Matxin-Informatika: versión del traductor Matxin adaptada al dominio de la informática.

Matxin-Informatika, version of Matxin translation system adapted to the computer science domain

Iñaki Alegria, Unai Cabezón, Gorka Labaka, Aingeru Mayor, Kepa Sarasola
Euskal Herriko Unibertsitatea
Manuel Lardizabal 1, -20018 Donostia
aingeru.mayor@ehu.es

Resumen: Presentamos Matxin-Informatika, una versión del traductor automático Matxin (de castellano a euskera) adaptada al dominio de la informática a partir de corpus bilingüe y recursos diccionariales. Esta versión va a ser utilizada para una tarea de postedición manual en un entorno colaborativo, a partir de la cual se obtendrá un corpus que servirá para obtener una nueva mejora del traductor mediante postedición estadística.
Palabras clave: Traducción automática, euskera, castellano, corpus, léxico

Abstract: We present Matxin-Informatika, a new version of Matxin translation system (from Spanish to Basque) that has been adapted to the domain of computer science using bilingual corpus and lexical resources. This version is being used in a manual post-editing task to allow further improvement of the translator by means of statistical post-editing.
Keywords: Machine translation, Basque, Spanish, corpus, lexicon.

Este trabajo se enmarca dentro del proyecto OpenMT-2[1] entre cuyas líneas de investigación en Traducción Automática (TA) se encuentran las siguientes:

- Recogida, anotación y explotación de corpus multilingüe.
- Mejora en los sístemas de TA gracias a la colaboración con comunidades de posteditores (Web 2.0).

En ese contexto un objetivo final de este proyecto es crear una versión mejorada del traductor automático basado en reglas Matxin (Mayor, 2007; Alegria et al., 2008) para el dominio de la informática. Esto se va a realizar en dos etapas: (1) la adaptación del léxico a partir de recursos diccionariales y corpus bilingüe del dominio, y (2) el uso de postedición estadística (Simard et al., 2007; Labaka, 2010).

En esta demo presentamos la versión obtenida tras el trabajo de adaptación léxica de la primera etapa. Esta versión ha sido integrada en una interfaz basada en OmegaT[2] para ser utilizada, en la segunda etapa, en el desarrollo de un corpus de postedición que servirá para la mejora del traductor mediante postedición estadística.

Adaptación del léxico a partir de corpus

Se ha recopilado corpus paralelo del dominio de la informática combinando los textos surgidos en la localización de distintas herramientas libres como Firefox o Thunderbird (138.000 segmentos, 600M palabras en castellano y 440M en euskara). Este tipo de textos puede no ser adecuado para su uso en traducción estadística, ya que en gran parte están compuestos por menús de aplicaciones informáticas (donde suelen aparecer sólo verbos no conjugados), pero si son útiles para extraer relaciones léxicas.

Este corpus ha sido analizado y lematizado antes de ser procesado con Giza++ (Och and Ney, 2003). Segidamente, basándonos en los alineamientos de Giza++, hemos extraído para

[1] http://ixa.si.ehu.es/openmt2

[2] http://www.omegat.org/

cada lema en el corpus la lista de sus posibles traducciones al otro idioma y su probabilidad.

Como los alineamientos de Giza++ pueden contener muchas posibles traducciones que no son adecuadas (sobre todo para las palabras con pocas apariciones), hemos limitado el uso de estas listas a reordenar las equivalencias léxicas definidas en el diccionario del sistema Matxin.

Así, en los casos en los que la traducción más probable según los alineamientos del corpus no es la primera acepción de nuestro diccionario (que es la que se usa en el proceso de traducción), hemos modificado el orden de los equivalentes de traducción (y por tanto la traducción usada). Mediante este proceso hemos modificado 444 entradas del diccionario. Por ejemplo, la primera acepción de 'dirección', que era 'norabide' (con el significado de 'rumbo') a sido sustituida por 'helbide' (en el sentido de la identificación de una ubicación)

Adaptación del léxico a partir de recursos diccionariales

Usando los diferentes diccionarios español-euskara accesibles en la red, se ha efectuado una búsqueda sistemática de acepciones de entradas léxicas marcadas como palabras del dominio de la *Informática* o similares, tales como *Electricidad*. Este proceso ha producido la inclusión de 1623 entradas en el léxico de castellano del Matxin original. Los términos que se han incorporado de este modo son sobre todo multi-palabra, como 'base de datos' y 'lenguaje de programación'. Aunque también hemos conseguido identificar nuevas palabras como 'iterativo', 'ejecutable' o 'ensamblador'.

Además, para 184 palabras (como 'rutina') se les ha modificado su equivalente de traducción ('ohitura', en el significado de 'hábito'), por una traducción específica del dominio ('errutina', como 'función-procedimiento').

Uso de postedición estadística

Varios experimentos han mostrado la mejora en traducción automática mediante el uso de postedición estadística (Simard et al., 2007; Labaka, 2010). Su uso exige disponer de un corpus de traducciones manualmente posteditadas de al menos 100.000 palabras. Con objeto de recopilar ese corpus durante el año 2011 vamos a utilizar Matxin-Informatika para obtener, en colaboración con la comunidad *eu.wikipedia*[3], la traducción posteditada de 50 artículos largos de Wikipedia que actualmente existen en castellano pero no en euskara.

Se ha creado una interfaz basada en OmegaT[4] para facilitar el trabajo de postedición de los colaboradores. En su implementación se ha preferido OmegaT frente a *World Wide Lexicon Translator*[5] (WWL3), y a *Google Translation Toolkit*[6] debido que es de código abierto y esto nos es indispensable para poder adaptarlo a otros motores de traducción.

Aparte de las mejoras obtenibles a partir de la adaptación al dominio de la informática, y del uso de postedición estadística, dentro del proyecto OpenMT2, también estamos trabajando en la confección de un sistema híbrido que combine métodos basados en reglas y métodos estadísticos (Labaka, 2010).

Referencias

Alegria I., Arregi X., Díaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., Sarasola K. 2008. Strategies for sustainable MT for Basque: incremental design, reusability, standardization and open-source. Proceedings of the IJCNLP-08, pp: 235-243. Hyderabad, India.

Labaka, G. 2010. EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation. PhD Thesis. (UPV-EHU). Donostia.

Mayor, A. 2007. Matxin: Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz. PhD Thesis. (UPV-EHU). Donostia.

Simard, M., Ueffing, N., Isabelle, P., and Kuhn, R. 2007. Rule-based translation with statistical phrase-based post-editing Proceedings of the Second Workshop on Statistical Machine Translation. pp:203-206. Prague, Czech Republic.

Franz Josef Och, Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, volume 29, number 1, pp. 19-51

[3] http://eu.wikipedia.org/wiki/Wikiproiektu:OpenMT2_eta_Euskal_Wikipedia

[4] http://www.omegat.org/

[5] https://addons.mozilla.org/en-US/firefox/addon/13897

[6] http://translate.google.com/toolkit