Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 321-322 recibido 03-05-2011 aceptado 24-05-2011

Gestión de la información morfológica para la creación de un nuevo par de lenguas con distintos dialectos en un sistema de traducción automática de código abierto

Morphological information management for the creation of a new pair of languages with different dialects in an open-source machine translation system

Garbiñe Aranbarri Ariztondo
Eleka Ingeniaritza Linguistikoa S.L.
Zelai Haundi kalea 3, Osinalde industrialdea,
20170 Usurbil, Gipuzkoa
garbine@eleka.net

Itziar Cortés Etxabe
Eleka Ingeniaritza Linguistikoa S.L.
Zelai Haundi kalea 3, Osinalde industrialdea,
20170 Usurbil, Gipuzkoa
itziar@eleka.net

Resumen: Se presenta una demostración del trabajo realizado para crear un sistema de traducción automática entre el euskera *batua* estándar y el vizcaíno estándar (un dialecto del euskera). La estandarización y la generación de los datos lingüísticos de ambas variantes, y los progresos tecnológicos realizados en *Apertium* han facilitado el manejo de datos de forma más eficiente para crear este traductor de código abierto.
Palabras clave: morfología, bases de datos lexicales, dialecto, *Apertium*, código abierto, *Foma*.

Abstract: This is a demonstration of the work procedure to create a machine translation system between standard Basque and Biscayan (a Basque dialect). The standardization and creation of linguistic data of both variants, and the technological progress made on *Apertium*, have helped to manage data on a more efficient way to create this open-source software translator.
Keywords: morphology, lexical databases, dialect, *Apertium*, open-source, *Foma*.

1 Introducción

El vizcaíno es un dialecto del euskera con un estándar definido (*Instituto Labayru Ikastegia*), con el cual ya se han realizado trabajos relacionados con PNL, como es el caso del corrector ortográfico para el vizcaíno *XuxenB* (Alegria, I. et al., 2010). Gracias a este último trabajo, existen recursos morfológicos tanto para el vizcaíno como para el euskera estándar. Además, contamos con una herramienta adecuada y en código abierto para crear traductores entre lenguas cercanas llamada *Apertium* (Ramírez-Sánchez, G. et al., 2006), cuya comunidad está trabajando en estos momentos en la integración de la tecnología *Finite State Transducers*, y para lo cual ha utilizando, entre otras tecnologías, *Foma* (Hulden, M., 2009). A continuación se presentan las fuentes que se han utilizado para crear un traductor entre el euskera estándar (*euskara batua*) y el vizcaíno (*bizkaiera*).

2 Objetivo

El objetivo es el siguiente: crear un traductor de código abierto entre el par de lenguas euskera batua (*eu*) y vizcaíno (*eu_bis*) utilizando la tecnología *Apertium* de la forma más eficiente posible.

Se estudiaron dos posibilidades para el manejo de datos: una consistiría en la modificación de los diccionarios morfológicos existentes para el euskera estándar en *Apertium,* y en la creación de nuevos diccionarios para el vizcaíno; la otra opción consistiría en la utilización de *Apertium* y *Foma* conjuntamente. Hemos desarrollado la segunda opción.

Aprovechando la posibilidad que ofrece Apertium para trabajar con *Foma,* y la experiencia que tenemos a la hora de tratar la información morfológica en estos formatos, se ha considerado que ésta podría ser una buena oportunidad para explotar las bases de datos lexicales del vizcaíno (*BIDBL*) (usado

ISSN 1135-5948

anteriormente para el proyecto *XuxenB*) y del euskera *batua* (*EDBLO*) (usado para las diversas herramientas PLN que se han implementado para el euskera *batua*). La ventaja de este método es que ofrece la posibilidad de comenzar a trabajar con un léxico completo y de reducirlo a la hora de liberar los datos para publicarlos en el repositorio de *Apertium*, ya que dichos datos, por ahora, son privados.

3 Desarrollo

Con la información de las bases de datos lexicales se ha creado la información necesaria para que la transferencia léxica esté integrada en el análisis morfológico. Así, se han creado relaciones entre las dos bases de datos (vizcaíno y euskera *batua*) de forma que se consigue unir la entrada de una base de datos con el análisis de la otra, obteniendo así un análisis en la lengua de destino. Es decir, para traducir en la dirección *eu* → *eu_bis* (*batua* → vizcaíno) el análisis morfológico de la forma en *eu* dará como resultado su correspondiente análisis en *eu_bis*. A través de ese procedimiento se integra la transferencia léxica con el análisis morfológico.

Tomemos como ejemplo la forma en *eu* 'eraman', cuyo análisis después de haber sido desambiguado es el siguiente: *^eroaN<vblex>+0<AMM><PART>+0<post><ABS><MG>$*. Este análisis se corresponde con el del vizcaíno, y teniendo en cuenta la información de la base de datos de esta variante, se ha generado la forma léxica correcta: 'eroan'.

Puede ocurrir que en *BiDBL* tengamos una referencia a una forma del euskera *batua* que no aparece en *EDBL*, ya que el enriquecimiento lexical de las bases de datos se ha realizado de manera independiente. Es por ello que para generar las formas en la lengua de destino se ha utilizado una combinación de transductores creados con el operador *priority union* de *Foma*. Se ha combinado el transductor de generación (lengua de destino) con el de la información morfológica para el análisis (lengua de origen). Así, aseguraremos que siempre habrá una forma en la lengua de destino.

Otros problemas que se han encontrado en la implementación de este proyecto son los relacionados con las características propias del euskera y su integración en un sistema como *Apertium*, además de los derivados de los datos que hemos obtenido de nuestras bases de datos. Si nos centramos en estos últimos, cabe mencionar que las bases de datos utilizadas fueron creadas, en un principio, para el análisis morfológico del euskera *batua* y del vizcaíno. Es decir, en el momento de su creación, no se había previsto utilizar las dos bases de datos cruzadas para un único proyecto.

4 Mejoras y trabajo futuro

La revisión de los datos creados es uno de los trabajos 'no automáticos' que ha de hacerse después de terminar con el tratamiento automático de los mismos. Las bases de datos *BiDBL* y *EDBLO* carecen de cohesión en una serie de aspectos, y es por ello que resulta imprescindible que sean revisadas. Por consiguiente, se prevé un replanteamiento de la estructura y de los datos de las bases de datos.

Tras la revisión, se procederá a reducir el léxico para poder liberar parte de los datos generados.

En cuanto a la implementación del sistema, se cambiarán los autómatas en formato *Foma* para convertirlos al formato *HFST* (Lindén, K. et al., 2009), gracias al cual se logrará una mejor integración de las bases de datos en *Apertium*.

5 Referencias

Alegria, I. et al. “A morphological processor based on foma for Biscayan (a Basque dialect).”. 2010.

Hulden, M. “Foma: a finite-state compiler and library.” *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*. 2009. 29–32.

Lindén, K., M. Silfverberg, and T. Pirinen. “HFST Tools for Morphology–An Efficient Open-Source Package for Construction of Morphological Analyzers.” *State of the Art in Computational Morphology* (2009)

Ramírez-Sánchez, G. et al. “Opentrad Apertium open-source machine translation system: an opportunity for business and research.” *Proceedings of the 28th International Conference on Translating and the Computer*. 2006.