La Traducción Automática y sus implicaciones en la representación del conocimiento

Mª José Ayuso Sánchez
Universidad Carlos III de Madrid

Resumen:

Reflexión teórica sobre las distintas tendencias que existen en los sistemas de traducción automática. Se proponen cuáles son los aspectos más significativos de los componentes y procesos implicados en las distintas fases de un proceso de traducción automática, con especial énfasis en el análisis lingüístico. Se describen los enfoques cognitivos presentes en los sistemas de traducción automática basados en la adquisición del conocimiento y la superación de técnicas para el procesamiento del lenguaje natural. Se plantean las incursiones de este área de conocimiento interdisciplinar con los sistemas expertos y las teorías procedentes del conectismo.

Palabras clave: Traducción automática. Procesamiento del lenguaje natural. Enfoques basados en el conocimiento. Lingüística Computacional. Desarrollo metodológico.

Abstract:

A theoretical approach about the different tendencies of the Machine Translation. The more significant aspects of components and processes involve in the different phases of the Machine Translation are proposed, special emphasis in the linguistic analysis. We will describe the different cognitive views that are present in Knowledge-Based Machine Translation and the development of techniques for natural language processing. We will show the incursions of this interdisciplinary knowledge area in the expert systems and the theory from the connectionism.

Keywords: Machine translation. Natural Language Processing. Knowledge-Based Approaches. Computational Linguistic. Methodological Development.

Evolución de las distintas tendencias en los sistemas de Traducción Automática.

Las investigaciones sobre sistemas de traducción automática desarrolladas en Japón, USA y Europa han dado lugar a la aparición de grandes paquetes de software [1] como Meteo, Systran y Logos que tuvieron entre los años 70 y 80 una gran difusión en el mercado. Estas investigaciones contribuyen de forma decisiva al avance en este área de conocimiento interdisciplinar, avances procedentes también del Programa de Investigación Básica de DARPA (Defense Advanced Research Projects Agency) que ha dedicado importantes esfuerzos a los enfoques basados en el conocimiento como una aportación de los sistemas expertos, una superación de los sistemas basados en reglas y una orientación hacia modelos estadísticos que disponen de analizadores basados en preferencias y la implementación de corpus etiquetados[2]. Los avances experimentados en los 90 por EAGLES (Expert Advisory Group on Language Engineering Standards) proyecto sustentado por la Unión Europea, han puesto de manifiesto por medio de un informe preliminar la importancia de la evaluación de los sistemas para el procesamiento del lenguaje natural (PLN) vinculando el desarrollo de la traducción automática con la industria de las tecnologías de la información [3].

Los factores presentes en los sistemas de traducción automática son transferencia sintáctica, semántica y conocimiento lingüístico suficiente para analizar el texto. Los sistemas de traducción automática basados en un enfoque directo requieren un análisis sintáctico y semántico muy profundo, aunque el principal problema radica en determinar ese conocimiento no lingüístico presente en todo análisis léxico y que puede determinar el dominio abarcado por el texto. La consolidación de las reglas lingüísticas presentes en la transferencia a todos los niveles, supone condensar el análisis morfológico y léxico de las fases de generación para un

estudio del propio proceso de traducción automática presente en las estructuras lingüísticas de las lenguas analizadas.

Los sistemas centrados en el estudio de los aspectos lingüísticos para condensar las fases de generación y transferencia empiezan a coexistir con sistemas que se centran en los factores condicionantes presentes en el corpus tratado. Surgen sistemas de traducción automática que se basan en orientaciones y métodos estadísticos, mientras que otros prefieren hacer uso de bases de conocimiento recurriendo a técnicas de unificación y consolidación de formalismos presentes en algunas teorías lingüísticas [4]. Un ejemplo claro de un sistema de traducción automática basado en un método de transferencia es el prototipo de concepción avanzada Eurotra.

La formación de equipos japoneses de investigación, que analizan un texto a partir de una serie de ejemplos traducidos con la aplicación de indicadores estadísticos, supone la aproximación a una nueva técnica para la selección de términos en las fases de generación y transferencia aunque la presencia de reglas morfológicas, sintácticas y semánticas es limitada. El enfoque más significativo de los sistemas de transferencia es la subdivisión mediante reglas lingüísticas, vinculadas a los niveles necesarios para superar el proceso de descomposición entre la lengua fuente/objeto en la configuración del interfaz que mejor se adecue a las fases más significativas [5]. El estudio de los contrastes gramaticales es fundamental en sistemas de traducción automática que precisan de una fase intermedia para trasladar las estructuras divergentes de la lengua fuente/objeto hacia un esquema convergente. La especificación de las variantes, presentes en las reglas que condicionan la transferencia en las posibilidades implícitas en los distintos niveles es símbolo de su aplicación a circunstancias específicas. Se puede mantener, por tanto, el mismo formalismo para las fases de análisis, generación y transferencia con los condicionantes bidireccionales oportunos.

El desarrollo de fuentes de información alternativas para un estudio del léxico con mayor exhaustividad, ha llevado a la consolidación de nuevos estudios en procesamiento del lenguaje natural, determinación de frecuencias en la selección y adquisición de información léxica y conceptual sin olvidar la Pragmática. Algunos equipos de investigación internacionales centrados en el desarrollo de diccionarios técnicos especializados, diccionarios bilingües y bancos de datos terminológicos, son en definitiva realizaciones que pueden contribuir a un desarrollo todavía más interdisciplinar en el campo de la traducción automática [6].

El uso de sistemas estadísticos para la alineación de palabras sobre ejemplos traducidos, la menor integración de reglas lingüísticas y la percepción que se deriva de los modelos basados en el conocimiento, son tendencias en un estudio del lenguaje natural de cara a la inserción de las estructuras morfosintácticas y su correspondiente análisis lingüístico en el contexto de la recuperación de la información. La complejidad de las reglas lingüísticas existentes en los primeros sistemas de traducción automática, necesarias para un equilibrio en la sistematización de relaciones de dependencia a nivel semántico y sintáctico, se van a desplazar hacia la construcción de reglas de dominio específico sobre corpus seleccionados previamente para evitar la ambigüedad en la interrelación semántico-léxica.

Las aportaciones de los métodos basados en frecuencias estadísticas y los modelos conectistas de inducción de reglas contribuyen al desarrollo de sistemas que ponderan el análisis lingüístico pertinente, parámetros estadísticos, información procedente de bases de conocimiento con aplicación de algoritmos y la superación de perspectivas directas e indirectas en el estudio del corpus.

Implicaciones en la representación del conocimiento.

La representación del conocimiento implica un proceso de captación de datos válidos de los corpus, objeto de estudio y descripción, representativos en la etapa de formación de reglas a partir de textos traducidos como un sistema de organización del conocimiento, como un modelo de percepción de estructuras de extracción del conocimiento lingüístico. La asignación de categorías como elementos identificadores de funciones léxicas en la configuración de redes semánticas y la simplificación de niveles informativos en la fase de generación y transferencia, son aspectos relevantes para una organización del conocimiento sistémica entre diversos niveles. La consolidación de sistemas para la representación del conocimiento conlleva la asignación de formalismos para el desarrollo de contenidos lingüísticos en un proceso de informatización de los mismos.

El proceso de diseño de un modelo conceptualizado para la traducción automática necesita una fase de asimilación de esquemas lógico-conceptuales fundamentales para cualquier estrategia necesaria en el procesamiento del lenguaje natural y la descripción de contrastes presentes en las condiciones que sustentan reglas de inferencia desde y hacia el contexto con el fin de lograr su sistema macroestructural [7].

La asimilación de las características semánticas y enlaces sintácticos unidos a las estructuras cognitivas establecidas para los algoritmos que identifican equivalencias entre palabras, disponen de elementos y características propias de los sistemas expertos.

La delimitación de conocimiento es una tarea primordial para una adecuada integración de las fases de generación hacia la de traducción y para que la adquisición del conocimiento automático contribuya a un mayor acercamiento al discurso. Una representación del significado mediante conjuntos lingüísticos que interactuen con una representación generalizada que tenga como objetivo principal un conocimiento progresivo del corpus y de las partes del discurso, con la finalidad de simplificar en la medida de lo posible una superposición innecesaria de conocimiento.

El objetivo que persigue la traducción automática en relación con el procesamiento del lenguaje natural es conseguir un conocimiento profundo de la estructura lingüística de las frases y un conocimiento real de las palabras [8].

Los progresos que se pueden realizar desde los estudios de alineación de palabras en dominios concretos aumentan la probabilidad de realizar traducciones razonables. Las posibles interpretaciones de un análisis que vaya desde el significado al texto, en un entorno más amplio, el de la adquisición de conocimiento automático, plantea el uso de modelos conceptuales y léxicos junto a reglas de combinación para una incorporación adecuada de conocimiento.

El paradigma conceptual de los sistemas de traducción automática, basados en el conocimiento, implica una aproximación a una descripción del lenguaje para representar el significado del texto con reglas que emulan el comportamiento cognitivo en tratamientos automatizados, desde la selección léxica en un amplio estudio del conocimiento de las estructuras textuales. Las fuentes de conocimiento deben ser particularmente revisadas en los procesos de adquisición automática de conocimiento, sobre todo en la etapa de generación y determinación de la funcionalidad del texto [9]. Las diversas realizaciones que se pueden atribuir a un significado favorecen una gran diversidad de dominios y significados del discurso.

Las reglas de planificación del texto contribuyen de otra forma al proceso de adquisición de conocimiento mediante el análisis regular de algunos significados y reglas para el análisis de la lengua fuente [10]. La interpretación semántica del texto y la formulación de una sintaxis contribuyen al diseño de un formalismo con capacidad suficiente para la adquisición del conocimiento. El análisis de la significación pragmática dibuja la dimensión de la situación del discurso.

Todos los aspectos relativos a un estudio detallado del léxico y reglas de procesamiento diversas, desde las centradas en la planificación del texto hasta el uso de diccionarios informatizados, son algunos de los retos más interesantes de los enfoques basados en el conocimiento para un procesamiento del lenguaje natural.

La adquisición de conocimiento en sistemas de traducción automática es una de las metas más prometedoras en la investigación lingüística, lo que supondrá además incursiones en el campo de la inteligencia artificial y los sistemas expertos. La orientación metodológica que guía esta nueva generación da lugar a un momento trascendental en los estudios sobre traducción automática. El estudio del proceso de traducción exige una revisión de las correspondencias en una comparación de las estructuras de los textos fuente hacia la lengua objeto.

El desarrollo de sistemas para el procesamiento del lenguaje natural que permitan identificar la función y categoría que una palabra desempeña como unidad lingüística en un discurso, ha supuesto un gran impulso para los distintos tipos de sistemas destinados a cumplir esta finalidad. Los enfoques estadísticos han sido usados para la generación de gramáticas y analizadores. Una gramática generada de forma automática se adapta de forma más sencilla a los cambios y su revisión es más ágil. El sistema propone las reglas gramaticales necesarias como resultado de un examen minucioso del corpus a partir de las frases más cortas. El control de las reglas se realiza por medio de un conjunto de condiciones [11].

Los enfoques estadísticos también permiten determinar la distribución de palabras y frases en el texto. La identificación de secuencias de palabras susceptibles de ser tratadas como frases contribuye a una recuperación de la información más pertinente. Un proceso de

evaluación de sistemas para el procesamiento del lenguaje natural implica trasladar las estructuras del texto resultado del tratamiento lingüístico con respecto al texto fuente. Es importante considerar en qué medida el sistema satisface las premisas para un procesamiento correcto en una doble vertiente: de los datos respecto del proceso y de las estructuras respecto de un análisis de contenido.

Las técnicas documentales contribuyen al desarrollo de sistemas eficaces para el procesamiento del lenguaje natural, como la integración de una teoría que contemple los factores que intervienen en la composición de un texto, aspectos semántico-léxicos y tratamiento de documentos pertinentes.

Los enfoques cognitivos en sistemas de traducción automática han planteado la asimilación de estructuras en investigación sobre recuperación de información, en paralelismo con estructuras cognitivas de organización [12] que implican en traducción automática los procesos de captación y entendimiento del conocimiento sobre el dominio cubierto por el texto y los factores de tipo lingüístico.

Las técnicas actuales para el procesamiento del lenguaje natural hacen uso de estrategias apropiadas de indexación y recuperación constituidas sobre un análisis lingüístico preciso. Este aspecto implica en traducción automática el estudio de términos compuestos y unidades descriptivas similares [13].

Las relaciones de dependencia entre términos para una representación del significado destinada a un procesamiento del lenguaje natural, supone la utilización de un lenguaje de indización que reduzca el universo del significado del término debido a la presencia de relaciones implícitas o explícitas entre los mismos.

Un proceso de recuperación de la información que suscite la determinación de estructuras conceptuales representativas de procesos lingüísticos puede plantear especificar el tipo de analizadores utilizados, reglas de procesamiento para determinar la competencia lingüística en un estudio de las categorías sintácticas, especialmente en sistemas de traducción automática para el procesamiento de diversas lenguas y que además están basados en la transferencia [14].

La recuperación del conocimiento va a estar condicionada por las propiedades distintivas que se consideran en una indización parcial de los componentes del lenguaje. Hay que proporcionar una mayor consistencia a los sistemas de traducción automática que parten de una adquisición automática del conocimiento, que precisan de unas reglas de formación basadas en contrastes durante la etapa de selección de equivalencias léxicas.

La transferencia de un nivel a otro cumple la aceptación de condiciones como enunciados de destino, especialmente en la fase de síntesis una vez confirmada la aceptabilidad de las relaciones establecidas. Se pretende superar los contrastes durante la generación del texto en la lengua objeto teniendo en cuenta que existen series de representaciones de múltiples niveles para procesos de transferencia léxica simple o en representaciones a un único nivel. La simplificación de reglas en las representaciones complejas de niveles múltiples incrementa la reducción de componentes gramaticales a un único estadio.

Las diferencias que se derivan de los formalismos supone unificar los elementos que se basan en estructuras lingüísticas con respecto a los que mantienen una relación exclusiva con los componentes semántico-léxicos. El desarrollo de un enfoque básico que consolide el sistema de traducción en su proceso de adaptación a las distinciones procedentes de las propiedades semántico-léxicas, es retomar una orientación que facilite el diseño de sistemas que permitan, sobre un estudio de las categorías, una aproximación a los modelos algorítmicos de sintaxis. La formalización puede contribuir a asociar parámetros de evaluación fundamentados sobre un mecanismo que determine la relevancia de los procesos asociados a la transferencia [15].

Existen múltiples formas de selección léxica en el entorno de un análisis conceptual dirigido al proceso de generación con intención de contribuir a la realización de frases en la lengua objeto o de destino. La consolidación de este proceso es una de las técnicas más interesantes en una teoría para el procesamiento del lenguaje natural.

Las futuras orientaciones pretenden eliminar las reglas de transferencia que median entre los diferentes niveles. El proceso de adquisición automática de conocimiento debe perfeccionar la información semántica y sintáctica en un ámbito superior que cubra los argumentos de representaciones consensuadas en la automatización inicial de estructuras conceptuales.

La formación de métodos estadísticos ha fundamentado el proceso de evaluación numérica como medida de la probabilidad basada en el número de veces que una palabra concurre, de forma que su análisis pormenorizado va a permitir cuantificar la operatividad del corpus en relación con el tratamiento inicial. Los componentes de un modelo estadístico proponen una concordancia entre las lenguas, como una herramienta para equilibrar el proceso de extracción automática de conocimiento y como un sistema de representación de aspectos multilingües en las tecnologías orientadas al aprendizaje del lenguaje.

Los sistemas de traducción automática basados en un proceso de adquisición del conocimiento se han consolidado en áreas de actuación propias de la Lingüística Computacional, es una etapa de adquisición de conocimiento léxico que se puede organizar sobre la determinación de técnicas estadísticas, un estudio sintáctico profundo, un análisis alternativo basado en los principios y en las iniciativas de las teorías que proponen la unificación y la aplicación de métodos exploratorios para la obtención de datos procedentes de las bases del conocimiento, importante confluencia con los sistemas expertos.

La recuperación de la información como técnica de exploración de modelos cognitivos ha permitido llevar a cabo estudios sobre procesamiento eficiente de funciones lingüísticas y relaciones de dominio proporcionales [16]. Las técnicas de representación textual necesarias en un proceso de extracción de significado para un adecuado procesamiento del lenguaje natural, han llevado a algunos expertos a analizar los procedimientos presentes en una teoría para la representación textual vinculada a la recuperación de documentos [17].

Conclusiones.

Coincidimos con las propuestas de Gómez Guinovart [18] en cuanto a las diversas líneas de investigación en sistemas de traducción automática, en el contexto de los trabajos en lingüística informática, aspecto que ha dado lugar a la reciente aparición de estudios en lingüística de corpus con aplicaciones en áreas de conocimiento muy diversas sobre todo en procesamiento del lenguaje natural .

El análisis lingüístico en el futuro puede precisar de estructuras categoriales para una indización pertinente de las condiciones lógicas sobre una red funcional, como punto de confluencia entre la vertiente artificial de las redes neuronales y las tendencias actuales en conectismo.

Es importante el desarrollo de sistemas que parten de un análisis lingüístico entre varias lenguas como medida de la distribución de los elementos léxicos, determinación de las condiciones sintácticas, especificación de las reglas de formación y representación de las condiciones y contrastes aceptados en el nivel procedente. El descender al estudio detallado de las categorías gramaticales es prever su integración en estructuras sintácticas asociadas a reglas lingüísticas con mecanismos que faciliten el reconocimiento de las estructuras superficiales y aceptación de las relaciones de dependencia, en particular cuando se trata de sistemas de traducción automática basados en el conocimiento.

La evaluación de los procesos de traducción automática es un aspecto fundamental en una aceptación de los enfoques que cubren necesidades específicas de sistemas informatizados integrados en la formación de un corpus para un posterior tratamiento. Con este propósito se mejora la selección de equivalencias durante la transferencia léxica facilitando el análisis sintáctico y semántico.

Las estructuras divergentes contribuyen a la formalización de categorías sobre la base de una relación sistemática entre los componentes sintácticos y aspectos semántico-léxicos. En realidad, las propiedades más importantes del sistema son la resolución de las divergencias y la combinación de estructuras desde una perspectiva general orientada hacia la traducción automática, proporcionando consistencia en la descripción del conocimiento como fuente básica en los procesos simplificados. La consolidación necesaria para una representación adecuada del conocimiento lingüístico supone aceptar algunas interdependencias en un dominio infomativo del que forman parte los valores característicos de los argumentos y predicados[19] expresados para un análisis entre varias lenguas. El estudio de los argumentos y predicados ha sustentado la base de una teoría para la transferencia en Eurotra, formada por cuatro niveles básicos de unificación en una representación de los diversos módulos contemplados en el proceso de estratificación.

Referencias.

[1] Díez Carrera, C. *Las industrias de la lengua: panorámica para los gestores de la información*. Madrid: Biblioteca Nacional, 1994.

[2] Church, K. W.; Mercer, R.L. Introduction to the special issue on Computational Linguistics using large corpora. *Computational Linguistics*, 19(1), 1993, 1-24.

[3] Lewis, D. Machine translation today: a critical look at current desktop systems. *The Linguist*, 37(2), 1998, 38-43.

[4] Hutchins, J. A new era in machine translation research. *Aslib Proceedings*, 47(10),1995, 211-219.

[5] *Ibid.*,p.211-212.

[6] *Ibid.*,p. 213-215.

[7] Moreiro González, J.A. Implicaciones documentales en el procesamiento del lenguaje natural. *Ciencias de la Información*, 24(1), 1993, 48-54.

[8] Knight, K. Automating knowledge acquisition for machine translation. *Artificial Intelligence Magazine*, 18(4), Winter 1997, 81-96.

[9] *Ibid.*, p.92-93.

[10] Defrise, C. The treatment of discourse in knowledge-based machine translation. En: Ramm, W.,ed. *Text and context in machine translation: aspects of discourse representation and discourse processing*. Luxembourg: Office for Official Publications of the EC, 1994, 53-75. (Studies in MT and NLP; vol. 6).

[11] Haas, S.W. Natural language processing: toward large-scale, robust systems. *Annual Review of Information Science and Technology*, 31, 1996, 83-119.

[12] Se trata del concepto "organizing cognitive structures". Véase Sugar, W. User-Centered perspective of IR research. *Annual Review of Information Science and Technology*, 30, 1995, 77-109.

[13] Lewis, D.D.; Sparck Jones, K. Natural language processing for information retrieval. En: Gilchrist, A., ed. *From classification to knowledge organization: Dorking revisited or "past is prelude"*. The Hague: International Federation for Information and Documentation, 1997, 49-61. (FID occasional paper; 14).

[14] Riabtseva, N.K. Metadiscourse collocations in scientific texts and translations problems. Conceptual analysis. *Terminologie et Traduction*, 2-3, 1992, 375-385.

[15] Dorr, B.J. Machine translation divergences: a formal description and proposed solution. *Computational Linguistics*, 20(4), December 1994, 598-633.

[16] Sugar, W.: *op. cit.*, p.80-81.

[17] Lewis, D.L.; Sparck Jones, K.: *op.cit.*

[18] Gómez Guinovart, J. Fundamentos de Lingüística Computacional: bases teóricas, líneas de investigación y aplicaciones. En: Baró i Queralt, J.; Cid Leal, P., ed. *Anuari SOCADI de Documentació i Informació 1998*. Barcelona: Societat Catalana de Documentació i Informació, 1998, 135-146.

[19] El estudio de estos componentes ha sido ampliamente desarrollado por equipos de investigación especializados, en un análisis lingüístico de gran profundidad. Véase Allegranza, V., et al. Linguistic for machine translation: the Eurotra linguistics specifications. En: Copeland, Ch.,et al.,ed. *The Eurotra linguistics specifications*. Luxembourg: Office for Official Publications of the EC, 1991, 15-124. (Studies in MT and NLP; vol.1).