Procesamiento del Lenguaje Natural, Revista nº 52 marzo de 2014, pp 93-96 recibido 15-11-13 revisado 05-02-14 aceptado 06-02-14

Etiquetación y desambiguación automáticas en gallego: el sistema XIADA

Automatic tagging and disambiguation in Galician: the XIADA system

Eva María Domínguez Noya
Universidad de Santiago de Compostela
Centro Ramón Piñeiro para a Investigación en Humanidades
Estrada Santiago-Noia, Km 3, A Barcia
15896 Santiago de Compostela
edomin@cirp.es

Resumen: Tesis doctoral en Lingüística realizada por Eva María Domínguez Noya en la Universidad de Santiago de Compostela (USC) bajo la dirección del Dr. Guillermo Rojo (USC) y la Dra. María Sol López Martínez (USC). El acto de defensa de la tesis tuvo lugar el lunes 25 de noviembre de 2013 ante el tribunal formado por la Dra. Rosario Álvarez Blanco (USC), la Dra. María Inês Pedrosa da Silva Duarte (Universidad de Lisboa), el Dr. Jorge Graña Gil (Universidad de A Coruña), la Dra. María Taulé Delor (Universidad de Barcelona) y la Dra. María Paula Santalla del Río (USC). La calificación obtenida fue Sobresaliente *Cum Laude* por unanimidad.
Palabras clave: Etiquetador estadístico, anotación automática, desambiguación, etiquetario, lexicón, corpus de entrenamiento, reglas lingüísticas, criterios de etiquetación, corpus en gallego.

Abstract: PhD thesis in Linguistics, written by Eva María Domínguez Noya at the University of Santiago de Compostela (USC), under the supervision of Dr. Guillermo Rojo and Dr. María Sol López Martínez (USC). The viva examination was held on the 25th of November 2013. The Examining Board consisted of the following members: Dr. Rosario Álvarez Blanco (USC), Dr. María Inês Pedrosa da Silva Duarte (University of Lisbon), Dr. Jorge Graña Gil (University of A Coruña), Dr. María Taulé Delor (University of Barcelona) and Dr. María Paula Santalla del Río (USC). The unanimously awarded grade was Excellent *Cum Laude*.
Keywords: Statistical tagger, automatic tagging, disambiguation, tag-set, lexicon, training corpus, linguistic rules, tagging criteria, corpus of Galician.

1 Introducción

La construcción de recursos lingüísticos, entre los cuales se encuentran los corpus o bases de datos textuales, es necesaria en toda lengua para continuar profundizando en su conocimiento, pero también es fundamental para el procesamiento del lenguaje natural. En los últimos años, las nuevas tecnologías surgen como un parámetro más de clasificación de las lenguas en función de su presencia o ausencia en ellas, y el gallego no es ajeno a esta corriente.

En la actualidad no contamos con ningún estudio general de lingüística aplicada al análisis automático de corpus en gallego, a pesar de que existen varios con distinto grado de anotación, por lo que este trabajo cubre un espacio vacío y, al tiempo, ayuda a la defensa y promoción de la lengua gallega.

Los objetivos principales que pretende cumplir la presente tesis son dos. Por un lado, dar cuenta de los diversos recursos lingüísticos que necesita un etiquetador automático de tipo estadístico-probabilístico para, sin normativizar de modo alguno los textos ni prescindir del análisis de elementos marginales, etiquetar

ISSN 1135-5948

automáticamente el *Corpus de Referencia do Galego Actual* (CORGA)[1] y obtener la caracterización morfosintáctica de cada unidad léxica presente en él mediante las etiquetas y lemas que respectivamente les correspondan.

Por otra parte, se persigue también dar a conocer los múltiples problemas que crea trabajar con lengua real, describir las soluciones que se adoptaron para resolver esos problemas y explicar con detalle los criterios que subyacen en la etiquetación y lematización aplicadas al corpus de entrenamiento.

2 Organización de la tesis

La tesis se organiza estructuralmente en cuatro capítulos que describen, respectivamente, el etiquetario utilizado, la estructura interna del lexicón, las características del etiquetador automático y la herramienta utilizada en la desambiguación manual, y, finalmente, las peculiaridades lingüísticas del corpus sobre el que se ejecuta el etiquetador así como los criterios que guían la anotación.

El Capítulo 1 acoge la determinación del etiquetario que se utiliza en el sistema XIADA[2]. Partiendo de las recomendaciones del grupo EAGLES, teniendo en cuenta el camino recorrido ya en etiquetación de corpus por lenguas como el portugués, el catalán o el español y basándonos en las principales gramáticas existentes para el gallego, se explica el establecimiento de cada categoría gramatical y de los diferentes atributos que se consideran pertinentes en cada una de ellas. Se proporcionan ejemplos prácticos de etiquetación para las distintas clases gramaticales establecidas y, asimismo, siempre en relación con los atributos, se explican ya las peculiaridades que afectan a la caracterización morfológica de algunos o de todos los elementos que se incluyen en una categoría dada. De la combinación de las 18 categorías gramaticales delimitadas y los atributos que se consideran pertinentes para ellas resultan un total de 383 etiquetas diferentes, que en teoría dan cuenta de la caracterización morfológica plena de cualquier palabra gallega.

En el Capítulo 2 se describe la formalización e implementación del diccionario léxico que le sirve al etiquetador de recurso principal para identificar las formas ortográficas presentes en un texto y caracterizarlas morfosintácticamente. Al trabajar con una lengua con morfología muy rica como es el gallego, no es aconsejable introducir en una base de datos todas las formas flexionadas que se subsumen bajo un paradigma, por lo que se crearon modelos formales para reducir la flexión en las categorías variables y facilitar así la introducción en el lexicón de las 100.000 formas más frecuentes del CORGA y las cerca de 50.000 entradas que posee el *Vocabulario Ortográfico da Lingua Galega* (VOLGa) (González y Santamarina, 2004). Por otra parte, se explica también el tratamiento aplicado sobre la gran cantidad de contracciones y las ingentes combinaciones de forma verbal con segundas formas del artículo y/o pronombres enclíticos que posee el gallego, y que permite el reconocimiento y caracterización diferenciados de tantos constituyentes como unidades léxicas simples conforman la unidad amalgamada, sin perder esta de vista. Por último, se describe el sistema de reglas lingüísticas que ayuda al etiquetador tanto en la segmentación y etiquetación de pronombres enclíticos como en la segmentación y etiquetación de las formas verbales cuando van acompañadas de enclíticos.

En el Capítulo 3 se describe el etiquetador XIADA y se justifica la elección del modelo estadístico. Se da cuenta, asimismo, de los requisitos que exige el tipo estadístico, ocupando un lugar destacado entre ellos la construcción y caracterización del corpus de entrenamiento o *gold standard* que le sirve al etiquetador de modelo para inferir la gramática del gallego. Se explica la actuación del etiquetador ante las palabras desconocidas, y se incluye una breve tipología de las ambigüedades que potencialmente disminuyen su capacidad de acierto, como son las categoriales, las atributivas y las segmentales, pero a las que, pese a ello, debe enfrentarse. Termina el capítulo con la descripción de la herramienta que se emplea para desambiguar manualmente: un editor XML de software libre, en concreto el *XMLmind XML Editor Standard Edition V2.9*, que ofrece unas funcionalidades de representación y tratamiento de los datos tales, que garantizan que no van a existir etiquetas imposibles ni contenido vacío de elementos, con lo que se incrementa la consistencia del corpus de entrenamiento y permite *a posteriori* test de comprobación automáticos sobre los datos finales.

Finalmente, el Capítulo 4 se centra en el análisis aplicado al corpus de entrenamiento,

[1] http://corpus.cirp.es/corga/

[2] http://corpus.cirp.es/xiada/etiquetario.html

del cual se puso a disposición pública mediante la consulta *on line* una versión con el nombre de *Corpus de Referencia do Galego Actual etiquetado* (CORGAetq)[3]. En este capítulo se demuestran las dificultades de trabajar con lengua real, dificultades que se incrementan más aun por trabajar con una lengua como el gallego, cuyo proceso de normalización choca con el español en numerosas interferencias de tipo léxico y morfosintáctico. En estas se centra la primera parte del capítulo, donde se analizan las peculiaridades lingüísticas que caracterizan a los textos sobre los cuales se ejecuta el etiquetador y se expone el tratamiento que reciben las erratas, las variaciones gráficas, los dialectalismos, interferencias, ultracorrecciones, etc. La segunda parte del capítulo constituye, en términos generales, una descripción gramatical parcial del gallego, pues en ella se refieren los criterios que guían la etiquetación aplicada en el CORGAetq. Ante la imposibilidad de dar cuenta en la tesis de todos los detalles, se prefirió optar por tratar aquellos aspectos referidos a la etiquetación y lematización de:

- Elementos que en general no se incluyen en los manuales de consulta: abreviaciones, silabeos, acortamientos, direcciones, identificadores, etc.
- Unidades en las que la ambigüedad es muy alta: *que*, *como*.
- Elementos pertenecientes a categorías muy próximas: adjetivo frente a participio o sustantivo frente a infinitivo.
- Formas comunes a dos lemas diferentes: *pór* vs. *poñer* y derivados.
- Combinaciones que reciben una caracterización *ad hoc* para facilitar la recuperación de información: tipo *boca arriba* o tipo *anos despois*.
- Criterios de delimitación de los tres tipos de locuciones reconocidos en XIADA, adverbiales, conjuntivas y preposicionales, con ejemplos de análisis para las más frecuentes.

3 Contribuciones

Las principales contribuciones de esta tesis son las siguientes:

- Especificación de los criterios que subyacen en la elección del etiquetario utilizado en XIADA.
- Descripción detallada de los modelos formales creados para reducir la flexión en los elementos de las categorías gramaticales variables y facilitar así la implementación del lexicón.
- Descripción de los recursos secundarios que permiten la identificación y caracterización individual de los componentes de una amalgama constituida por una forma verbal a la que acompaña uno o más pronombres en enclisis (*dán<u>nolo</u>* [*nos lo dan*], *vái<u>sellenos</u>* [*se nos va*], etc.) o una segunda forma del artículo (*compra-<u>los</u> libros* [*comprar los libros*], *leréille-<u>las</u> historias ós nenos* [*les leeré las historias a los niños*], etc.), sin que la unidad amalgamada esté recogida en ninguna parte del sistema.
- Distinción práctica, y no solo teórica, entre *unidad y forma* o *unidad ortográfica* y *unidad léxica*. Por ejemplo, la unidad ortográfica *leréillelas* se desagrega en tres constituyentes compuestos, respectivamente, por las unidades léxicas *lerei*, *lles*, *las*, a las que se les proporciona la etiqueta que las caracteriza morfosintácticamente y el lema que les pertenece. La salida del etiquetador correspondiente a este análisis en un formato de texto es el siguiente:

```
<análise_unidade>
 <unidade>leréillelas</unidade>
 <alternativas>
  <alternativa válido="si">
   <constituínte>
    <forma>lerei</forma>
    <etq_lema válido="si">
     <etiqueta>Vfi10s</etiqueta>
     <lema>ler</lema>
    </etq_lema>
   </constituínte>
   <constituínte>
    <forma>lles</forma>
    <etq_lema válido="si">
     <etiqueta>Rad3ap</etiqueta>
     <lema>lle</lema>
    </etq_lema>
   </constituínte>
   <constituínte>
    <forma>las</forma>
    <etq_lema válido="si">
     <etiqueta>Ddfp</etiqueta>
     <lema>o</lema>
    </etq_lema>
   </constituínte>
```

[3] http://corpus.cirp.es/corgaetq/

```
    </alternativa>
  </alternativas>
</análise_unidade>
```

Lo que se interpreta del siguiente modo:

lerei	Vfi10s [Verbo, futuro de indicativo, 1.ª persona singular, del lema *ler*]
lles	Rad3ap [Pronombre átono, dativo, 3.ª persona, masculino/femenino, plural, lema *lle*]
las	Ddfs [Artículo determinado femenino plural, lema *o*]

- Descripción pormenorizada de la DTD que define la estructura y sintaxis del documento etiquetado.
- Estudio lingüístico de las peculiaridades relativas a variación que caracterizan los textos etiquetados y su tratamiento.
- Conjunto de criterios lingüísticos que guían la anotación morfosintáctica del corpus de entrenamiento.

4 Conclusiones

Dada la diversidad de los recursos que se describen, así como su aplicación ya en un corpus público (CORGAetq), las conclusiones que se pueden extraer son de dos tipos.

Por un lado están aquellas que inciden en aspectos que es preciso mejorar en el sistema, como son:

- Ampliación del corpus de entrenamiento para documentar etiquetas que todavía no han surgido.
- Continuar alimentando el lexicón para cubrir todos los posibles análisis de las unidades implementadas en él.
- Incorporar las unidades multipalabra pertenecientes a las categorías variables.
- Incluir un módulo para los sustantivos propios.
- Crear reglas lingüísticas que corrijan errores sistemáticos de etiquetación en los que la estadística falla.

Por otra parte, están las que se refieren a la valoración del sistema en sí, en las que sin duda, sin olvidarnos de las zonas oscuras de la etiquetación que provocan inseguridad en el anotador, así como la necesidad de contar con más estudios gramaticales que ayuden en el proceso de etiquetación, destacan dos resultados muy concretos que ubican al gallego en el grupo de lenguas que cuentan con recursos computacionales robustos:

- Etiquetador morfológico automático estadístico de alta precisión, pues su tasa de acierto se sitúa ya en el 96%.
- Corpus etiquetado morfológicamente de casi 600.000 unidades léxicas, que cuenta con una anotación minuciosa y un sistema de recuperación de información flexible y cómodo.

Bibliografía

Barcala, F. M., M. A. Molinero y E. Domínguez. 2007. XML rules for enclitic segmentation. En *Computer Aided Systems Theory. Lecture Notes in Computer Science*, 4739:273-281.

Domínguez Noya, E. 2008. O Corpus de Referencia do Galego Actual *(CORGA)*: presente e futuro. En *A lexicografía galega moderna: Recursos e perspectivas*, 139-151.

Domínguez Noya, E., Fco M. Barcala Rodríguez y M. A. Molinero. 2009. Avaliación dun etiquetador automático estatístico para o galego actual: Xiada. *Cadernos de Lingua*, 30/31:151-193.

Domínguez Noya, E. y X. M. Mosquera Carregal. 2011. Corrector ortográfico especializado para o proxecto IANUS. En *Lingua e Sanidade. VII Xornadas sobre Lingua e Usos*, 91-123.

Domínguez Noya, E. 2012. Partículas exceptivas: problemas de delimitación e proposta de análise. *Cadernos de Lingua*, 34:5-64.