Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 443-444 recibido 20-05-2005; aceptado 01-06-2005

El proyecto METIS-II

Toni Badia, Gemma Boleda, Maite Melero, Antoni Oliver
GLiCom, Universitat Pompeu Fabra
Passeig de Circumval·lació, 8 08003 Barcelona
{toni.badia,gemma.boleda,maite.melero,antonio.oliver}@upf.edu

Resumen: Presentamos el proyecto METIS-II, dirigido a la creación de un sistema de traducción automático estadístico que utiliza como recursos principales un corpus monolingüe (de la lengua destino) y un diccionario bilingüe, eliminando la necesidad de disponer de un corpus paralelo para entrenar el sistema.
Palabras clave: traducción automática estadística

Abstract: We present the METIS-II project, aimed at creating a Statistical Machine Translation system which uses only a monolingual corpus of the target language and a bilingual dictionary, thus eliminating the need for parallel corpora to train the system.
Keywords: statistical machine translation

1. Motivación

Las técnicas actuales de traducción automáticas basadas en corpus se basan en corpus paralelos bilingües para entrenar los sistemas de traducción automática. Dentro de estas técnicas hay dos aproximaciones principales: por un lado, la traducción automática estadística, que se basa en la teoría de la probabilidad (Yamada y Knight, 2001). Por otro, la traducción automática basada en ejemplos, inspirada en el razonamiento analógico: las traducciones se computan en analogía a un conjunto de traducciones extraídas de corpus bilingüe (Carl y Way, 2003).

La necesidad de disponer de corpus paralelos adecuados es un inconveniente común a todos los sistemas de traducción automática estadística actuales, pues son difíciles de obtener incluso para los pares de lenguas mayoritarias. Además, los corpus existentes suelen estar limitados a un dominio determinado, como es el caso de EUROPARL (Koehn, 2002), que contiene las actas del Parlamento Europeo desde 1997 hasta 2003 en once lenguas europeas. En cambio, hay corpus monolingües de gran tamaño para un número muy elevado de lenguas.

2. El proyecto METIS-II

El proyecto europeo METIS-II (IST-FP6-003768) responde a la situación que acabamos de describir, y se propone construir un sistema de traducción automática basado únicamente en un diccionario bilingüe y un corpus monolingüe de la lengua destino. Este proyecto tuvo como antecesor METIS-I (Dologlou et al., 2003), en el que se implementó un prototipo para griego-inglés que operaba sólo a nivel de la oración. En METIS-II se ampliará el sistema para que opere con unidades más pequeñas que la oración, se incluirán más lenguas (neerlandés, alemán, español), y se integrará el sistema con herramientas de post-edición, de manera que se pueda adaptar a entornos reales de uso de traducción automática.

Los participantes en el proyecto son cuatro grupos de investigación, cada uno responsable de una lengua: ILSP (Atenas) para griego, KUL (Lovaina) para neerlandés, IAI (Saarbrücken) para alemán, y GLiCom-UPF (Barcelona) para español. Actualmente el proyecto está en su fase inicial (fecha de inicio: 1 octubre 2004, duración: 3 años). Se ha definido las especificaciones generales del sistema a partir de un estudio con usuarios potenciales, así como la arquitectura, representada en la figura 1. Se está desarrollando asimismo el sistema de traducción propiamente dicho (la parte etiquetada como *núcleo del sistema* en la figura 1). Actualmente cada grupo está experimentando con un prototipo de sistema que será evaluado en verano del 2005.

3. Aproximación actual en GLiCom

El grupo GLiCom está estudiando una estrategia basada en n-gramas. En primer lugar, creamos un modelo de lenguaje a partir

ISSN: 1135-5948

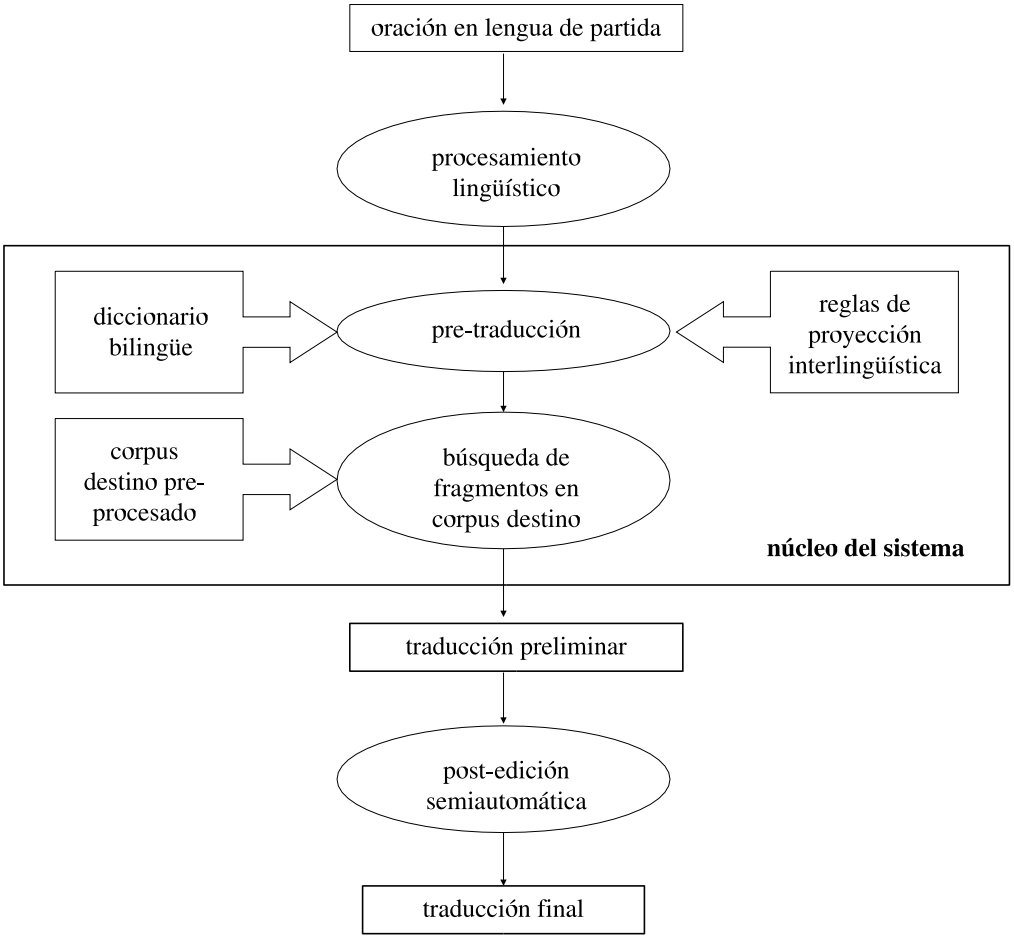

Figura 1: Arquitectura de METIS-II

de un corpus destino lematizado y etiquetado (*British National Corpus*), extrayendo los n-gramas (tanto los n-gramas de lemas como los que contienen combinaciones de lemas y etiquetas). Las oraciones de entrada en español se lematizan y etiquetan y se buscan las traducciones correspondientes a cada lema en el diccionario bilingüe.

Se procede a traducir, empezando por el valor más alto de n (por ahora, 4), de izquierda a derecha. Para tratar el problema del orden de palabras, los n-gramas traducidos se consideran conjuntos (o *bolsas*) y no secuencias ordenadas de palabras. A partir de la frecuencia de cada n-grama traducido en el corpus monolingüe, se asigna una probabilidad a cada candidato. En caso de que no se encuentren traducciones para un n-grama determinado, se recurre a las etiquetas morfosintácticas en vez de a los lemas.

Cuando se han hecho todos los cálculos para un determinado valor de n, se traducen todos los fragmentos de la frase (empezando por el n-grama con mayor probabilidad) y se recalculan los n-gramas de los fragmentos no traducidos, para empezar de nuevo el proceso con $n = n-1$. Esta aproximación está en fase de implementación.

Bibliografía

Carl, M. y A. Way. 2003. *Recent Advances in Example-Based Machine Translation.* Kluwer Academic Publishers.

Dologlou, Y., S. Markantonatou, G. Tambouratzis, O. Yannoutsou, A. Fourla, y N. Ioannou. 2003. Using monolingual corpora for statistical machine translation: The METIS system. En *Proceedings of EAMT-CLAW 03: Controlled Language Translation*, páginas 61–68.

Koehn, P. 2002. Europarl: a multilingual corpus for evaluation of machine translation. Draft.

Yamada, K. y K. Knight. 2001. A syntax-based statistical translation model. En *Proceedings of the 39th ACL*, páginas 5–11.