ESTRUCTURACIÓN DE ÍNDICES GRAMATICALES Y LÉXICOS PARA LA EXTRACCIÓN Y RECUPERACIÓN DE INFORMACIÓN

Couto, Javier (*) jcouto@fing.edu.uy
Crispino, Gustavo (*) crispino@fing.edu.uy
Grassi, Mariela (*) mgrassi@seciu.edu.uy
Skorodynski, Mónica (*) mskorodynski@bps.gub.uy

(*) Instituto de Computación
Facultad de Ingeniería
Universidad de la República
Uruguay

Resumen

Basta consultar un diccionario más o menos completo del castellano o cualquier otra lengua natural para concluir que las unidades lingüísticas son generalmente polisémicas. Esta es la razón principal de numerosos casos de ambigüedad lingüística. Teniendo en cuenta este hecho, el Método de Exploración Contextual provee el marco necesario para identificar información semántica específica contenida en los textos así como también mecanismos que conducen a resolver indeterminaciones semánticas. Este método puede ser utilizado en distintas aplicaciones que trabajan con lenguaje natural para la extracción y recuperación de información. En este artículo presentamos una plataforma informática para este método y algunas reflexiones surgidas en el proceso de definición de los elementos que deben integrarse a la base de conocimientos de la plataforma informática para poder realizar el tratamiento de textos escritos en castellano.

1. Introducción

La lingüística computacional ha buscado durante mucho tiempo construir representaciones semánticas apoyándose en análisis sintácticos previos, los cuales necesitan, a su vez, análisis morfológicos. Para asegurar la coherencia textual, varios equipos de investigación han intentado introducir consideraciones pragmáticas, haciendo uso de conocimientos cada vez más numerosos relativos a los dominios tratados (Pazienza 1997). Esos métodos movilizan importantes recursos lingüísticos y son difíciles de poner en práctica sobre textos heterogéneos. Además, necesitan conocimientos lingüísticos y ontologías no siempre disponibles en los dominios tratados.

Sin embargo, tiende a imponerse la necesidad de introducir más semántica en las herramientas de búsqueda y extracción de información. La cuestión a resolver es la de introducir las nociones semánticas, de manera razonable, sin pasar por los métodos costosos o parciales que han sido propuestos durante un cierto período por la Inteligencia Artificial.

El Método de Exploración Contextual (Desclés et al. 1991, Desclés 1996, Desclés et al.1997) identifica los conocimientos lingüísticos ubicándolos en sus contextos y organizándolos en tareas especializadas. En este enfoque, los lingüistas analizan los textos identificando indicadores e índices gramaticales y léxicos pertinentes para la resolución de un problema, y luego conciben y escriben las reglas de exploración del contexto de los índices identificados en los textos. Este método no está limitado a tratamientos específicos, sino que ofrece un marco de trabajo realista.

En este trabajo presentamos una plataforma informática para este método y algunas reflexiones surgidas en el proceso de la definición de los elementos que deben integrarse a la base de conocimientos de la plataforma informática para poder realizar el tratamiento de textos escritos en castellano.

Este artículo se organiza de la siguiente manera. En la sección 2 hacemos una breve presentación del Método de Exploración Contextual y sus

aplicaciones. En la sección 3 describimos una plataforma informática capaz de soportar las distintas aplicaciones del método.
En la sección 4 presentamos una propuesta de organización conceptual de la base de conocimientos, a partir de nuestro trabajo específico para el castellano. En la sección 5 exponemos las conclusiones de nuestra investigación.

2. *Método de Exploración Contextual*

El Método de Exploración Contextual (MEC) fue desarrollado por el equipo LaLIC (UMR 8557 du CNRS, EHESS, Université Paris-Sorbonne) que dirige el Prof. Jean-Pierre Desclés. El objetivo de este método consiste en proveer el marco necesario para identificar información semántica específica contenida en los textos. El MEC parte de la hipótesis que establece que **todo texto posee unidades lingüísticas que permiten resolver indeterminaciones semánticas, en algunos casos, y tomar ciertas decisiones para construir el sentido, en otros.**

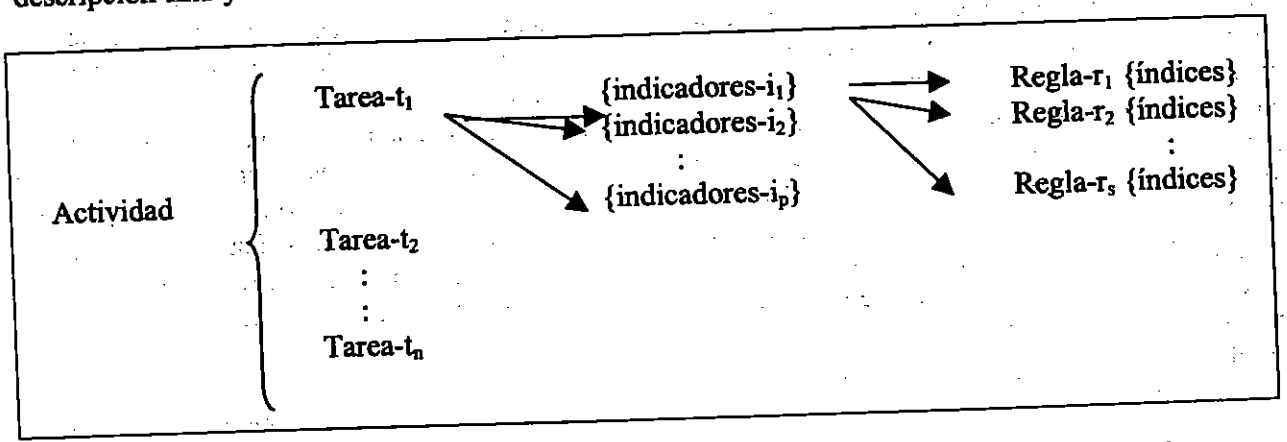
El método se implementa informáticamente bajo la forma de *bases de conocimiento lingüístico*. Este sistema emplea conocimiento exclusivamente lingüístico y presente en el texto. El método requiere entonces de una descripción fina y detallada de ciertas unidades lingüísticas llamadas **indicadores** y de otras denominadas **índices**, complementarias de las primeras.

Los indicadores son expresiones lingüísticas que disparan la ejecución de ciertas reglas de exploración contextual encargadas de determinar el *valor semántico* del indicador para cierta **tarea** específica, por ejemplo, **reconocer una conclusión** o **filtrar una definición del texto.** De manera que los indicadores están asociados a ciertas tareas, son específicos de cada tarea.

Por otra parte, cada indicador tiene asociado un conjunto de reglas de exploración contextual, heurísticas unas y lingüísticas otras. La aplicación de una regla, invocada por un indicador, explora el contexto de ese indicador buscando índices lingüísticos con el objetivo de resolver la tarea, esto es, determinar el valor semántico del indicador.

Todos estos elementos –indicadores, índices y reglas- componen la base de conocimiento lingüístico que el método emplea para realizar las **actividades** requeridas (Desclés 1996)

A modo de síntesis y para comprender mejor el funcionamiento del MEC, damos a continuación una representación gráfica de los componentes de este método y sus relaciones:



En este diagrama la llave grande se lee como "se compone de", las flechas indican "asociada a" y las llaves chicas denotan conjuntos. De manera que el esquema completo debe leerse así:

- Una **actividad** se compone de una o más tareas.
- Cada **tarea** está asociada a uno o más conjuntos de indicadores.
- A cada **conjunto de indicadores** se asocia una o más reglas.
- Las **reglas** tienen la forma de un condicional:

si <condición>
entonces <conclusiones >

- En la **condición** aparecen los **conjuntos de índices** que permiten tomar una decisión, es decir que permiten desambiguar el valor semántico del indicador.
- Una **conclusión** consiste en asignar una etiqueta semántica a un segmento textual.

A modo de ejemplo, algunas actividades pueden ser: *hacer un resumen, filtrar definiciones de un texto, filtrar citas, identificar relaciones de causalidad.*
Algunas tareas: *buscar anuncios temáticos, conclusiones, recapitulaciones, definiciones.*
Algunas etiquetas semánticas: *anuncio temático, definición, conclusión, relación causal.*

Entonces, a modo de ejemplo, la actividad de *hacer un resumen* se compone de las siguientes tareas: *buscar anuncios temáticos, buscar recapitulaciones* y *buscar conclusiones*. A su vez, cada una de esas tareas está asociada a un conjunto de indicadores.

El MEC se basa en los siguientes supuestos básicos:

- El procesamiento de textos necesita identificar y estudiar la semántica de ciertos fragmentos de texto (oraciones, párrafos, etc.) que son independientes del dominio (médico, económico, técnico), pero que no son independientes del tipo de texto con el que se trabaje.

- El estudio de esos fragmentos o categorías textuales implica identificar indicadores lingüísticos específicos, los cuales son claves importantes para la estructuración del conocimiento semántico. Pero esto no es suficiente para la identificación léxica. El procesamiento semántico de una unidad lingüística depende de otras claves lingüísticas –los índices- que deben estar presentes en el mismo contexto para resolver la ambigüedad causada por el fenómeno de la *polisemia.*

- Para cada categoría textual, el MEC sugiere la misma metodología:

  i) Identificar información semántica relevante y sus indicadores lingüísticos, en función de una tarea.

  ii) Identificar, en el texto analizado, los límites del contexto C necesarios para resolver las indeterminaciones semánticas.

  iii) Determinar los pasos para encontrar las claves lingüísticas más importantes explorando el contexto C con el fin de resolver posteriores problemas de ambigüedad.

- Muchas tareas de procesamiento de texto, como la extracción de conocimiento o el resumen automático, pueden ser resueltas analizando exclusivamente las unidades léxicas del texto, siempre y cuando tengamos en cuenta su contexto lingüístico. Ahora bien, en ciertos casos también es útil y necesario recurrir a indicadores textuales más que lingüísticos, es decir, a información del tipo de ubicación de una palabra en la oración, ubicación de oraciones y de párrafos en el texto, signos gráficos utilizados en los títulos, y otros. Así es que el método puede utilizar diversos indicadores –léxicos, temáticos, textuales y estructurales- en sus estrategias de búsqueda, con el objetivo de asignar etiquetas semánticas a las oraciones u otras unidades textuales (párrafos, secciones, etc.).

Entonces el sistema de exploración contextual se compone de:

i) una base de datos de indicadores lingüísticos semánticamente relevantes;

ii) una base de datos con índices lingüísticos para resolver la ambigüedad que afecta a los indicadores en su contexto;

iii) una base de datos de reglas de exploración contextual. La tarea de estas reglas es identificar indicadores lingüísticos con el fin de asignar etiquetas semánticas.

3. *Una plataforma informática para el MEC*

En el equipo LaLIC del CAMS se han desarrollado diversas investigaciones destinadas a identificar ciertas informaciones semánticas a partir de marcas de superficie:

- identificar lás acciones en textos técnicos (García 1998) ;
- identificar las relaciones causales entre situaciones (Jackiewicz 1998) ;
- identificar las definiciones de términos propuestos explícitamente o implícitamente por un autor (Cartier 1998) ;
- identificar los anuncios temáticos puestos en evidencia por un autor (Cartier 1998) ;

Actualmente se está trabajando sobre una plataforma informática (ContextO) capaz de soportar las diferentes aplicaciones del MEC. Este trabajo se está desarrollando en el marco del programa ECOS (Francia - Uruguay, Acción n° U97E01) para el desarrollo de proyectos conjuntos de investigación científica entre Uruguay y Francia. La colaboración es llevada a cabo entre el equipo LaLIC del CAMS y el grupo de TALN del Instituto de Computación de la Facultad de Ingeniería de la Universidad de la República (Uruguay).
ContextO (Crispino et al. 1999) está constituida por un motor de exploración contextual, un conjunto de agentes especializados para orientar y posteriormente explotar el trabajo del motor, y por un sistema de gestión de la base de conocimientos.

Como respuesta a invocaciones determinadas en función de parámetros fijados por el usuario, **el motor de exploración contextual** dispara, para una o varias tareas especializadas, el proceso de reconocimiento de *indicadores* e *índices* presentes en un segmento textual. Este proceso es realizado por el sistema de gestión de conocimientos lingüísticos, el cual proporciona al motor de exploración contextual el conjunto de reglas potencialmente aplicables.
Un lenguaje de descripción permite al lingüista constituir su base de conocimientos especificando las *tareas*, los *indicadores*, los *índices* y las *reglas* de exploración contextual asociadas. Estas últimas se expresan en un lenguaje formal de tipo declarativo. Cada regla comprende una parte de *Declaración de un Espacio de Búsqueda*, una parte de *Condición* y una parte *Acción*, la cual es ejecutada solamente si se verifica la *Condición*. Como resultado de la aplicación de las reglas, se colocan etiquetas semánticas que "decoran" la jerarquía del texto a diversos niveles; por ejemplo, una regla puede atribuir una etiqueta semántica a una oración.

Los agentes especializados tienen por objetivo explotar las "decoraciones semánticas" del texto en función de las necesidades definidas por el usuario. Hay entonces un agente que construye un resumen compuesto de oraciones del texto de entrada que corresponden a un perfil tipo y un agente que construye diferentes extractos del texto de entrada en función de perfiles seleccionados por el usuario.
Estos agentes especializados permiten desarrollar tratamientos específicos para un usuario explotando el modelo genérico de tratamiento de conocimientos lingüísticos.

El sistema de gestión de conocimiento lingüístico tiene por objetivo agrupar en una base de datos general las diferentes tareas definidas en los sistemas de exploración contextual. Se trata de un sistema no sólo capaz de permitir el acceso a los datos lingüísticos, sino también de facilitar la adquisición, la modelización, la explotación y la posibilidad de compartir esos datos lingüísticos.
Este sistema tiene tres componentes:

- una base de datos lingüísticos de la exploración contextual
- una capa de servicios de búsqueda y extracción de datos lingüísticos de la exploración contextual
- herramientas de ayuda a la adquisición y la modelización de conocimientos lingüísticos de la exploración contextual

4. *Organización conceptual de los marcadores*

La base de conocimientos para el tratamiento de textos en francés cuenta actualmente con aproximadamente 11.000 marcadores (indicadores e índices) y unas 250 reglas de exploración contextual.
Nosotros hemos comenzado un trabajo de construcción de bases para el tratamiento de textos en castellano, apoyándonos en las ya existentes para el francés. Hasta el presente

hemos completado la información necesaria para captar anuncios temáticos, conclusiones, e identificación de acciones en textos técnicos.
En este trabajo hemos definido algunos criterios para la organización conceptual de índices e indicadores que presentaremos en lo que sigue de esta sección.

Los *índices* y los *indicadores* que se emplean en el MEC son **unidades léxicas y sintagmáticas**, es decir, *verbos*, *sustantivos*, *adjetivos*, *adverbios* y *conjunciones*, para los primeros, y *sintagmas nominales*, *verbales*, *preposicionales* y otros, para los segundos. Cada una de estas unidades posee propiedades que las caracterizan.

Por ejemplo, el **verbo** presenta flexión de *tiempo*, *número* y *persona*. Estas son propiedades gramaticales de la categoría verbal. Así, en el caso de la forma verbal *canto*, el morfema *-o* indica *persona*:primera, *número*:singular y *tiempo*:presente del modo indicativo. En cambio, en la forma verbal *cantaron* la desinencia verbal *-aron* indica *persona*:tercera, *número*:plural y *tiempo*:pretérito perfecto simple del modo indicativo. (Real Academia Española 1973)
El **adjetivo**, en cambio, presenta flexión de *género* y *número*. Por ejemplo, en *blancos*, el morfema *–o* indica género *masculino* y el morfema *–s* indica número *plural*.

Por otro lado, los indicadores llevan asociado un *tipo* que definimos como *simple*, *continuo* o *discontinuo* dependiendo de la composición de los mismos. Si el indicador consta de una sola unidad léxica, es decir que se compone de una sola palabra, decimos que es un indicador **simple**. En cambio, si la unidad léxica consta de dos o más palabras adyacentes, decimos que el indicador es **complejo continuo**. Finalmente, si el indicador está compuesto por dos o más palabras entre las cuales pueden intercalarse otros elementos, se trata de indicadores **complejos discontinuos**. Dado que la continuidad o discontinuidad se aplica sólo a indicadores complejos, podemos simplificar la expresión diciendo que el valor asociado al argumento *tipo* del atributo será *simple*, *continuo* o *discontinuo*. Son ejemplos de los casos mencionados los siguientes:

- **Indicadores simples:** demostrar, desarrollar, exponer, explicar, presentar
- **Indicadores continuos:** en resumen, en síntesis, en suma, para resumir
- **Indicadores discontinuos:** tratar...de, no es...sino, poner...el acento en

Con el objetivo de representar formalmente estas propiedades, definimos atributos. Los atributos se asocian a cada una de las propiedades consideradas relevantes con el fin de caracterizar cada tipo de unidad léxica. Así tendremos que la categoría gramatical es un atributo, en tanto que el género, tiempo, número y persona también lo son. A su vez, las unidades léxicas se agrupan en conjuntos. Y una misma unidad léxica puede pertenecer a varios conjuntos a la vez.

Por su parte los conjuntos también tienen propiedades que los caracterizan. Hay conjuntos homogéneos y conjuntos heterogéneos. Un conjunto homogéneo es aquel en el que todos sus elementos pertenecen a la misma categoría gramatical, mientras que en un conjunto heterogéneo los miembros pertenecen a distintas categorías gramaticales.
De manera que hay atributos asociados al conjunto y atributos asociados a los elementos del conjunto. Por lo tanto, a grandes rasgos, un conjunto se compone de:

- atributos[1] del conjunto
- atributos de los elementos del conjunto
- elementos del conjunto

A continuación presentamos un ejemplo de conjunto:

[1] En estos casos cuando decimos *atributos* significa el par (*nombre del atributo: valor del atributo*)

- Conjunto de *SUSTANTIVOS*:

Conjunto

Atributos del conjunto:

Nombre:	introduce-tema
Descripción:	sustantivos que introducen la temática del texto. contexto: *el X de este texto*
Tipo conjunto:	*homogéneo*
Nivel:	1 (los elementos del conjunto son unidades léxicas)

Atributos de los elementos del conjunto:

tipo del indicador:	{*simple*}
categoría gramatical:	*sustantivo*
atributos:	
género:	{*femenino, masculino*}
número:	{*singular, plural*}

Cuerpo: {hipótesis, motivo, motivos, objetivo, objetivos, planteo, planteos, premisa, premisas, problema, problemas, problemática, problemáticas propósito, propósitos, sujetos, tema, temas, temática, temáticas, tesis}

Fin-Conjunto

5. *Conclusiones*

La plataforma ContextO está actualmente operativa con una base de conocimientos que permite un tratamiento bastante completo de textos en francés. Pensamos que su arquitectura, que privilegia el concepto de componentes de software y de agentes especializados, la hace apta para representar diferentes tipos de tratamiento lingüístico ya que es posible definir nuevas bases de conocimiento para nuevas tareas de etiquetado semántico. En particular, la incorporación de marcadores para el tratamiento del castellano nos está mostrando su flexibilidad y su potencialidad para constituir una base multilingüe.

Por otra parte, la organización de marcadores que presentamos en este trabajo, permite un manejo no sólo productivo para el tratamiento de textos, sino que también brinda al lingüista condiciones para realizar un trabajo conceptual y sistemático que permitirá enriquecer la calidad de los trabajos basados en el método.

Referencias Bibliográficas

Cartier, Emmanuel. (1998). Analyse automatique des textes : l'exemple des informations définitoires. *RIFRA '98, Rencontre Internationale sur l'Extraction le Filtrage et le Résumé Automatiques*. Sfax, Tunisie.

Crispino, Gustavo; Ben Hazez, Slim; Minel, Jean-Luc (1999). Architecture logicielle de ContextO plate-forme d'ingénierie linguistique. *TALN 1999, Cargèse, France, 12-17 juillet 1999*

Desclés, Jean-Pierre, Christophe Jouis, Hum-Ghum Oh, Danièle Maire Reppert. (1991). Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte. In *Knowledge modeling and expertise transfer*, pp.371-400, D. Herin-Aime, R. Dieng, J-P. Regourd, J.P. Angoujard (éds), Amsterdam.

Desclés, Jean-Pierre. (1996). Systèmes d'exploration contextuelle. *Actes du colloque sur le Calcul du sens et contexte*. Université de Caen.

Desclés, Jean-Pierre, Emmanuel Cartier, Agata Jackiewicz, Jean-Luc Minel. (1997). Textual Processing and Contextual Exploration Method. In *CONTEXT'97*, Rio de Janeiro, Brasil.

García, Daniela. (1998). Analyse automatique des textes pour l'organisation causale des actions. Réalisation du système informatique COATIS. *Thèse de Doctorat, Université Paris-Sorbonne*.

Jackiewicz, Agata. (1998). L'expression de la causalité dans les textes. Contribution au filtrage sémantique par une méthode informatique d'exploration contextuelle. *Thèse de Doctorat, Université Paris-Sorbonne*.

Pazienza, M.T. (1997) (éd.). Information extraction (a multidisciplinary approach to an emerging information technology*), International Summer School, SCIE'97*, Springer Verlag (Lectures Notes in Computer Science).

Real Academia Española (1973). Esbozo de una nueva gramática de la lengua española, 1973, Madrid:Espasa-Calpe