View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repositorio Institucional de la Universidad de Alicante

Procesamiento del Lenguaje Natural, Revista nº 42, marzo de 2009, pp. 123-124 recibido 28-11-08, aceptado 02-03-09

Empleo de métodos no supervisados basados en corpus para construir traductores automáticos basados en reglas*

Using unsupervised corpus-based methods to build rule-based machine translation systems

Felipe Sánchez-Martínez
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant. E-03071, Alacant, Spain
fsanchez@dlsi.ua.es

Resumen: Tesis doctoral en Informática realizada en la Universitat d'Alacant por Felipe Sánchez Martínez bajo la dirección de los doctores Juan Antonio Pérez Ortiz y Mikel L. Forcada. La defensa de la tesis tuvo lugar el 30 de junio de 2008 ante el tribunal formado por los doctores Rafael C. Carrasco (Univ. d'Alacant), Lluís Padró y Lluís Màrquez (Univ. Politècnica de Catalunya), Harold Somers (Univ. of Manchester) y Andy Way (Dublin City Univ.). La calificación obtenida fue Sobresaliente *Cum Laude* por unanimidad, con mención de Doctor Europeo.
Palabras clave: Traducción automática, desambiguación léxica categorial, inferencia de reglas de transferencia, modelado del lenguaje.

Abstract: PhD thesis in Computer Engineering written by Felipe Sánchez-Martínez at Universitat d'Alacant under the joint supervision of Dr. Juan Antonio Pérez-Ortiz and Dr. Mikel L. Forcada. Author was examined on June 30^{th}, 2008 by the committee formed by Dr. Rafael C. Carrasco (Univ. d'Alacant), Dr. Lluís Padró and Dr. Lluís Màrquez (Univ. Politècnica de Catalunya), Dr. Harold Somers (Univ. of Manchester) and Dr. Andy Way (Dublin City Univ.). The grade obtained was *Sobresaliente Cum Laude* (highest mark), with the European Doctor mention.
Keywords: Machine translation, part-of-speech tagging, language modeling, transfer rules inference.

1. Introducción

Recientemente los enfoques basados en corpus para el desarrollo de sistemas de traducción automática (TA) han visto incrementada la atención recibida; sin embargo, los sistemas de TA basados en reglas siguen siendo desarrollados dado que no todos los pares de lenguas para los cuales existe demanda tienen a su disposición la gran cantidad de textos paralelos necesarios para entrenar sistemas de TA de propósito general basados en corpus; y también porque los sistemas basados en reglas son más fácilmente diagnosticables y los errores que producen suelen tener una naturaleza más repetitiva y previsible, lo cual ayuda a los profesionales que tienen que corregir su salida.

Esta tesis se centra en el desarrollo de sistemas de TA basados en reglas y más concretamente en sistemas de TA por transferencia estructural superficial (Hutchins y Somers, 1992) para la traducción entre lenguas emparentadas.

De todos los recursos que son necesarios para construir un sistema de TA por transferencia (estructural) superficial esta tesis se centra en la obtención de forma no supervisada, a partir de corpus, de:

- los desambiguadores léxicos categoriales empleados para resolver la ambigüedad léxica de los textos a traducir, y
- el conjunto de reglas de transferencia que se emplean para adecuar la traducción a la reglas gramaticales de la lengua meta.

2. Desambiguadores léxicos categoriales para TA

En TA, la correcta elección de la categoría léxica de las palabras a traducir es crucial dado que la traducción de una palabra en lengua origen (LO) a la lengua meta (LM) puede diferir de una categoría léxica a otra.

* Tesis financiada por el Ministerio de Educación y Ciencia y el Fondo Social Europeo a través de la ayuda a la investigación BES-2004-4711.

 © 2009 Sociedad Española para el Procesamiento del Lenguaje Natural

De entre los diferentes enfoques existentes para la obtención de desambiguadores léxicos categoriales, esta tesis se centra en el desarrollo de desambiguadores léxicos categoriales basados en modelos ocultos de Markov (MOM) (Cutting et al., 1992). Éstos pueden entrenarse de forma *supervisada* mediante el empleo de textos desambiguados (o etiquetados) a mano, o bien de forma *no supervisada* mediante el uso del algoritmo de Baum y Welch con texto no etiquetado. Estos métodos sólo emplean información de la lengua que pretenden desambiguar. Sin embargo, cuando el desambiguador léxico categorial resultante se integra en un sistema de TA hay que tener en consideración:

- que un modelo estadístico de la LM puede utilizarse de forma no supervisada para obtener mejores desambiguadores léxicos categoriales, y
- que en TA lo que realmente importa es la calidad final de la traducción, no la precisión del desambiguador.

Se propone un nuevo método, inspirado en los dos hechos arriba mencionados, para el entrenamiento de desambiguadores léxicos categoriales de la LO basados en MOM, mediante el empleo de información de la LM, así como del resto de módulos del sistema de TA en el que el desambiguador se integra. Los experimentos realizados con tres pares de lenguas de Apertium (`http://www.apertium.org`) muestran que el sistema de TA ofrece mejores resultado cuando el desambiguador léxico categorial es entrenado usando este nuevo método que cuando es entrenado con el algoritmo de Baum y Welch.

3. Inferencia automática de reglas de transferencia estructural

Esta tesis también propone un método no supervisado para la inferencia de reglas de transferencia estructural superficial. Esta reglas se basan en plantillas de alineamiento (Och y Ney, 2004) como las usadas en TA estadística. Para su empleo en sistemas de TA basados en reglas las plantillas de alineamiento han tenido que ser adaptadas y extendidas con un conjunto de restricciones que controlan su aplicación como reglas de transferencia.

Una vez obtenidas, las plantillas de alineamiento son filtradas atendiendo a su frecuencia de aparición en la colección de textos paralelos. Finalmente las plantillas de alineamiento seleccionadas se emplean para la generación de reglas de transferencia en el formato usado por el ingenio de TA Apertium.

Para evaluar las reglas inferidas se han realizado experimentos con tres pares de lenguas de Apertium. Las reglas inferidas ofrecen mejores resultados que la traducción palabra por palabra, y resultados próximos a los obtenidos cuando las reglas de transferencia son codificadas a mano por lingüistas.

En cuanto a la cantidad de corpus paralelos necesarios para obtener un conjunto de reglas de transferencia que proporcionen una calidad de traducción aceptable, los experimentos realizados con distintos tamaños de corpus demuestran que con un corpus de medio millón de palabras la calidad de las reglas inferidas es satisfactoria, incluso para algunos pares de lenguas la calidad es similar a la obtenida cuando las reglas de transferencia se obtiene a partir de un corpus de entrenamiento de dos millones de palabras.

Información adicional

Los métodos descritos en esta tesis han sido liberados como código abierto y pueden descargarse desde `http://sf.net/projects/apertium/`; paquetes `apertium-tagger-training-tools` y `apertium-transfer-tools`. Estos paquetes se integran perfectamente en el proceso de desarrollo de nuevos pares de lenguas para Apertium. La tesis está disponible en `http://www.dlsi.ua.es/~fsanchez/pub/thesis/thesis.pdf`.

Bibliografía

Cutting, D., J. Kupiec, J. Pedersen, y P. Sibun. 1992. A practical part-of-speech tagger. En *Proceedings of the Third Conference on Applied Natural Language Processing*, páginas 133–140.

Hutchins, W. J. y H. L. Somers. 1992. *An Introduction to Machine Translation*. Academic Press.

Och, F. J. y H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.