

Problemática de la interlingua como estrategia de traducción automática

J.Abaitua, J.Soler, J.Vivaldi
Fujitsu España S.A
I+D Barcelona

Resumen

Esta ponencia presenta un aspecto de la problemática de la estrategia interlingüa en traducción automática (TA). En conexión con ella, se analizan los principales factores que motivan la investigación sobre la representación del conocimiento en inteligencia artificial (IA), así como los estudios en torno a los primitivos semánticos, tanto en el área del procesamiento del lenguaje natural (PLN) como de la lingüística teórica (LT). Esta discusión sirve para situar en contexto la estrategia de traducción empleada por el sistema de TA ATLAS-II, en concreto, su uso de las redes semánticas y la función del diccionario de palabras.

PROBLEMATICA DE LA INTERLINGUA COMO ESTRATEGIA DE TRADUCCION AUTOMATICA

Joseba ABAITUA, Josep SOLER y Jorge VIVALDI
Fujitsu España S.A.
I-D Barcelona
Sabino de Arana 34, 2
08028 Barcelona

Resumen

Esta ponencia presenta un aspecto de la problemática de la estrategia interlingua en traducción automática (TA). En conexión con ella, se analizan los principales factores que motivan la investigación sobre la representación del conocimiento en inteligencia artificial (IA), así como los estudios en torno a los primitivos semánticos, tanto en el área del procesamiento del lenguaje natural (PLN) como de la lingüística teórica (LT). Esta discusión sirve para situar en contexto la estrategia de traducción empleada por el sistema de TA ATLAS-II, en concreto, su uso de las redes semánticas y la función del diccionario de palabras.

1. Estrategias en Traducción Automática

Actualmente, dos grandes estrategias compiten en el desarrollo de nuevos sistemas de traducción automática: la de transferencia y la interlingual (o pivot). Las diferencias entre ambas son bien conocidas. A grandes rasgos, empero, podemos recordar que en la aproximación interlingual la traducción se realiza a través de una representación del significado neutra e independiente de las peculiaridades lingüísticas de las lenguas particulares. La traducción se realiza en dos fases: análisis-interlingua-generación. Por el contrario, con la aproximación de transferencia la traducción se efectúa entre tantas representaciones abstractas como lenguas haya para traducir. Cada una de estas representaciones es dependiente de cada una de las lenguas y la traducción se realiza en tres fases: análisis-RLO-RLD-generación (R:representación L:lengua O:origen D:destino). La fase que media entre RLO y RLD se conoce como fase de transferencia.

Existen disparidad de criterios sobre las respectivas adecuaciones de estas dos estrategias y abundan defensores y retrectores en ambos casos. Teóricamente, por ejemplo, la interlingua es superior en cuanto a economía en la traducción, si se diseña

un sistema de traducción multilingüe, por cuanto economiza de forma exponencial el número de transferencias a realizar. (El número de módulos de transferencia para n lenguas es de n elevado a $n-1$, mientras que son sólo n , elevados a uno, los módulos con interlingua).

De todas maneras es necesario reconocer que esta distinción es en gran medida una distinción maximalista. Sirve para marcar dos polos entre los que poder ubicar los sistemas de traducción automática. En la práctica, la mayor parte de los sistemas tienden a beneficiarse de ambas estrategias, situándose en un punto intermedio entre los dos polos. Dentro de la estrategia de transferencia, por ejemplo, existe la propensión natural a definir rasgos lingüísticos lo más universales posibles, de forma que estos resulten útiles en más de una aplicación. Adviértase que si se lograra definir una colección de rasgos lingüísticos universales estaríamos muy cerca de lo que persigue la estrategia interlingual. Por otro lado, en todo sistema operativo interlingual existe algún dispositivo complementario de transferencia por regla para afinar las correspondencias entre las estructuras de dos lenguas concretas. Por esto decimos que no siempre se puede extrapolar completamente en la clasificación de sistemas de TA. Quizá el criterio más fiable para poder caracterizar un sistema como interlingua o de transferencia está en saber si éste contiene diccionarios bilingües o no. Esto es así porque mientras en la aproximación de transferencia se necesitan tantos diccionarios bilingües como pares de lenguas se vayan a traducir, además de los respectivos diccionarios monolingües, en la aproximación interlingual sólo son necesarios los diccionarios monolingües (en la interlingua) para cada una de las lenguas.

En lo que queda de esta ponencia vamos a exponer los fundamentos y problemática de la estrategia interlingual en general para pasar luego a hablar del sistema ATLAS-II como ejemplo concreto de sistema interlingual. Antes, a modo de inciso ilustrativo, podemos citar algunos de los sistemas de TA más conocidos. Por el lado de la estrategia de transferencia pueden destacarse, entre otros muchos, sistemas como METAL (Slocum et al., 1984), EUROTRA (Johnson et al., 1985), MC (Nagao et al., 1985), ROSETTA (Landsbergen, 1987). Por otra parte, sistemas como TRANSLATOR (Kirenburg et al., 1985), MOPTRANS (Lytinen & Schank, 1982) y ATLAS-II (Uchida et al., 1982) optan por una aproximación interlingual. Para una clasificación más completa consúltese Slocum ed., 1987.

2. La Estrategia Interlingual

Decir que de entre las dos estrategias apuntadas la interlingua es sobre el papel la más ambiciosa y la que presenta a nivel científico las características más interesantes no debe plantear ningún motivo de discordia. La disputa y las críticas pueden plantearse, sin embargo, a nivel de implementación

práctica. Concretamente, ¿es viable o no la interlingua?

La estrategia interlingua viene precedida por las investigaciones en torno a la representación del conocimiento en IA, especialmente durante la década de los setenta. Concretamente deben destacarse los estudios de Schank 1975 que junto con los de Wilks 1973 representan uno de los primeros intentos serios en IA para conseguir representaciones abstractas de contenidos y significados en forma de primitivos semánticos. El problema de los primitivos semánticos es todavía hoy uno de los grandes temas en IA.

En la actualidad la investigación más reciente de Schank (Schank 1982) está siendo puesta en práctica en el sistema MOPTRANS. Este sistema toma como base la teoría de los MOPs (paquetes de organización de la memoria). Los MOPs sirven para organizar la información sobre determinadas situaciones formando un conjunto de unidades de representación del tipo script. Los MOPs sirven como forma de representar el conocimiento y, en definitiva, como constituyentes básicos de la interlingua. Las investigaciones de Carbonell & Tomita 1987 han tomado una línea parecida en la elaboración de la interlingua, con el formalismo de "orientación al objeto" (entity-oriented grammar) de Hayes 1984, al igual que lo han hecho los experimentos de Niremburg et al., 1987 con las frames del sistema TRANSLATOR. El factor común a todas estas investigaciones es que parten de la idea de que es necesaria la comprensión del texto si se quiere conseguir una traducción de calidad. Para ello la traducción debe estar basada a la vez en el conocimiento lingüístico y del dominio concreto de la aplicación, intentando la modularidad entre ambos aunque marcando el énfasis sobre este último.

Slocum 1987 ha mencionado algunos de los problemas inherentes a esta estrategia basada en el conocimiento. El problema principal lo constituye la representación del conocimiento especializado en sistemas que son concebidos para uso general. Mientras que dominios limitados pueden ser codificados con cierta facilidad para sistemas experimentales o de uso muy localizado, esta misma codificación se convierte en un problema más que delicado cuando el corpus de materia a traducir adquiere otras dimensiones, como es el ejemplo clásico, citado por Slocum, de manuales de centrales telefónicas de más de 100.000 páginas. La cantidad de conocimiento que tales documentos poseen es una tarea teóricamente hartamente complicada de codificar, con el gravamen además de que difícilmente pueden llegar a ser rentables en la práctica porque requieren inversiones fuertes a plazos muy largos. Esto plantea la duda de si realmente es necesaria la "comprensión" de un texto para realizar traducciones de calidad. Johnson 1983, por ejemplo, ha señalado que él ha hecho traducciones, consideradas como buenas por especialistas, en materias sobre las que no poseía ningún conocimiento previo y que apenas si llegaba a comprender.

En resumen, podemos decir que se ha criticado, no sin cierta

lógica, primero, la dificultad para representar grandes cantidades de conocimiento; segundo, y como consecuencia de lo primero, la baja rentabilidad de tales tareas para sistemas operativos con gran volumen de material a traducir, y tercero, y punto más cuestionable, la indispensabilidad de tales tareas. El lector interesado puede consultar Slocum ed., 1987 y también Nirenburg ed., 1987.

Otra crítica frecuente contra la interlingua es la cuestión de su misma factibilidad. Se ha apuntado que no existe todavía ninguna interlingua escrita y se ha cuestionado si alguna vez existirá una. En el terreno de la lingüística teórica, sin embargo, existe una antigua aspiración estructuralista de lograr una organización del "significado" válida para todas las lenguas (cf. Lyons 1968:470-481). Entre otras cuestiones se considera la posibilidad de un análisis decomposicional (en unidades de significado básicas) del vocabulario de todas las lenguas del mundo. Esta aspiración, que reincide en la noción de primitivos semánticos universales apuntada anteriormente, animó en la década de los setenta a los lingüistas de la corriente de la semántica generativa, como Lakoff o McCawley, y sigue siendo hoy en día la base de trabajos como los de Dowty 1979 o Jackendoff 1985.

Bajo este punto de vista, la interlingua proporcionará un sistema de rasgos lingüísticos de acuerdo con los cuales será posible codificar el vocabulario de cualquier lengua humana. Debemos reconocer, no obstante, que esta tarea de definir rasgos lo más universales posibles es compartida tanto por la estrategia interlingua como por transferencia. La diferencia es que la interlingua opta por esta solución de una forma mucho más decidida.

La elección de una estrategia interlingua trae consigo otras consecuencias interesantes. Para los sistemas de TA que optan por representaciones conceptuales es posible hablar de ontologías particulares, de submundos. En un sistema de este tipo existe una visión o conocimiento del mundo propio que generalmente se recoge en un enorme conjunto de teoremas. Estos se escriben en el formalismo lógico propio del sistema. Es precisamente en este ámbito donde empezamos a vislumbrar los problemas más atractivos, con motivo de la selección de un modelo de representación del mundo y de un formalismo en el que expresarlo. A este respecto podemos considerar varios modelos de semánticas formales, que sólo mencionaremos sin entrar en mayores detalles. Hoy en día están en liza modelos teóricos tales como el de la lógica intensional de Montague, las semánticas de Kamp 1982 -representación del discurso- y de Barwise & Perry 1982 -semántica de situaciones-, como modelos más destacados. (Ver con este motivo la polémica en el número 8 de Language and Philosophy). Estas semánticas de modelos-teóricos tienen el beneficio de poseer fundamentos lógicos sólidos, es decir, incorporan las ventajas que se derivan de los sistemas lógicos como formalismos de representación. Estas ventajas son básicamente su elegancia y consistencia formal, por un lado, y la facilidad con que soportan inferencias deductivas

por otro. En estos modelos la semántica juega también el papel de determinar las expresiones que están bien formadas - de acuerdo tanto con las reglas del modelo semántico como de la gramática para esa lengua concreta. Los valores de corrección suelen expresarse generalmente como valores booleanos, con la mejora respecto al cálculo de predicados en admitir un tercer valor de no-resolución o incertidumbre, además de un análisis más elaborado de la cuantificación. Los modelos semánticos formales constituyen en la actualidad un gran foco de atención en PLN y sus avances teóricos pueden jugar un papel primordial en el futuro de la estrategia interlingua.

3. La Estructura Conceptual de ATLAS-II

ATLAS-II es un sistema diseñado para la traducción multilingüe entre lenguas europeas (inglés, alemán, español) y lenguas asiáticas (japonés, coreano, chino) para el que se ha optado por una estrategia interlingual que básicamente recoge las características apuntadas más arriba. La expresión del contenido o estructura conceptual del texto a traducir se realiza mediante un lenguaje de representación intermedio independiente de las lenguas involucradas. La interlingua aporta tanto el vocabulario de símbolos con el que se construye esta estructura conceptual como su gramática. Esta última se materializa en ATLAS-II en un formalismo de representación conocido como red semántica.

La red semántica es un formalismo que ha tenido un considerable éxito como método para la representación del conocimiento en IA, particularmente durante la década de los setenta. Una red semántica puede definirse brevemente como un grafo direccionado compuesto por nodos y arcos. Por ello se encuadra en la tendencia a usar representaciones gráficas para la simbolización de las estructuras conceptuales que caracterizó a la pasada década (Woods, 1975). Esta representación del significado está en parte inspirada por la idea de que las interconexiones neuronales en el cerebro humano están organizadas precisamente en forma de red. La idea de que el acceso a la información semántica se realiza a través de una activación expansiva de la red es todavía hoy una idea atractiva.

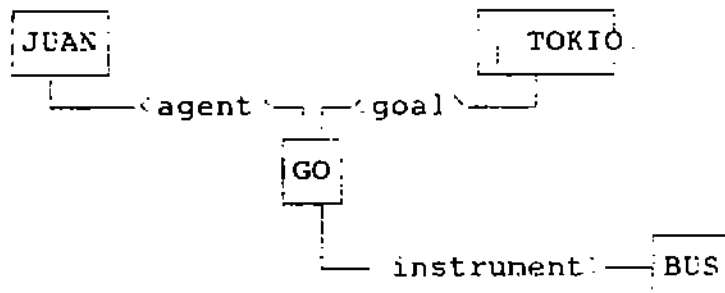
Entre las razones del éxito de las redes semánticas como grafos orientados y las ventajas prácticas que de ello se derivan, Tanimoto 1987 señala las siguientes:

- El hecho de representar objetos como nodos en la red y el poder representar las referencias cruzadas entre tales objetos mediante punteros permite un método de representación más eficaz que el tipo de representación plana de la lógica de predicados. Los punteros además permiten la eliminación de la redundancia en la representación de los objetos.

- La búsqueda de un objeto en una red se puede resolver con rapidez porque las asociaciones entre los nodos de la red son expresadas como arcos en el grafo y existen algoritmos de búsqueda muy eficaces para realizar esta tarea. Es fácil optimizar además el proceso de búsqueda implementando la red con un esquema de listas de adyacencia.

No existe sin embargo un formalismo estándar de redes semánticas. Así por ejemplo en algunos modelos las constantes e individuos lógicos vienen codificados como nodos, mientras que las funciones y predicados se representan directamente como arcos conectando los nodos. No obstante es más corriente que los predicados también se expresen como nodos, unidos por arcos con cada uno de sus argumentos. En ATLAS-II se ha optado por un tipo de red similar a la segunda citada. De esta manera la red semántica sirve para expresar el significado en términos de objetos y relaciones entre objetos. Los nodos indican tanto objetos (eg. individuos) como relaciones (eg. predicados) mientras que los arcos indican el tipo de conexión entre los nodos. Un ejemplo simplificado de red semántica viene dado por:

Juan fue a Tokio en autobús.



(Todos los los términos en la red indican primitivos semánticos; los expresados en mayúscula son del tipo objetos y relaciones lógicas, mientras que los términos en paréntesis angulares indican roles profundos, sobre los que trataremos seguidamente.)

El hecho de que no exista un modelo estandarizado de red semántica hace de ella un formalismo "vago", en el sentido de que no contiene restricciones expresadas para la representación del significado. Por otro lado, las redes semánticas son formalismos fácilmente "aumentables" y no es difícil conseguir su equivalencia lógica con modelos más restringidos. Un ejemplo es la "partición" llevada a cabo por Hendrix 1978, incorporando la noción de espacios o subconjuntos de nodos y arcos en la red. Estos espacios resultan útiles para marcar el alcance (scope) de cuantificadores y otros operadores lógicos, permitiendo así que la red semántica adquiriera el poder formal de una lógica de predicados. De esta forma una red semántica puede convertirse en una variante notacional de un formalismo lógico más elaborado.

Se puede decir, en resumidas cuentas, que una red semántica

puede llegar a considerarse formalmente equivalente a una lógica de predicados en sus propiedades declarativas con la ventaja de permitir algoritmos procedurales de búsqueda más eficaces.

Hemos visto que las reglas que rigen la formación de estructuras conceptuales son muy sencillas. Consisten simplemente en unir nodos mediante arcos formando una red. Los nodos en la red representan conceptos, esto es, relaciones y objetos. Los arcos representan el tipo de asociación o dependencia entre los nodos. En ATLAS-II la mayor parte de estas dependencias se expresan como casos profundos o roles temáticos adoptados de la gramática de casos, según la cual las asociaciones entre un predicado y sus argumentos se expresan en forma de casos profundos. Además se dispone de un número reducido de arcos unarios, que sirven para expresar valores de temporalidad, modalidad y otros atributos gramaticales.

En una red semántica como la de ATLAS-II, predicados y argumentos vienen representados como nodos y los casos profundos vienen representados como arcos. Las oraciones se analizan en forma de red semántica y en este proceso el diccionario de palabras cumple una función de especial importancia. El diccionario de palabras contiene prácticamente toda la información necesaria que hace posible la construcción de la red semántica.

En primer lugar, el diccionario contiene para cada palabra de una lengua el correspondiente símbolo semántico en la interlingua. Según la categoría gramatical de la palabra, este símbolo semántico se activará como un nodo o como un arco. En segundo lugar, el diccionario contiene información sobre las posibilidades de combinación de los símbolos semánticos. Es decir, el diccionario contiene información conceptual sobre la adecuación de una red semántica. En tercer lugar, el diccionario de palabras también contiene información de tipo gramatical, que comprueba la aceptabilidad de la oración analizada.

En realidad, el diccionario de palabras cumple una labor fundamental en todo el proceso de análisis y generación, tanto en el componente morfológico, sintáctico como semántico contemplados en el sistema ATLAS-II (para más información al respecto véase Uchida et al., 1982). El hecho que sea el diccionario el portador de la información hace de ATLAS-II un sistema versátil y fácilmente extendible.

Conclusión

Hemos defendido la estrategia interlingua como estrategia válida en TA. Hemos visto que sus postulados teóricos no están tan alejados de la técnica de transferencia, cuando ésta también se plantea como una búsqueda de rasgos lingüísticos universales. El problema más polémico es el de la representación del conocimiento, su alcance y formalización. Tanto ésta como los primi-

tivos de la interlingua son tareas a resolver más empírica que teóricamente y suponen el mayor reto que tiene planteado el procesamiento del lenguaje natural hoy.

Referencias

- Barwise, J. & Perry, J. **Situations and Attitudes**. Cambridge Mass.: MIT Press.
- Carbonell, J.G. & Tomita, M. 1987. Knowledge-based Machine Translation, the CMC Approach. In Nirenburg ed. 68-89.
- Dowty, D. 1979. **Word Meaning in Montague Grammar**. Dordrecht: Reidel.
- Hayes, P.J. 1984. Entity-oriented parsing. In **Proceedings of COLING-84**:213-217.
- Hendrix, G. 1978. Semantic Knowledge. In Walker, D.E. ed. **Understanding Spoken Language**. New York: North Holland, 121-226.
- Jackendoff, R. 1985. **Semantics and Cognition**. Cambridge Mass.: MIT Press.
- Johnson, R. L. 1983. Parsing -an MT Perspective. In Jones, K.S. & Wilks Y. eds. **Automatic Natural Language Parsing**. Ellis Horwood, Ltd., Chichester, West Sussex, England.
- Johnson, R.L., King, M., & des Tombe, L. 1985. EUROTRA: a multilingual system under development. In **Computational Linguistics** 11:155-169.
- Kamp, J.A.W., 1981. A Theory of Truth and Semantic Representation. In Groendijk, J. Janssen, T. & Stokhof, M. eds. **Formal Methods in the Study of Language**. Part 1. Amsterdam: Mathematical Centre Tracs, 227-321.
- Landsbergen, J. Isomorphic Grammars and their use in the Rosetta Translation System. 1986. In M. King ed. **Machine Translation today**. Edinburgh: Edinburgh University Press.
- Lyons, J. 1967. **Introduction to Linguistics**. Cambridge University Press.
- Lytinen, S., & Schank, R.C. 1982. **Representation and Translation**. Report 234. Department of Computer Science, Yale University, New Haven, CT.

- Nagao, M., Tsujii J., & Nakamura J. 1985. The Japanese Government Project for Machine Translation. *Computational Linguistics* 11:91-110.
- Nirenburg, S. ed. 1987. **Machine Translation: Theoretical and Methodological Issues**. Cambridge: University Press.
- Nirenburg, S., Raskin, V., Tucker, A.B. 1987. The Structure of Interlingua in TRANSLATOR. In Nirenburg ed. 90-113.
- Schank, R. 1975. **Conceptual Information Processing**. Amsterdam: North Holland.
- Schank, R. 1982. Reminding and Memory Organization: An Introduction to MOPs. In Lenhart W., & Ringle M., eds. **Strategies for Natural Language Processing**. Lawrence Erlbaum Assoc., Hillsdale, NJ.
- Slocum, J. ed. 1987. **Machine Translation Systems**. Cambridge: University Press.
- Slocum, J., Bennett, W., Bear, J., Morgan, M., & Root, R. 1984. **METAL: The LRC Machine Translation System**. Working Paper LRC-84-2, Linguistic Research Centre, University of Texas, Austin.
- Tanimoto, S. T. 1987. **The Elements of Artificial Intelligence**. Maryland: Computer Science Press.
- Uchida, H., Hayashi, T., Kushima, H. 1985. ATLAS: Automatic Translation System. *Fujitsu Scientific and Technical Journal* 21:317-329.
- Wilks, Y. A. 1973. An Artificial Intelligence Approach to Machine Translation. In Schank R. C. & Colby K. M. eds. **Computer Models of Thought and Language**. San Francisco: Freeman, 114-151.
- Woods, W.A. 1975. What's in a link: Foundations for Semantic Networks. In Bobrow, D. & Collins A., eds. **Representation and Understanding**. New York: Algorithmics Press.