documentos originales.

t_i el término i de la lista de términos TFSYNTAX. Estos términos poseen una buena cobertura sobre los términos que representan los conceptos clave extraídos manualmente (ver sección 7.2.3).

α el parámetro a estimar.

la función de estimación se expresa como:

$$\text{freq}(t_i, \text{informe}(D)) = \alpha * \text{frec}(t_i, D)$$

Dada esta función de estimación, α representa el ratio de la densidad de conceptos en informes y documentos originales. La idea es ajustar el parámetro α con el fin de obtener una distribución estimada de términos TFSYNTAX similar a la de los informes manuales. El resultado de la estimación será un vector v tal que:

$$v_i \equiv \text{freq}(t_i, \text{informe}(D))$$

Para que esta optimización de α sea posible, es necesario un criterio de evaluación del proceso de estimación. El marco de evaluación QARLA, mediante la medida $\text{QUEEN}_{M,x}(v)$, permite calcular la similitud de v respecto a un conjunto M de vectores de distribución modelo en relación a una determinada métricas de similitud x. Es decir, podemos optimizar las distribuciones v estimadas mediante QUEEN, esto es:

$$\text{QUEEN}_{x,M}(v) \equiv P(x(v, m) \geq x(m', m''))$$

donde el conjunto M incluye los vectores de distribución de términos TFSYNTAX en los informes modelo. QUEEN representa en este caso la probabilidad de que el vector de frecuencias estimado se asemeje tanto al vector de frecuencias de un informe modelo como dos de éstos entre sí.

Definimos estos vectores de frecuencia considerando los primeros 64 elementos de la lista TFSYNTAX. Esta cantidad de términos ofrece una cobertura en torno al 70 % sobre los conceptos clave anotados en ISCORPUS (ver sección 7.2.3). Como métrica x de similitud entre vectores tomamos como la inversa de la distancia euclídea entre vectores, es decir TFSYNTAXSim (ver sección 7.2.4).

7.3.2. Resultados

La figura 7.8 muestra la relación entre la calidad de la estimación ($\text{QUEEN}_{x,M}(v)$) y el parámetro a estimar α. Cada curva representa uno de los temas mono-evento. Podemos ver que para todos los temas, el valor óptimo para α está localizado entre 1,2 y 1,3.

La figura 7.7, ya descrita, muestra que la distribución de los términos TFSYNTAX en los documentos originales es muy similar a la distribución en los informes modelo, aunque sensiblemente inferior. La cuestión que nos planteamos es si se puede modelar mediante el parámetro α' esta relación. La figura 7.9 muestra que, fijando el parámetro α en 1.3, la distribución v estimada obtiene valores mayores en QUEEN que la distribución en los documentos originales, y por tanto, se asemeja más aún a la distribución en los informes modelo.

En resumen, existe una relación lineal entre la distribución de los conceptos clave en los documentos originales y su distribución en los informes modelo, aunque

sería necesario aplicar este mismo experimento en un conjunto mayor de temas, y para informes de diferente longitud. En cualquier caso, estos resultados sugieren que la estimación de la distribución de conceptos clave es una estrategia potencialmente útil en el desarrollo de un sistema de SI, dado que ofrece un criterio más de optimización del informe generado automáticamente.

7.4. Uso de conceptos clave en la evaluación automática de la Síntesis de Información

En esta sección describimos un experimento que muestra que la distribución de conceptos clave en los informes es un rasgo que puede aportar fiabilidad al proceso de evaluación automática.

En general, la evaluación automática de resúmenes se basa en el cálculo de la similitud entre el resumen evaluado y un conjunto de modelos (ver sección 4.1.2). Esta similitud puede ser calculada en relación a diversos rasgos de los resúmenes. El marco QARLA evalúa los informes en función de su semejanza a informes modelo, y permite combinar diferentes criterios de similitud con el fin de encontrar aquellos que caracterizan a los informes modelo respecto a informes automáticos (ver capítulo 6).

La distribución de conceptos clave no es información suficiente para evaluar un informe. Por ejemplo, la figura 7.1 muestra que algunas estrategias generan informes con una buena distribución de conceptos clave (representado por QUEEN sobre NICOS), pero con escaso solapamiento de frases o vocabulario respecto a los informes modelo. El objetivo es por tanto, comprobar si la consideración de conceptos clave puede contribuir al proceso de evaluación de informes.

7.4.1. Definición del experimento

En principio, el modo más directo de introducir un rasgo nuevo en el proceso de evaluación consiste en integrar dicho aspecto en la métrica de evaluación empleada. Para ello es necesario asignar un peso a cada uno de los aspectos integrados. Sin embargo, el marco de evaluación QARLA permite combinar sin necesidad de criterios de pesado distintas métricas de similitud. Podemos por tanto integrar dentro del conjunto de métricas empleadas la métrica TFS.64 calculada sobre conceptos clave extraídos automáticamente (ver sección 7.2.4), y estudiar la fiabilidad de las métricas mediante la medida KING (seción 6.4).

En el siguiente experimento abordaremos tres cuestiones.

1. Comprobar si el valor KING asociado a las métricas individuales se ve incrementado al añadirle TFS.64, es decir, al considerar la distribución en los informes de los 64 términos mas relevantes extraídos mediante TFSYNTAX. Dado que KING es una medida que cuantifica la fiabilidad de un conjunto de métricas, este resultado implica que desde la perspectiva del marco QARLA, considerar la distribución de conceptos clave en los informes evaluados, incrementa la fiabilidad del proceso de evaluación.

2. Comprobar si la mejor combinación de métricas de similitud incluye métricas basadas en la identificación de conceptos clave, como NICOS (sección 7.1.1), o variantes de TFS (sección 7.2.4). Métricas que aportan información en el proceso de evaluación incrementan el valor de KING. Por otro lado, métrica que no aportan información, hacen decrementar el valor de KING cuando se dispone de un número finito de modelos. Por tanto, el hecho de que la mejor combinación incluya métricas basadas en la distribución de conceptos clave es otro indicio de que estos rasgos aportan fiabilidad al proceso de evaluación.

3. Siguiendo la metodología propuesta en el modelo QARLA, es necesario comprobar que, para las métricas evaluadas, los informes automáticos de test conformen un conjunto lo suficientemente heterogéneo para asegurar la fiabilidad de la medida KING . Es decir, hemos de aplicar la medida JACK (ver sección 6.5). En este experimento comprobaremos que la unión de los diversos tipos de estrategias básicas de generación automática de informes conforma un conjunto heterogéneo de informes automáticos de muestra sobre el que aplicar KING.

Métricas de similitud empleadas

Para realizar este análisis hemos considerado un conjunto variopinto de métricas de similitud. Hemos empleado métricas ya descritas en este libro, como son R-1 (sección 6.7.1), NICOS (sección 7.1.1), variantes de TFS (sección 7.2.4) y precisión de frases (sección 4.1.2).

Además, introducimos en este experimento la métrica DocSim. Esta métrica consiste en calcular, análogamente a la precisión de frases, la precisión en los documentos empleados para la extracción de frases incluídas en el informe. Evidentemente esta métrica no es en sí misma fiable, y solo puede ser aplicada a informes de tipo extractivo, pero aporta información complementaria que no está cubierta por otras métricas.

En este experimento, no hemos introducido otras métricas como *AveragedSentenceLength* o GRAMSIM (sección 6.7.1), dado que se centran en aspectos de fluidez de los textos. Esta información no es relevante en este caso dado que se trata de resúmenes de tipo extractivo.

7.4.2. Resultados

Fiabilidad de las métricas: resultados de la medida KING

La figura 7.10 muestra los valores obtenidos en KING para diferentes combinaciones de métricas de similitud generadas a partir de TFS, R-1, precisión de frases, NICOS, y DocSim. El eje vertical representa las combinaciones de métricas empleadas, y el eje horizontal los valores KING obtenidos. Estos cálculos han sido realizados a partir de todos los temas incluídos en ISCORPUS, tanto temas mono-evento como multi-evento.

El resultado más importante es que si comparamos los KING obtenidos por las métricas individuales, y los obtenidos tras añadirles la métrica TFS.64, podemos comprobar que considerar la distribución en los resúmenes de los 64 primeros términos TFSYNTAX mejora la fiabilidad de la evaluación. Sin embargo, la métrica TFS.64 en solitario no obtiene un alto valor en KING. Esto se debe a que se basa solo en un aspecto concreto de los informes. Es decir, considerar conceptos clave es una estrategia de evaluación no fiable por si misma, pero que aporta fiabilidad a otras métricas basadas en otros rasgos.

La primera de las columnas de la figura muestra la combinación de mayor KING. Un dato relevante es que una combinación de 5 métricas supere al resto de combinaciones, incluyendo métricas aisladas. Este dato verifica la utilidad del marco QARLA como mecanismo de combinación de métricas de similitud. Por otro lado, tres de las métricas incluidas en la combinación están relacionadas con la distribución de conceptos clave: TFS.4, TFS.64 y NICOS. Es decir, la mejor combinación considera la distribución de conceptos anotados manualmente (NICOS), los 64 y los 4 términos más relevantes extraídos mediante TFSYNTAX (TFS.64).

En definitiva, los resultados de la medida KING sugieren que las métricas basadas en conceptos clave (NICOS y TFS) complementan a otras métricas basadas en la coselección de frases o vocabulario, a efectos de evaluación dentro del marco QARLA.

Fiabilidad de los informes automáticos de test: resultados de la medida JACK

Para validar los resultados generados por KING, es necesario comprobar que disponemos de un conjunto lo suficientemente heterogéneo de informes automáticos de muestra. La medida JACK (ver sección 6.5) estima esta heterogeneidad. La siguiente tabla muestra los valores JACK obtenidos para la mejor combinación de métricas de similitud aplicado sobre distintos subconjuntos de informes automáticos. Los cuatro primeros se corresponden con informes generados mediante estrategias basadas en la extracción de la primera, primera y segunda, etc. frases del documento. El último se corresponde con el JACK calculado sobre el conjunto total de informes automáticos.

Extracción de la primera frase	0,07
Extracción de la primera y segunda frase	0,18
Extracción de las tres primeras frases	0,16
Extracción de todo el documento	0
Todas las estrategias	0,68

Como puede verse, el JACK obtenido para subconjuntos de informes es muy inferior al JACK obtenido para el conjunto completo. Es decir, el conjunto total de informes automáticos es considerablemente más heterogéneo que los subconjuntos generados por distintos tipos de estrategias. Este resultado sugiere que el conjunto total de estrategias empleadas para la generación de informes automáticos da soporte a la medida KING, medida sobre la que hemos optimizado el conjunto de métricas de similitud.

7.5. Recapitulación

En este capítulo hemos descrito una serie de experimentos centrados en el estudio del papel que juegan los conceptos clave en la Síntesis de Información. En cada uno de los apartados anteriores se ha dado respuesta a diferentes cuestiones relativas a este asunto.

En primer lugar, considerar los conceptos clave es necesario para mejorar aproximaciones como la selección de la primera frase de los documentos. Los resultados sugieren que los conceptos clave son una característica común en los informes generados por sujetos de prueba. Esta característica no es compartida por informes generados mediante la extracción de primeras frases de documentos aunque consiguen sin embargo un buen solapamiento con los modelos en relación a palabras o frases.

En segundo lugar, es posible generar una lista de términos con una amplia cobertura sobre la lista de conceptos clave presentes en los documentos originales. Esta cobertura se obtiene considerando información sintáctica superficial, como es la posición de los términos en relación al verbo, es decir, empleando el criterio de pesado TFSYNTAX. Los términos extraídos mediante este criterio de pesado caracterizan los informes de manera similar a como lo hace una lista de conceptos extraída manualmente.

Muchos sistemas de resumen automático parten de la hipótesis de que un fragmento en donde aparecen términos relevantes es más susceptible de pertenecer a un resumen. Sin embargo, aumentar en exceso la frecuencia de estos términos en un informe puede decrementar la calidad del mismo, en cuanto a que es menos semejante a los informes modelo. En el tercer apartado de este capítulo se pudo comprobar que, por ejemplo, las primeras frases de los documentos originales contienen más conceptos clave de los que contiene un informe generado manualmente (sección 7.3). Es necesario por tanto estimar la distribución de conceptos clave. En este trabajo se ha podido comprobar que es posible estimar la frecuencia de un término TFSYNTAX en un informe manual a partir de su frecuencia en los documentos originales(sección 7.3). Es decir, es posible predecir la distribución de estos términos en un informe generado manualmente.

El último experimento sugiere que los conceptos clave son útiles en la evaluación de informes por similitud a informes modelo (sección 7.4). Todas las métricas estudiadas cobran más fiabilidad en QARLA si son combinadas con una métrica orientada a conceptos clave. Además, la mejor combinación de métricas de similitud incluye varias de ellas basadas en la distribución de conceptos clave. Estos datos suponen una evidencia a favor del uso de conceptos clave en la elaboración y evaluación de sistemas de Síntesis de Información.

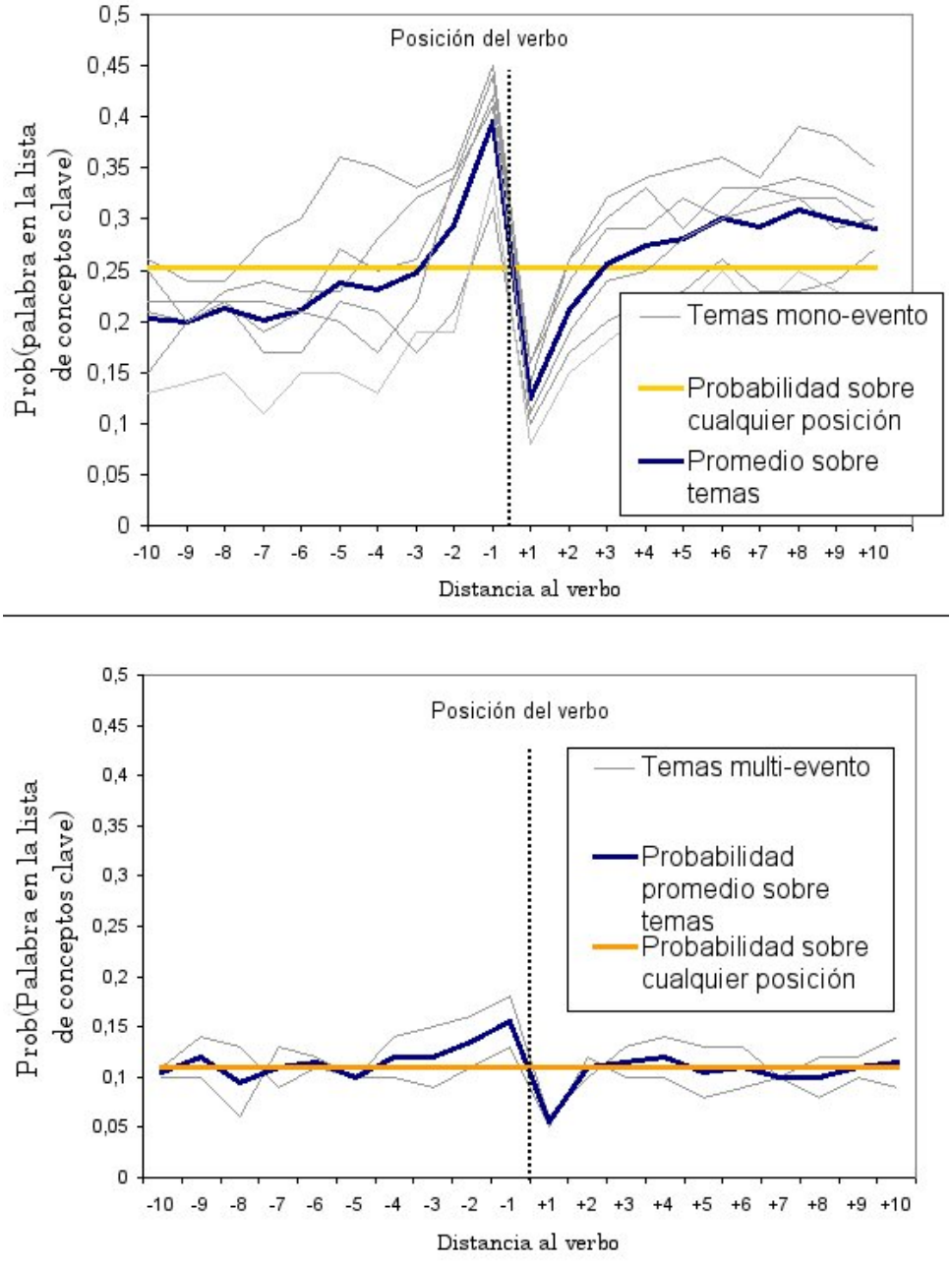


Figura 7.3: Probabilidad de encontrar conceptos clave en relación a la distancia al verbo

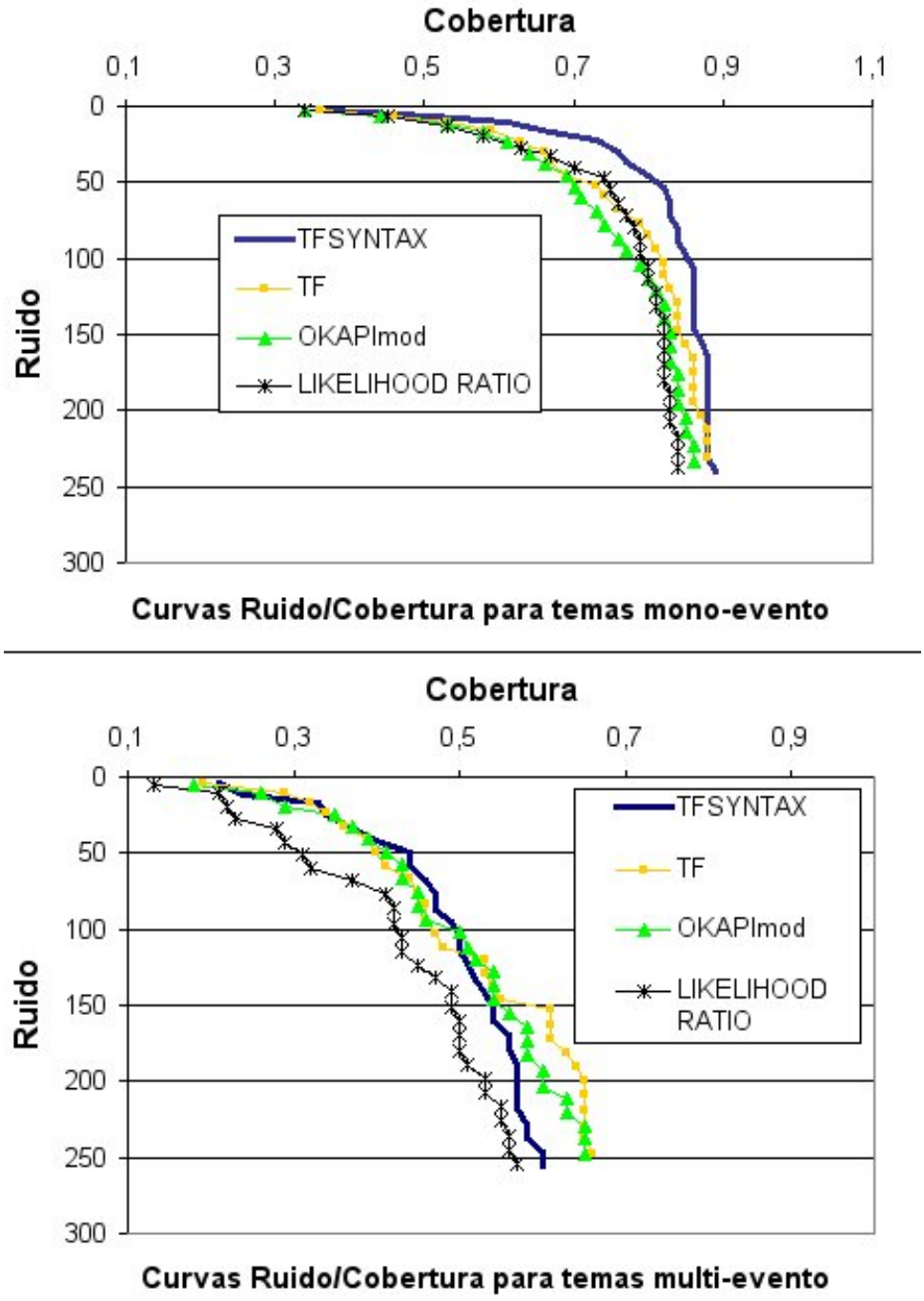


Figura 7.4: Comparación de criterios de pesado en la extracción de conceptos clave

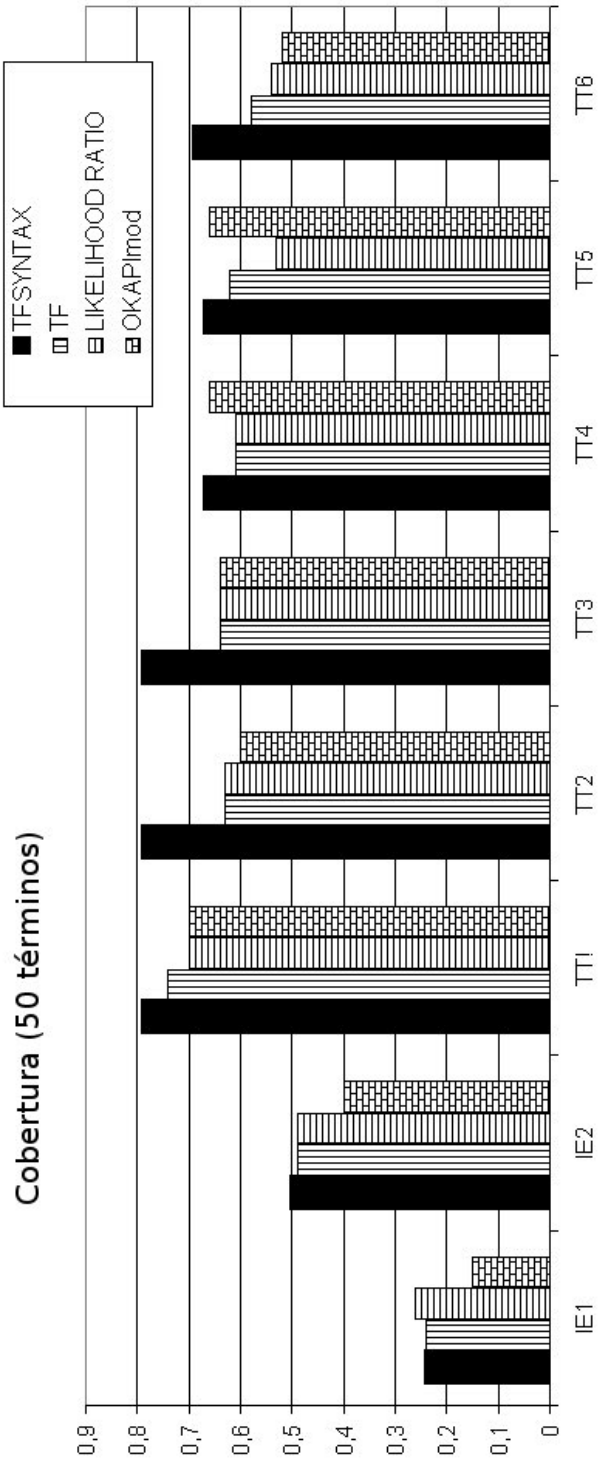


Figura 7.5: Comparación de estrategias de pesado en la extracción de conceptos clave en los distintos temas

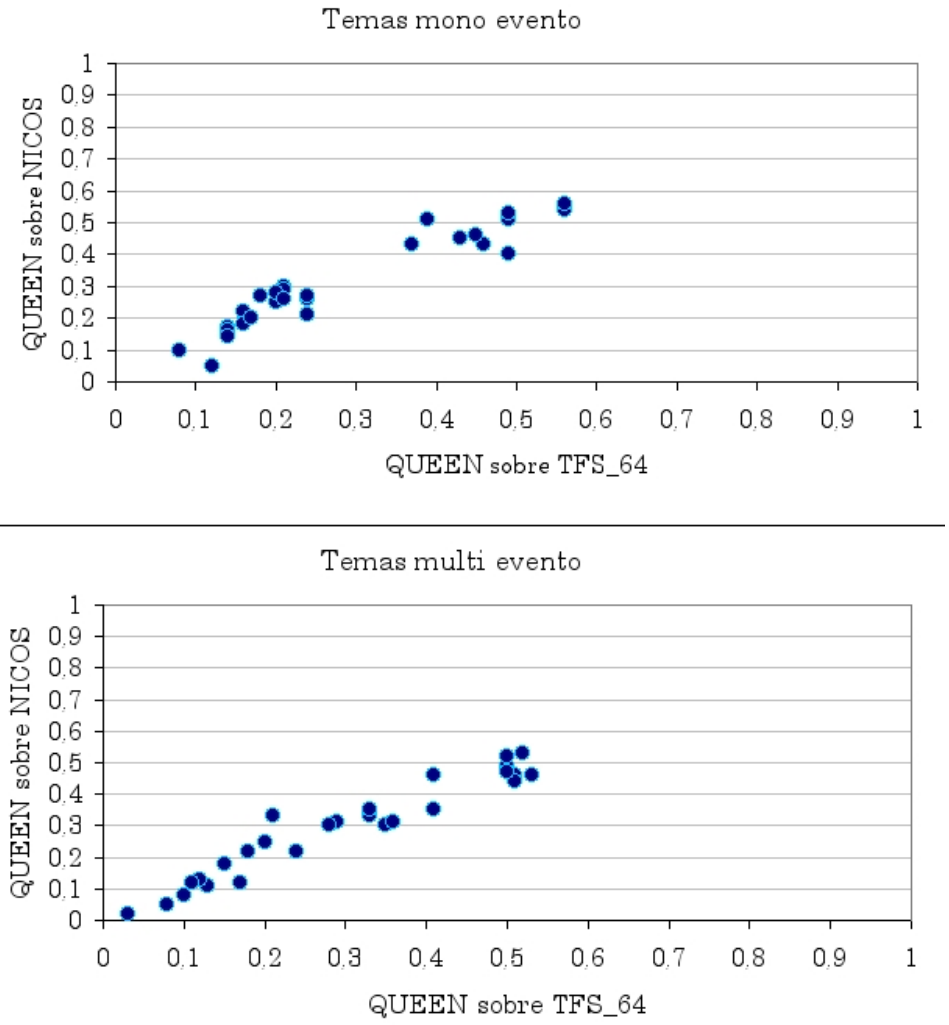


Figura 7.6: Comportamiento de la medida QUEEN sobre la métrica de similitud NICOS frente a la metrica de similitud TFS.64

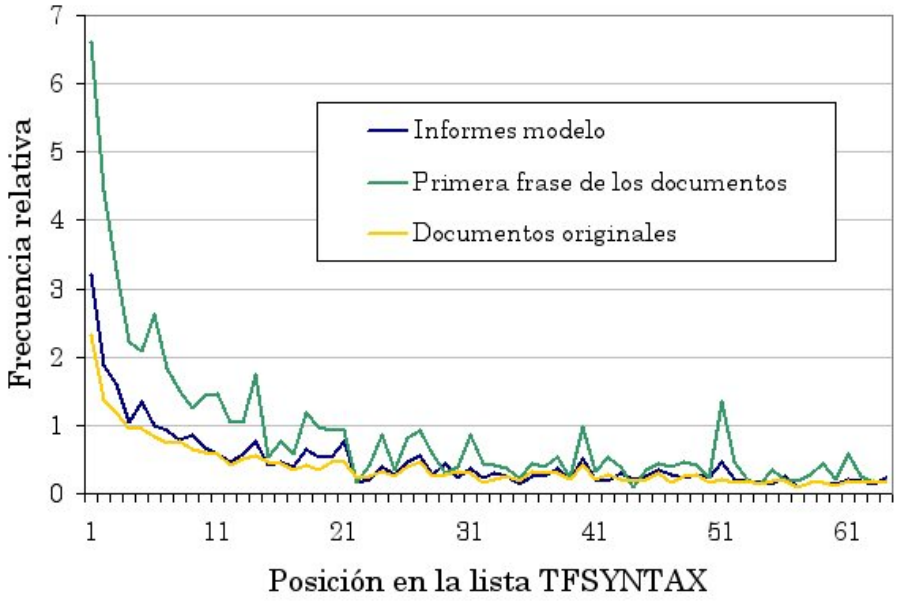


Figura 7.7: Frecuencias de términos TFSYNTAX en informes manuales, documentos originales y primeras frases de documentos

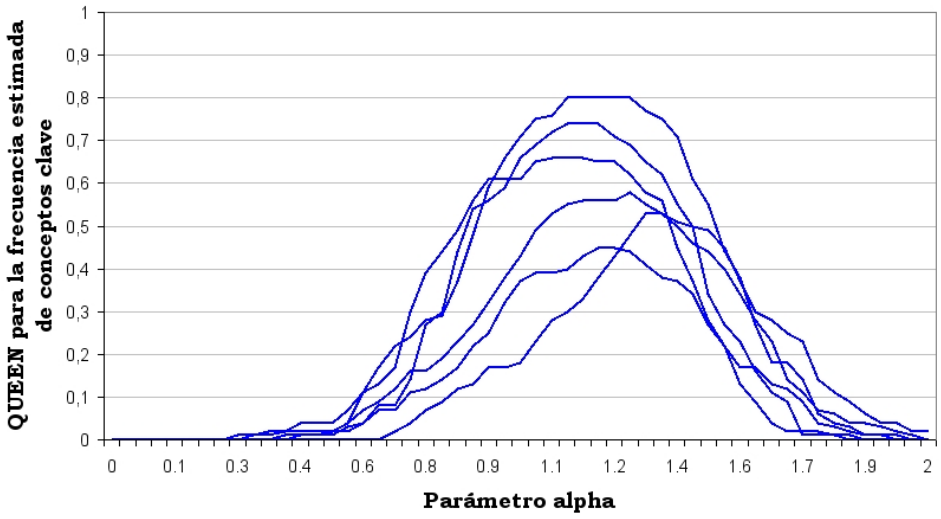


Figura 7.8: Estimación del ratio entre la frecuencia de conceptos clave en informes y documentos originales.

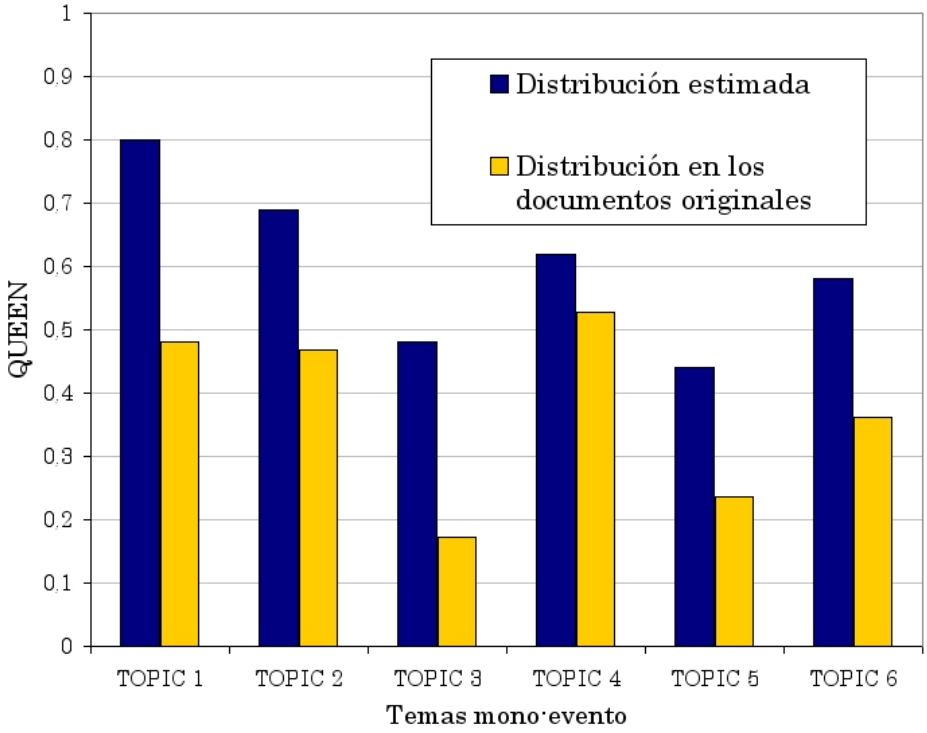


Figura 7.9: Valores QUEEN para la distribución de conceptos clave estimada frente a la distribución en los documentos originales

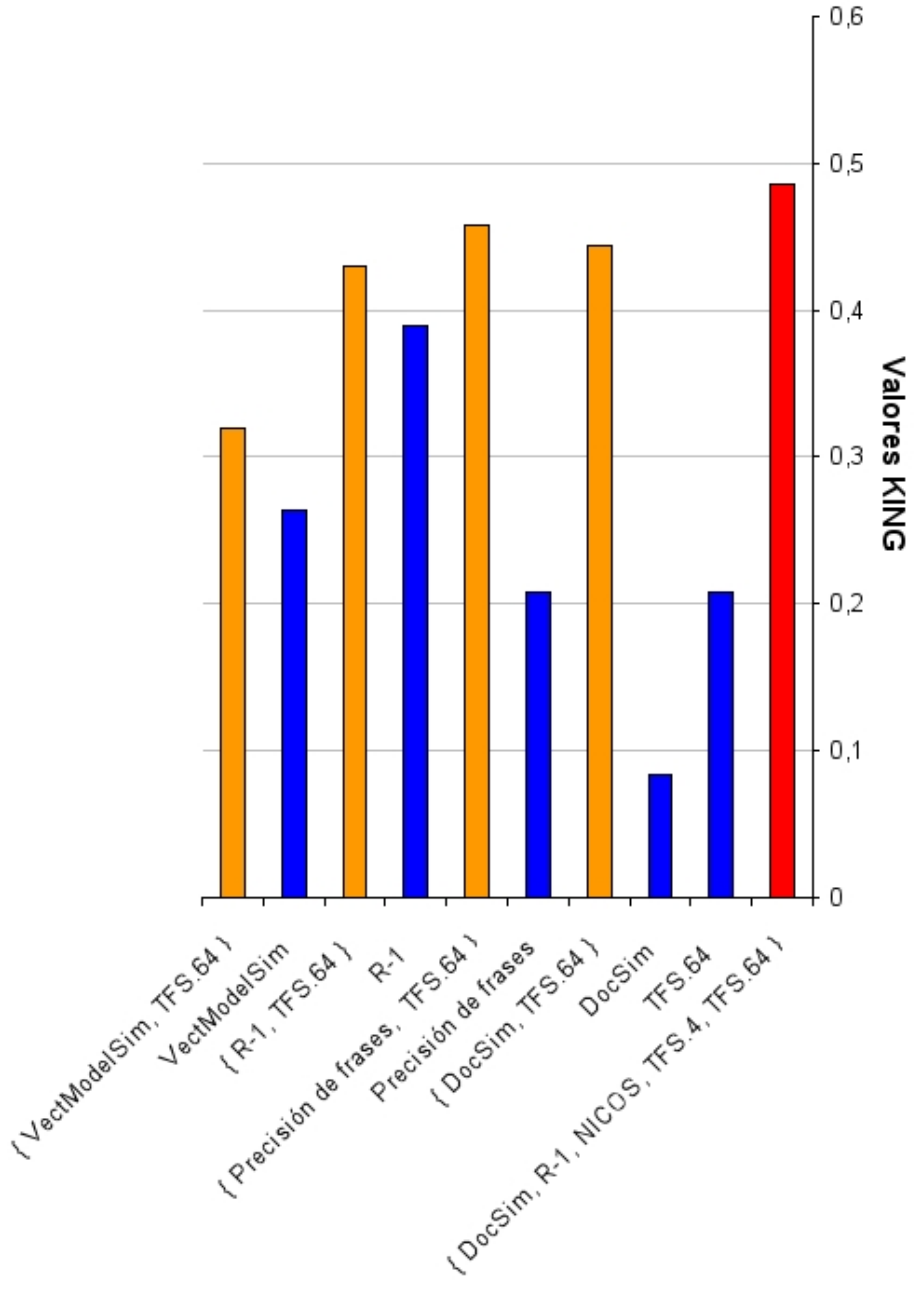


Figura 7.10: Valores KING para combinaciones de métricas aplicadas en ISCORPUS

Capítulo 8

Evaluación de estrategias de exploración de contenidos en un sistema interactivo de Síntesis de Información

La exploración de contenidos en general ha sido empleada como estrategia en el desarrollo de sistemas interactivos de resumen (sección 3.3). En estas aproximaciones, el sistema organiza la información textual de las fuentes de forma que el usuario pueda explorar los contenidos que desea resumir.

En este trabajo desarrollamos el modelo interactivo PRISMA orientado a Síntesis de Información (capítulo 9). En este modelo se propone el acceso a los contenidos en las fuentes por medio de una lista de términos representativos de los conceptos clave del tema tratado. Es este capítulo se realiza una pre-evaluación de dicha estrategia de interacción con el usuario.

Existen varias evidencias de que términos representativos de los conceptos clave presentes en las fuentes originales pueden ayudar a organizar y presentar la información al usuario con vistas a la elaboración de un informe. Por un lado, la mayoría de los sistemas interactivos de resumen ofrecen, entre otras funcionalidades, una lista de términos relevantes extraída automáticamente de las fuentes como elemento de interacción con el usuario (sección 3.3). Por otro lado, considerando los resultados obtenidos en los capítulos anteriores, sabemos que los conceptos clave juegan un papel importante en el proceso de Síntesis de Información (capítulo 7).

Sin embargo, los sistemas tradicionales de gestión de contenidos textuales, por ejemplo, buscadores en la WEB o las bibliotecas digitales, por lo general organizan la información al usuario por medio de títulos de documentos. La exploración por medio de títulos se corresponde con el uso de un sistema de recuperación de documentos tradicional como herramienta para la elaboración de informes. La exploración por medio de términos relevantes supone una alternativa al uso de sistemas de recuperación de información, dado que el usuario no explora documentos, sino

piezas de información asociadas a cada uno de estos términos. A pesar de que la mayoría de los sistema interactivos de resumen actuales emplean esta aproximación, no ha habido, que sepamos, una evaluación comparativa entre ambos tipos de aproximaciones. En este capítulo comparamos la estrategia de exploración de contenidos por medio de títulos, con estrategias basadas en exploración de términos relevantes en el contexto de la Síntesis de Información. Es decir, nos planteamos si acceder a la información contenida en los documentos vía una lista de términos relevantes es más eficiente que acceder por medio de los títulos.

Dentro de la exploración de términos, queda abierta la cuestión de qué piezas de información han de asociarse a cada uno de estos términos relevantes. Una posible aproximación es la selección automática de frases en las que el término aparece dentro del sujeto sintáctico. Esta estrategia permite visualizar las piezas de información asociadas a un término de forma más estructurada, es decir, a modo de tablas sujeto-acción-objeto (sección 9.2.1).

Sin usuarios de prueba no podemos evaluar la ganancia de cada estrategia en cuanto a visualización de contenidos, dado que no podemos predecir de qué modo un usuario asimila la información ofrecida por el sistema. Sin embargo, podemos cuantificar otros aspectos, como por ejemplo, la cobertura sobre piezas de información útiles en la elaboración de informes. Para comparar las diferentes estrategias de exploración de contenidos, medimos la calidad de los informes que se pueden generar a partir de las piezas de información más accesibles en cada estrategia. Para ello, aplicamos la metodología de evaluación QARLA particularizada para sistemas interactivos (sección 6.6).

8.1. Estrategias de exploración de contenidos

Una estrategia de exploración de contenidos puede entenderse como una lista de elementos a través de los que el usuario accede a piezas de información contenidas en los documentos originales. Estos elementos pueden ser, títulos de documentos, términos relevantes, conjuntos de términos, fechas, etc. En este capítulo, compararemos las siguientes estrategias en el contexto de la SI:

- **Exploración de títulos en orden cronológico**. En este esquema, el usuario accede a los documentos por medio de una lista de títulos ordenados según su fecha de edición. Este criterio de ordenamiento es posible dado que trabajamos con documentos periodísticos.
- **Exploración de títulos ordenados por relevancia**. De nuevo, los usuarios acceden a los documentos por medio de los títulos, pero en este caso, éstos están ordenados por relevancia. La relevancia de los documentos es calculada mediante el motor de búsqueda INQUERY [ACS+98]. Esta aproximación se corresponde directamente con el uso de un sistema tradicional de recuperación de documentos como asistente para la generación de un informe. Podemos definir diferentes configuraciones de esta estrategia de exploración, dependiendo del número i de documentos mostrados al usuario ($i = 20$, $i = 50$, etc.).

- **Exploración de términos relevantes**. El usuario accede a frases por medio de una lista de términos sugeridos por el sistema. En nuestro caso, estos términos son extraídos automáticamente mediante la estrategia de pesado TFSYNTAX (sección 7.2.3). La frases asociadas a dichos términos se ordenan teniendo en cuenta su localización en los documentos originales, de forma que fragmentos que aparecen en la primera posición de un documento tienen preferencia sobre fragmentos que aparecen en posiciones posteriores. De nuevo, dependiendo del número de términos mostrados al usuario podemos definir distintas configuraciones en esta estrategia de exploración.

- **Exploración de términos clave y sus ocurrencias como sujeto sintáctico.** Esta estrategia es análoga a la anterior, solo que se muestra al usuario únicamente ocurrencias del término como parte del sujeto sintáctico. Esta estrategia permite presentar al usuario los fragmentos recuperados organizados según estructuras sujeto-acción-objeto, facilitando así el proceso de visualización (ver sección 9.2.1).

Procesamiento lingüístico superficial de los documentos

Con el fin de estudiar la estrategia basada en exploración de términos clave y sus ocurrencias como sujeto, es necesario previamente procesar los textos, identificando los componentes básicos de la frase. Esto es: sujeto, verbo y complementos del verbo. Para ello, se ha implementado una herramienta sencilla, rápida y robusta para la extracción de esta información lingüística. Esta herramienta permite procesar corpus grandes, del orden de 100.000 documentos en un tiempo razonable.

Inicialmente, los documentos son procesados con un análisis sintáctico superficial, particionando los elementos de la frase y asignando estas unidades al siguiente conjunto de categorías:

- `[N]`: sintagmas nominales, que se corresponde con nombres o adjetivos precedidos de un determinante, principio de frase o signo de puntuación:

- `[V]`: formas verbales.

- `[Mod]`: sintagmas o adverviales o preposicionales, identificados como sintagmas nominales precedidos de un adverbio o preposición.

- `[Sub]`: términos que introducen una nueva frase subordinada (*que, cuando, mientras*, etc.).

- `[P]`: Signos de puntuación

Esto es un ejemplo de la salida de este analizador sintáctico superficial:

Previamente [Mod] , [P] el presidente Bill Clinton [N] había dicho [V] que [Sub] tenemos [V] la obligación [N] de cambiar la política estadounidense [Mod] que [Sub] no ha funcionado [V] en Haití [Mod]. [P]

Aunque la precisión de este procesamiento es limitado, es suficiente para el estudio estadístico que planteamos en este capítulo.

Identificamos estructuras sujeto de la siguiente forma. En primer lugar, dividimos las frases por nexos de subordinación (unidades tipo `[Sub]`). En segundo lugar, extraemos patrones del tipo `[N][Mod]`$\star$. Asumimos que patrones `[N][Mod]`$\star$ que aparecen inmediatamente antes del verbo, o antes de un nexo de subordinación, contienen un sujeto de oración. Por ejemplo:

El presidente [N] en funciones [Mod] de Haití [Mod] ha afirmado [V] *que [Sub]...*

El presidente [N] en funciones [Mod] de Haití [Mod] *que [Sub] llegó [V] ayer [MOD]* ***ha afirmado [V]*** *que [Sub]* ***...***

El resto de unidades tipo `[N]` y `[Mod]` son consideradas como parte de los complementos del verbo. Esta heurística, aunque en la mayoría de los casos identifica el sujeto de la oración, no captura, por ejemplo, sujetos posteriores al verbo.

Como puede verse en el segundo ejemplo, si aparece una oración subordinada tras el sujeto, se asocia a este sujeto el primer verbo que aparece tras la oración subordinada. Esta estrategia permite identificar, con una precisión relativamente pobre pero razonable, estructuras sujeto-acción-objeto, dentro y fuera de oraciones subordinadas.

Las ocurrencias de términos que aparecen dentro de algún sujeto de oración, son seleccionadas como piezas de información asociadas al término en la estrategia de exploración de términos con selección de ocurrencias como sujeto sintáctico.

8.2. Definición del experimento

La evaluación de estas estrategias de exploración se basa en la comparación de los informes modelo generados en ISCORPUS con los informes que pueden ser generados empleando cada una de las estrategias.

Los informes manuales en ISCORPUS han sido generados por medio de la exploración de títulos de documentos (sección 5.2). Es decir, en realidad las muestras de informes modelo están sesgadas. Sin embargo, este sesgo favorece a la evaluación de estrategias basadas en exploración de títulos, dado que es de esperar que los resúmenes que se pueden generar mediante la exploración de títulos tiendan a ser semejantes a estos informes modelo, generados mediante esa misma estrategia. La cuestión es si a pesar de este sesgo, los informes que se pueden generar a partir de una estrategia basada en exploración de términos logran parecerse más a los informes modelo.

Esquemáticamente, los pasos seguidos en este experimento son:

1. Definimos un conjunto de estrategias dadas por el tipo de elemento de exploración (títulos o términos), el número de elementos y el criterio de asociación de listas de frases a cada elemento.

2. Generamos una lista con las n frases más accesibles para cada estrategia.

3. Generamos para cada lista y estrategia un conjunto de informes potenciales mediante la extracción aleatoria de 50 frases de la lista.

4. Evaluamos el conjunto de informes potenciales asociados a cada estrategia en relación a informes modelo mediante QUEEN generalizada para dominios interactivos.

El modo en que se definen las estrategias de exploración ya han sido descritas en el apartado 8.1.

Identificación de la lista de frases accesibles

Para definir la lista de n frases más accesibles para el usuario, partimos de la base de que cada estrategia de exploración hace más accesible un conjunto diferente de frases. Por ejemplo, como pudo verse en el estudio de ISCORPUS (capitulo 5), la mayor parte de las frases extraídas de los documentos se encontraban en la primera o segunda posición del mismo. Es decir, los sujetos seleccionaron las frases más accesibles desde los títulos. En este experimento asumimos que el conjunto de frases más accesibles en una estrategia de exploración determina el conjunto de informes potenciales asociado a dicha estrategia. Por ello, necesitamos definir una medida concreta de accesibilidad.

Denominamos "elemento de exploración" a la pieza de información que el usuario emplea para acceder a los contenidos. En nuestro caso, estos elementos son títulos o términos sugeridos por el sistema. Entonces, asumimos la hipótesis de que las primeras frases asociadas a un "elemento de exploración" son más accesibles que el resto. Por ejemplo, en la exploración de títulos las frases más accesibles son las primeras de cada uno de los documentos mostrados al usuario. Si tomamos solo los 50 documentos más relevantes y las 200 frases más accesibles se corresponden con las primeras 4 frases de cada documento. En la exploración de términos, consideramos que son más accesibles las primeras frases asociadas a cada término. Las 100 frases más accesibles dada una lista de 20 términos se corresponderán con las cuatro primeras frases asociadas a cada término. Cabe resaltar que en el caso de las estrategias basadas en exploración de ocurrencias de términos, una misma frase puede estar asociadas al más de un término. En este experimento, consideramos que esta frase aparecería dos veces en la lista de frases accesibles.

Generación del conjunto de informes potenciales

La cuestión es entonces cómo definir el conjunto de informes potenciales asociado a cada una de las estrategias de exploración. La tarea propuesta a los sujetos en la elaboración de ISCORPUS consistía en la generación de un informe extractivo mediante la selección de un máximo de 50 frases. Tomamos por tanto la frase como unidad de trabajo. Generamos informes potenciales mediante la extracción aleatoria de frases a partir del conjunto explorado por un hipotético usuario.

El conjunto de informes potenciales viene dado por todos los informes que se pueden generar mediante la extracción aleatoria de 50 frases del conjunto, que es el tamaño máximo permitido a los sujetos en ISCORPUS. Por tanto, para cada número fijo n de elementos de exploración mostrados al usuario, y cada estrategia de exploración, tenemos un conjunto de informes potenciales.

Evaluación de informes potenciales

El siguiente aspecto a abordar es cómo evaluar los informes potenciales. Para ello, aplicamos el marco QARLA. Concretamente, aplicamos la medida QUEEN generalizada para dominios interactivos (sección 6.6). Esta medida permite evaluar la semejanza a los informes modelo de un conjunto de informes. QUEEN valora el que todos los informes modelo, es decir, sujetos de prueba, se vean satisfechos por al menos uno de los informes del conjunto.

El conjunto total de posibles combinaciones de frases es demasiado grande para que puedan ser evaluados cada uno de ellos. Sin embargo, dado que QUEEN se define en términos probabilísticos, podemos simplemente reducir el número de muestras de informes potenciales. De esta forma, es posible comparar distintas estrategias siempre que empleemos el mismo número de muestras para cada una de ellas.

Para evaluar mediante la medida QUEEN es necesario definir un conjunto de métricas de similitud. Empleamos para ellos el conjunto formado por NICOS (sección 7.1.1) , TFS.4, TFS.64 (sección 7.2.4), R-1 (sección 6.7.1) y DocSim (sección 7.4). Esta combinación es precisamente la que obtiene mayor KING, es decir, la más fiable según el marco QARLA (sección 7.4).

8.3. Resultados

Las figuras 8.1 y 8.2 muestran los resultados obtenidos para el experimento descrito. El eje horizontal representa el número de fragmentos explorados por el hipotético usuario. El eje vertical representa los valores QUEEN obtenidos para el conjunto de informes potenciales generados para cada estrategia de exploración y número de fragmentos explorados.

Temas mono-evento frente a multi-evento

La primera conclusión que podemos obtener es que, como muestran las figuras, la calidad QUEEN de los informes potenciales es mayor en el caso de los temas mono-evento que en los temas multi-evento. Esto se debe a que en el segundo caso, existe más acuerdo entre resumidores humanos. Este hecho pudo observarse en los resultados obtenidos en el solapamiento de frases entre informes descrito en el capítulo 5. Por tanto, es más difícil que un informe potencial se asemeje a un informe manual tanto como dos informes manuales entre sí, que es el principio de la medida QUEEN.

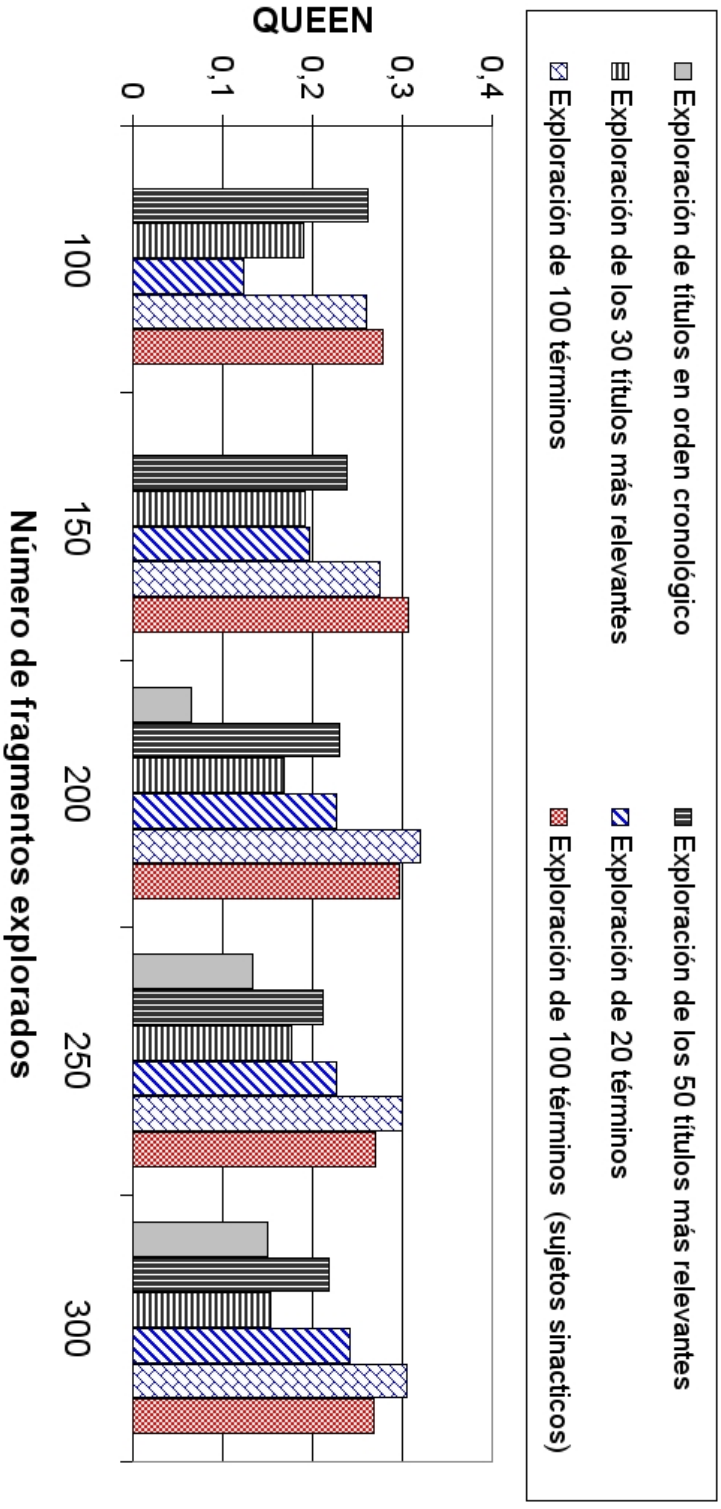


Figura 8.1: Evaluación de estrategias interactivas de exploración de contenidos en temas mono-evento

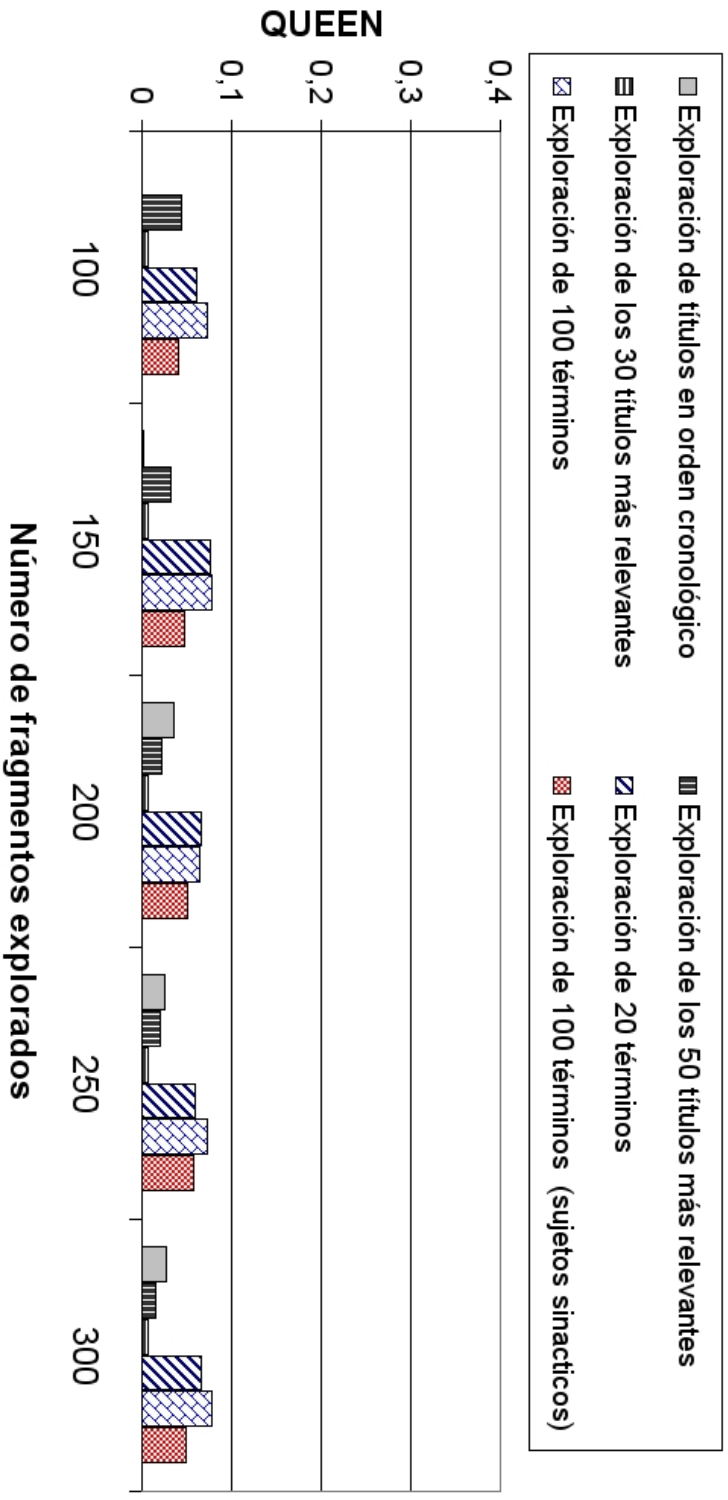


Figura 8.2: Evaluación de estrategias interactivas de exploración de contenidos en temas multi-evento

Exploración por títulos frente a exploración por términos

En segundo lugar, puede verse que la exploración por medio de términos extraídos automáticamente supera a la exploración por títulos para cualquier número de documentos explorados. Por ejemplo, para el caso de los temas mono-evento, explorar las dos primeras frases de cada documento (200 fragmentos en la estrategia de recorrido de títulos), no supera el 0.1 en QUEEN, a pesar de que más de la mitad de los fragmentos seleccionados en los informes manuales se encuentran en la primera o segunda frase del documento. Sin embargo, explorar las dos primeras frases asociadas a 100 términos llega al 0.3 en QUEEN (200 fragmentos en la estrategia de exploración de 100 términos). Aunque a menor escala, estas diferencias son aun más notables en el caso de los temas multi-evento.

Otro aspecto que cabe resaltar es que en general, los resultados mejoran cuando cubrimos un número mayor de términos en la estrategia de exploración. Posiblemente, abarcando más términos, se consigue una mayor cobertura sobre los temas tratados en la colección. Estas diferencias son especialmente notables en el caso de los temas mono-evento.

Exploración de títulos relevantes

En los temas mono-evento, resulta eficiente seleccionar los documentos relevantes, frente a seleccionar todos los documentos en las estrategias de exploración basadas en títulos. Esto no ocurre en el caso de los temas multi-evento. Probablemente, la causa de este hecho es que en los temas multi-evento, el conjunto de documentos debe tener una amplia cobertura sobre todos los eventos tratados en la colección. Sin embargo, el motor de búsqueda INQUERY no contempla esta necesidad. Hay que tener en cuenta que es posible que los resultados basados en la exploración de documentos relevantes pudieran optimizarse aplicando motores de búsqueda que cubran en unos pocos documentos todos los temas tratados en la colección. Esto podría hacerse por ejemplo, mediante algún mecanismo de agrupación de documentos.

Ocurrencias del término como sujeto sintáctico

Por último, la selección de ocurrencias en donde el término aparece como parte del sujeto sintáctico pierde calidad QUEEN aproximadamente en un 10 % en relación a seleccionar todas las ocurrencias del término, a excepción de los 100 y 150 fragmentos más accesibles en temas mono-evento, en donde la consideración de sujetos sintácticos hace aumentar la cobertura. En cualquier caso, los valores de QUEEN siguen manteniéndose por encima de la exploración por títulos.

8.4. Conclusiones

Evaluar objetivamente la usabilidad de una estrategia interactiva sin usuarios de prueba es difícil, dado que existen aspectos como la visualización que no son tratables desde esta perspectiva. Por contra, una evaluación con usuarios de prueba

es muy costosa. Por ejemplo, tomando un mínimo de 10 usuarios de prueba por estrategia, serían necesarios $10 \times 6 \times 8$ casos de uso, para evaluar las 6 estrategias propuestas sobre los 8 temas. En este capítulo hemos definido y empleado una metodología sin usuarios de prueba. Esta metodología permite obtener información útil que puede ser empleada para seleccionar estrategias interactivas sobre las que, posteriormente, evaluar de forma objetiva la usabilidad del sistema.

Los resultados sugieren que una exploración por términos ofrece una mayor cobertura de contenidos que una exploración por títulos, con vistas a la elaboración de un informe. En base a este resultado definiremos el modelo PRISMA como una estrategia de interacción con el usuario basada en la exploración de términos relevantes extraídos automáticamente de las fuentes.

Dentro de las estrategias basadas en listas de términos, en estos experimentos hemos comparado dos posibles estrategias de selección de frases asociadas a cada uno de estos términos. Por un lado, todas aquellas frases en las que aparece el término, y por otro lado, sólo aquellas frases en las que el término aparece como núcleo de sujeto sintáctico. Los resultados sugieren que esta criba no afecta excesivamente a la cobertura sobre contenidos con vistas a la generación del informe. Como veremos en capítulo siguiente, considerar únicamente frases en las que el término aparece como núcleo de sujeto, permite una visualización más estructurada de los fragmentos asociados al término, es decir, permite mostrar los fragmentos en una tabla sujeto-acción-objeto, y reducir cada una de estas frases al contexto del término empleando criterios sintácticos. Considerando estos resultados, fijamos en el modelo PRISMA la estrategia de exploración de términos y sus ocurrencias como núcleo del sujeto como esquema de interacción con el usuario.

Capítulo 9

Modelo PRISMA

En este capítulo describimos el modelo PRISMA, una aproximación interactiva al problema de la Síntesis de Información. La interacción entre usuario y sistema se establece de forma acorde con el comportamiento de una serie de sujetos de prueba descrito en la sección 5.6. En este estudio pudimos comprobar que, en el contexto abordado en este libro, el proceso de SI es fundamentalmente extractivo. Los sujetos de prueba extrajeron la misma cantidad de frases a lo largo de todo el proceso de SI. Es decir, en general, la cantidad de fragmentos recopilados no es más reducido al comienzo o al final del proceso. Además, los sujetos consideraron autocontenido el informe resultante, dado que en la gran mayoría de los casos, no sintieron la necesidad de introducir anotaciones en el informe. El modelo PRISMA plantea la SI como un proceso de recopilación de información.

Por otro lado, hemos podido comprobar también que los sujetos dedicaron tiempo al análisis de documentos concretos de los que finalmente no se extrajo información alguna. Es por ello por lo que uno de los aspectos en los que se centra el modelo PRISMA es en facilitar la lectura de documentos completos.

En tercer lugar, hemos podido comprobar (capítulo 7) que la distribución de términos representativos de los conceptos clave tratados en el asunto, es un rasgo compartido por los informes en mayor medida que otros rasgos como las frases seleccionadas o el vocabulario presente en los informes. Este hecho sugiere que estos términos clave pueden ser útiles en el análisis de contenidos con vistas a la elaboración de un informe.

En el capítulo 8 hemos comparado dos tipos de estrategias de exploración de contenidos: vía una lista de términos clave y vía la lista de títulos de documentos. Los resultados muestran que los informes que se pueden generar mediante la primera estrategia son más semejantes a los informes modelo que los que se pueden generar mediante la segunda estrategia. Teniendo en cuenta estos resultados, en el modelo PRISMA se ofrece al usuario una lista de términos clave a través de los cuales se accede a piezas de información distribuídas en las fuentes.

Por medio de cada uno de los términos clave ofrecidos al usuario se accede a frases o documentos en los que dicho término aparece. Los resultados obtenidos en el capítulo 8 sugieren también que reducir estas frases a aquellas en las que el término aparece formando parte del sujeto sintáctico no reduce excesivamente la

cobertura sobre contenidos para la elaboración de un informe. Por contra, esta reducción de las frases accesibles desde un término permite, como se describirá más adelante, mostrar al usuario la información de manera más estructurada.

La descripción del modelo PRISMA en este capítulo se estructura como sigue. En la primera sección se describe cada uno de los niveles de acceso a la información que componen el modelo PRISMA. Es decir, la piezas de información que el usuario visualiza y selecciona durante el proceso de Síntesis. En la segunda sección se detalla cada una de las fases que componen el modelo interactivo PRISMA, describiendo el papel que juega el usuario y el sistema en cada una de las fases. En la tercera sección se describe algunos detalles de implementación del prototipo implementado según el modelo PRISMA. Posteriormente, analizamos las semejanzas y diferencias ente nuestro modelo y otros modelos interactivos de resumen. Finalmente, en la última sección, hacemos un análisis en relación a los requisitos que satisface PRISMA como modelo interactivo de Síntesis de Información, y revisamos los aspectos evaluados cuantitativamente a lo largo de este libro.

9.1. Niveles de acceso a la información

La figura 9.1 muestra los distintos niveles que estructuran el modelo PRISMA y las unidades de información implicadas. A continuación se describe las unidades de información que componen cada uno de los niveles:

Términos de búsqueda Una de las ventajas de aplicar paradigmas interactivos en el desarrollo de sistemas de acceso a la información consiste en que el usuario no tiene que especificar mediante una consulta de manera exacta y a priori sus necesidades de información, dado que es el usuario quien guía el proceso de acceso a la información de acuerdo a sus necesidades. En este caso, el punto de partida no es una consulta explícita sino un conjunto de términos clave que ayudan al sistema a discriminar inicialmente información potencialmente relevante. En la figura estos términos son *conflicto* y *Palestina*, que representan una necesidad de información consistente en obtener información acerca del dicho conflicto.

Documentos relevantes. Se entiende como documentos relevantes aquellos que contengan información potencialmente útil para el proceso de síntesis. Esto incluye, bien información que será integrada en el informe, bien información útil para otros procesos como la organización o la contextualización de piezas de información.

Conceptos clave Son aquellos conceptos que adquieren protagonismo en el asunto tratado en las fuentes, expresados en forma de términos. En el ejemplo de la figura 4.1, entidades como “EEUU”, “Arafat” o “Israel” acometen acciones que repercuten en el asunto. También se incluye conceptos más abstractos como “atentado” o “proceso de paz”. Estos términos clave conforman una dimensión sobre la que organizar la información contenida en los documentos e independiente de los títulos.

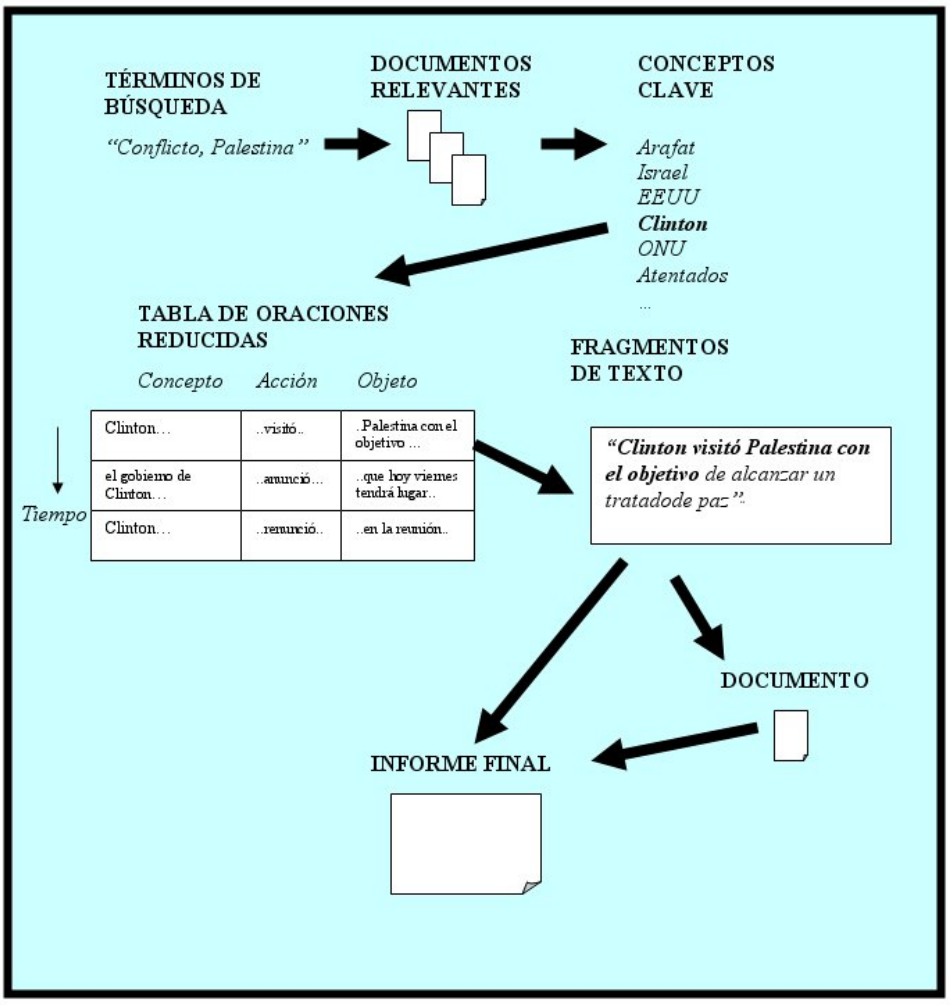


Figura 9.1: Niveles intermedios de acceso a la información en PRISMA

En el contexto de PRISMA, las expresiones *"concepto clave"* y *"términos clave"* se prestan a confusión. La extracción automática de términos clave mediante el criterio de pesado TFSYNTAX ha sido optimizada en función de la cobertura y precisión sobre los conceptos clave anotados manualmente por sujetos en ISCORPUS (ver sección 7.2). A lo largo de este capítulo emplearemos la expresión *"concepto clave"* para referirnos a estos términos extraídos automáticamente.

Tabla de oraciones reducidas. Esta tabla contiene oraciones reducidas a sus componentes principales, es decir, estructuras agente-acción-objeto. Estas oraciones representan piezas de información que muestran de qué modo los conceptos clave actúan en el asunto. Estas estructuras se organizan verticalmente sobre una dimensión temporal, basada en la fecha de edición de su documento. Horizontalmente, las oraciones se organizan marcando las tres partes de la oración. En el ejemplo de la figura aparecen oraciones reducidas y descompuestas en tres elementos como por ejemplo *Clinton.. ..visitó.. ..Palestina con el objetivo...* Los componentes agente, acción y objeto, no son necesariamente contiguos en la frase original. En los trabajos realizados por Guo y Stylios ([Y. Guo y G. Stylios, 2003]), ya se ha explorado una estrategia de organización de la información semejante. En concreto, estos autores consideran cuatro dimensiones: agente, tiempo, lugar y acción, para la elaboración automática de resúmenes.

Fragmentos de texto. Consideramos como fragmentos de texto a las piezas de información que el usuario selecciona durante el proceso de recopilación de información. La frase es la unidad lingüística más habitual en estos casos. En el capítulo 5 se muestra que un proceso de extracción de frases permite la elaboración de un informe, por lo menos en un primer estadio del proceso de Síntesis de Información. Además, la mayoría de los sistemas de Resumen Automático de tipo extractivo las emplean para generar el resumen. El fragmento de texto (frase completa) y el documento son las unidades de información que en este modelo ayudan a contextualizar una oración reducida cuando la tabla no aporta información suficiente.

Informe final. Es el resultado del proceso de síntesis. Contiene las piezas recopiladas organizadas en un único documento. Este elemento es además un objeto software que contiene meta-información asociada a los fragmentos recopilados, como concepto clave y oración reducida desde la que se ha accedido, documento original o fecha de edición. Esta meta-información supone la posibilidad de ofrecer al usuario diferentes vistas de su informe.

9.2. El proceso de Síntesis de Información en PRISMA

Al igual que la mayoría de los sistemas interactivos de resumen actuales, el proceso de acceso a la información en el modelo PRISMA se compone de dos fases: Vista global de los contenidos y contextualización (apartado 3.3). En el modelo

PRISMA se incluye además una tercera fase de elaboración del informe final, en la que el usuario recopila fragmentos de textos y edita finalmente un informe acorde a las necesidades de información.

Cada una de estas fases se compone de una serie de tareas. Éstas son:

Vista global de la información

1. Expresión de términos de búsqueda en relación al asunto que se desea tratar.
2. Recuperación de documentos potencialmente relevantes mediante términos de búsqueda.
3. Identificación automática de los conceptos clave implicados en el asunto tratado en los documentos.
4. Selección interactiva de los conceptos clave. En esta fase, el usuario únicamente tiene que elegir aquellos conceptos sugeridos por el sistema que le parezcan más pertinentes.
5. Creación automática de la tabla de oraciones reducidas asociada al concepto. Las casillas de la columna contendrán, de acuerdo con la componente temporal de la tabla, fragmentos reducidos de texto extraídos y tratados automáticamente por el sistema, que estén relacionados con el concepto seleccionado.

Contextualización

6. Selección de fragmentos reducidos de la tabla.
7. Visualización del fragmento original completo.
8. Visualización del documento original completo subrayado automáticamente.

Elaboración del informe final

9. Selección de fragmentos para la elaboración del informe final.
10. Organización automática de los fragmentos seleccionados en el informe final.
11. Edición del informe final.

En cada una de estas tareas el usuario y el sistema adquieren dos roles diferenciados. El sistema es el encargado de generar las piezas que conforman los niveles de acceso a la información, que requiere identificación, selección y organización de grandes cantidades de información en poco tiempo, por lo que debe ser el sistema quien las realice. Sin embargo, el sistema no conoce ni entiende con exactitud las necesidades del usuario, por lo que es el usuario el que debe tomar decisiones

de alto nivel sobre las piezas mostradas por el sistema. La tabla de la figura 9.2 muestra el reparto de tareas en cada una de las etapas implicadas en el proceso de síntesis.

ETAPAS	USUARIO	SISTEMA
Vista global de la información	1. Definición de términos de búsqueda	
		2. Recuperación de documentos relevantes.
		3. Identificación de conceptos clave
	5. Selección de conceptos clave	
		6. Generación de la tabla de oraciones reducidas
Contextualización	7. Selección de oraciones reducidas	
		8. Visualización del fragmento original
		9. Visualización del documento original
Elaboración de la información	10.Selección de fragmentos	
		11. Organización de los fragmentos seleccionados
	12. Edición del informe final	

Figura 9.2: Rol del usuario y del sistema en PRISMA

A lo largo de los siguientes apartados se describe cada uno de los componentes del modelo PRISMA.

9.2.1. Vista global de la información

El análisis de los contenidos requiere en primer lugar una vista general de la información contenida en las fuentes. Disponiendo de esta vista, el usuario podrá acceder a piezas de información para la recopilación de fragmentos. Por otro lado, se establece en esta etapa criterios organizativos de la información que ayudan al usuario a organizar y relacionar piezas para la elaboración del informe final. Esta vista debe de estar centrada en aspectos relacionados con las necesidades de información, que han de ser interpretadas por el usuario, por lo que debe construirse mediante un proceso interactivo.

El punto de partida es un conjunto de términos de búsqueda definidos por el usuario y el resultado es una tabla de oraciones reducidas descrita en el apartado anterior. A continuación se describe cada una de las tareas implicadas en esta etapa. La figura 9.3 muestra parte del interfaz de un prototipo implementado siguiendo el modelo PRISMA.

Expresión de términos de búsqueda

El usuario introduce a modo de consulta una serie de términos de búsqueda que ayudan al sistema a recopilar documentos susceptibles de contener información relevante para el proceso de SI (parte superior de la figura 9.2).

El usuario no tiene que expresar de antemano con precisión sus necesidades de información. Esto es un aspecto importante dado que en muchas ocasiones el usuario, bien no sabe de antemano con exactitud qué información necesita, o bien desconoce la terminología apropiada dentro del dominio, y más concretamente, la que se emplea en los documentos de los que se dispone.

Por ejemplo, supongamos que el usuario necesita elaborar un informe en relación a la tentativa de invasión de Haití en 1994. Sabe que el informe debe aportar información sobre el desarrollo del conflicto en los últimos meses. Sin embargo, desconoce cuales son los aspectos más determinantes en relación al asunto en este tiempo. El usuario será capaz de introducir términos como "invasión", o "Haití", que permiten discriminar un conjunto de documentos relevantes. Sin embargo, posiblemente le sería mucho más complicado elaborar una consulta del estilo de las empleadas en metodologías de evaluación de sistemas de recuperación de documentos. Por ejemplo:

"Generar un resumen con la información más importante en relación a la invasión de Haití por los soldados de la ONU y de los EEUU, tanto acerca de la discusión sobre la decisión de la ONU de enviar tropas americanas a Haití, como la invasión misma. Se hablará también de sus consecuencias directa"

Esta concreción de las necesidades del usuario se resolverá a lo largo del proceso interactivo de SI.

En el caso del ejemplo de interacción con el prototipo mostrado en la figura 9.3 el usuario ha introducido los términos de búsqueda *Invasión* y *Haití* con el objetivo de elaborar un informe que atienda a esta necesidad de información.

Recuperación de documentos

El proceso de recuperación de documentos consiste en una búsqueda de aquellos documentos susceptibles de contener información relevante para el proceso de SI. El proceso de recuperación requiere el procesamiento de grandes cantidades de información, por lo que el sistema juega un papel importante en esta fase.

Existen multitud de técnicas desarrolladas en el ámbito de sistemas de Recuperación de Información que permiten, de forma automática, identificar un conjunto de documentos que satisfaga las necesidades del usuario. Estas técnicas son implementadas en el motor de búsqueda del sistema de recuperación de documentos. En el caso de la Síntesis de Información es necesario un motor de búsqueda de con una buena cobertura sobre documentos relevantes, dado que la información requerida para la elaboración del informe no se encuentra en un único documento.

En este trabajo no hemos profundizado en este aspecto. Como primera aproximación y para asegurar la cobertura, hemos aplicado en el prototipo implementado (figura 9.3) una búsqueda booleana sobre los términos de consulta. Es decir, se recuperan todos aquellos documentos que contengan simultáneamente todos los

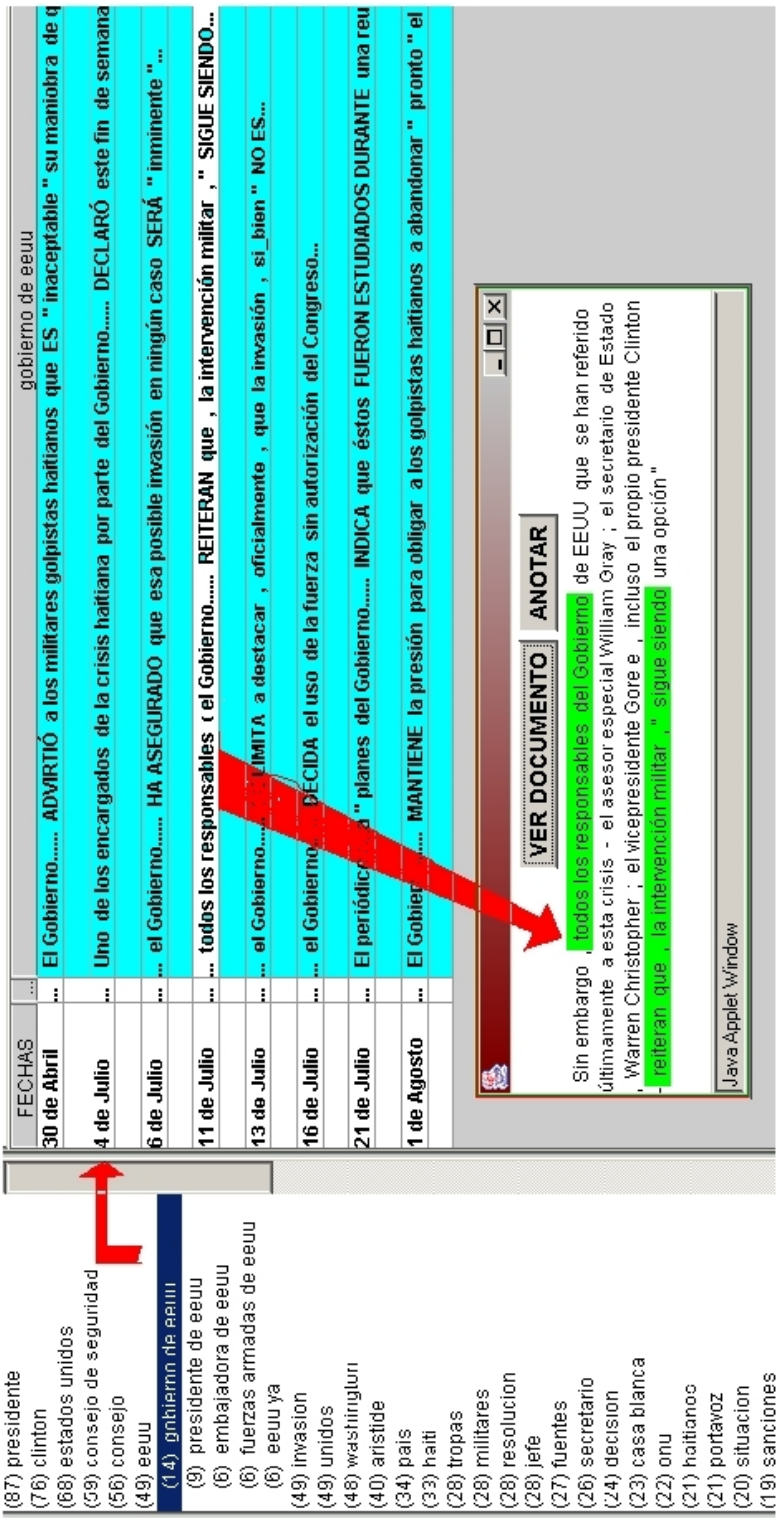


Figura 9.3: Interfaz del prototipo PRISMA

términos de la consulta. El único procesamiento de la consulta consiste en la eliminación de términos sin contenido (determinantes, nexos, etc.) y la reducción de los términos a su raíz.

Identificación de conceptos clave

En diferentes modelos interactivos de resumen se han planteado diferentes alternativas en cuanto al tipo de términos mostrados al usuario como conceptos relevantes. Algunos tipos de términos extraídos son: de sintagmas nominales [BKB+98, JLP02], palabras [BGMP01], unigramas y bigramas [LLS03] o nombres propios, términos multi-palabra y abreviaturas [NC99]. En el modelo PRISMA nos centramos en sintagmas nominales, identificados mediante el procesamiento lingüístico superficial descrito en la sección 8.1.

Una vez que se han identificado los términos candidatos es necesario ordenarlos por relevancia en una lista, de forma que la lista resultante ha de ser representativa de los conceptos clave del asunto tratado en las fuentes (parte izquierda de la figura 9.3). Para la ordenación de esta lista de terminos PRISMA emplea el criterio de pesado TFSYNTAX descrito en la sección 7.2.3. Como se mostró en la sección correspondiente, este criterio de pesado permite extraer una lista de términos con una buena cobertura respecto a los conceptos clave identificados manualmente en ISCORPUS. El criterio de pesado ha sido adaptado al sistema PRISMA de forma que se extraen, no palabras, sino sintagmas nominales que aparecen en una posición inmediatamente anterior al verbo.

En el prototipo implementado los conceptos clave se muestran ordenados por frecuencia de aparición en los documentos originales, y organizados jerárquicamente por relaciones de subsunción. Es decir, un término contiene al término superior en la jerarquía. Las jerarquías de subsunción permiten al usuario acceder, a partir de un concepto general, a conceptos más precisos. En la figura 9.3 vemos como *EEUU* subsume a *Gobierno de EEUU, presidente de EEUU, embajadora de EEUU y fuerzas armadas de EEUU*.

Selección de conceptos clave

Una vez preseleccionados de forma automática una lista de conceptos clave, el usuario selecciona aquellos que considere más relevantes en relación a sus necesidades de información.

Esta selección podría hacerse en un único paso. Esta es la estrategia seguida por sistemas interactivos de resumen basados en el control de parámetros del sistema (sección 3.3) . Sin embargo, diversos estudios muestran que las necesidades de información se definen a lo largo de todo el proceso de acceso a la información (sección 3.1). Esto implica que el usuario posiblemente no tenga la capacidad de definir a priori el conjunto óptimo de conceptos clave. Parece más adecuado que el usuario pueda explorar progresivamente la información asociada a distintos conceptos.

Generación de la tabla de oraciones reducidas

Es el sistema el encargado de generar una tabla de oraciones a partir de los conceptos seleccionados por el usuario. La propuesta se apoya en la utilidad del procesamiento lingüístico a nivel sintáctico en la presentación de la información y generación de niveles intermedios. El análisis sintáctico permite extraer estructuras del tipo agente-acción-objeto (AAO) *"Clinton [sujeto] dio[acción] una conferencia[objeto]..."*. La descripción de la herramienta empleada en el sistema PRISMA para el procesamiento sintáctico superficial ya ha sido descrita en la sección 8.1.

La tabla se organiza en función de la fecha de publicación del documento que contiene el fragmento de texto. La tabla contiene cuatro columnas correspondientes a la fecha, el sujeto, la acción y el objeto. Esta organización de las piezas de información asociadas a un concepto permite, en primer lugar, seleccionar aquellas piezas en las que el concepto, al formar parte del sujeto, adquiere un papel relevante en la oración, y en segundo lugar, mostrar de forma organizada las piezas de información, facilitando su lectura.

El acceso a estructuras AAO desde el sujeto sintáctico posee características que lo hace interesante como mecanismo de interacción con el usuario:

1. Atendiendo a los resultados del experimento descrito en el capítulo 8, podemos limitar los fragmentos asociados a un concepto clave a sus ocurrencias como sujeto sintáctico. Esta criba no decrementa excesivamente la cobertura de contenidos en relación al número de fragmentos explorados, con vistas a la realización del informe. Además, la cobertura sigue siendo mayor que la que ofrece una exploración por títulos de documentos.

2. El rol sintáctico de sujeto está asociado a la estrategia de extracción automática de conceptos basada en TFSYNTAX. En este criterio de pesado se escogen términos ocurrentes inmediatamente antes que el verbo. Esto hace que, a partir de una lista de términos extraídos mediante TFSYNTAX obtengamos una mayor cantidad de ocurrencias de estos términos como componentes del sujeto.

3. Una estructura AAO por sí misma puede ofrecer información a cerca de un tema o asunto, mientras que un término o un sintagma no lo hace *"Clinton dio una conferencia..."*.

4. La identificación de estructuras AAO supone un mecanismo de reducción en relación a la frase completa. Por ejemplo, *"Clinton, el presidente, dio una conferencia sobre el estado del conflicto"* se reduce a *"Clinton ... dio una conferencia..."*. Este procesamiento permite reducir la cantidad de información que el usuario debe recorrer durante el proceso de síntesis.

5. Las estructuras AAO son fáciles de organizar de forma automática a partir de sus contenidos, por ejemplo, por sujetos o verbos. Podemos alinear piezas de información como *"Clinton dio una conferencia.." "Clinton advirtió de las*

consecuencias..." , "Clinton volvió satisfecho...", etc. Es decir, el rol sintáctico de los componentes permite establecer un mecanismo de organización de las estructuras.

Por otro lado dos de los problemas que conlleva esta organización de la información son:

- Los sujetos sintácticos pueden ser distintos y referirse al mismo concepto clave. Sería necesario por lo tanto incorporar en PRISMA técnicas de resolución de correferencias que permitan agrupar sujetos como *"el presidente de EEUU"* y *"Clinton"*, o *"los atentados"* y *"los actos de terrorismo"*. Existen estrategias automáticas para la resolución de este tipo de problemas, pero otra alternativa consiste en que sea el propio usuario el agrupe conceptos de forma interactiva. Esto, aunque más costoso para el usuario, permite aunar conceptos como *"Israel" y "Sharón"*, que a efectos de la tabla puede considerarse como un mismo concepto, mientras que el sistema nunca será capaz de asociarlos de forma automática.

- La fecha de edición del documento no corresponde siempre con la fecha del evento representado en la oración. Este asunto se ha abordado en alguna aproximación al problema del Resumen Automático [BME99] mediante la identificación y normalización de fechas en los fragmentos de texto ("El año pasado"≡1993). Sería necesario evaluar la necesidad real de aplicar este tipo de estrategias en la generación de la tabla o si basta con considerar la fecha de edición del documento.

En el ejemplo de la figura 9.3 vemos como el usuario selecciona el agente *"Gobierno de EEUU"*, accediendo a todos aquellos fragmentos en los que dicho término participa como núcleo o complemento del sujeto sintáctico.

En este punto, aparecen una serie de cuestiones que es necesario analizar para confirmar la eficiencia de la tabla de oraciones reducidas como nivel intermedio de acceso a la información. Por ejemplo, ¿adquieren sentido para el usuario las oraciones reducidas fuera de su contexto original y en el contexto de la tabla? Se espera que las oraciones adquieran sentido en relación a otras oraciones asociadas al mismo concepto en momentos distintos o a oraciones asociadas a otros conceptos en el mismo intervalo de tiempo. Por ejemplo, la oración reducida *"EEUU ha criticado la decisión de la ONU.."* adquiere sentido si en un punto próximo de la tabla aparece una oración del tipo *"La ONU ha decidido que EEUU..."*, aunque no pertenezca necesariamente al mismo documento.

Otra cuestión pendiente es si es posible eliminar información redundante y qué técnicas son la más adecuadas. Una de las ventajas de las estructuras AAO como elementos intermedios de acceso a la información consiste en que éstas permiten alinear dos fragmentos no solo en función de los términos que aparecen en éstos, sino también en función del rol sintáctico de dichos términos. Por ejemplo, la diferencia sustancial entre las oraciones *"EEUU ha criticado la decisión de la ONU.."* y *"La ONU ha decidido que EEUU..."* se puede identificar por medio del

rol sintáctico que adquieren los términos en las respectivas oraciones, a pesar de que los términos empleados sean prácticamente los mismos. Algunas aproximaciones al problema del Resumen Automático han empleado este tipo de estrategias para el alineamiento de fragmentos [MB97, HL02a].

9.2.2. Contextualización

Al igual que en otros modelos interactivos de resumen, el usuario contextualiza las piezas de información mostradas como niveles intermedios. La contextualización es un requisito necesario en un esquema de interacción orientado a Síntesis de Información (sección 3.5). En PRISMA, el proceso de contextualización se realiza mediante la elección de oraciones reducidas y el acceso a las oraciones completas o al documento al que pertenece.

Selección de oraciones reducidas

El usuario identifica dentro de la tabla de oraciones reducidas aquellas en las que desee profundizar, bien para adquirir conocimiento que le pueda ser útil en el proceso de síntesis o bien para confirmar que la pieza de información debe de incluirse en el informe final.

Visualización de la frase completa

Una vez seleccionada una oración reducida, el sistema muestra al usuario la frase completa. En este punto, el usuario ya puede añadir la frase al informe final o bien puede acceder al documento completo.

La oración reducida aparece subrayada en la oración completa (margen derecho inferior de la figura 9.3). Como puede verse en la figura, el sistema resalta la partes más importantes de la frase —el núcleo de sujeto, el verbo y algunos complementos— con el fin de facilitar la lectura.

Visualización del documento

En muchos casos, la contextualización de una pieza de información seleccionada por el usuario requiere una lectura del contenido completo del documento original. Por otro lado, puede ser necesario recopilar fragmentos vecinos a la pieza preseleccionada. Es posible ayudar al usuario a realizar dicho recorrido, por ejemplo, resaltando mediante subrayado automático puntos de referencia en el texto.

Trabajos preliminares cubren técnicas como subrayado de términos, frases o párrafos relevantes. El sistema iNeast [LLS03] permite visualizar documentos subrayándose aquellos fragmentos considerados relevantes por el sistema. Los criterios empleados corresponden con algunas técnicas aplicadas en Resumen Automático. Es decir, fragmentos que contienen términos relevantes, localización de los fragmentos etc. En otros casos, el subrayado automático está orientado hacia el marcado de términos claves que constituyen puntos de referencia en el documento ([NC99], Google, Altavista etc.)

En esta aproximación el subrayado tiene como objetivo facilitar la lectura de cada uno de los fragmentos que componen el texto, subrayándose los elementos principales de la oración principal dentro de cada fragmento. Para ello, hemos realizado un estudio cualitativo previo de las características sintácticas del subrayado manual dentro de cada frase. Pudimos comprobar que gran parte de las estructuras sintácticas marcadas por los sujetos de prueba dentro de cada oración compuesta se ajustan a patrones sencillos fácilmente extraíbles de forma automática. La hipótesis de la que parte esta aproximación consiste en que el subrayado ayuda al usuario a la lectura no solo identificando fragmentos relevantes o puntos clave del documento, sino aportando una vista estructurada de cada oración compuesta.

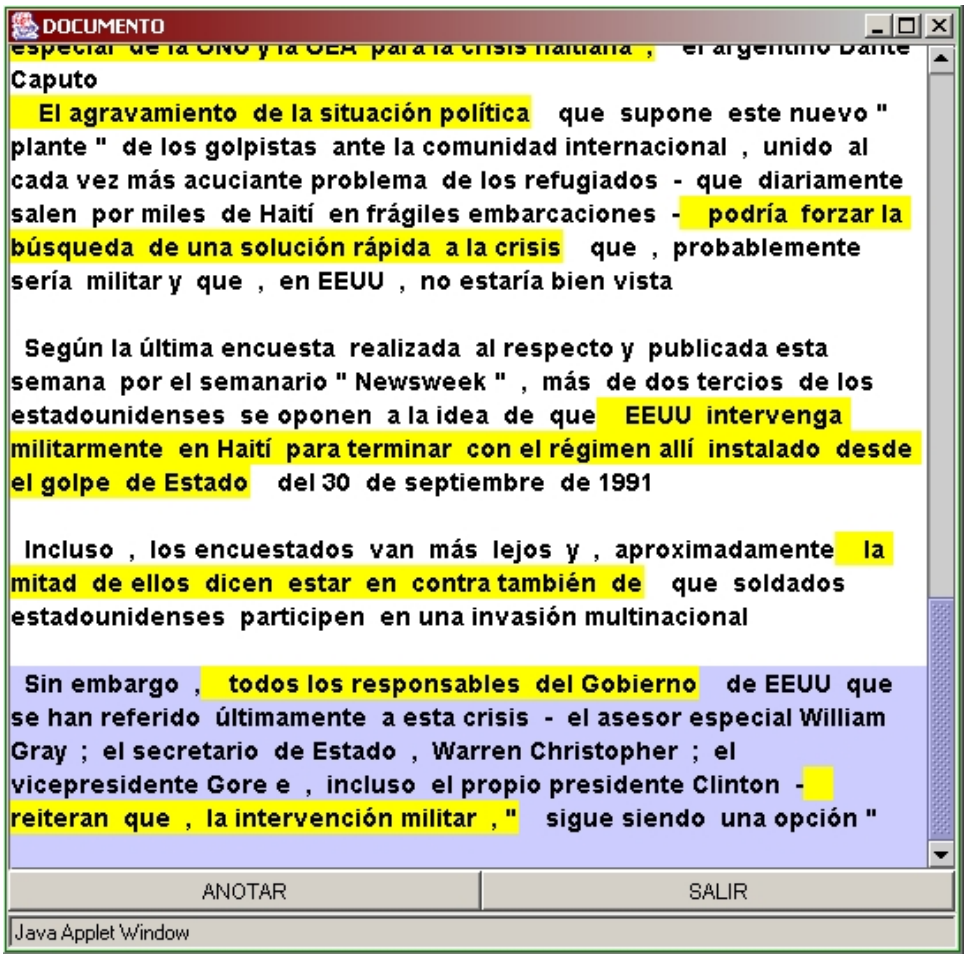


Figura 9.4: Visualización de un documento en el prototipo PRISMA

La figura 9.4 refleja la forma en que PRISMA muestra el contenido de un documento completo. Los criterios de PRISMA para la elección de la oración principal dentro de una oración compuesta son: proposiciones principales frente a oraciones subordinadas, estructuras completas frente a proposiciones sin sujeto o sin complementos del verbo, y localización de la proposición. Una vez seleccionada la

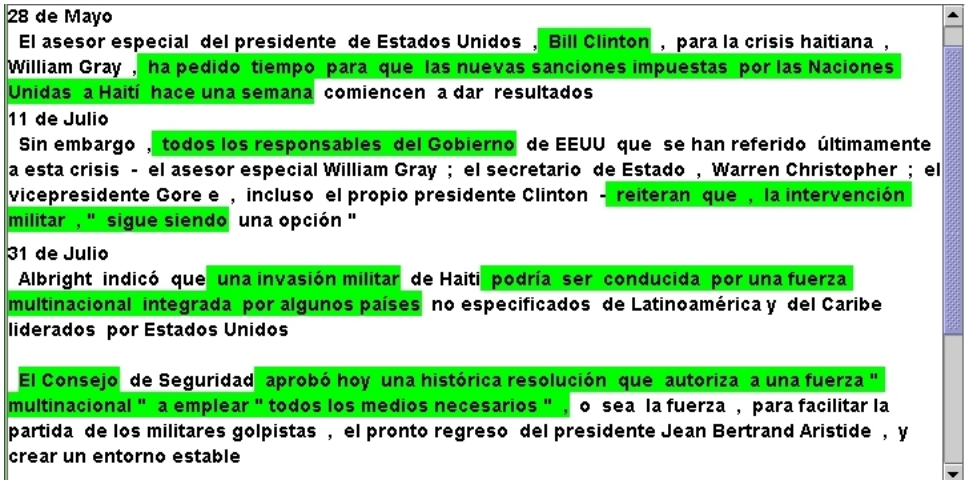


Figura 9.5: Visualización del informe generado en el prototipo PRISMA

proposición, se resalta únicamente el núcleo de sujeto, el verbo y los primeros complementos del éste. A partir del marco que muestra el documento completo, el usuario puede seleccionar también oraciones que considere relevantes para ser incluídas en el informe final.

En cualquier caso, se plantea como trabajo futuro evaluar cuantitativamente en qué medida el subrayado de las oraciones principales dentro de oraciones complejas ayuda al usuario en el proceso de lectura.

9.2.3. Elaboración del informe

Esta fase es añadida a las fases que constituyen el proceso de acceso a la información en otros modelos interactivos de resumen, en los que no se ofrece ninguna funcionalidad que refleje físicamente el proceso de recopilación de información llevado a cabo durante la síntesis de los documentos.

Selección de los fragmentos

Es el usuario el que determina qué fragmentos participan en el informe. Lo hace seleccionando piezas a partir de las frases completas mostradas por el sistema en la fase de exploración de la información preseleccionada. Es decir, a partir de los contextos mostrados al usuario.

En el modelo PRISMA se accede a un fragmento por medio de un concepto clave seleccionado por el usuario. El sistema subraya la oración reducida en donde el concepto aparece como sujeto dentro de la frase. Un aspecto interesante del modelo PRISMA es que esta oración reducida y el concepto desde el que se ha accedido se mantiene subrayada en el informe final, facilitando su legibilidad.

Organización de los fragmentos seleccionados

Es el usuario también el que toma las decisiones finales a este respecto. Sin embargo, el sistema puede, de forma automática sugerir una organización de los fragmentos seleccionados en el informe final. La figura 9.5 muestra como se organizan en el prototipo implementado las piezas de información seleccionadas por el usuario para constituir el informe final.

El mismo proceso de organización de las piezas de información sobre una tabla bidimensional nos puede servir para organizar de forma automática las piezas recopiladas por el usuario. Por ejemplo, el sistema puede ordenar los fragmentos por tiempo o por conceptos mostrando al usuario distintas vistas de la misma información. Es interesante, como ya hemos apuntado, el hecho de que el resultado del trabajo realizado por el usuario durante la fase de recopilación no es en realidad un texto compuesto por fragmentos recopilados, sino un objeto software constituido por piezas de información organizadas de forma que pueden plasmarse en un documento en formato tradicional de múltiples formas.

Edición del informe final

Independientemente de que el usuario pueda refinar la organización del informe propuesta por el sistema, quedan pendientes otras tareas de más alto nivel que no se abordan en este trabajo, en las que se ve implicado el problema de la generación de lenguaje. El sistema podría refinar aspectos formales del informe como la fluidez o la legibilidad de forma automática. Este aspecto ya ha sido abordado en algunas aproximaciones al problema del Resumen Automático.

Sin embargo, llegamos aquí a un punto en el que difícilmente podrá asistir el sistema al usuario. Ésta es la elaboración de nuevas piezas de información en base a inferencias realizadas a partir de la información recopilada. Por ejemplo *¿debe Sharón retirar sus tropas?*. Este tipo de preguntas requiere conocimiento experto, sentido común, y en algunos casos, llegando al extremo, características tan individuales de la persona como juicios de valor.

9.3. Implementación de PRISMA

Describiremos en esta sección brevemente algunas características de la implementación del sistema. PRISMA se compone fundamentalmente de tres módulos. Un módulo de indexación, un servidor y un interfaz de usuario.

9.3.1. Módulo de indexación

El módulo de indexación es el encargado de generar tablas con información extraída de la colección de documentos que requiere PRISMA. Para el prototipo PRISMA implementado se han indexado más de 200.000 documentos correspondientes a noticias en formato digital aportadas por la agencia EFE. Concretamente, el conjunto de noticias en español editadas en 1994.

El módulo de indexación genera una serie de tablas en formato Berkeley DB. El conjunto de tablas extraídas es:

- **docInfo.db:** Contiene información básica de los documentos, como su identificador, su fecha de edición, su título, su cabecera o el número de frases que contiene.
- **docTermPreVerb.db:** Contiene el conjunto de sintagmas nominales ocurrentes antes de un verbo identificados dentro de cada documento. Esta información se emplea fundamentalmente para la aplicación del criterio de pesado TFSYNTAX
- **fragChunks.db:** Contiene para cada frase de cada documento, su contenido con las etiquetas gramaticales anotadas.
- **termChunks.db** Contiene para cada término lematizado ocurrente en la colección, los identificadores de las frases y la posición con la que aparece. Esta tabla es empleada fundamentalmente para la recuperación de documentos mediante términos de búsqueda.

9.3.2. Módulo servidor

Tanto el módulo servidor como el interfaz de usuario ha sido programado en Java. El servidor es el responsable de acceder a la base de datos para ofrecer al interfaz de usuario la información que requiera. Ambos se ejecutan en máquinas distintas. La comunicación entre interfaz y servidor sigue un protocolo RMI. Este protocolo permite implementar una llamada al servidor como una simple llamada a un procedimiento de un objeto, facilitando así el proceso de implementación.

El servidor atiende a las siguientes peticiones:

Recuperación de documentos Dado un conjunto de términos de búsqueda, devuelve los identificadores de documentos en los que aparecen simultáneamente dichos términos. Los términos de búsqueda y los resultados son almacenados en el servidor, por lo que en cualquier otra llamada al servidor puede especificarse el conjunto de documentos según los términos de búsqueda sin coste adicional.

Identificación de conceptos clave Dado un conjunto de términos de búsqueda, lo que corresponde a un conjunto de documentos, el servidor devuelve una lista de conceptos clave extraídos por medio del criterio de pesado TFSYNTAX.

Recuperación del contenido de un documento Dado un identificador de documento, el servidor devuelve la información asociada al documento, es decir, su título, su fecha de edición, y su contenido con las etiquetas gramaticales anotadas.

Recuperación de frases asociadas a un término Dado un término, el servidor devuelve el conjunto de fragmentos en los que el término aparece como parte

del sujeto. Esta respuesta contiene, para cada fragmento, su contenido etiquetado, su identificador, y la oración reducida en la que aparece el término como sujeto.

Tanto los documentos relevantes como la lista de conceptos clave son recuperados de forma instantánea. La recuperación de fragmentos asociados a un término es algo más lenta, del orden de segundos. Esto se debe a que es necesario un procesamiento de orden lineal sobre todas las frases en donde aparece el concepto seleccionado con el fin de identificar su localización en la estructura sintáctica de la frase.

Este procesamiento sintáctico en tiempo de ejecución reduce la cantidad de información sintáctica a extraer en relación a un procesamiento completo en tiempo de indexación. Dicho de otro modo, dado un término que aparece como sujeto en una oración, requiere menos procesamiento lingüístico la identificación del verbo y objeto asociado a dicho sujeto que analizar sintácticamente la frase completa.

9.3.3. Módulo de interfaz de usuario

Este módulo, realiza las llamadas pertinentes al servidor y muestra al usuario la información obtenida en el formato descrito en los apartados anteriores.

Los componentes gráficos han sido implementados mediante objetos “Swing”. Cada uno de las secciones que componen el interfaz, así como sus constituyentes, se implementa como un objeto Java tipo “Panel”. Estos componentes son básicamente: el campo de consulta, la jerarquía de conceptos clave, la lista de títulos de documentos recuperados, la tabla de oraciones reducidas, la visualización de fragmentos, la visualización de documentos y el informe final.

Existe un objeto denominado internamente “PiezaInformación” que es compartido por los objetos descritos en el apartado anterior. Este objeto puede instanciarse como una frase, una oración reducida, una fecha, etc. Una instancia de este tipo puede contener la oración reducida dependiente del término desde el que se ya llegado a dicho fragmento, la frase completa y el conjunto de términos subrayados dentro de la frase. Este subrayado puede haberse generado mediante criterios puramente sintácticos, como ocurre en la visualización de documentos, o mediante la identificación de la oración reducida correspondiente al término desde el que se ha accedido a la oración. Este objeto es compartido por la la visualización del documento, la visualización del fragmento, la tabla de oraciones reducida, y el informe final. Mediante esta compartición del objeto “PiezaInformación”, el subrayado se mantiene en cualquiera de los paneles en los que se muestre el fragmento completo.

9.4. PRISMA frente a otros modelos interactivos de resumen

Como ya hemos apuntado en otros apartados del libro, las listas de conceptos relevantes extraídos automáticamente ya han sido empleadas en varias aproximaciones interactivas de elaboración de resúmenes.

En algunas de ellas [NC99, JLP02, LLS03] el resumen es generado automáticamente, tomando como entrada los conceptos clave seleccionados por el usuario a partir de una lista sugerida por el sistema. Es decir, se basan en control sobre parámetros del sistema.

En otros casos [BKB+98, BGMP01, RPH+95] el usuario accede a piezas de información organizadas por conceptos clave, elaborándose progresivamente el resumen. Es decir, se apoyan en un esquema de interacción basado en niveles intermedios de acceso a la información. La diferencia fundamental con el esquema anterior, consiste en que es el propio usuario el que, en última instancia, decide uno a uno qué fragmentos de textos son relevantes. Los niveles intermedios permiten generar un resumen de forma progresiva, mientras que el control de parámetros del sistema requiere una revisión del resumen completo en cada iteración usuario/sistema. Dado que el informe resultante de la SI es relativamente largo, hemos ajustado el modelo PRISMA a un esquema de interacción basado en niveles intermedios de acceso a la información.

En el caso de PRISMA, abordamos la tarea de SI a partir de un conjunto voluminoso de documentos donde es posible aplicar medidas estadísticas no sólo sobre ocurrencias de palabras, sino también sobre el rol sintáctico que desempeñan. Precisamente, la particularidad de PRISMA respecto a otros modelos reside en el uso de conocimiento sintáctico superficial para extraer los conceptos clave y mostrar al usuario la información asociada a cada concepto.

PRISMA frente al modelo de Boguraev

En este sentido, PRISMA mantiene varias semejanzas con el modelo propuesto en [BKB+98]:

- El conjunto de candidatos en el proceso de identificación de conceptos clave viene dado por los sintagmas nominales que aparecen en el documento.
- El rol sintáctico que desempeña el sintagma nominal es un criterio de selección aplicado en el proceso de extracción de conceptos clave.
- Se muestra al usuario el contexto de los conceptos clave en unidades de información más pequeñas que la frase completa.

Sin embargo, PRISMA y el modelo de Boguraev difieren en la tarea de alto nivel para la que se destina el sistema. En el caso de la aproximación de Boguraev, el objetivo consiste en identificar información relevante contenida en un único documento, mientras que en el caso de PRISMA tratamos el problema de la síntesis de información a partir de un conjunto de documentos. Este hecho deriva en una seria de diferencias entre ambos modelos:

Uso de estructuras AAO En el modelo de Boguraev se muestra por orden de aparición todos los contextos en los que aparece el concepto clave. En el caso de PRISMA al partir de un conjunto voluminoso de textos, es necesario seleccionar y organizar las piezas de información asociadas al concepto. En

concreto, se muestran aquellas en las que el concepto aparece como sujeto de una oración. Esto posibilita la visualización de los fragmentos por medio de la tabla de oraciones reducidas (agente-acción-objeto)

Visualización de documentos completos En PRISMA se ofrece además una vista del contenido completo de un documento resaltando la proposición principal de cada una de las frases del texto.

Vistas del informe final En PRISMA se dispone de un panel que contiene las oraciones seleccionadas por el usuario. Estas oraciones se estructuran en base al documento y al término desde el que se ha seleccionado el fragmento. El modelo de Boguraev no contempla estas funcionalidades.

9.5. Evaluación del modelo PRISMA

Con vistas a la evaluación de PRISMA, considerando los trabajos existentes en modelos interactivos de acceso a la información (capítulo 3), podemos identificar una serie de requisitos a tener en cuenta en la definición de un modelo interactivo de SI. Éstos son:

- **Mostrar al usuario una visión global de los contenidos.** La lista de conceptos clave y la tabla de oraciones reducidas ofrece al usuario una vista global de los contenidos de los documentos sobre los que se realiza un proceso de Síntesis de Información. La elaboración de esta vista global depende de la interpretación del usuario de sus necesidades de información, por lo que en PRISMA se realiza de forma interactiva.

- **Contextualización de los fragmentos de información mostrados al usuario.** La interpretación de los fragmentos de texto extraídos en PRISMA pueden requerir contextualización. El sistema PRISMA ofrece la posibilidad de acceder a oraciones completas o al documento completo a partir de una oración reducida.

- **Generación de informes extensos.** El usuario debe tener la posibilidad de construir progresivamente su informe, abordando sucesivamente diferentes aspectos del tema tratado en los documentos. El modelo PRISMA satisface esta condición.

- **Independencia de dominio de los documentos originales.** La adaptabilidad a diferentes dominios es una característica deseable en un sistema de SI, o por lo menos la adaptabilidad a diferentes tipos de contenido dentro de un mismo dominio. Tanto los criterios de identificación de conceptos clave como los de organización de la tabla de oraciones reducidas se basan en información sintáctica, que es, en principio, independiente del dominio tratado.

- **Adaptabilidad a distintos usuarios y necesidades de información.** En el modelo PRISMA, el usuario tiene la posibilidad de guiar el proceso en relación a sus necesidades de información. Es decir, la relevancia de las distintas piezas de información es un elemento flexible, no prefijado por el sistema.

Por otro lado, a lo largo de este trabajo hemos evaluado cuantitativamente diferentes mecanismos de acceso a la información integrados en el modelo PRISMA. Estas evaluaciones se apoyan en ISCORPUS y la metodología QARLA. Los aspectos de PRISMA ya evaluados son:

- **I. ¿Es necesario abordar el acceso a la información en la SI desde una perspectiva interactiva?** El estudio de los informes generados por distintos sujetos en ISCORPUS (capítulo 5) muestra un pequeño grado de acuerdo en cuanto a selección de fragmentos. Por otro lado, el análisis de distintos modelos cognitivos de acceso a la información sugieren que la SI es una tarea altamente subjetiva, en la que las necesidades de información se definen a lo largo de todo el proceso de acceso a la información. Es necesario por tanto la incorporación de componentes interactivos en un sistema de SI.
- **II. ¿Es más apropiado aplicar un esquema de interacción basado en niveles intermedios de acceso a la información?** El estudio del comportamiento de los sujetos de prueba durante la elaboración de informes en ISCORPUS, muestra que la SI es un proceso fundamentalmente extractivo. Es decir, es la recopilación de información la que embebe el trabajo de lectura y análisis de contenidos. Un esquema basado en niveles intermedios de acceso a la información ofrece la posibilidad al usuario de recopilar piezas de texto a medida que asimila y analiza los contenidos.
- **III. ¿Juegan los conceptos clave participantes en el asunto un papel importante en el proceso de la Síntesis de Información?** La distribución de conceptos clave, en nuestro caso, el conjunto de personas, organizaciones y factores relevantes en el asunto, son un rasgo característico de los informes manuales (sección 7.5). Por otro lado, según los experimentos descritos en el capítulo 8, la exploración de los contenidos de los documentos por términos permite una mayor cobertura con vistas a la elaboración de informes que la exploración por títulos.
- **IV. ¿Es TFSYNTAX un criterio de pesado adecuado para identificar y preseleccionar conceptos clave participantes en el asunto?** Es posible generar de forma automática una lista de términos con una amplia cobertura sobre la lista de conceptos clave presente en los documentos originales mediante el criterio de pesado TFSYNTAX (sección 7.5).
- **V. ¿Es útil considerar el rol de sintáctico para identificar piezas de información asociadas a un término?** No disponemos de muestras de interacción entre sujetos y PRISMA. Sin embargo, en el capítulo 8 se pudo comprobar que reducir los fragmentos asociados a cada término tomando únicamente aquellas ocurrencias de los mismos como parte del sujeto sintáctico, sigue

manteniendo una mejor cobertura sobre contenidos en relación a los fragmentos más accesibles que la exploración por títulos. Aunque esta reducción pueda hacer que se decremente sensiblemente la cobertura en relación a tomar todas las ocurrencias, la evaluación cualitativa de PRISMA muestra que puede facilitar la visualización de contenidos.

Existen sin embargo otros aspectos del modelo relacionados con la visualización de la información que requieren una evaluación mediante sujetos de prueba. Estos aspectos representan cuestiones abiertas que serán abordadas en trabajos futuros:

- **VI. ¿Facilitamos la exploración de piezas de información al reducir estructuras sintácticas?** Es necesario evaluar objetivamente con usuarios de prueba si la reducción de oraciones facilita el proceso de lectura sin perder excesiva cantidad de información. Esto incluye que el usuario pueda identificar con facilidad información redundante y que no exista confusión con las fechas en la que acontencen los sucesos descritos en las oraciones.
- **VII. ¿Facilitamos la lectura de un documento resaltando las proposiciones principales?** Es necesaria una evaluación formal del subrayado automático de los documentos basado en sintaxis.
- **VIII. ¿Cuál es el mejor criterio para organizar automáticamente los fragmentos recopilados por el usuario?** Una vez recopilados los fragmentos en un informe final, la cuestión que queda pendiente es cuáles son los criterios óptimos de organización de estos fragmentos, y si estos criterios pueden automatizarse. En cualquier caso, un sistema interactivo deja abierta la posibilidad de que el usuario escoja el criterio de organización sobre el que refinar su informe final.

Capítulo 10

Conclusiones y resultados del trabajo

En este capítulo se describen los resultados obtenidos a lo largo del trabajo de investigación descrito en esta monografía, se enumeran los productos resultantes, las publicaciones que ha generado y, por último, se especifican algunas líneas de investigación en marcha que parten de los resultados presentados en esta monografía.

10.1. Resultados del trabajo de investigación

Hemos obtenido una serie de conclusiones relativas al problema de la Síntesis de Información que permiten ubicar el problema (sección 10.1.1), establecer una metodología de evaluación (sección 10.1.2), y definir criterios de desarrollo de un modelo de SI (sección 10.1.3).

10.1.1. Acotación del problema

Hemos definido la Síntesis de Información como **el proceso mediante el cual, dada una necesidad de información compleja, se extraen, organizan e interrelacionan las piezas de información contenidas en un conjunto de documentos relevantes, con el fin de obtener un informe completo y sin redundancias que satisfaga esa necesidad de información.**

Desde un punto de vista computacional, hemos analizado las relaciones existentes entre la Síntesis de Información y tareas de acceso a la información ya abordadas desde un punto de vista computacional, como son la Recuperación de Información, la Búsqueda de Respuestas, la Extracción de Información o el Resumen Automático. La tarea de Resumen es la más semejante a la Síntesis de Información, en particular, cuando resume un conjunto de documentos y existe una necesidad de información predeterminada, es decir, el Resumen Multi-documento Orientado a Consulta.

Mediante el estudio del estado del arte en cuanto a estrategias automáticas de

resumen, hemos podido constatar que la SI plantea nuevos retos no abordados hasta el momento por los sistemas actuales de acceso a la información. Por ejemplo, el análisis de la información de los sistemas de resumen actuales se mantiene en un nivel muy superficial. Algunas de estas técnicas son el análisis de la cohesión, coherencia, identificación del tema tratado en los textos y extracción de fragmentos potencialmente relevantes. Estas técnicas son insuficientes para cubrir por sí solas el problema de la SI. Una forma de compensar estas limitaciones es introducir interacción entre usuario y sistema.

Hemos analizado y catalogado los esquemas de interacción entre sistema y usuario en sistemas de acceso a al información textual. A partir de este análisis concluémos que el esquema de interacción más apropiado para la SI es la presentación de niveles intermedios de información. Este esquema ofrece al usuario una visión global de los contenidos, posibilidad de contextualizar piezas de información, generar informes extensos, y puede adaptarse además a diferentes dominios, usuarios y necesidades de información.

En cuanto a la metodologías de evaluación, concluímos que para abordar la SI es necesario cubrir algunos aspectos no tratados por las metodologías existentes. La principal limitación de estas metodologías es la necesidad de combinar criterios de similitud entre el informe por evaluar y los informes modelo de referencia. Es decir, cómo aplicar y validar un conjunto de métricas de similitud, cada una de ellas con sus propiedades de escala y posibles redundancias entre métricas.

10.1.2. Metodología de evaluación: QARLA

Mediante la definición del marco QARLA hemos abordado la necesidad de considerar múltiples métricas de evaluación, múltiples informes modelo, y múltiples informes potenciales que se pueden obtener mediante la aplicación de un mismo modelo interactivo. Mediante un método no paramétrico es posible, siguiendo la metodología QARLA, combinar distintas métricas sin necesidad de ponderación, con independencia de las propiedades de escala de dichas métricas, y con independencia de métricas redundantes en el conjunto. Curiosamente, es la propia multiplicidad de informes modelo la que resuelve el problema de la multiplicidad de métricas de evaluación. QARLA resuelve este aspecto bajo el principio de que un buen informe debe asemejarse tanto a un modelo como dos modelos entre sí respecto a cualquier métrica que se considere. Dado que se dispone de múltiples informes modelo, es posible medir esta condición en términos de probabilidad.

Por otro lado, en QARLA se considera que el mejor conjunto de métricas es aquel que agrupa en el espacio a los informes modelo en relación a informes automáticos. QARLA se centra, por tanto, en identificar rasgos de los informes modelo que no aparecen en los informes evaluados. Dicho de otro modo, la evaluación en QARLA se orienta a la capacidad de los sistemas de emular un informe generado manualmente. Esto tiene la ventaja de ser un criterio objetivo de evaluación, y la desventaja de que no necesariamente replica los juicios subjetivos emitidos por jueces humanos. En cualquier caso hemos podido comprobar que, por lo menos para el caso de la evaluación de ciertas tareas de resumen, existe una estrecha

relación entre ambas perspectivas.

Hemos podido comprobar, además, que es posible medir en términos absolutos la fiabilidad del conjunto de informes automáticos sobre los que se evalúan las combinaciones de métricas. Es decir, podemos comprobar en términos de QARLA si el conjunto de sistemas de generación de informes es suficientemente representativo para afirmar que los resultados de la evaluación son estables.

Por último, hemos visto que QARLA es generalizable para la evaluación de sistemas interactivos. La metodología QARLA no requiere usuarios de prueba, por lo que aspectos como la usabilidad del sistema difícilmente pueden ser evaluados. Sin embargo, otros aspectos, como la cobertura sobre contenidos de una estrategia de exploración, sí pueden ser tratados desde QARLA, lo que solventa algunos problemas de otras metodologías de evaluación de sistemas interactivos. Para ello, la metodología introduce el concepto de evaluación de informes/resúmenes potenciales.

10.1.3. Desarrollo de un modelo de SI: PRISMA

En esta monografía se ha presentado un modelo de asistente para la Síntesis de Información que se basa en una serie de evidencias empíricas recogidas sobre el corpus de prueba (ISCORPUS) utilizando básicamente la metodología de evaluación QARLA.

Características generales de la SI

Mediante el análisis del comportamiento de los sujetos de prueba, hemos podido identificar algunas características del proceso de Síntesis de Información en el contexto de ISCORPUS que guían el desarrollo de un modelo interactivo:

- La tarea de SI es fundamentalmente extractiva, al menos en un primer orden de aproximación. Los sujetos no sintieron la necesidad de reescribir o anotar información sobre los informes elaborados de forma extractiva. Es decir, independientemente de su fluidez y coherencia, los informes generados de forma extractiva, son autocontenidos (sección 5.6). Además, los sujetos extrajeron en promedio la misma cantidad de fragmentos a lo largo de todo el proceso de síntesis. Es decir, no hemos podido identificar una fase de análisis previa o posterior a la selección de piezas de información.

- El proceso de SI varía según el tipo de necesidad de información y el tipo de documentos de partida. Hemos establecido un distinción empírica entre el caso de documentos que tratan diferentes aspectos de un mismo asunto (temas mono-evento) del caso en el que aparecen distintas instancias de un mismo tipo de evento (temas multi-evento). En temas mono-evento la SI requiere un proceso de análisis mayor que en temas multi-evento. Además, aparece un menor grado de acuerdo entre informes modelos en cuanto a cuáles son las frases, documentos o conceptos clave que deben quedar reflejados (sección 5.6).

Síntesis de Información y conceptos clave

La hipótesis principal sobre la que se definimos el modelo PRISMA es que los conceptos clave del tema tratado en los documentos originales juegan un papel importante en el proceso de Síntesis de Información. Los experimentos descritos a lo largo de este libro han permitido extraer las siguientes conclusiones:

- La distribución de conceptos clave en un informe es un aspecto por considerar en la elaboración y evaluación de informes. Por un lado, una estrategia tan básica como la extracción de primeras frases de los documentos se asemeja a los informes modelo en cuanto a las frases o al vocabulario que contiene, pero no en cuanto a su distribución de términos clave. Por otro lado, las métricas de evaluación basadas en otros rasgos adquieren mayor fiabilidad cuando se compaginan con métricas basadas en conceptos clave (secciones 7.1 y 7.4).

- Es posible extraer de forma automática conceptos clave a partir de los documentos originales. Criterios de selección estadísticos elaborados no superan a criterios sencillos como la frecuencia del término. Esto se debe a que partimos de un volumen de información considerable. Sin embargo, los resultados sí mejoran cuando consideramos información sintáctica superficial, como es la distancia al verbo (sección 7.2).

- No necesariamente un informe con más densidad de conceptos clave es de mejor calidad. Por ejemplo, la selección de primeras frases de documentos produce informes con una mayor densidad de conceptos clave que un informe modelo. Por otro lado, es posible estimar la frecuencia de los conceptos clave en un informe modelo en función de su frecuencia en los documentos originales(sección 7.3).

Estrategias de exploración de contenidos en SI

Finalmente, hemos comparado diferentes técnicas de exploración de contenidos en el contexto de la Síntesis de Información. Concretamente, hemos estudiado la cobertura de contenidos con vistas a la realización de un informe semejante a los informes modelo en ISCORPUS. A partir de estos experimentos hemos obtenido los siguientes resultados:

- La exploración de fragmentos asociados a conceptos clave es un esquema de interacción que supera en cuanto a cobertura de contenidos a la exploración de términos por títulos en el contexto de la SI (sección 8.3).

- La selección de documentos relevantes mejora en cuanto a cobertura de contenidos a la exploración de documentos en orden cronológico. Este hecho es más notable en el caso de los temas mono-evento y cuando se consideran pocas piezas de información accesibles desde el sistema (sección 8.3).

- Mostrar exclusivamente piezas de información en las que el concepto clave aparece formando parte del sujeto sintáctico no reduce en exceso la cobertura de contenidos en relación a mostrar todas las ocurrencias. En algunos casos concretos (pocas piezas de información exploradas en temas mono-evento) puede aumentar. Además, un análisis cualitativo del prototipo PRISMA muestra que presentar al usuario dichos fragmentos en tablas concepto-acción-objeto mejora el proceso de visualización de piezas de información (sección 8.3).

10.2. Productos resultantes

El desarrollo del trabajo descrito en este libro ha generado diferentes productos reutilizables en futuras investigaciones y aplicaciones. El primer producto es la propia definición del problema de la SI. En el foro de evaluación DUC2005 [1] (Document Understanding Conferences), se ha tomado como referencia uno de nuestros trabajos [APG+04] para definir la tarea sobre la que evaluar sistemas de resumen en, al menos, 2005 y 2006. El principal interés por la SI surge de combinar el problema de la generación de resúmenes con la búsqueda de respuestas, además de partir de un conjunto más amplio de documentos y elaborar resúmenes de mayor longitud.

El segundo producto es ISCORPUS, un corpus de SI. Éste contiene una cantidad representativa de informes generados en condiciones controladas por un conjunto de sujetos de prueba. Contiene además listas de conceptos clave anotadas manualmente por los mismos sujetos de prueba tras la elaboración del informe. Además ha quedado registrada la serie de acciones realizadas por los sujetos durante el proceso de elaboración del informe. Esta información es potencialmente útil para futuros estudios sobre el comportamiento de los humanos durante el proceso de SI. Por último, se añade en el corpus una serie de informes generados mediante estrategias básicas, como es la selección de primeros fragmentos de documentos o la identificación de documentos relevantes.

El tercer producto resultante es el marco de evaluación QARLA. Este marco tiene como entrada un conjunto de elementos modelo, elementos por evaluar, y métricas de similitud. QARLA es, por tanto, aplicable a tareas en las que se disponga de multiplicidad de modelos y métricas de evaluación. Es decir, tareas en las que se genera un producto elaborado y con un componente subjetivo o creativo. En el campo de la lingüística computacional este marco es aplicable a tareas como la traducción automática o la búsqueda de respuestas de tipo definición. Es decir, tareas de carácter subjetivo en las que el sistema genera un producto elaborado, con diferentes aspectos susceptibles de evaluación.

Finalmente, estre trabajo se ha materializado en el modelo y prototipo PRISMA. Dado que el modelo se basa en la extracción de conceptos relevantes y la presentación de la información aplicando únicamente conocimiento estadístico y sintáctico superficial, no se restringe en principio a ningún dominio. PRISMA pue-

[1] http://www-nlpir.nist.gov/projects/duc

de ser aplicado por tanto en otros contextos como por ejemplo, el dominio médico o de comercio. Por otro lado, el prototipo implementado PRISMA ha sido probado sobre una colección voluminosa de artículos periodísticos. Mediante el uso de una base de datos eficiente y herramientas de procesamiento sencillas, el acceso a la información es rápido y robusto. La implementación por medio de un protocolo de comunicación tipo cliente-servidor, permite el acceso simultáneo desde distintas máquinas. La arquitectura empleada permite ejecutar todos los procesos de interacción con el usuario en una máquina remota, mientras que el servidor se limita a gestionar los accesos a la base de datos.

10.3. Publicaciones del autor relacionadas con el trabajo

En [APnGV04] (Lecture Notes in Computer Sciences) se describe un prototipo antecesor de PRISMA orientado a recuperación interactiva de documentos. En este prototipo se accede a la información en el contexto periodístico por medio de una lista de entidades nombradas (personas, organizaciones, lugares, etc.) categorizadas.

En [AGPnV04] (SEPLN) se describe el sistema PRISMA y la metodología de desarrollo empleada. Incluye una descripción de cada una de las funcionalidades de PRISMA, una metodología de desarrollo basada en la evaluación de aspectos parciales del modelo y una comparación de PRISMA con modelos interactivos de resumen.

En [APG$^+$04] (ACL2004) se describe un estudio sobre la necesidad de considerar conceptos clave en sistemas de SI. Este trabajo incluye la descripción de ISCORPUS y su desarrollo, y un análisis sobre los conceptos clave como rasgo común en informes generados manualmente.

En [AGP$^+$04] (COLING2004) se presenta y evalúa el criterio de pesado TFSYNTAX para la extracción automática de conceptos clave. Este trabajo incluye un análisis sobre la distribución de conceptos clave en distintas partes de la oración, como sujeto sintáctico, complementos del sujeto, predicados, etc. Finalmente, se compara el criterio de pesado TFSYNTAX con criterios basados en técnicas estadísticas.

En [AGPnV05b](ACL2005) se describe y evalúa el marco de evaluación de resúmenes QARLA, aplicado sobre ISCORPUS. En [APnGV05] y [AGPnV05a] se aplica QARLA sobre el corpus de resúmenes DUC2004.

En [GAH05] se describe un trabajo de aplicación del modelo QARLA en el contexto de los sistemas de traducción automática. En él, se combinan dentro del marco QARLA un conjunto representativo de métricas de evaluación, obteniéndose altas correlaciones con juicios humanos.

Finalmente, en [GA06] se describe el sistema IQ, una implementación del modelo QARLA orientada a la evaluación de sistemas de traducción.

10.4. Trabajos en desarrollo y líneas futuras

De este trabajo se derivan fundamentalmente dos líneas: la aplicación del modelo PRISMA en entornos reales y la adaptación de la metodología QARLA a nuevos problemas, no necesariamente relacionados con la evaluación de sistemas.

En cuanto al modelo QARLA, se ha estudiando su aplicación en el dominio de los sistemas de traducción automática. Los primeros resultados obtenidos se describen en [GAH05]. En principio, QARLA ha sido aplicado sobre dos corpus de traducciones de poca longitud (8 y 6 palabras en promedio). En este contexto, con tan poca información por traducción, resulta muy complicado establecer métricas de similitud capaces de identificar los rasgos comunes de traducciones modelo, por lo que se ha definido una variante de la medida QUEEN. Esta variante mantiene las propiedades primitivas del modelo, aunque con la desventaja de que emplea menos cantidad de información para estimar la calidad de una traducción. Como resultado de este trabajo, la gran mayoría de las métricas individuales elevan su correlación con juicios humanos al aplicarse dentro de QARLA.

En cuanto a la aplicación del modelo PRISMA en entornos reales, ésta permitiría la evaluación de aspectos relacionados con la interacción entre usuario y sistema, no contemplados en los experimentos descritos en este libro (sección 9.5). Por otro lado, sería interesante evaluar la aplicación en PRISMA de alguna herramienta de procesamiento lingüístico más precisa, o la adecuación del modelo en otros dominios como el médico o la enseñanza. Este tipo de cuestiones será abordado en futuros experimentos que ayudarán a entender con más claridad en qué consiste el proceso de Síntesis de Información.

En definitiva, la Síntesis de Información es un problema de alto coste que surge en multitud de dominios, y sobre el que no se han desarrollado por el momento trabajos en profundidad desde un punto de vista computacional. Pero, ¿es ésta una línea de investigación susceptible de ofrecer beneficios a nivel práctico? En las últimas décadas la capacidad de almacenar y recuperar grandes cantidades de información textual ha supuesto en muchos sentidos una transformación en la industria. La cuestión que se plantea entonces es si los sistemas son capaces de asistir al usuario en el análisis y elaboración de información a partir de todos esos datos en formato textual, esto es, la Síntesis de Información.

Bibliografía

[ABBN00] R. Ando, B. Boguraev, R. Byrd, and M.Ñeff. Multi-document Summarization by Visualizing Topical Content. In *Proceedings of ANLP/NAACL 2000 Workshop on Automatic Summarization*, 2000.

[ACS+98] James Allan, James P. Callan, Mark Sanderson, Jinxi Xu, and Steven Wegmann. INQUERY and TREC-7. In *Text REtrieval Conference*, pages 148–163, 1998.

[AGP+04] E. Amigó, J Gonzalo, V. . Peinado, A. Peñas, and F. Verdejo. Using Syntactic Information to Extract Relevant Terms for Multi-Document Summarization. In *Proceedings of the 36th Annual Conference on Computational Linguisticsion for Computational Linguistics (Coling'04)*, Geneva, August 2004.

[AGPnV04] E. Amigó, J. Gonzalo, A. Peñas, and F. Verdejo. PRISMA: Un Modelo Interactivo de Síntesis de Información. *Sociedad Española para el Procesamiento de Lenguaje Natural*, 33, 2004.

[AGPnV05a] E. Amigó, J. Gonzalo, A. Peñas, and F. Verdejo. Evaluating DUC 2004 with QARLA Framework. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Michigan, June 2005. Association for Computational Linguistics.

[AGPnV05b] E. Amigó, J. Gonzalo, A. Peñas, and F. Verdejo. QARLA: a Framework for the Evaluation of Automatic Sumarization. In *43th Annual Meeting of the Association for Computational Linguistics*, Michigan, June 2005. Association for Computational Linguistics.

[APG+04] E. Amigó, V. Peinado, J. Gonzalo, A. Peñas, and F. Verdejo. An Empirical Study of Information Synthesis Task. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, July 2004.

[APnGV04] E. Amigó, A. Peñas, J. Gonzalo, and F. Verdejo. *Suggesting Named Entities for Information Access*, volume 2588 of *Lecture Notes in Computer Sciences*, pages 557–561. Springer Verlag, July 2004.

[APnGV05] E. Amigó, A. Peñas, J. Gonzalo, and F. Verdejo. Evaluación de Resúmenes Automáticos mediante QARLA. *Sociedad Española para el Procesamiento de Lenguaje Natural*, 34, 2005.

[Bar03] R. Barzilay. Information Fusion for Multidocument Summarization: Paraphrasing and Generation, 2003. PhD Thesis, Columbia University.

[Bat90] M. Bates. Where Should the Person Stop and the Information Search Interface Start. *Information Processing and Management*, 26(5):575–591, 1990.

[Bax58] P. B. Baxendale. Man-Made Index for Technical Literature - an Experiment. *IBM Journal of Research and Development*, 2(4):354–361, 1958.

[BCD02] M. Brunn, Y. Chali, and B. Dufour. U of L Summarizer at DUC2002. In *Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July 2002.

[BCV+01] Burger, C. Cardie, Chaudhri V., Gaizauskas R., Harabagiu S., D. Israel, C. Jacquemin, C. Lin, S. Maiorano, G. Miller, D. Moldovan, B. Ogden, J. Prager, E. Riloff, A. Singhal, R. Shrihari, T. Strzalkowski, and R. Voorhees, E.and Weischedel. Structures to Roadmap Research in Question & Answering (Q&A), 2001.

[BDH+00] B. Baldwin, R. Donaway, E. Hovy, E. Liddy, I. Mani, D. Marcu, K. McKeown, V. Mittal, M. Moens, D. Radev, K. Sparck-Jones, B. Sundheim, S. Teufel, R. Weischedel, and M. White. An Evaluation Road Map for Summarization Research. The Summarization Roadmap, 2000.

[Bel80] N.J. Belkin. Anomalous States of Knowledge as a Basis for Information Retrieval. *Canadian Journal of Information Science*, 5:133–143, 1980.

[Bel93] N. J. Belkin. Interaction with Texts: Information Retrieval as Information-Seeking behavior. In *Information Retrieval*, pages 55–66, 1993.

[Bel02] N.J. Belkin. Evaluation Methods for Interactive Information Retrieval. Invited presentation at LIDA 2002, (Libraries in the Digital Age), Dubrovnik, Croatia, Mayo 2002.

[BGMP01] O. Buyukkokten, H. García-Molina, and A. Paepcke. Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld

Devices. In *Proceedings of 10th International WWW Conference*, Hong-kong, 2001.

[BKB+98] B. Boguraev, C. Kennedy, R. Bellamy, S. Brawer, Y. Wong, and J. Swartz. Dynamic Presentation of Document Content for Rapid On-line Skimming. In *Proceedings of the AAAI Spring 1998 Symposium on Intelligent Text Summarization*, 1998.

[BM98] B. Baldwin and T. S. Morton. Dynamic Co-Reference Based Summarization. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP-3)*, Granada, Spain, June 1998.

[BME99] R. Barzilay, K. R. McKeown, and M. Elhadad. Information Fusion in the Context of Multi-Document Summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 550–557, College Park, Maryland, USA, June 16–20 1999.

[BP02] C. Blake and W. Pratt. Collaborative Information Synthesis. In *Proceedings of Annual Conference of the American Society for Information Science and Technology (ASIST 2002), Philadelphia, PA*, 2002.

[Chu97] G. Churcher. Dialogue Management Systems: a Survey and Overview. Report 97.6, School of Computer Studies, University of Leeds, 1997.

[CKPT92] D.D. Cutting, J. Karger, Pedersen, and J. Turkey. Scatter/gather: A Cluster-Based Approach to Browsing Large Document Collections. In *Proceedings of the Fifteenth Interntional Conference on Research and Development in Information Retrieval*, pages 318–329, 1992.

[CKSZ03] Y. Chali, M. Kolla, N. Singh, and Z. Zhang. The university of lethbridge text summarizer at duc 2003. In *Workshop on Text Summarization*, Edmonton, Canadá, 2003.

[Cou03] D. Coughlin. Correlating Automated and Human Assessments of Machine Translation Quality. In *In Proceedings of MT Summit IX*, New Orleans,LA, 2003.

[CR03] C. Culy and S. Riehemann. The Limits of N-Gram Translation Evaluation Metrics. In *Proceedings of MT Summit IX*, New Orleans,LA, 2003.

[CS04] T. Copeck and S. Szpakowicz. Vocabulary Usage in Newswire Summaries. In *Proceedings of ACL-04 Workshop on Text Summarization*, Barcelona, Spain, 2004.

[Dae93] Walter Daelemans. Memory-based lexical acquisition and processing. In *EAMT Workshop*, pages 85–98, 1993.

[Der77] B. Dervin. Useful Theory for Librarianship: Communication, not Information. *Drexel Library Quarterly*, 13(3):16–32, 1977.

[Dun93] T. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74, 1993.

[Edm69] H. P. Edmundson. New Methods in Automatic Extracting. *Journal of the Association for Computing Machinery*, 16(2):264–285, April 1969.

[GA06] Jesús Giménez and Enrique Amigó. IQMT: A Framework for Automatic Machine Translation Evaluation. In *Proceedings of the 5th LREC*, 2006.

[GAH05] Jesús Giménez, Enrique Amigó, and Chiori Hori. Machine Translation Evaluation Inside QARLA. In *Proceedings of the International Workshop on Spoken Language Technology (IWSLT'05)*, 2005.

[GHW03] J. Ge, X. Huang, and L. Wu. Approaches to Event-Focused Summarization Based on Named Entities and Query Words. In *Workshop on Text Summarization*, Edmonton, Canadá, 2003.

[GMCC00] J. Goldstein, V. O. Mittal, J. G. Carbonell, and J. P. Callan. Creating and Evaluating Multi-Document Sentence Extract Summaries. In *CIKM*, pages 165–172, 2000.

[Hea99] *Modern Information Retrieval*, chapter User Interfaces and Visualization, pages 257–224. Addison-Wesley Longman Publishing Company, 1999.

[HH76] M.A.K. Halliday and R. Hasan. *Cohesion in English*. Longman, London, 1976.

[Hir94] Litman D. J. Hirschberg, J. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics*, 19(3):501–530, 1994.

[HL02a] S. Harabagiu and F. Lacatusu. Generating single and multi document summaries with GISTEXTER. In *Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July 2002.

[HL02b] Eduard Hovy and Chin-Yew Lin. Manual and Automatic Evaluation of Summaries. In Udo Hahn and Donna Harman, editors, *ACL02-WS*, July 11–12 2002.

[HM98] E. Hovy and D. Marcu. Coling/acl-98 tutorial on automated text summarization, 1998.

[HO02] W. Hersh and P. Over. Trec 2002 interactive track report. In *Proceedings of the Text Retrieval Conference (TREC) 2002*, 2002.

[Hou97] D. House. Interactive Text Summarization for Fast Answers, 1997.

[HSIM02] T. Hirao, Y. Sasaki, H. Isozaki, and E. Maeda. NTT's Text Summarization System for DUC 2002 . In *Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics*, 2002.

[HT00] W. Hersh and A. Turpin. Why batch and user evaluations do not give the same results. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2000.

[HT03] H. V. Halteren and S. Teufel. Examining the consensus betwen human summaries: initial experiments with factoids analysis. In *HLT/NAACL-2003 Workshop on Automatic Summarization*, 2003.

[J.98] McKenzie. J. Searching for the Grail: Power Searching with Digital Logic. *The Educational Technology Journal*, 7 (4), 1998.

[JBME98] H. Jing, R. Barzilay, K. McKeown, and M. Elhadad. Summarization Evaluation Methods Experiments and Analysis. In *In AAAI Intelligent Text Summarization Workshop*, pages 60–68, 1998.

[JLP02] S. Jones, S. Lundy, and G. W. Paynter. Interactive Document Summarization Using Automatically Extracted Keyphrases. In *Proceedings of the 35th Hawaii International Conference on System Sciences*, 2002.

[KCC+98] B. Ken, Y. Chali, T. Copeck, S. Matwin, and S. Szpakowicz. The Design of a Configurable Text Summarization System. citeseer.ist.psu.edu/barker98design.html, 1998.

[KPC95] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *Proceedings of the 18 ACM SIGIR conference on research and development in information retrieval*, 1995.

[KSH02] W. Kraaij, M. Spitters, and A. Hulth. Headline Extraction based on a Combination of Uni- and Multi-Document Summarization Techniques. In *Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July 2002.

[Kuh91] C. C. Kuhlthau. Inside the Search Process: Information Seeking from the User's Perpective. *Journal of the American Society for Information Science*, 42(5):361–371, 1991.

[Kuh01] C. Kuhlthau. Information Search Process of Lawyers: A Call for Just for Me Information Services. *Journal of Documentation*, 57:25–43, 2001.

[LH02] C. Lin and E. Hovy. NeATS in DUC 2002. In *Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July 2002.

[LH03a] C. Lin and E. H. Hovy. Automatic Evaluation of Summaries Using N-gram Co-ocurrence Statistics. In *Proceeding of 2003 Language Technology Conference (HLT-NAACL 2003)*, 2003.

[LH03b] Chin-Yew Lin and Eduard Hovy. The Potential and Limitations of Automatic Sentence Extraction for Summarization. In Dragomir Radev and Simone Teufel, editors, *HLT-NAACL 2003 Workshop: Text Summarization (DUC03)*, Edmonton, Alberta, Canada, May 31 - June 1 2003. Association for Computational Linguistics.

[Lin04a] C. Lin. Orange: a Method for Evaluating Automatic Metrics for Machine Translation. In *Proceedings of the 36th Annual Conference on Computational Linguisticsion for Computational Linguistics (Coling'04)*, Geneva, August 2004.

[Lin04b] Chin-Yew Lin. Rouge: A Package for Automatic Evaluation of Summaries. In Marie-Francine Moens and Stan Szpakowicz, editors, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[LLS03] A. Leuski, C. Y. Lin, and S. Stubblebine. iNEATS: Interactive Multidocument Summarization. In *Proceedings of the 4lst Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan, 2003.

[Mah99] K. Mahesh. Hypertext Summary Extraction for Fast Document Browsing. In *Natural Language Processing for the World Wide Web. Papers from the 1997 AAAI Spring Symposium*, pages 95–104, Stanford, CA, 1999.

[Man01a] I. Mani. *Automatic Summarization*, volume 3 of *Natural Language Processing*. John Benjamins Publishing Company, Amsterdam/Philadelphia, 2001.

[Man01b] I. Mani. Summarization Evaluation: an Overview. In *Proceedings of the NTCIR Workshop 2 Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization*, 2001.

[Mar95] D. Marcu. Discourse Trees Are Good Indicators of Importance in Text. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 123–136, Cambridge, MA, 1995. MIT Press.

[Mar97] D. Marcu. From Discourse Structures to Text Summaries. In *The Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 82–88, Madrid, Spain, July 11 1997.

[May95] M. Maybury. Generating Summaries from event Data . *Information Processing and Management*, 31(5):735–751, 1995.

[MB97] I. Mani and E. Bloedorn. Multi-Document Summarization by Graph Search and Matching. In *Proceedings of the 14th National Conference on Artificial Intelligence*, pages 622–628, Providence, Rhode Island, 1997.

[MBE+01] K. R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, M. Yen Kan, B. Schiffman, and S. Teufel. Columbia Multi-Document Summarization: Approach and Evaluation. In *Proceedings of the Document Understanding Conference (DUC-2001)*, 2001.

[MG01] D. Marcu and L. Gerber. An Inquiry into the Nature of Multidocument Abstracts, Extracts, and Their Evaluation. In Jade Goldstein and Chin-Yew Lin, editors, *Proceedings of the Workshop on Automatic Summarization at the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–8, Pittsburgh, PA, June 2001.

[MHK+98] I. Mani, D. House, G. Klein, L. Hirschman, L. Obrst, T. Firmin, M. Chrzanowski, and B. Sundheim. The Tipster Summac Text Summarization Evaluation: Final report. Technical report, DARPA, 1998.

[MKA92] A.H. Morris, G.M. Kasper, and D.A. Adams. The effects and limitations of automated text condensing on reading comprehension performance. *Information Systems Research*, 2(1):17–35, 1992.

[MR95] K. R. McKeown and D. R. Radev. Generating Summaries of Multiple News Articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 1995.

[MS02] R. Barzilay D. Evans V. Hatzivassiloglou J. L. Klavans A. Nenkova C. Sable B. Schiffman McKeown, K. R. and S. Sigelman. Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. In *Proceedings of Human Language Technology Conference*, CA, USA, 2002.

[MSB97] M. Mitra, A. Singhal, and C. Buckley. Automatic Text Summarization by Paragraph Extraction. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, pages 39–46, Madrid, Spain, July 1997. Association for Computational Linguistics.

[NC99] M. S. Neff and J. W. Cooper. Ashram: active summarization and markup. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS-32): Understanding Digital Documents*, 1999.

[NM97] D.Ñicholas and H. Martin. Assessing information needs: a case study of journalists. In *Aslib Proceedings*, volume 49,2, pages 43–53, 1997.

[NP04] A.Ñenkova and R. Passonneau. Evaluating content selection in summarization: The pyramid method. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 145–152, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[ORL02] J. Otterbacher, D. R. Radev, and A. Luo. Revisions that Improve Cohesion in Multi-Document Summaries: a Preliminary Study. In Udo Hahn and Donna Harman, editors, *Proceedings of the Workshop on Text Summarization at the 4Oth Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July 11–12 2002.

[OSM94] K. Ono, K. Sumita, and S. Miike. Abstract Generation Based on Rhetorical Structure Extraction. In *Proceedings of the International Conference on Computational Linguistics*, pages 344–348, Kyoto, Japan, 1994.

[PnGV01] A. Peñas, J. Gonzalo, and F. Verdejo. Browsing by phrases: terminological information in interactive multilingual text retrieval. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 253–254, 2001.

[PnVG02] A. Peñas, F. Verdejo, and J. Gonzalo. Terminology retrieval: Towards a synergy between thesaurus and free text searching. In *IBERAMIA 2002*, pages 684–693, 2002.

[PW94] K. Preston and S. Williams. Managing the Information Overload. *Physics in Business*, June 1994.

[RBGZR91] D.R. Radev, S. Blair-Goldensohn, Z. Zhang, and R.S. Raghavan. Interactive, Domain-Independent Identification and Summarization of Topically Related News Articles. citeseer.ist.psu.edu/507083.html, 1991.

[Red90] G. Redeker. Ideational and Pragmatic Markers of Discourse Structure. *Journal of Pragmatics*, 14:367–381, 1990.

[RG00] P. Rayson and R. Garside. Comparing Corpora Using Frequency Profiling. In *Proceedings of the workshop on Comparing Corpora, pages 1–6, 2000.*, 2000.

[RHB00] D. R. Radev, J. Hongyan, and M. Budzikowska. Centroid-Based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies. In Udo Hahn, Chin-Yew Lin, Inderjeet Mani, and Dragomir R. Radev, editors, *Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, WA, April 2000.

[ROQT03] D.R. Radev, J. Otterbacher, H. Qi, and D. Tam. MEAD ReDUCs: Michigan at DUC 2003. In *Workshop on Text Summarization*, Edmonton, Canadá, 2003.

[RPH+95] R. Rao, J. Pedersen, M. A. Hearst, J. D. Mackinlay, S. K. Card, L. Masinter, P. Halvorsen, and G. G. Robertson. Rich Interaction in the Digital Library. *Communications of the ACM*, 38(4):29–39, 1995.

[RRS61] G. Rath, A. Resnick, and R. Savage. The Formation of Abstracts by the Selection of Sentences: Part 1: Sentence Selection by Man and Machines. *American Documentation*, 12(2):139–141, 1961.

[RWHB+92] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC. In *Text REtrieval Conference*, pages 21–30, 1992.

[Sal89] G. Salton. *Automatic Text Processing*. Addison-Wesley, 1989.

[SBW00] G. C. Stein, A. Bagga, and G. B. Wise. Multi-Document Summarization: Methodologies and Evaluations. In *Proceedings of the 7th Conference on Automatic Natural Language Processing TALN*, pages 337–346, Lausanne, Switzerland, October 2000.

[Sch94] Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, 1994.

[Sch02] B. Schiffman. Building a Resource for Evaluating the Importance of Sentences. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain, May–June 2002.

[Sch03] J.C. Scholtz. Evaluation of Intelligent Information Access Systems. In *Proceedings of the American Association of Artificial Intelligente (AAAI) Conference*, Acapulco, Mexico, August 2003.

[Sei88] *Non-parametric Statistics for the Behavioral Sciences*. McGraw-Hill, Berkeley, CA, 2 edition, 1988.

[SL00] H.o Saggion and G. Lapalme. Concept Identification and Presentation in the Context of Technical Text Summarization. In Udo Hahn, Chin-Yew Lin, Inderjeet Mani, and Dragomir R. Radev, editors, *Proceedings of the Workshop on Automatic Summarization at the 6th Applied Natural Language Processing Conference and the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, WA, USA, April 30 2000. Association for Computational Linguistics.

[SNM02] B. Schiffman, A.Ñenkova, and K. McKeown. Experiments in multi-document summarization. In *In Proceedings of the Second Conference on Human Language Technology (HLT-2002).*, San Diego, CA., 2002.

[SOC+02] J. D. Schlesinger, M. E. Okurowski, J. M. Conroy, D. P. O'Leary, A. Taylor, J. Hobbs, and H. Wilson. Understanding Machine Performance in the Context of Human Performance for Multi- Document Summarization. In *Proceedings of the Workshop on Multi-Document Summarization Evaluation of the 2nd Document Understanding Conference at the 4Oth Meeting of the Association for Computational Linguistics*, Philadelphia, PA, July 2002.

[SSMB97] G. Salton, A. Singhal, M. Mitra, and C. Buckley. Automatic text Structuring and Summarization. In *Information Processing and Management*, number 2, pages 193–207, 1997.

[TM97] S. Teufel and M. Moens. Sentence extraction as a classification task. In *Workshop 'Intelligent and scalable Text summarization',ACL/EACL*, 1997.

[TM98] S. Teufel and M. Moens. Sentence Extraction and Rhetorical Classification for Flexible Abstracts. In Eduard Hovy and Dragomir R. Radev, editors, *Proceedings of the AAAI Symposium on Intelligent Text Summarization*, pages 16–25, Stanford, California, USA, March 23–25 1998. The AAAI Press.

[TM03] L. Turian, J. P. Shen and I. D. Melamed. Evaluation of Machine Translation and its Evaluation. In *In Proceedings of MT Summit IX*, New Orleans,LA, 2003.

[VH00] E. M. Voorhees and D. Harman. Overview of the sixth text REtrieval conference (TREC-6). *Information Processing and Management*, 36(1):3–35, 2000.

[WLKA97] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. PARADISE: A framework for evaluating spoken dialogue agents. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Somerset, New Jersey, 1997. Association for Computational Linguistics.

[WN97] P. Williams and D.Ñicholas. Journalist, News, and the Internet. *New Library World*, 98:217–223, 1997.

[YCG+99] J. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. Topic Tracking in a News Stream. In *Proceedings of the DARPA Broadcast News Workshop*, pages 133–138, 1999.

[YW03] C.C. Yang and F.L. Wang. Fractal summarization: summarization based on fractal theory. In *SIGIR 2003*, pages 391–392, 2003.