Área: Inteligencia Artificial

La traducción automática estadística usando paráfrasis de traducción

Statistical Machine Translation Using Translation Paraphrases

por Francisco Guzmán y Leonardo Garrido

RESUMEN

En este trabajo de investigación proponemos una metodología para mejorar la calidad de los sistemas de traducción automática cuando existe escasez de recursos, utilizando paráfrasis de traducción. El propósito es estimar qué tantas mejoras se pueden hacer a un sistema de traducción utilizando paráfrasis de traducción para traducir entre inglés y español empleando el francés como lengua intermedia.

Palabras claves: Traducción automática, Sistemas Inteligentes, Métodos estadísticos, Inteligencia Artificial.

ABSTRACT

In this research, we propose a methodology for enhancing the quality of statistical machine translation systems when data scarcity is present by using translation paraphrases. The purpose of the research presented in this document is to find out how much extra information (i.e. improvements in translation quality) can be found when using Translation Paraphrases (TPs) when translating between English and Spanish using French as an intermediary language.

Key words: Machine Translation, Intelligent Systems, Statistical Methods, Artificial Intelligence

Francisco Guzmán es estudiante del Doctorado en Tecnologías de la Información y Comunicaciones, del Campus Monterrey. Tiene el grado de Ingeniero Físico Industrial (2004) por el mismo instituto. Actualmente se encuentra realizando su estancia de investigación en Carnegie Mellon University. Su correo electrónico es: A00774831@itesm.mx

Leonardo Garrido es Doctor en Inteligencia Artificial, por el Tecnológico de Monterrey, Campus Monterrey (2001). De la misma institución obtuvo los títulos de Maestro en Ciencias de Sistemas Computacionales e Ingeniero en Sistemas Electrónicos. Desde 1993 forma parte del Centro de Sistemas Inteligentes y sus principales intereses de investigación son los agentes autónomos, Sistemas Multiagentes, aprendizaje y razonamiento automático. Su página web es http://homepages.mty.itesm.mx/lgarrido y su correo electrónico es leonardo.garrido@itesm.mx.

Conforme la tecnología ha ido avanzando, nos ha permitido incorporarla cada vez más en nuestras vidas. Así, hemos empezado a cederle a las computadoras ciertas tareas que implican que ésta "entienda" el contenido de la información. Sin lugar a dudas, una de las tareas más difíciles y en la cual ha sido más difícil efectuar avances es la traducción automática o traducción asistida por computadora [1].

La permeabilidad de los traductores en nuestra sociedad ha sido gradual; desde los diccionarios incluidos en las primeras agendas electrónicas que permitían una traducción palabra por palabra, hasta los traductores de los motores de búsqueda como Google o Yahoo!, que permiten traducir páginas de Internet completas. Son innumerables los esfuerzos que a través de las últimas décadas se han realizado para lograr traducir de una lengua a otra. Sin embargo, no fue sino hasta los años 90, que la gran disponibilidad de colecciones de textos bilingües o corpus de traducciones, aunado a los avances de la infraestructura de cómputo, permitieron el surgimiento de un nuevo paradigma en la traducción automática: los métodos estadísticos.

LA TRADUCCIÓN AUTOMÁTICA ESTADÍSTICA Y LA ESCASEZ DE RECURSOS

Se puede visualizar a la traducción estadística como una metodología que se enfoca en el resultado y no en el proceso [2]. Es decir que mientras otros paradigmas de traducción se enfocan en crear reglas para traducir de una lengua a otra, los métodos estadísticos buscan encontrar la mejor traducción bajo dos criterios: la fluidez y la fidelidad. La fluidez nos da una idea de qué tan gramatical es una construcción en la lengua objetivo, y está contenida dentro de lo que se conoce como modelo del lenguaje. El modelo de lenguaje p(e) es la probabilidad de que cierta frase de la lengua objetivo (tradicionalmente ***e*** por el inglés) exista o esté bien formada, de acuerdo a estimaciones hechas. Por otra parte, la fidelidad nos dice qué tanto del significado original se conserva. Ésta es representada en el modelo de traducción; p(f|e) que nos dice cuál es la probabilidad de que una palabra en la lengua origen (tradicionalmente ***f*** por el francés) haya sido generada por una frase en la lengua objetivo. Así, al tratar de maximizar ambos objetivos simultáneamente, tenemos la ecuación característica de la traducción estadística [3]:

$$\hat{e} = \arg\max_{e} p(f|e)\,p(e)$$

Así es que, para obtener la mejor traducción ***ê,*** sólo bastaría buscar entre todas las posibles traducciones aquella que maximice las probabilidades.

Cabe resaltar que aunque los métodos estadísticos han probado su efectividad en comparación con otros métodos de traducción, presentan ciertos problemas. Uno de ellos es la dependencia de los corpus (o grandes colecciones de documentos) de entrenamiento. Recordemos que tanto el modelo de lenguaje como el modelo de traducción son estimados sobre corpus de texto.

LA ESCASEZ Y LOS RECURSOS MULTILINGÜES

La escasez de recursos es una de las mayores limitantes para la construcción de sistemas de traducción. Sobre todo cuando nuestro par de lenguas (origen y objetivo) son lenguas minoritarias. Muchas son las estrategias que se han ido empleando para sobrellevar esta exigencia, una de ellas es el uso de información extraída de otros lenguajes diferentes al origen u objetivo. Por ejemplo Gispert y Mariño [4] usan al español como una lengua intermedia para traducir entre inglés y catalán. En su trabajo, analizan dos diferentes maneras de lograr tal objetivo. Su primer método implica la traducción indirecta. Es decir, hacen uso de dos sistemas de traducción;

Número 84, octubre, 2008, pp. 30-31

uno para traducir de inglés a español y otro para traducir de español a catalán. La segunda alternativa que proponen es usar su sistema de traducción español-catalán, para traducir el lado español de un corpus inglés-español. De tal manera, obtienen un corpus inglés-catalán que les permite crear un sistema de traducción inglés-catalán.

Por otra parte, Callison-Burch et al. [5] usa un método distinto para incrementar la cobertura o cantidad de traducciones conocidas, de su sistema español-inglés. Su trabajo está basado en el uso de paráfrasis. Es decir, si para cierta frase en español no existe traducción conocida, entonces se intenta traducir alguna de las paráfrasis de la frase original.

UNA LENGUA INTERMEDIARIA: PARÁFRASIS DE TRADUCCIÓN

Lo que ambas metodologías tienen en común es que hacen uso de recursos extraídos de lenguas distintas a la origen y objetivo. En el trabajo que presentamos en [6], planteamos una metodología que abona a esta dirección de investigación. En nuestra propuesta hacemos uso de una lengua intermedia para obtener lo que hemos denominado "paráfrasis de traducción". Las paráfrasis de traducción son pares de frases bilingües que comparten significado con otros pares de frases. Por ejemplo el par inglés-francés *(the child wants a candy, l'enfant veut un bonbon)* y el par italiano-español *(Il bambino vuole un dolce, el niño quiere un dulce)* contienen pares de traducciones cuyo significado es similar, aunque las lenguas en las que está expresado este significado sean distintas.

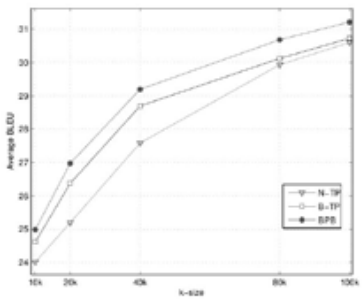
Ahora bien, las paráfrasis de traducción nos sirven para generar nuevos pares de frases en las lenguas origen-objetivo, mediante el uso de una lengua intermediaria. Por ejemplo el par inglés-francés *(the child wants a candy, l'enfant veut un bonbon)* y el par francés-español *(l'enfant veut un bonbon, el niño quiere un dulce)* tienen en común la frase francesa *l'enfant veut un bonbon.* En este caso, la paráfrasis de traducción nos permite generar un nuevo par inglés-español *(the child wants a candy, el niño quiere un dulce)*, obtenido indirectamente a través del francés. Ahora bien, para calcular la probabilidad de que *the child wants a candy* se traduzca en *el niño quiere un dulce,* necesitaríamos hacer una marginalización de probabilidades. La manera en que lo calculamos es:

$$p(f|e) \approx \sum_i p(i|f)p(e|i)$$

Donde p(f|e) es nuestro modelo de traducción, p(f|i) es la probabilidad de que la frase en la lengua intermedia se traduzca en la frase en la lengua origen y p(i|e) es la probabilidad de que la frase en la lengua objetivo se traduzca en la frase en la lengua intermediaria.

EXPERIMENTACIÓN Y RESULTADOS

Para verificar que los pares de frases obtenidos mediante las paráfrasis de traducción fueran útiles para el sistema de traducción, hicimos la comparación de dos sistemas: El primero, es un sistema de traducción que fue entrenado con datos de inglés y español. El segundo sistema fue entrenado con los mismos datos, pero además utilizando las frases extraídas por las paráfrasis de traducción inglés-francés-español. En la siguiente gráfica se puede apreciar el desempeño del sistema. En el eje horizontal se puede ver el tamaño del corpus de entrenamiento utilizado medido en número de líneas. En el eje vertical podemos ver la calificación BLEU [7] que cada sistema obtuvo. La calificación BLEU es una métrica automática de la calidad de la traducción que ha demostrado estar correlacionada con las calificaciones otorgadas por jueces humanos. En la gráfica vemos tres curvas: BPB simboliza el *best-practical bound* o mejor posible, es decir, el promedio de las mejores traducciones para cada caso; N-TP representa el promedio BLEU de las traducciones hechas por el sistema base; y B-TP representa el promedio BLEU de las traducciones hechas por el sistema que utiliza las paráfrasis de traducción.



Gráfica de resultados experimentales.

Como se puede ver, el sistema que utiliza las paráfrasis de traducción se comporta significativamente mejor que el sistema que no las utiliza en todos los casos estudiados. Sin embargo las mejoras son estadísticamente significativas únicamente cuando los datos de entrenamiento son bajos (escasez de entrenamiento: 10k, 20k y 40k). Esto nos sugiere que el uso de paráfrasis de traducción sería benéfico en estos casos, ya que hay más pares de frases que se pueden descubrir.

CONCLUSIONES

La traducción automática usando métodos estadisticos está limitada a situaciones en las cuales la cantidad de recursos para entrenamiento es grande. Sin embargo, muy pocos pares de lenguas poseen tales cantidades de recursos de entrenamiento. La investigación que estamos realizando está enfocada a la mejora de dichos sistemas utilizando pares de frases extraídos mediante paráfrasis de traducción. En este artículo hemos presentado los resultados experimentales de tal metodología comparada con un sistema de base. De nuestro estudio, podemos concluir que las paráfrasis de traducción presentan una alternativa para obtener nuevos pares de frases cuando se presenta un escenario de escasez de recursos. Con nuevos pares de frases, nuestra traducción se hace más robusta, lo cual hace que el sistema sea más competitivo.

REFERENCIAS

[1] Philipp Koehn and Christof Monz, editors. *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, New York City, June 2006.

[2] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice-Hall, 2000.

[3] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Laerty, Robert L. Mercer, and Paul S. Roossin. *A statistical approach to machine translation*. Comput. Linguist., 16(2):79-85, 1990.

[4] Adria de Gispert and Jose B. Marino. *Catalan-English statistical machine translation without parallel corpus: Bridging through Spanish*. In LREC 5th Workshop on Strategies for developing Machine Translation for Minority Languages, 2006.

[5] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. *Improved statistical machine translation using paraphrases*. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pp. 17-24, 2006.

[6] Francisco Guzmán and Leonardo Garrido. *Translation paraphrases in phrase-based machine translation*. In Computational Linguistics and Intelligent Text Processing, Vol. 4919/2008, pp. 388-398. 2008.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. *Bleu: a method for automatic evaluation of machine translation*. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311-318, 2002.