Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 153-160 recibido 02-05-2005; aceptado 01-06-2005

Algoritmo de stemming para el gallego

Marisa Moreda Leirado
Dep. Computación
Fac. de Informática
Univ. da Coruña
Campus de Elviña
15071 A Coruña
lmoreda@udc.es

Ángeles S. Places
Dep. Computación
Fac. de Informática
Univ. da Coruña
Campus de Elviña
15071 A Coruña
asplaces@udc.es

Eloy Vázquez Fontenla
Dep. Computación
Fac. de Informática
Univ. da Coruña
Campus de Elviña
15071 A Coruña
evazquez@udc.es

Miguel R. Penabad
Dep. Computación
Fac. de Informática
Univ. da Coruña
Campus de Elviña
15071 A Coruña
penabad@udc.es

Resumen: La cantidad y calidad de los recursos y herramientas para el procesamiento del lenguaje natural que existan para un idioma dado depende de dicho idioma. En la Península Ibérica, el gallego es una de las lenguas para la que no existen, hasta el momento, suficientes herramientas y recursos apropiados. Para contribuir al desarrollo de estas herramientas, este artículo presenta un algoritmo de *stemming* para el gallego. Aunque fue presentado por primera vez en 2002, en años sucesivos ha sido optimizado, completado y probado con corpora de distinta naturaleza con el objetivo de ser usado principalmente en servicios de búsqueda por contenido en bibliotecas digitales.

Palabras clave. Bibliotecas Digitales, Recuperación de Textos, stemming

Abstract: The quantity and quality of the resources and tools for natural language processing for a given language depend on such a language. In the Iberian Peninsula, Galician is one of the languages that lack this type of tools and resources. To contribute to their development, this paper shows a stemmer specifically designed for the Galician language. It was first introduced in 2002, but since then it has been optimized, completed and tested with several different corpora, with the final aim of embedding this stemmer in a content search service for digital libraries.

Keywords: Digital libraries, text retrieval, stemming

1 Introducción

El nivel de desarrollo de las herramientas y recursos de Procesamiento del Lenguaje Natural depende fuertemente del idioma de que se trate. El inglés es, sin duda, el idioma que cuenta con más herramientas. El portugués y el español, aunque tienen un largo camino por recorrer, cuentan ya con diccionarios electrónicos, herramientas de lematización y/o stemming, etc. En cuanto al gallego, está casi todo por hacer, aunque hay equipos de investigación en la Universidade de Vigo (http://webs.uvigo.es/sli/) y en el Centro Ramón Piñeiro para a Investigación en Humanidades (http://www.cirp.es/) trabajando en distintos frentes. También, desde el año 2002, investigadores en Filoloxía Galego-Portuguesa e Informática (http://rosalia.dc.fi.udc.es/lbd) de la Universidade da Coruña están trabajando en el desarrollo de herramientas y recursos de recuperación de textos para el gallego.

En este trabajo se presenta un algoritmo de *stemming* para el gallego. Una primera versión del mismo fue presentado en el congreso SPIRE'2002 (Brisaboa et al, 2002), y en los últimos años ha sido optimizado, completado y probado con corpora de distinta naturaleza, para poder ser usado en servicios de búsqueda por contenido en bibliotecas digitales. El objetivo es que, una vez el usuario haya introducido la palabra de búsqueda, sea la biblioteca digital la que, usando la herramienta de stemming,

Este trabajo está parcialmente financiado por MCYT (PGE y FEDER) ref. TIC2003-06593.

automáticamente amplíe la búsqueda a todas las variaciones morfológicas de la palabra buscada.

El resto del artículo se organiza como sigue: en la sección 2 se describe brevemente el proceso de stemming, sus principales problemas y las aproximaciones existentes en lingüística computacional, para finalmente describir el algoritmo que hemos desarrollado para el gallego. La sección 3 explica las características especiales del gallego, que han influido en la configuración del algoritmo. La sección 4 describe los resultados obtenidos en la aplicación del algoritmo desarrollado a un corpus formado por textos de diferentes áreas, y la última sección ofrece algunas conclusiones sobre el trabajo desarrollado.

Finalmente se presentan dos anexos en los que aparecen algunas de las reglas del algoritmo y las tablas de las pruebas realizadas, respectivamente.

2 Lematización y stemming

De forma muy sencilla, podemos definir la lematización como el proceso de representar mediante un único término (el lema) todas las posibilidades flexivas de una palabra. Desde el punto de vista lingüístico, un lema es un término que representa y unifica todos los elementos de un conjunto de palabras morfológicamente similares (Crystal, 2000). De forma similar, el stemming reduce un conjunto de palabras a su *stem* o raíz común. Así, *camion-* sería la raíz de *camioneiro, camións, camiós, camiois*, etc., y *garraf-* la de *garrafón, garrafa, garrafiña*, etc.

Es posible realizar el stemming mediante un algoritmo que use reglas gramaticales de derivación morfológica para el idioma en cuestión, o usando un diccionario informatizado que asocie a cada forma su lema (palabra) representante. Dentro de estos últimos, MACO (Carmona et al., 1998) es el lematizador de referencia para el español.

Una diferencia importante entre estas dos aproximaciones es la dificultad de elaboración, ya que mientras que para la creación de un diccionario es necesario un gran esfuerzo de recopilación para introducir cada palabra y su raíz de forma manual (a lo sumo semiautomática), la técnica basada en un algoritmo permite realizar el stemming declarando simplemente una serie de reglas lingüísticas.

Los resultados de ambas técnicas son similares dentro de un margen de error razonable. El principal problema de los diccionarios informatizados es la ambigüedad semántica, que sólo podría ser evitada realizando un análisis sintáctico y/o semántico de cada entrada. La aproximación basada en reglas presenta, aparte de la inevitable ambigüedad semántica, dos problemas básicos: el *overstemming* (reducir demasiado, representando con un mismo stem formas que deberían ser representadas con varios) y el *understemming* (reducir poco, al obtener stems distintos para formas que se corresponderían únicamente con uno).

Uno de los primeros algoritmos de stemming, desarrollado para el inglés en 1980, se debe a Martin Porter, por lo que se conoce como "algoritmo de Porter" (Porter, 1980). Actúa como un autómata de estados finitos que incluye un grupo de reglas que se emplean para eliminar terminaciones morfológicas y flexivas de palabras en inglés. Su idea básica era la reducción del plural al singular para normalizar los términos, por lo que básicamente eliminaba las "s" finales. Evidentemente, esto no es suficiente cuando hablamos de lenguas románicas, con mayores variaciones morfológicas y flexivas. Por ello, han ido apareciendo adaptaciones del algoritmo de Porter para otras lenguas como el español, el portugués o el francés, que lo ajustan en lo posible a las reglas del idioma concreto. En http://www.udlap.mx/~is112924/IS346/Tarea1.html puede verse la adaptación al español. En el caso del portugués, se realizaron modificaciones sustanciales, teniendo en cuenta todos los sufijos que esta lengua usa, y añadiendo también listas de excepciones que incluyen palabras que no deben ser reducidas (Moreira e Huyck, 2001). Para el gallego se partió de este algoritmo, adaptándolo a las características de esta lengua y reduciendo así al máximo el overstemming y el understemming.

2.1 Algoritmo de stemming para el gallego

El algoritmo de stemming para el gallego está construido por reglas, tantas como sufijos existen en dicha lengua. Cuando una palabra debe ser reducida a su raíz, el algoritmo comprueba qué regla debe aplicarse, teniendo en cuenta sus sufijos. Para la construcción de

estas reglas hemos seguido la *Gramática da Lingua Galega* (vol. II, III) (Freixeiro, 1999, 2000) y el *Vocabulario Ortográfico da Lingua Galega* (VOLGA) (http://www.linux-galicia.org/diccionario/volga_revisado.zip). La Tabla 1 presenta la sintaxis de estas reglas, así como un ejemplo. El listado completo de las reglas usadas está disponible en http://bvg.udc.es/recursos_lingua/stemming.jsp.

Sintaxis	Ejemplo
"Sufijo a cambiar",	"eiro",
"Tamaño mínimo de la raíz",	"3"
"Sufijo sustituto",	""
"Lista de excepciones"	{canteiro, mareiro, peleiro}

Tabla 1. Sintaxis de las reglas del algoritmo

A continuación se describen los 4 componentes de cada regla:

Sufijo a cambiar: es la terminación que se sustituye o, en ocasiones, se elimina. Este componente sirve para realizar la primera comprobación sobre el término a reducir. En el ejemplo, se aplicaría la regla si el término acaba en *–eiro*.

Tamaño mínimo de la raíz: especifica el tamaño mínimo que puede tener la raíz una vez se ha eliminado el sufijo. Si la raíz es menor, la regla no se aplicará. Por ejemplo, esta regla se usaría para reducir *palleiro* (pajar) ya que su raíz (*pall–*) tiene 3 caracteres, pero no se aplicaría a *abeiro* (amparo) ya que produciría una raíz (*ab–*) de 2 caracteres. Este componente nos permite evitar la eliminación de terminaciones que no son sufijos sino que forman parte de la propia raíz.

Sufijo sustituto: es el sufijo que sustituye al "sufijo a cambiar". Si se especifica una cadena vacía, esto indica que el sufijo no se cambia por nada, sino que se elimina.

Lista de excepciones: contiene una relación de palabras para las que la regla no se debe aplicar. En el ejemplo, la regla no se aplicaría a *canteiro*, para evitar que se generase como raíz *cant–*, lo que provocaría *overstemming* (coincidiría con la raíz del verbo *cantar*).

Las reglas se organizan en etapas, dependiendo del tipo de sufijos que traten. Dentro de cada etapa, las reglas se examinan secuencialmente y sólo se aplica una de ellas. Incluso en los casos en los que la regla no se aplica sobre un término por estar éste incluido en su lista de excepciones, deja de ejecutarse dicha etapa y se pasa a la siguiente. Por ello, el orden de las reglas dentro de cada etapa es muy importante, ya que hace que se comprueben antes, y como consecuencia se cambien o eliminen, sufijos más largos, asegurando que se aplica la regla más adecuada a la terminación de la palabra. Por ejemplo, en las reglas del plural el sufijo *–ais* se comprueba antes que *–s*, y en las de sufijos apreciativos se comprueba *–deiro* antes que *–eiro*. De no hacerlo así, *panadeiro* se reduciría a "*panad–*" y no a *pan–* como debe ser (tras la eliminación de la vocal final en una etapa posterior).

El algoritmo consta de 8 etapas que se ejecutan según el diagrama flujo de la Figura 1. A continuación se describe cada una de estas etapas.

Etapa 1. Reducción de desinencias de plural. En esta etapa sólo se comprueban las palabras terminadas en *–s*, verificando si el vocablo acaba en alguna de las terminaciones de alguna regla, cambiando el sufijo por el indicado cuando corresponda. En el caso del morfema *–s*, este es simplemente eliminado. Así, si el término original es *normais*, el algoritmo cambia *–ais* por *–al*, produciendo la forma singular de la palabra (*normal*), que será posteriormente tratada a través de reglas de otras etapas.

Etapa 2. Unificación gráfica. Mediante esta etapa se pretende unificar en lo posible las distintas variaciones ortográficas de un mismo sufijo para evitar, en etapas posteriores, el tener diferentes reglas para la misma terminación escrita en las distintas variantes existentes en gallego. Estas variaciones morfológicas son debidas, como se describe en la siguiente sección, a la ausencia o coexistencia de varias normativas ortográficas. Así, en esta etapa, palabras como *bendición* y *bendiçom* o *irmão, irmao, irmau,* e *irmán* convergen a un único sufijo. Hay que destacar que esta es una etapa especial, ya que la palabra no se reduce a su raíz, sino que se altera el sufijo, que será tratado posteriormente.

Etapa 3. Reducción de sufijos adverbiales. Esta etapa sólo contiene una regla, puesto que existe un único sufijo capaz de formar adverbios: *–mente* (*xentil>xentilmente*).

Etapa 4. Reducción de sufijos apreciativos. Esta etapa trata el conjunto de sufijos apreciativos (diminutivos, aumentativos, peyorativos e intensificadores), es decir, aquellos que no tienen capacidad para cambiar

la categoría de la palabra sobre la que actúan. Una característica especial de este tipo de sufijos es su morfología recursiva, que permite acumular sufijos sobre una misma base, como sucede con la palabra *gordochiño*. En este caso, se sustituye *–iño* por *–o* (*gordochiño>gordocho*), y seguidamente se vuelve a revisar la palabra, eliminando la terminación *–ocho* (*gordocho>gord*) y obteniendo así la raíz final. Hasta ahora, el algoritmo de stemming para el gallego es el único capaz de reconocer más de un sufijo de este tipo en una misma palabra.

Etapa 5. Reducción de sufijos nocionales. Estos sufijos muestran una gran tendencia a la lexicalización, resultando con frecuencia dificultosa la distinción entre la base y el sufijo. Existen palabras con sufijo nocional que no deberán ser reducidas a la raíz original, ya que el significado varió, y de hacerlo se produciría overstemming. Por ejemplo, en la palabra *lanzal* (esbelto) la diferencia entre la base (*lanza*) y el sufijo se ha olvidado por completo en el habla, y en una posible búsqueda de *lanzal* los usuarios probablemente no estarán interesados en documentos que hablen de lanzas. Por este motivo, las reglas de esta etapa suelen tener un elevado número de excepciones.

Etapa 6. Reducción de desinencias verbales. En esta etapa se obtiene la raíz del verbo, después de la eliminación de la vocal temática y los morfemas de tiempo y número. Así, por ejemplo, *cantaban* se reduce a *cant–* .

Etapa 7. Reducción vocálica. Esta fase elimina las vocales que todavía se mantienen tras las etapas anteriores. Es el caso, por ejemplo, de *movedizo*, que después de la extracción del sufijo nocional *–dizo* queda como *move*, cuando la raíz es en realidad *mov–*. La otra alternativa, considerar *–edizo* como sufijo, ampliaría demasiado el número de terminaciones de cada etapa. Dentro de este grupo se incluyen también las palabras que varían la raíz si van con las vocales “a, o” o con “e, i”, es decir, los dígrafos “gu” y “qu” que se deben sustituir por “g” y “c” (por ejemplo, *cheguemos>chegu–>cheg–*; *marquei>marqu–>marc–*).

Etapa 8. Eliminación de la tilde. Este último paso es necesario para uniformar, ya que las raíces en las palabras en gallego llevan tilde o no dependiendo del sufijo. Por ejemplo, palabras como *práctica* y *practicamente* (ya que en gallego las palabras acabadas en *–mente* no llevan tilde) darían lugar a dos raíces distintas. Tras esta fase, ambas se reducirían a *pract–* .

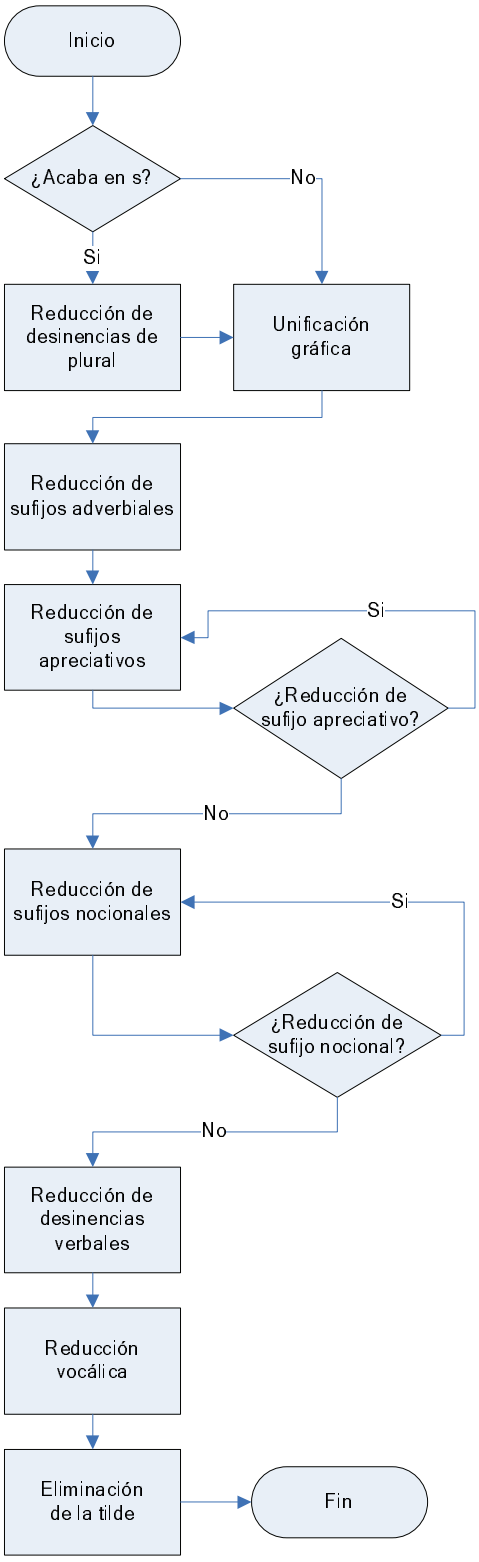

Figura 1. **Diagrama de flujo del algoritmo de stemming.**

3 Particularidades del gallego

El idioma gallego presenta una serie de características que hacen que el algoritmo de stemming diseñado para él sea especialmente complejo. Una de las más importantes es el número inusualmente elevado de sufijos, que hace que la cantidad de reglas del stemmer aumente (de forma proporcional a este número).

Los motivos que explican la existencia de este tipo de características en el gallego son principalmente históricos. La marginación sufrida por el gallego, que durante siglos fue transmitido básicamente de forma oral, propició la aparición de muchas variantes dialectales para una misma palabra. Por ejemplo, las formas acabadas en *–ón* tienen terminaciones diferentes para la formación del plural: *–óns* (en la parte occidental de Galicia), *–ós* (en la parte central) y *–ois* (en la parte oriental), e incluso formas que coinciden con el portugués, como *–ões*. También encontramos en los textos gallegos gran cantidad de interferencias del español, no sólo léxicas, sino también morfológicas, como la realización del plural de las palabras terminadas en *–l* con el sufijo *-les* (*animales*, *normales*, en vez de *animais, normais*). Estas variaciones morfológicas y flexivas no pueden omitirse en la construcción de una herramienta que va a trabajar sobre textos gallegos, debido a la frecuencia con la que estas aparecen en ellos.

Además, durante mucho tiempo no existió una institución pública ni autoridad individual que, con su prestigio, lograse imponer una normativa. Por ello, cuando en el siglo XIX los escritores comienzan de nuevo a escribir en gallego, después de los *Séculos Escuros,* estos mantuvieron un comportamiento libre de ataduras normativas, resolviendo los problemas que se les manifestaban con criterios propios. Aún hoy en día, con una normativa de 18 años de edad y revisada el 12 de julio de 2003, continúa habiendo discrepancias ortográficas. Por este motivo, si sólo tuviésemos en cuenta la normativa oficial (por ejemplo, aceptando sólo *camións* como en el caso anterior), serían demasiadas las formas que quedarían sin reducir, o que se representarían con una raíz incorrecta.

A consecuencia de esta situación, nos encontramos con que el algoritmo de stemming para el gallego debe ser capaz de tratar más de una variante morfológica de la misma palabra, con lo que el número de sufijos (y por tanto, de reglas) usados en el algoritmo es significativamente superior a otras lenguas romances, como el castellano o el portugués. Aunque esto dificulta la elaboración del stemmer, creemos que es necesario, pues en cualquier corpus de textos gallegos, especialmente los literarios, encontraremos estas variaciones.

Sin embargo, no siempre es posible reducir correctamente todas las variantes. Por ejemplo, el algoritmo no es capaz de reconocer *faisão* como otra forma gráfica de *faisán*.

En la construcción del stemmer hemos realizado una serie de estudios estadísticos que nos dan la frecuencia de aparición de las palabras en textos gallegos. Para estos estudios hemos usado principalmente 2 corpus: el *Tesouro Informatizado da Lingua Galega* (http://www.ti.usc.es/TILGA/) y la Biblioteca Virtual Galega (http://www.bvg.udc.es).

Estos corpus se han utilizado tanto para asegurar que el algoritmo reduce la gran mayoría de las formas existentes en el idioma gallego, como para decidir qué forma tratar correctamente cuando se presenta el dilema de elegir entre dos variaciones gráficas de la misma forma.

Comprobamos, por ejemplo, que debíamos incluir el plural en *–ás* para las palabras acabadas en *–al*, pues aunque no es una terminación muy común hoy en día, sí tuvo un uso frecuente en el siglo XIX, especialmente en la forma *reás* (reales), como demuestra su aparición en 1197 documentos del TILGA y en 21 obras de la BVG.

4 Resultados empíricos

Para comprobar la efectividad de este algoritmo lo hemos aplicado a un corpus monolingüe en gallego compuesto por documentos de diferentes géneros (literarios, periodísticos y jurídicos) extraídos de la BVG, del periódico *A Nosa Terra* y del *Diario Oficial de Galicia*. El corpus resultante tiene un tamaño de 42'1 MB, siendo mayoritarios los documentos de tipo literario (26'8 MB).

El hecho de no existir otras herramientas semejantes para el gallego, imposibilita la comparación de nuestro algoritmo para medir su efectividad, por ello sólo hemos comparado los resultados mostrados en el apéndice (Tabla 3) con los obtenidos con stemmers para el español o el portugués con corpus similares. Los resultados de estas comparaciones se

muestran en el apéndice, en la Tabla 3 y la Tabla 5, respectivamente (Brisaboa, 2002).

Debemos destacar que, proporcionalmente, el número de formas antes de reducir es significativamente superior en el corpus de gallego. Esto no sorprende si se consideran las posibles variaciones léxico-morfológicas de las palabras en gallego.

Una vez realizado el proceso de stemming, el tamaño de los vocabularios se redujo significativamente en todos los casos. Sin embargo, los algoritmos para el gallego y el portugués se comportan de forma distinta dependiendo de la naturaleza de los textos, mientras que la respuesta del algoritmo para el español es indiferente a esta naturaleza. Esto es debido a que la complejidad de los algoritmos para el gallego y el portugués es mayor que la del algoritmo para el español.

Tanto en el portugués como en el gallego, los textos jurídicos son los que menos se reducen, aunque el gallego da porcentajes más altos (63'28%), es decir, más raíces diferentes. Esto es debido a que los documentos de partida, extraídos del DOG, contienen gran cantidad de fechas, abreviaturas y nombres propios que esta herramienta no reduce. El algoritmo para el español es el que más reduce, en un porcentaje de 40'50%. Esto puede ser una indicación de que se produce overstemming, ya que en este tipo de documentos el número de palabras reducidas a su raíz es normalmente mucho menor, por la existencia de símbolos como números o códigos, fechas, etc., que normalmente no se reducen a ninguna raíz. En el caso del portugués, al igual que el gallego, se produce un menor stemming en los textos jurídicos con relación a otros géneros.

En cuanto a la comparación de nuestro método con otros que usen una técnica diferente, como el uso de diccionarios electrónicos, hemos elegido MACO (Carmona et al., 1998) para tal comparación.

Tomando como partida el corpus CLiC+TALP (con un fuerte sesgo en su contenido hacia los documentos periodísticos), con alrededor de un millón de palabras, MACO genera algo más de 130.000 lemas, es decir, en torno al 15% (Civit y Martí, 2002). Sin embargo, esta comparación no es realmente objetiva, ya que los valores que produce MACO se refieren a palabras totales en el corpus, no palabras *distintas* como hemos considerado en nuestro método. Además, MACO obtiene lemas para las palabras, y nuestro algoritmo obtiene sus raíces (stems).

La Tabla 2 muestra una prueba realizada con nuestro stemmer usando un fragmento de un texto del coruñés Ramón Armada Teixeiro, sacado de su obra teatral *Non máis emigración,* del año 1886.

TEXTO ORIGINAL	TEXTO REDUCIDO
Pedide cabritiños	ped cabr
Á Virxen d'o Cristál,	A virx d'o cristal
Qu'o meu amor non fuxa,	Qu'o meu am non fux
N-a vida, d'o lugár.	N-a vid d'o lugar

Tabla 2. Stemming sobre un texto gallego.

Una vez realizado el stemming sobre este fragmento podemos observar que el sistema inspeccionó correctamente los sufijos, procediendo a su eliminación (*cabritiños>cabrito>cabr–*), y trató correctamente las desinencias verbales (*pedide>ped*, *fuxa>fux*). Las palabras que pertenecen a categorías cerradas, como las preposiciones (*d'o, N-a)* o los pronombres (*meu*) no son reducidas.

5 Conclusiones y trabajo futuro

En este artículo hemos presentado un algoritmo de stemming para el gallego, herramienta de gran utilidad para el desarrollo de otras herramientas de procesamiento de lenguaje natural, e igualmente importante para desarrollo de servicios de búsqueda por contenido en bibliotecas digitales.

El algoritmo desarrollado está basado en reglas y listas de excepciones, y presenta una mejora significativa con relación al prototipo creado en 2002, ya que ahora es capaz de tratar tanto la recursividad morfológica, que suponía un problema por estar nuestros textos repletos de construcciones de este tipo, como las diferentes formas flexivas de una misma palabra, que se adscriben a una u otra área lingüística de Galicia. También hay que destacar los buenos resultados obtenidos en relación con los stemmers existentes para otras lenguas de la Península cuando se ejecutan sobre textos de características semejantes. Esto es debido a la mayor precisión con que el algoritmo para el gallego detecta y elimina sufijos para la obtención de la raíz.

Finalmente, es necesario hacer referencia a la gran riqueza léxica del gallego. Aunque para evitar el overstemming existen listas de

excepciones, debido a esta abundancia léxica es imposible señalar exhaustivamente todos los casos. Por ello, no es posible eliminar al cien por cien este tipo de errores. Sin embargo, el uso de este stemmer está orientado principalmente a un servicio de recuperación de información, por lo que consideramos que dentro de un margen razonable de errores, el algoritmo aquí presentado funciona de forma correcta.

Como trabajo futuro inmediato nos planteamos, en primer lugar, hacer una evaluación más completa de nuestro algoritmo, obteniendo valores de precisión y cobertura del mismo. Eso nos permitirá realizar comparaciones más objetivas con otros métodos, así como ver el grado de mejora en la futura evolución del algoritmo.

Bibliografía

Brisaboa, N. 2002. Compresión de textos en Lenguas Romances. En *Ingeniería del Conocimiento*. Colombia.

Brisaboa, N. y C. Fernández. 2001. Introdución ás Bibliotecas Dixitais. *Revista Galega de Filoloxía 2:*27-51. Baía Edicións, A Coruña.

Brisaboa, N., C. Callón, J.R. López, A. Places y G. Sanmartín. 2002. Stemming Galician Texts. En *Proceedings of the 9th String Processing and Text Retrieval (SPIRE'2002)*, volumen 2476 de *Lecture Notes in Computer Science*, Lisboa.

Carmona, J., S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, and J. Turmo. An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. In *Proceedings of the First Conference on Language Resources and Evaluation*. LREC'98, pages 915–922, Granada, 1998.

Civit, M. y M.A. Martí (2002) Design principles for a Spanish Treebank. *The first Workshop on Treebanks and Linguistic Theories (TLT2002)*, Bulgaria.

Crystal, D. 2000. *Diccionario de Lingüística y Fonética*. Ediciones Ocataedro.

Fernández, C. y A. Places. 2004. *As bibliotecas dixitais*. Laiovento, Santiago de Compostela.

Freixeiro Mato, X. R. 1999. *Gramática da Lingua Galega III. Semántica*. A Nosa Terra, Vigo.

Freixeiro Mato, X. R. 2000. *Gramática da Lingua Galega II. Morfosintaxe*. A Nosa Terra, Vigo.

Moreira, V. y C. Huyck. 2001. A Stemming Algorithm for the Portuguese Language. En *Proceedings of the 8th International Symposium on String Processing and Information Retrieval (SPIRE'2001)*. IEEE Computer Society, Laguna de San Rafael, Chile.

Porter, M. (1980): An algorithm for suffix stripping. En Program 14 (3):130-137.

Anexo 1: Análisis de stemmers en gallego, español y portugués

Archivos	Tamaño (MB)	Palabras diferentes	Raíces obtenidas	Porcentaje
LITERATURA	26'8	231.291	90.443	39'10%
PERIODISMO	7'47	56.452	23.928	42'38%
JURÍDICO	7'83	68.510	43.355	63'28%
TOTAL	42'1	356.253	157.726	44'27%

Tabla 3. Vocabulario de textos en gallego antes y después del stemming.

Archivos	Tamaño (MB)	Palabras diferentes	Raíces obtenidas	Porcentaje
LITERATURA	88	305.309	129.437	42'40%
PERIODISMO	7'6	61.966	26.520	41'80%
JURÍDICO	9'6	49.312	19.965	40'50%
TOTAL	105	416.587	175.922	42'22%

Tabla 4. Vocabulario de textos en español antes y después del stemming.

Archivos	Tamaño (MB)	Palabras diferentes	Raíces obtenidas	Porcentaje
LITERATURA	15	116.838	40.495	34'65%
PERIODISMO	35	136.573	56.263	41'20%
JURÍDICO	1'4	10.765	5.590	51'90%
TOTAL	51'4	264.176	102.348	38'74%

Tabla 5. Vocabulario de textos en portugués antes y después del stemming.

Anexo 2: Extracto de reglas del algoritmo de stemming para el gallego

ETAPA 1. REDUCCIÓN DE DESINENCIAS DE PLURAL (20 reglas en total)				
SUFIJO	TAM.	SUST.	EJEMPLO	EXCEPCIONES
ns	1	n	bons→ bon	luns, furatapóns
ais	1	al	normal→normais	cais, mais, pais, ademais, namais, lapis
s	2		casas→casa	barbadés, xoves, martes, aliás, pires, mas, férias, …

ETAPA 2. UNIFICACIÓN GRÁFICA (27 reglas en total)				
SUFIJO	TAM.	SUST.	EJEMPLO	EXCEPCIONES
íssimo	4	ísimo	facilíssimo→facilísimo	

ETAPA 3. REDUCCIÓN DE SUFIJOS ADVERBIALES				
SUFIJO	TAM.	SUST.	EJEMPLO	EXCEPCIONES
mente	4		felizmente→feliz	experimente, vehemente

ETAPA 4. REDUCCIÓN DE SUFIJOS APRECIATIVOS (46 reglas en total)				
SUFIJO	TAM.	SUST.	EJEMPLO	EXCEPCIONES
dísimo	5		cansadísimo→cansa	
án	3		charlatán→charlat	ademán, agremán, alcavarán, alcorán, astracán, bambán, bardán , barragán, barregán, capitán, , …

ETAPA 5. REDUCCIÓN DE SUFIJOS NOCIONALES (61 reglas en total)				
SUFIJO	TAM.	SUST.	EJEMPLO	EXCEPCIONES
eira	3		marisqueira→marisqu	cabeleira, canteira, cocheira, folleira, milleira
dade	3		lealdade→leal	acridade, calidade

ETAPA 6. REDUCCIÓN DE DESINENCIAS VERBALES (169 reglas en total)				
SUFIJO	TAM.	SUST.	EJEMPLO	EXCEPCIONES
aba	2		amaba→ am	
ar	2		cantar→ cant	azar, bazaar, patamar
ara	2		cantara→ cant	arara, prepara

ETAPA 7. REDUCCIÓN VOCÁLICA (13 reglas en total)				
SUFIJO	TAM.	SUST.	EJEMPLO	EXPEPCIONES
gue	2	g	segue→seg	
a	3		pana→pan	amasadela