Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 135-138 recibido 14/03/2016 revisado 13/04/2016 aceptado 03/05/2016

Integración de Paradigmas de Traducción Automática (IMTraP)

Integration of Machine Translation Paradigms (IMTraP)

Marta R. Costa-jussà
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
marta.ruiz@upc.edu

Resumen: La Traducción Automática (TA) es un campo altamente interdisciplinar y multidisciplinar porque en él trabajan: ingenieros, informáticos, estadísticos y lingüístas. El objetivo de este proyecto es acercar los diferentes perfiles de la comunidad de la TA para plantear un paradigma integrado de TA que incluya tecnologías lingüísticas y estadísticas. Básicamente, nuestra investigación se centra en el problema de integrar dinámicamente dos de los paradigmas de traducción más populares: el basado en reglas y el estadístico. Una de las principales ideas es usar tecnologías lingüísticas desarrolladas para los sistemas basados en reglas o en el contexto del procesamiento del lenguaje natural. El nuevo paradigma proporcionará soluciones a los retos actuales de la TA como palabras desconocidas, reordenamiento y ambigüedades semánticas. El proyecto se focaliza en tres de las lenguas más hablados en el mundo: Chino, Castellano e Inglés; y todas las combinaciones de traducción entre ellas. Estos pares de lenguas no solo involucran intereses económicos y culturales, sino que además tienen importantes retos de TA como el morfológico, sintáctico y semántico.
Palabras clave: Traducción Automática Híbrida, Morfología, Sintaxis, Semántica, Chino, Castellano

Abstract: Machine Translation (MT) is a highly interdisciplinary and multidisciplinary field approached from the point of view of engineering, computer science, informatics, statistics and linguists. The goal of this research project is to approach the different profiles in the MT community by providing a new integrated MT paradigm which mainly includes linguistic technologies and statistical algorithms. Our research focuses on the problem of dynamically integrating the two most popular MT paradigms: the rule-based and the statistical-based. We will use linguistic technologies developed either for the rule-based MT systems or other natural language processing tasks into statistical MT systems. The new paradigm will provide solutions to current MT challenges such as unknown words, reordering and semantic ambiguities. The project focuses on the three most spoken languages in the world: Chinese, Spanish and English; and all translation combinations among them. These language pairs do not only involve many economic and cultural interests, but they also include some of the most relevant MT challenges such as morphological, syntactic and semantic variations.
Keywords: Hybrid Machine Translation, Morphology, Syntax, Semantics, Chinese, Spanish

1 Introducción

La Traducción Automática (TA) es un campo interdisciplinario y multidisciplinario, que incluye profesionales como: traductores, ingenieros, informáticos, matemáticos y lingüístas. Pero la cooperación y la interacción entre ellos es todavía baja. El objetivo de este proyecto es proponer y validar un paradigma de TA completamente nuevo, capaz combinar de manera eficiente el conocimiento lingüístico con los recursos y algoritmos estadísticos. Se pretende combinar las arquitecturas de la TA basada en reglas y la estadística. El resultado del proyecto será un sistema de TA de tecnología híbrida más allá del estado del arte. El sistema resultante deberá ser, en la medida de lo posible, independiente de la lengua y de código abierto.

ISSN 1135-5948

2 Objetivos

Los objetivos del proyecto se resumen de la siguiente manera:

- Desarrollar aproximaciones para integrar la información estructural y lingüística (morfológica, sintáctica y semántica) en TA estadística y formular una nueva arquitectura híbrida.
- Integrar las comunidades de TA (especialmente, lingüístas, informáticos e ingenieros) con el fin de resolver los problemas más comunes de TA incluyendo la morfología, la sintaxis y la semántica.
- Analizar y comparar en detalle la estructura y el funcionamiento de los sistemas basados en reglas y los estadísticos.
- Desarrollar un sistema híbrido de TA que sea entrenable (en la medida de lo posible) en cualquier par de lenguas y de código abierto.

3 Resultados de la primera fase del proyecto

Durante la primera fase (12 meses) del proyecto IMTraP, se han logrado los objetivos principales que se detallan a continuación:

- Revisión de los trabajos previos relacionados y revisión exhaustiva de la TA basada en reglas, que incluye la descripción del estado del arte sobre TA híbrida teniendo en cuenta los diferentes niveles lingüísticos: ortografía, morfología, léxico, semántica y sintaxis (Costa-jussà y Farrús, 2014).
- Definición de los sistemas de TA y corpus experimental, que incluye la recopilación de corpus para el chino-castellano e inglés-castellano y experimentos realizados con TA estadística (Costa-jussà, Henríquez Q, y Banchs, 2012).
- Análisis detallado y comparación de los sistemas basados en reglas y estadística, que incluye: el desarrollo del sistema basado en reglas chino-a-castellano construido en el marco de código abierto Apertium[1]; la descripción y análisis de los sistemas basados en reglas y estadísticos; y propuesta de varias arquitecturas híbridas a estudiar más a fondo en la segunda fase del proyecto (Costa-jussà y Centelles, 2015).

Asimismo, en términos de desarrollo de sistemas de TA, en el marco del proyecto IMTraP:

- Se ha construido con éxito los sistemas de TA estadística de referencia para el chino-castellano e inglés-castellano utilizando los datos recogidos y los parámetros sintonizados (Costa-jussà, Henríquez Q, y Banchs, 2012).
- Se ha desarrollado el primer sistema de código abierto basado en reglas chino–castellano, que ha sido construido utilizando técnicas híbridas mediante la combinación de los conocimientos humanos y las técnicas estadísticas. En particular, el conocimiento humano se ha utilizado para los diccionarios monolingües y bilingües así como para la definición de reglas de transferencia estructural. El conocimiento estadístico ha complementado todos los pasos mencionados. Además, el conocimiento estadístico ha sido la única fuente para las reglas de transferencia léxicas. La mejora del conocimiento estadístico en la TA basada en reglas se ha evaluado y se ha demostrado que proporcionan mejoras notables en la salida de la traducción. En este sentido, se han mostrado eficaces técnicas de construcción de un sistema basado en reglas usando técnicas híbridas. Por otra parte, el sistema basado en reglas supera el sistema estadístico en los experimentos fuera del dominio. El nuevo sistema basado en reglas, así como las metodologías para su construcción se han evaluado de forma automática y con un análisis manual. Por otra parte, la salida de la última versión del sistema basado en reglas ha sido contrastada en esos términos con un sistema estadístico estado del arte y usando un texto fuera del dominio. El sistema basado en reglas supera el sistema estadístico a todos los niveles lingüísticos excepto a nivel sintáctico. Hay una gran mejora en términos de cobertura léxica (Costa-jussà y Centelles, 2015).
- Se ha construido un traductor chino-castellano que está disponible como ser-

[1] https://www.apertium.org/

vicio web[2] y como aplicación en Android (Costa-jussà, Centelles, y Banchs, 2014).

4 Nuevas perspectivas en híbridización

Relacionado con el estado de arte en TA estadística con información lingüística, podemos argumentar que la investigación en el campo de la TA hace que el concepto de hibridación no sea estático (Costa-jussà, 2015a).

- Por un lado, los sistemas híbridos son vistos como una combinación de los sistemas estadísticos con sistemas basados en reglas (sentido estricto).
- Por otra parte, existe un creciente interés en la combinación de los conocimientos lingüísticos en todas sus formas (por ejemplo, morfológico, sintáctico y semántico) en los sistemas estadísticos existentes (sentido amplio de la hibridización). Algunos de los problemas encontrados en TA, específicamente en TA basada en segmentos, han sido superados por la incorporación de técnicas que usan conocimiento morfológico (Toutanova, Suzuki, y Ruopp, 2008), sintáctico (Khalilov y Fonollosa, 2011) y semántico (Banchs y Costa-jussà, 2011). El rendimiento de los sistemas de TA puede ser claramente mejorado mediante el uso de tales conocimientos lingüísticos. Sin embargo, la TA todavía no es capaz de cubrir correctamente todas las variedades problemáticas. Alternativamente, en lugar de ser general, cada extensión a TA tiende a centrarse en un desafío particular para lograr la mejora deseada.

5 Planteamiento de arquitecturas híbridas

Existe un largo camino por recorrer hacia una arquitectura con un mayor nivel de hibridación/integración de paradigmas.

5.1 En sentido estricto

Hemos identificado estrategias, interesantes para la comunidad, de construir una arquitectura híbrida dado un sistema basado en reglas y uno estadístico (Costa-jussà, 2015a):

- A partir de un sistema basado en reglas, existe la necesidad de extraer reglas de transferencia de corpus paralelo. Esto permitiría a la construcción de sistemas basados en reglas por un lingüista monolingüe. Por el momento, los sistemas basados en reglas tienen que ser desarrollados por lingüistas nativos bilingües o por lo menos la gente que son competentes en el idioma de origen y destino. Trabajos en el tema incluyen (Sánchez-Cartagena, Pérez-Ortiz, y Sánchez-Martínez, 2015) y se pueden tomar como punto de partida.
- Con el fin de mejorar los sistemas basados en reglas ser más fluido y natural, sería bueno integrar un modelo de lenguaje en la etapa de generación. El modelo de lenguaje puede ser a base de n-gramas com se propone en (Labaka et al., 2014), o la formación basada en el neuronal basada en la sintaxis. En cada caso, se requere una decodificación diferente para ser integrado en el sistema.
- Comenzando con el núcleo de un sistema basado en la estadística, existe la necesidad de integrar reglas de transferencia y acoplarlas en el modelo de traducción. Se pueden añadir reglas de transferencia jerárquicas a los sistemas estadísticos, asegurándose de que se adopta algún formalismo gramatical compatible. En esta línea de investigación, es importante identificar la mejor manera de hacer que la integración: dé prioridad a las reglas de transferencia sobre los estadísticos o extraiga una puntuación para hacer competir la transferencia de reglas en igualdad de condiciones a los estadísticos. En esta línea también se encuentran trabajos precedentes a los que se podría dar continuidad (Labaka et al., 2014).

5.2 En sentido amplio

Además, si ampliamos el alcance del concepto de hibridización como hemos visto en la sección 4, otros enfoques para mejorar los sistemas estadísticos a nivel de la morfología puede realizar la traducción en dos etapas, haciendo más amplia post-edición automática morfológica (Costa-jussà, 2015b). A nivel de la sintaxis, la gramática otro formalismo puede ser experimentado, así como varias extensiones de los sistemas jerárquicos. Por último, en el campo de la semántica, las proyecciones del vector en el espacio, las reducciones de espacio y redes neuronales abre una nueva vía.

[2] http://www.chispa.me

El proyecto IMTraP, en su segunda fase, se está focalizando en estas últimas direcciones (Gupta et al., 2016; Costa-Jussà y Fonollosa, 2016).

Estas son líneas de investigación alentadoras que pueden dar lugar a un acoplamiento más natural de la lingüística y la estadística. Hay muchas preguntas que quedan por resolver incluyendo la forma correcta de aplicación de la investigación mencionada. Sin embargo, las perspectivas son prometedoras y, sin duda, los grandes avances en TA se pueden derivar de colaboraciones multidisciplinares. Al final, esto es de lo que la hibridación trata.

6 Detalles del proyecto

El proyecto IMTraP ha sido financiado por el *Seventh Framework Program of the European Commission* en el contexto del programa denominado *International Outgoing Fellowship Marie Curie Action (IMTraP-2011-29951)*, siendo la *investigadora* la Dr. Marta R. Costa-jussà. La primera fase del proyecto (12 meses) se ha desarrollado en el Institute for Infocomm Research bajo la supervisión del Prof Haizhou Li y el Dr. Rafael E. Banchs. La segunda fase del proyecto (12 meses) se está desarrollando en la Universitat Politècnica de Catalunya bajo la supervisión del Prof. José A. R. Fonollosa. Más detalles sobre el proyecto se pueden encontrar en su correspondiente página web[3].

Bibliografía

Banchs, R. E. y M. R. Costa-jussà. 2011. A semantic feature for statistical machine translation. En *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, SSST-5, páginas 126–134.

Costa-jussà, M. R. 2015a. How Much Hybridization Does Machine Translation Need? *Journal of the American Society for Information Technology*, 6(10):2160–2165.

Costa-jussà, M. R. 2015b. Ongoing study for enahncing chinese-spanish translation with morphology strategies. En *Proceedings of the Fourth Workshop on Hybrid Approaches to Translation*, HyTra.

Costa-jussà, M. R. y J. Centelles. 2015. Description of the chinese-to-spanish rule-based machine translation system developed using a hybrid combination of human annotation and statistical techniques. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 15(1):1:1–1:13.

Costa-jussà, M. R., J. Centelles, y R E. Banchs. 2014. A client mobile application for chinese-spanish statistical machine translation. En *Proc. of the Interspeech 2014 Demo Track*.

Costa-jussà, M. R. y M. Farrús. 2014. Statistical Machine Translation enhancements through linguistic levels: A survey. *ACM Computing Surveys*, 46(3):42.

Costa-Jussà, M. R. y J. A. R. Fonollosa. 2016. Character-based neural machine translation. En *Proceedings of the Annual Conference of the Association of Computational Linguistics (ACL)*.

Costa-jussà, M. R., C. A. Henríquez Q, y R. E. Banchs. 2012. Evaluating indirect strategies for chinese-spanish statistical machine translation. *Journal of Artificial Intelligence Research*, 45(1):761–780.

Gupta, P., M. R. Costa-jussà, P. Rosso, y R. E. Banchs. 2016. A deep source-context feature for lexical selection in statistical machine translation. *Patter Recognition Letters*, 75:24–29.

Khalilov, M. y J.A.R. Fonollosa. 2011. Syntax-based reordering for statistical machine translation. *Computer Speech and Language*, 25(4):761–788.

Labaka, G., C. España-Bonet, L. Màrquez, y K. Sarasola. 2014. A hybrid machine translation architecture guided by syntax. *Machine Translation*, 28:91–125.

Sánchez-Cartagena, V., J. A. Pérez-Ortiz, y F. Sánchez-Martínez. 2015. A generalised alignment template formalism and its application to the inference of shallow-transfer MT rules from scarce bilingual corpora. *Computer Speech and Language. Special Issue on Hybrid MT: Integration of Linguistics and Statistics*.

Toutanova, K., H. Suzuki, y A. Ruopp. 2008. Applying morphology generation models to machine translation. En *Proc. of the conference of the Association for Computational Linguistics and Human Language Technology (ACL-HLT)*, páginas 514–522.

[3] http://cordis.europa.eu/project/rcn/103170_en.html