http://dx.doi.org/10.12795/PH.1997.v11.i02.04

BENEFICIOS QUE APORTA LA SEMÁNTICA A LA TRADUCCIÓN AUTOMÁTICA

Mª Gabriela Fernández Díaz

This paper deals with a particular method designed to incorporate semantic information to a Machine Translation (MT) system. Our claim is that in MT and Natural Language Processing it is possible to obtain accurate results when we incorporate semantic information. We try to prove the benefits we get when dealing with Semantics. The paper outlines part of a research project which goal consists of implementing a bidirectional MT system between English and Spanish with voice input and output for the restricted domain of currency exchange in a banking environment. At the beginning, the MT system could only process syntactic information. In this way, sentences syntactically correct but semantically anomalous were analysed and translated instead of being rejected as incorrect. In order to solve this problem, it was necessary to incorporate semantic patterns based on the semantic relation between a verb and its complements (this has been represented by means of ISA relations). Sentences which do not follow these semantic patterns will be disregarded by the MT system.

1. Introducción

Este artículo presenta un método utilizado en un proyecto de investigación para incluir información de tipo semántico en un prototipo de Traducción Automática (TA). Del mismo modo se exponen las ventajas y beneficios que supone incluir información semántica en los sistemas de TA. En un principio, se ofrece una perspectiva generalizada, tratando de ver el interés que muestran por la semántica los investigadores que se dedican a campos como el Procesamiento del Lenguaje Natural (PLN) y la TA. Por último, se desarrolla el método específico que se diseñó con el fin de incorporar información de tipo semántico en un prototipo de TA.

2. La Traducción Automática

La TA forma parte de una esfera mayor que engloba la investigación práctica relacionada con el PLN la Lingüística Computacional y la Inteligencia Artificial, todas ellas ciencias encargadas de explorar el mecanismo básico del lenguaje y la mente, con el propósito de modelarlos y simularlos por medio de programas informáticos. La principal tarea dentro de la TA puede definirse de un modo muy simple: el ordenador debe ser capaz de obtener como entrada un texto en una lengua (lengua fuente) y producir a su vez como salida un texto en otra lengua (lengua destino), de manera que el texto obtenido como

I.S.S.N. 1132-0265

lengua destino venga a ser el mismo que el texto con el que se contaba como lengua fuente. Esta afirmación es susceptible de muchos matices, dado que existen muchos puntos de vista y opiniones en torno al texto que se obtiene en lengua destino.

Existen dos tipos de sistemas de traducción, los que se conocen como sistemas directos y los sistemas de traducción indirecta, que engloban a los sistemas basados en transferencia y a los sistemas de interlingua (Hutchins, J. y H. Somers, 1992). La característica fundamental de los sistemas de traducción directos es que traducen palabra por palabra de una lengua a otra. Estos sistemas contienen un léxico y una lista de modelos típicos de palabras y frases en la lengua fuente y las palabras y frases correspondientes de la lengua destino. No existe un análisis sintáctico de las oraciones, a lo máximo que se llega a analizar es a nivel sintagmático y no todos los sistemas directos traducen a este nivel. El componente morfológico es el único con el que cuentan estos sistemas. Hoy en día los sistemas de traducción directa se consideran inadecuados en muchos sentidos y se trabaja con los tipos de sistemas indirectos. No obstante, los sistemas directos pueden resultar apropiados para determinados tipos de aplicaciones en los que se traducen textos con vocabulario limitado y un estilo definido.

Existe todo un debate con respecto a cual de los dos enfoques de sistemas indirectos es el más acertado. Común a ambos es la idea de una representación intermedia que capture el "significado" de la frase en lengua fuente con el fin de generar la frase en lengua destino con un significado equivalente. La naturaleza y el nivel lingüístico de esta representación intermedia es lo que distingue el diseño de los sistemas de transferencia de los de interlingua.

A pesar de las diferencias, ambos tipos de sistemas se caracterizan por el hecho de intentar "capturar el significado" de la frase de origen para poder traspasarlo a la frase destino. Hay muchos métodos a la hora de intentar reflejar en la traducción el conocimiento que contiene el texto en lengua fuente.

Hay sistemas que se basan en las sublenguas para tratar la traducción de textos. Al tratar con sublenguas se hace más fácil para el traductor captar el contenido del texto en lengua fuente, ya que la variedad del lenguaje que se usa en una ciencia determinada no sólo es mucho más reducida que el conjunto de toda una lengua, sino que también es claramente más sistemática en su estructura y significado. Esto ha llevado a lingüistas y científicos relacionados con el campo de la computación a colaborar en el estudio de las propiedades de estos lenguajes especiales, o adaptados a propósitos específicos, a los que se les puede denominar *sublenguas*. Lo que califica a una variedad de una lengua como sublengua, parafraseando a un estudioso de la materia llamado Kittredge (1987:63), "*no es su medida o complejidad, sino su adherencia a un uso sistemático*". El afirma que "*es en realidad el 'grado' de sistematicidad lo que determinará lo apropiado que resulte una sublengua para la traducción automática*". Si un analizador del lenguaje fuente se basa en una gramática de sublengua en vez de en una gramática de toda una lengua será posible obtener un alto grado de eficacia. En primer lugar, el tiempo de análisis queda reducido dado que las gramáticas de las sublenguas son siempre más pequeñas que las gramáticas de toda una lengua. En segundo lugar, el problema de la ambigüedad, tanto léxica como estructural, se reduce en

gran medida a consecuencia de que muchos de los análisis e interpretaciones posibles en el lenguaje estándar no tienen sentido alguno en una sublengua específica. La aplicación semántica incorporada al prototipo de TA con el que trabajamos también se basa en una lengua específica.

Hasta ahora se ha ofrecido una visión global de lo que es la TA, los distintos tipos de enfoque que se utilizan comúnmente en el campo de la TA, así como diversos mecanismos que emplean los sistemas de TA para capturar el significado, el contenido del texto en lengua fuente, para traspasarlo al texto en lengua destino. A continuación, vamos a tratar de ver el papel que juega la semántica en el campo de la TA.

3. La Semántica y la Traducción Automática

Con respecto a esta cuestión, existe un artículo muy interesante titulado "How much Semantics is Necessary for MT Systems?" publicado por Bennett, (1990). Bennett afirma que la semántica es un tema que ha estado presente en la TA desde que comenzó la investigación en ese campo. Las cuestiones que surgen siempre hacen referencia a la centralidad de la semántica en la TA, así como la forma y la extensión que toman los componentes de carácter semántico.

Hoy en día se reconoce que la semántica aporta grandes ventajas a los sistemas de TA, ya que ayuda a resolver muchas de las ambigüedades tanto léxicas como estructurales que se dan en los lenguajes naturales. El problema surge a la hora de decidir qué forma debe tomar esa semántica y así es como lo señalan Lehrberger y Bourbeau (1988:103): "*The need for semantic analysis is generally recognized; the big question is how to do it.*"

Bennett expone en su artículo que la creencia común a este respecto apunta a que debe realizarse un análisis semántico profundo si se quiere conseguir un sistema de TA viable. Sin embargo, él cuestiona esta idea. Aunque es obvio que mientras más poderoso sea el análisis semántico, más profundo será el análisis, Bennett hace referencia a la naturaleza del coste que ha de suponer para el sistema de TA la incorporación de un componente semántico. Así, señala que al tratar con el componente semántico existen tres elementos que repercuten en el coste y que son: la eficacia del sistema, el léxico o la base de conocimiento y la rentabilidad.

En cuanto a la eficacia del sistema, Bennett observa que cualquier componente semántico reducirá la velocidad del sistema, por lo que mientras más sofisticado sea el componente semántico, mayor será también la pérdida de velocidad del sistema. Para él, el truco consiste, por lo tanto, en tener un componente semántico que sea lo más efectivo posible y que produzca, a su vez, una carga mínima en el sistema.

En lo que respecta al léxico o la base de conocimiento, Bennett hace referencia al tiempo que se tarda en construir, siendo éste un factor que también afecta al coste del sistema. Para él, un análisis semántico profundo requerirá más tiempo en la adquisición y mantenimiento del léxico o base de conocimiento que un sistema más sencillo y efectivo.

En cuanto a la rentabilidad del sistema, Bennett se refiere a la relación de los gastos con la capacidad del sistema para actuar tal y como es de esperar.

Estos son los tres aspectos que deben tenerse en cuenta para elegir entre lo que Bennett denomina: "a semantic feature system" (un sistema con rasgos o características semánticas) y, "a deep semantic analysis" (un análisis semántico profundo).

En un sistema con rasgos semánticos la información semántica está codificada en las entradas léxicas y es tratada en la fase de análisis, del mismo modo que la información morfológica o la sintáctica. Por el contrario, en un sistema basado en un análisis semántico profundo, el componente semántico representa una parte fundamental y distintiva del sistema.

Otra diferencia entre ambos tipos de sistemas se refiere al coste computacional. El coste de un sistema de rasgos es considerablemente menor que el de un análisis semántico profundo. La cuestión, para Bennett, es si está justificado el coste computacional adicional necesario en un sistema basado en el análisis semántico profundo en relación a los resultados que se obtienen.

Con todo esto, Bennett concluye que el coste económico que supone desarrollar, correr y mantener un sistema basado en el análisis semántico profundo frente a los logros que se obtienen en el análisis deberían impedir tal inversión. Bennett expone que un sistema con características semánticas no genera gastos adicionales en cuanto a su desarrollo u operatividad.

El prototipo de TA utilizado para nuestro estudio mostraba ciertas carencias en cuanto al componente semántico. No obstante, los problemas encontrados podían solucionarse dentro de la fase de análisis, sin repercutir en modo alguno, en el diseño y la arquitectura global del sistema. Es decir, era necesario subsanar ciertos problemas generados en la fase de análisis del sistema. Sin embargo, no era necesario crear un componente semántico independiente que se encargara de realizar un análisis semántico profundo. Dicha labor, además de costosa, hubiera repercutido negativamente, del modo que apuntaba Bennett en su artículo, además de que en nuestro caso era innecesaria, dada la naturaleza de la sublengua empleada.

Con todo lo expuesto hasta ahora, se puede concluir diciendo que existen muchas maneras de acercarse al tratamiento de un texto en el campo de la TA. El hecho de que unas personas se inclinen por un enfoque u otro, o decidan incorporar en el sistema módulos independientes de naturaleza sintáctica o semántica profunda depende en gran medida de la finalidad del sistema. Es decir, dependiendo de para qué queramos la aplicación, nos inclinaremos por un sistema u otro, o por un mecanismo u otro. Creemos que no se debe abogar por un sistema cerrado, sino que por el contrario, una combinación de diferentes técnicas integradas de una manera razonada puede probar ser bastante más acertada que aplicar un paradigma cerrado.

4. Incorporación de información semántica a un prototipo de TA

En esta última parte del artículo se desarrolla el prototipo de TA con el que hemos trabajado. En concreto, nos centraremos en un mecanismo diseñado para incorporar información de tipo semántico en el sistema.

El objetivo perseguido consistía en implementar un sistema de TA bidireccional entre el inglés y el castellano, con entrada y salida por voz, para el dominio de cambio de moneda en un entorno bancario. En un principio, el sistema de traducción sólo procesaba información sintáctica. En este sentido, oraciones que eran sintácticamente correctas pero semánticamente anómalas eran analizadas por el parser del sistema, y consecuentemente eran traducidas, en vez de ser rechazadas como incorrectas. Con el fin de solucionar este problema era necesario incluir en el parser información de tipo semántico, de manera que todas aquellas oraciones de naturaleza agramatical fueran rechazadas. Así, se mostrará un método elaborado con el fin de incorporar información de tipo semántico en el sistema. La información que ofrecen los verbos del entorno del cambio de moneda desempeña un papel fundamental en esta tarea.

El prototipo de TA consta de tres módulos fundamentales: el módulo del reconocimiento del habla, el módulo de la traducción y el módulo de la síntesis del habla. Dada una frase de entrada, el reconocedor del habla genera una serie de candidatos, ordenados del más al menos probable. Como se trata de un reconocedor de habla continua, es posible que la frase reconocida no coincida con la frase proporcionada como entrada. Esto es, dos palabras como 'pasaporte' y 'peseta', por motivos acústicos, pueden confundirse. Cuando esto sucede, el módulo que se encarga de la traducción, debe ser capaz de rechazar aquellos candidatos que no sean válidos por un motivo u otro.

El módulo de la traducción consta de tres fases: la fase de análisis, la fase de transferencia y la fase de generación. Cuando el módulo del reconocedor del habla envía un candidato al módulo de traducción, el parser o analizador gramatical del sistema debe decidir si el candidato es o no correcto. De ahí, que el parser deba incluir suficiente información sintáctica y semántica que le ayude en su tarea de seleccionar al candidato idóneo. Una vez que se elige un candidato, el parser lo analiza. Ya en la fase de transferencia, el léxico y las estructuras de la lengua fuente se transfieren a las estructuras de la lengua destino. En la última fase, la fase de generación, la frase en lengua destino es generada. De todo esto se deduce que el parser del sistema debe actuar como un filtro, de manera que si encuentra un fallo, es decir, si el candidato que le envía el reconocedor de voz no es el correcto, el parser, haciendo uso de la información que posee, debe ser capaz de rechazar la frase y pedir a la persona que la formule de nuevo.

Por último, el módulo de la síntesis del habla es el encargado de convertir en voz el texto generado por el módulo de traducción en la lengua destino que corresponda.

El analizador del sistema, llamado "LEKTA", debía contener información que le ayudase a seleccionar el candidato correcto. Sin embargo, en un primer momento, LEKTA recibía sólamente información de tipo sintáctico. La gramática de LEKTA está basada en el formalismo léxico-funcional conocido como LFG, del inglés *Lexical Funcional Grammar*

(Bresnan ed, 1982). La LFG es una teoría conocida lingüística y computacionalmente. Una gramática léxico-funcional asigna dos niveles de análisis. Por un lado, existe un nivel superficial que incluye sólamente información sintáctica y que se conoce como "estructura-c" o estructura de constituyentes. Por otro lado, la "estructura-f" o estructura funcional actúa como entrada al componente semántico. Es en este nivel donde se debía incluir la información de tipo semántica que el parser no poseía en un principio.

Como hemos señalado anteriormente, se encontraron casos anómalos en los que la información sintáctica proporcionada al parser no bastaba para solucionarlos. Así, la oración 'Me gustaría depositar un pasaporte' era aceptada como correcta y, consecuentemente, era analizada y traducida. Sin embargo, se sabe que desde un punto de vista semántico ésto no es correcto, por lo que el parser no debería de haber analizado esa oración.

Para prevenir este tipo de fenómenos, era necesario incluir en el parser información de carácter semántico. Con tal fin, se diseñaron unos patrones semánticos basados en los verbos de la sublengua del entorno bancario. Para crear dichos patrones había que clasificar las palabras y las oraciones del corpus atendiendo a características semánticas.

En primer lugar, se clasificaron todas las palabras del corpus. Así, tales palabras como 'cheque', 'cheques de viaje' y 'cheques personales' fueron englobadas bajo la categoría mayor "CHEQUE". Palabras como 'libras', 'peseta', 'franco' y 'dólar' pertenecen al grupo denominado "DIVISA". Adjetivos tales como 'español', 'japoneses', 'alemán' e 'inglés' son miembros del grupo "NACIONALIDAD".

En segundo lugar, una vez finalizada la agrupación de palabras atendiendo a campos semánticos, se procedió a clasificar las oraciones del corpus. A modo de ejemplo, para aquellas frases que proporcionan información relativa a la tasa de cambio, se construyó un patrón denominado "Declarativas de la Tasa". Frases típicas de este patrón son: 'La tasa de cambio entre dólares y pesetas es de cien pesetas por dólar', 'El tipo de cambio actual es bajo' y 'La tasa de cambio actual entre libras inglesas y francos franceses está a novecientas doce libras por franco'. La finalidad de todos estos patrones oracionales era agrupar las oraciones del corpus en base a criterios semánticos.

Una vez finalizada la clasificación de las oraciones del corpus atendiendo a la información semántica que proporcionaban se procedió a construir los patrones semánticos verbales. El motivo fundamental por el que se diseñaron estos patrones semánticos basados en los verbos del dominio bancario se basa en los beneficios que se obtienen en la traducción de las oraciones cuando se incorpora en el parser información de tipo semántico y no simplemente sintáctica. El objetivo consiste en mostrar las relaciones semánticas sostenidas entre los verbos y los objetos regidos por los verbos. Así, por poner un ejemplo, dado el verbo 'cambiar', sabemos que en el dominio bancario sólo podemos cambiar dinero o cosas relacionadas con el dinero, tales como un cheque, y que opcionalmente, podemos cambiar ésto en otro tipo de dinero. Algunas frases utilizadas frecuentemente en el dominio bancario son: '¿Cuánto dinero puedo cambiar?', 'Deseo cambiar setecientas ochenta y cuatro libras', 'Quiero cambiar estas libras en dólares' y '¿Es posible cambiar cuatrocientos

cincuenta francos?'. Por ello, necesitabamos un grupo semántico que recoja todos los tipos de dinero que se pueden cambiar. Este es el grupo semántico que se denominó "TIPO_DE_DINERO".

El método elegido para desarrollar los modelos semánticos verbales está inspirado en lo que se conoce como gramática de dependencias. En este modelo de gramática, los complementos de un grupo semántico que dependen de un verbo se representan de manera que los primeros dependen de los últimos por medio de ramas. Los patrones semánticos se utilizan para representar la relación entre los verbos (o elementos gobernantes) y sus complementos (o elementos gobernados). Con los modelos de dependencia creados para los verbos del dominio bancario se pretende demostrar la importancia que conlleva la información que estos verbos ofrecen, así como el papel fundamental que juega esta información en nuestro objetivo, que no es otro que lograr un análisis correcto de los diferentes tipos de oraciones que conforman el corpus del dominio bancario. Este tipo de información es fundamental para que el parser logre obtener unos resultados acertados.

Los ejemplos 1(a) y 1(b) suponen una muestra patente de dos oraciones que si bien su realización sintáctica es idéntica, su realización semántica no lo es.

1(a) Me gustaría depositar un dólar.

1(b) Me gustaría depositar un pasaporte.

El verbo principal en ambas oraciones ('depositar') implica de su significado que el objeto que recibe la acción verbal se refiera a un nombre cuyo campo semántico guarde relación con el dinero. De esta afirmación se deduce que la frase 1(b) no tiene sentido alguno. Sin embargo, la gramática diseñada en base a la sintaxis no era capaz de recoger dicha apreciación. El patrón semántico creado para el verbo 'depositar' muestra cómo su objeto debe ser obligatoriamente un nombre que pertenezca al grupo semántico denominado "TIPO_DE_DINERO". El nombre del primer ejemplo 'dólar' es un miembro de ese grupo, pero el nombre del segundo 'pasaporte' no lo es. Gracias al modelo semántico del verbo 'depositar', las oraciones que se asemejen al tipo 1(b) serán rechazadas automáticamente, mientras que aquellas oraciones que se parezcan a la frase 1(a) serán aceptadas y analizadas correctamente.

Una vez finalizado el desarrollo de los patrones semánticos verbales, simplemente quedaba verificar los logros que se obtienen de la incorporación de los modelos semánticos al parser, de manera que se comprobase cómo el sistema rechazaba realmente las oraciones semánticamente anómalas. La implementación de los patrones semánticos se llevó a cabo por medio de redes semánticas. El propósito de una red semántica consiste en organizar el conocimiento del mundo real a través de una taxonomía que presenta la propiedad de la herencia. Una taxonomía sitúa los individuos en clases y especifica qué clases son a su vez subclases de otras clases. Cada entrada está definida por su posición en la red y por la(s) relación(es) que sostiene con los nudos de su alrededor. Las redes semánticas en nuestro sistema están basadas en la relación "ISA" (del inglés 'is_a'), 'es_un', utilizada

normalmente para describir relaciones de calidad de miembro. Por herencia, un subtipo hereda todas las propiedades de su tipo.

Se han seleccionado tres oraciones para mostrar los resultados de esta aplicación. Las oraciones son: 'Me gustaría cambiar un pasaporte', 'Me gustaría cambiar pasaportes a pesetas' y 'Me gustaría cambiar dólares a pesetas'. Desde un punto de vista sintáctico esas tres frases son correctas, por lo que el parser no debería de encontrar ningún tipo de anomalía.

El parser analizó la oración 'Me gustaría cambiar un pasaporte' antes de incorporar la información semántica al sistema de TA sin encontrar ninguna anomalía.

Al incorporar al parser la información semántica proporcionada por el patrón semántico creado para el verbo 'cambiar', el mensaje generado por el sistema fue: "No". La palabra 'pasaporte' no es un miembro del patrón semántico del verbo 'cambiar'. No es un miebro del grupo "TIPO_DE_DINERO", sino que pertenece al grupo "DOCUMENTO". La unificación falla y el parser no analiza la oración.

Para mostrar la implementación de una frase correcta en todos los sentidos, le dimos al parser la oración "Me gustaría cambiar dólares a pesetas". El éxito en el análisis de esta frase se debió a que el parser fue capaz de instanciar con éxito la información con la que contaba. El verbo 'cambiar' requiere que su objeto sea un tipo de dinero tal y como se recoje en la entrada léxica. En la fase de unificación la variable debe instanciarse con su valor. La unificación trata de unir la información del sintagma nominal con la información que se especifica en la entrada léxica. Una vez que la variable se instancia, el algoritmo de la unificación consultará la jerarquía isa para confirmar la compatibilidad semántica. El efecto final de tal interacción es el que conduce al correcto análisis de dicha oración. De este modo, la oración, una vez analizada, será enviada al módulo de traducción.

5. Conclusión

Al implementar estas frases antes y después de incorporar los modelos semánticos verbales se ha intentado probar la eficacia del método elegido. De este modo, ha quedado demostrado cómo el parser puede realmente distinguir las frases semánticamente correctas de las que no lo son.

En esta última parte del artículo hemos tratado de demostrar cómo la información semántica que los verbos proporcionan es decisiva a la hora de resolver ciertos problemas localizados en la fase de análisis del módulo del traductor del sistema. Como hemos podido ver el traductor necesita todo tipo de información con el fin de poder generar una oración correcta en todos los sentidos. Mientras más información reciba el parser mayor será el grado de acierto, así como mayores serán los resultados obtenidos en la traducción. Al tratar con los campos de la TA y el PLN la información proporcionada por cualquiera de las áreas de la Lingüística desempeña un papel fundamental en el objetivo perseguido en este terreno, es decir, en la generación de traducciones perfectas en la medida de lo posible.

BIBLIOGRAFÍA

Bennett, W. S. (1990): *How much Semantics is Necessary for MT Systems?*. Siemens Communications Systems, Inc. y The Linguistics Research Center, University of Texas, Austin.

Hutchins, J. y H. Somers (1992): *An Introduction to Machine Translation*. London: Academic Press.

Kaplan, R. y J. Bresnan (1982): "Lexical-Functional Grammar: a formal system for grammatical representation", en J. Bresnan, ed. *The Mental Representation of Grammatical Relations*. Cambridge, Mass: MIT Press.

Kittredge, R. I. (1987): "The significance of sublanguage for automatic translation", en Nirenburg, S. ed. Machine Translation: Theoretical and Methodological Issues. Cambridge: Cambridge University Press. 59-68.

Lehrberger, J. & L. Bourbeau (1988, 103). En Bennett (1990).