MENTOR: un proyecto en Traducción Automática

P. Rodríguez Marín

L. de Sopeña

Centro de Investigación UAM-IBM

Universidad Autónoma de Madrid

RESUMEN

El propósito de esta comunicación es presentar brevemente el proyecto de Traducción Automática MENTOR, recientemente iniciado en el Centro de Investigación UAM-IBM, en colaboración con otros centros de investigación. Su objetivo es construir un prototipo multilingüe que lleve a cabo la traducción de textos fuente escritos en inglés a diversos lenguajes objeto: castellano y hebreo son las dos primeras lenguas consideradas. Se exponen los objetivos fundamentales y las líneas generales de diseño del sistema proyectado.

INTRODUCCIÓN

La Traducción Automática (TA) constituye un área de investigación de gran interés, ya que, a la vez que representa una importante aplicación del procesamiento del Lenguaje Natural (LN), proporciona un marco excelente para la investigación en otras áreas del LN.

Es ciertamente una aplicación importante por derecho propio: la traducción de documentos es imprescindible para cualquier organización con intereses internacionales, y su mecanización llega a convertirse, en muchos casos, en una necesidad. Al mismo tiempo, el camino hacia la construcción de un sistema de TA implica la revisión y síntesis de investigaciones en numerosos subcampos del procesamiento del LN. Puede así servir para comprobar la validez de las teorías, en términos de su aplicabilidad a un amplio subconjunto de un idioma, y la universalidad de las mismas, en el sentido de su adaptabilidad a otros idiomas. Al mismo tiempo, el proyecto va a servir como base para el desarrollo de herramientas integradas para el procesamiento del LN.

Es cierto que hoy se sigue cuestionando la posibilidad de que un ordenador lleve a cabo automáticamente una traducción: recordemos, por ejemplo, el artículo de M. Kay: "Machine Translation will not work" [1]. Sin embargo, la controversia se reduce, posiblemente, a ponerse de acuerdo sobre qué se entiende por "Traducción Automática". Evidentemente, conseguir una traducción de cualquier tipo de texto, de forma totalmente automatizada y con calidad para su publicación, es todavía impensable; pero exigir tanto no es necesario para que un sistema de TA sea útil (y, no lo olvidemos, ¿sería un traductor humano capaz de proporcionar los resultados antes requeridos al ordenador?). Basta con que su uso aumente de forma significativa el rendimiento del traductor, la cantidad de material traducido en un tiempo dado. De hecho, muchos de los sistemas de TA hoy existentes necesitan de la intervención humana en alguno de los estadios del proceso: pre-edición (para reescribir el texto fuente eliminando ambigüedades y construcciones problemáticas), interacción durante la traducción (variable, según se trate de sistemas de Traducción Automática asistida por el usuario, o de Traducción Manual asistida por el ordenador), y post-edición (revisión del resultado por un traductor humano, fase de todos modos necesaria también tras un proceso manual de traducción). Cada tipo de intervención presenta una serie de dificultades e inconvenientes que pueden hacer inutilizable en la práctica un sistema que se apoye en un uso excesivo de cualquiera de ellos.

En los últimos años se ha vuelto a producir un renovado interés en investigación y desarrollo en TA, con el nacimiento o resurgimiento de diversos proyectos en Europa, Japón y los Estados Unidos. Ello se basa fundamentalmente

en el reconocimiento de que los avances de la lingüística y las técnicas del procesamiento del LN pueden ser aprovechados y aplicados con ventaja en esta área. Buena prueba de ello la constituyen los números especiales de la revista Computational Linguistics dedicados al tema recientemente [2].

El objetivo de esta comunicación es describir brevemente el proyecto de Traducción Automática MENTOR (Multilingual ENGLISH Translator), que ha sido iniciado el presente año en el Centro de Investigación UAM-IBM, Madrid, en colaboración con el Centro Científico de IBM en Haifa (Israel). Este proyecto tiene como objetivo la construcción de un prototipo multilingüe que lleve a cabo la traducción automática de textos escritos en inglés a dos lenguas terminales: castellano y hebreo, inicialmente. Evidentemente, el tipo de textos para traducir debía restringirse para hacer factible el proceso: el dominio elegido por nosotros es el de los manuales de productos IBM. La característica de multilingüidad se considera uno de los objetivos prioritarios de diseño para hacer posible la incorporación al sistema de otras lenguas objeto en fases futuras.

La arquitectura se basa en el modelo *transfer*, que divide el proceso de traducción en tres etapas: 1) análisis de la oración fuente y representación interna de la misma, 2) transferencia, es decir, conversión de dicha representación interna en una representación interna de la oración objeto deseada, 3) generación de la oración en lenguaje objeto a partir de esta última. Se pretende posponer y reducir al máximo la fase de *transfer* para que una buena parte del proceso sea válida para cualquier lenguaje objeto, ya que las decisiones dependientes de éste empiezan a tomarse en dicha fase.

DESCRIPCIÓN GENERAL

Clasificación

De acuerdo con la clasificación que de los sistemas de TA hace Tucker [3], dividiéndolos en tres generaciones con respecto a la cantidad de conocimientos lingüísticos que incorporan, MENTOR pertenece a la segunda generación (en cuanto que se basa en una representación intermedia alrededor de la cual se organiza el análisis y la síntesis, manteniéndose así independiente en buena medida de los lenguajes fuente y objeto), pero también incluye elementos de la tercera (manejo de información semántica y utilización de estructuras de representación del conocimiento).

Si nos atenemos al nivel de interacción hombre-máquina que se requerirá, criterio seguido en la clasificación de Slocum [4], el sistema es de tipo automático, aunque, como es habitual, se considera imprescindible la post-edición. En el camino hacia un sistema "automático" se contempla también la posibilidad de un proceso intermedio de traducción "asistida": está previsto el estudio de herramientas que permitan una cierta intervención humana durante la traducción en aquellos aspectos que resultan más conflictivos; de este modo, el análisis será más preciso y se facilitará la tarea a las fases subsiguientes.

Dominio de aplicación

El sistema va a ceñirse a un dominio de textos fuente muy específico, pero suficientemente general desde el punto de vista lingüístico. Los expertos en TA están de acuerdo en que el problema principal de la traducción se encuentra en la semántica, en elegir, para una palabra del texto fuente, su correcto equivalente en la lengua objeto entre todos los posibles significados en cada contexto concreto. Para vencer este obstáculo, la primera posibilidad es restringir la semántica a base de escoger un dominio de textos limitado. En nuestro caso, como ya hemos adelantado, el ámbito elegido es el de los manuales de IBM. De este modo, el vocabulario que se va a utilizar es más reducido, la semántica es limitada y la ambigüedad menor, a la vez que se facilita la elaboración del diccionario, una de las fases cruciales y laboriosas de una traducción.

Esta elección va a ser uno de los factores que contribuyan a la calidad del texto generado. Se acepta, sin embargo, que este sea, como en otros sistemas de TA, sometido a un proceso manual de post-edición, que subsane posibles errores y mejore la calidad, el estilo y la legibilidad.

En cambio, la pre-edición del texto de entrada queda excluida, ya que, aunque es obviamente útil para obtener oraciones más fáciles de procesar, no se considera una alternativa viable debido a la cantidad de trabajo que supone, y al hecho de que el texto así obtenido resultaría artificialmente distorsionado, cuando el mayor interés de una TA es tratar material real pre-existente.

Características de diseño

Como se ha mencionado, el sistema va a seguir el modelo *transfer*; en consecuencia, el proceso de traducción se dividirá en tres fases:

1. Análisis del texto inglés
2. Transferencia
3. Generación del texto objeto

La fase de análisis incluye todo el tratamiento del texto fuente que sea independiente del lenguaje objeto: análisis, desambiguación, transformación. En ella utilizamos diversas herramientas ya existentes, desarrolladas en el laboratorio T.J. Watson de IBM en Yorktown [5]: un procesador, PLNLP [6], una gramática inglesa, PEG [7], y un léxico (basado en los diccionarios Webster-7 y Longman). El resultado es uno (o varios) árboles de análisis con diversas informaciones asociadas.

PLNLP es un analizador basado fundamentalmente en la sintaxis, y con un diseño que permite la adición de una o varias fases de procesamiento semántico, adaptadas a las necesidades y características específicas de cada aplicación. Así, pensamos en incluir procesos de desambiguación tanto automáticos como manuales. La desambiguación automática consiste en seleccionar las estructuras y significados más probables a partir de restricciones contextuales (al nivel de cada oración). De esta forma, un intérprete semántico generaría la representación del significado de las oraciones utilizando una base de conocimientos; el trabajo de Sowa [8] sobre los grafos conceptuales que se generan

a partir de los grafos canónicos asociados con cada término es un ejemplo de la dirección en la que vamos a orientar este proceso. El procedimiento permite elegir entre diversos posibles árboles de análisis cuando la oración de entrada es ambigua, decidir sobre la coherencia semántica de una oración, seleccionar una acepción específica para una palabra en caso de ambigüedad léxica, etc. Adicionalmente (o como alternativa parcial), esta desambiguación automática podrá complementarse, cuando la base de conocimientos no baste, con otra interactiva, mediante el diálogo con el usuario. Éste deberá ser competente en el lenguaje fuente, y el mecanismo consistirá fundamentalmente en pedir que seleccione entre paráfrasis alternativas de fragmentos del texto.

Debido a que buena parte de la post-edición necesaria actualmente se dedica a corregir traducciones erróneas derivadas de una desambiguación inadecuada, este método permitirá una reducción considerable del coste de la traducción.

Como resultado del proceso, el árbol de análisis desambiguado llevará asociadas explícitamente informaciones de la estructura profunda junto con la información superficial que caracteriza a la oración fuente, ambas necesarias para generar correctamente la salida. Ésta deberá expresar el mismo significado que aquélla, manteniendo, al mismo tiempo, tantas características estructurales y estilísticas del original como sea posible.

Esta primera fase será general y compartida, cualquiera que sea el lenguaje que se genere. Sólo la segunda fase, la de transferencia, empezará a utilizar información del lenguaje objeto. Durante ella, la estructura de análisis antes obtenida se modifica de modo que refleje la estructura de la lengua que se va a generar. También se ha de emplear un diccionario bilingüe que contenga las equivalencias de cada término, con el objeto de construir el esqueleto básico en el que se encuentre la información sintáctica y semántica necesaria para la última fase.

El proceso de transferencia consta fundamentalmente de dos partes: transferencia léxica (que selecciona el significado de cada palabra en el lenguaje objeto), y transferencia estructural (que selecciona las formas sintácticas). El diseño de estos procesos es muy dependiente de la representación que se haya utilizado; parece observarse que las representaciones que reflejan físicamente la forma de la fuente, tales como los árboles de análisis, aumentan el tamaño y la complejidad de la fase de *transfer* [9]. Por ello puede resultar útil codificar la información directamente, mejor que reformularla en forma de árbol, es decir, reducir las oraciones a especificaciones sobre su contenido básico (tal como se hace al reducir oraciones a formas lógicas), junto con especificaciones de las variantes sintácticas utilizadas para expresar dicho contenido.

La fase final de síntesis generará la oración de salida a base de recorrer la estructura previamente construida, flexionando las palabras en la forma adecuada, y utilizando un diccionario específico.

CONSIDERACIONES FINALES

El sistema se está desarrollando en un entorno VM/CMS y, en el tiempo transcurrido desde la iniciación del proyecto (marzo de 1987), se han realizado diversas actividades básicas (además de las de organización y formación): instalación del procesador del inglés PLNLP y PEG, estudio de la estructura

del diccionario, selección de un corpus básico de texto para traducir, estudio de las características de los manuales de IBM, a base de muestras de diversa procedencia.

Asimismo se han considerado los problemas teóricos que presenta la traducción de determinadas estructuras inglesas, tales como el subjuntivo y las palabras en '-ing', y se ha intentado definir reglas, cuando ello es posible, para facilitar y sistematizar su traducción.

El objetivo inmediato que se pretende alcanzar es la construcción de un primer prototipo rudimentario que sea capaz de procesar un texto sencillo previamente seleccionado; ello va a permitir tratar en profundidad algunos de los problemas que se van a presentar más adelante y crear una herramienta básica sobre la que añadir progresivamente nuevas características de creciente complejidad.

REFERENCIAS

[1] Key, M.: *Machine Translation Will Not Work*, Proc. of the 24th Annual Meeting of the ACL, June 1986.

[2] Computational Linguistics, Special Issues on Machine Translation, vol. 11, nos. 1 and 2, 1985.

[3] Tucker, A.B.: *A Perspective on Machine Translation: Theory and Practice*, Comm. of the ACM, vol. 27, no. 4, 1984.

[4] Slocum, J.: *A Survey of Machine Translation: Its History, Current Status and Future Prospects*, Computational Linguistics, vol. 11, no.1, 1985.

[5] Jensen, K., et.al.: *PLNLP, PEG and CRITIQUE: Three Contributions to Computing in the Humanities*, RC 11841, IBM Thomas J. Watson Research Center, 1986.

[6] Heidorn, G.: *Natural Language Inputs to a Simulation Programming System*, PhD Thesis, Yale University, 1972.

[7] Jensen, K.: *PEG 1986: A Broad-coverage Computational Syntax of English*, IBM Thomas J. Watson Research Center, draft, 1986.

[8] Sowa, J.: *Implementing a Semantic Interpreter using Conceptual Graphs*, IBM J. of Research and Development, vol. 30, no.1, 1986.

[9] Newman, P., et.al.: *LASC MT Project: Goals and Directions*, IBM Los Angeles Scientific Center, Report G320-2795, 1986.