Aplicación de transductores de estado-finito a los procesos de unificación de términos

Carmen Galvez
Doctora en documentación en 2003 por la Universidad de Granada.
E-mail: cgalvez@ugr.es

Resumen

Se presenta una aplicación basada en técnicas de estado-finito a los procesos de unificación de términos en español. Los algoritmos de unificación, o conflación, de términos son procedimientos computacionales utilizados en algunos sistemas de Recuperación de Información (RI) para la reducción de variantes de términos, semánticamente equivalentes, a una forma normalizada. Los programas que realizan habitualmente este proceso se denominan: stemmers y lematizadores. El objetivo de este trabajo es evaluar el grado de deficiencias y errores de los lematizadores en el proceso de agrupación de los términos a su correspondiente radical. El método utilizado para la construcción del lematizador se ha basado en la implementación de una herramienta lingüística que permite construir diccionarios electrónicos representados internamente en Transductores de Estado-Finito. Los recursos léxicos desarrollados se han aplicado a un corpus de verificación para evaluar el funcionamiento de este tipo de analizadores léxicos. La métrica de evaluación utilizada ha sido una adaptación de las medidas de cobertura y precisión. Los resultados muestran que la principal limitación del proceso de unificación de variantes de término por medio de tecnología de estado-finito es el infra-análisis.

Palabras clave

Unificación de términos. Lematización. Transductores de estado-finito.

Application of transducers of state-finite to unification processes of term variants

Abstract

An approach based on techniques of state-finite has applied to the processes of unification of terms in Spanish. The algorithms of conflation are computational procedures utilized in some Information Retrieval (RI) systems for the unification of term variants, semantically equivalent, to a normalized form. The programs that carry out habitually this process are called: stemmers and lematizadores. The objective of this work is to evaluate the deficiencies and errors of the lematizadores in the conflation of terms. The method utilized for the construction of the lematizador has been based on the implementation of a linguistic tool that permits to build electronic dictionaries represented internally in Finite-State Transducers (FST). The lexical resources developed have applied to a corpus of verification to evaluate the performance of these lexical parsers. The metric of evaluation utilized has been an adaptation of coverage and precision measures. The results show that the main limitation of unification processes of tem variants through technology of state-finite is the under-analysis.

Keywords

Term conflation. Lemmatization. Finite-state transducers.

INTRODUCCIÓN

En Recuperación de Información (RI), la reducción de las palabras que tienen la misma raíz bajo el mismo término de indización podría incrementar la eficacia en la equiparación entre los términos del documento y los términos de la pregunta del usuario (Harman, 1991). Para evitar la perdida de documentos relevantes, muchos sistemas agrupan las variantes de términos por medio de algoritmos de unificación, o también denominados *conflation algorithms*. Una variante se define como un término que está conceptualmente relacionado a otro término normalizado. Varios estudios se han realizado en relación a los problemas que presentan tanto las variantes de unitéminos como de multitérminos en RI (Sparck Jones y Tait, 1984; Jacquemin y Tzoukermann, 1999).

La unificación de términos se podría definir como un procedimiento computacional a través del cual se agrupan las variantes de un mismo término, que son semánticamente equivalentes, a una forma unificada. Hay diferentes clasificaciones de los métodos de unificación de variantes de término en RI (Lennon *et alii*, 1981; Frakes, 1992; Galvez *et alii*, 2005). En general, los programas que realizan esta función se denominan:

1. *Stemmers*, cuando este proceso se realiza aplicando *algoritmos de stemming*. Esta técnica tiene como objetivo agrupar las formas variantes de un término a una forma base, radical o *stem*.

2. *Lematizadores*, cuando este proceso se realiza aplicando *algoritmos de lematización*. El objetivo de esta técnica es agrupar las formas variantes de un término a un *lema*, definido como el conjunto de palabras con la misma *raíz* y la misma *categoría léxico-gramatical*, o etiqueta *part-of-speech* (*POS*).

PROCEDIMIENTOS DE UNIFICACIÓN DE TÉRMINOS EN RI

El método de unificación basado en técnicas de stemming implica la eliminación de afijos de acuerdo a un diccionario, que contiene listas de terminaciones de palabra, y a un conjunto de reglas. Los algoritmos se

aplican en el modo que se conoce como '*longest matching*', los más conocidos son los algoritmos de Lovins (1968), Dawson (1974), Porter (1980), y Paice (1990) y se suelen aplicar al idioma inglés. Aunque dentro de estos métodos también se encuentran algoritmos específicamente creados para otras lenguas (Popovic y Willett, 1992; Savoy, 1999). El algoritmo de Porter, disponible en *Snowball Web Site* (2006), ha sido implementado para el francés, español, italiano, portugués y alemán, entre otras lenguas.

Por otra parte, los algoritmos de *similaridad de cadenas* se basan habitualmente en la medida *n-gram*, calculada a partir de cualquier subcadena de longitud fija (donde *n* es el número de caracteres de la subcadena, siendo *n=1* en los *unigrams*, *n=2* en los *bigrams*, o *n=3* en los *trigrams*. Este método se ha aplicado frecuentemente en tareas relacionadas con RI (Adamson y Boreham, 1974; Lennon *et alii*, 1981; Robertson y Willett, 1998).

Frente a las técnicas anteriores, otros métodos se enfrentan al problema de la variabilidad del lenguaje desde una aproximación lingüística, por medio de técnicas cuyo objetivo es la reducción de las variantes léxicas a lemas. En esta línea, una de las implementaciones computaciones más importantes la constituyó el analizador *PC-KIMMO* (Karttunen, 1994) basado en tecnología de estado-finito, posteriormente utilizado en el Analizador Morfológico de Xerox desarrollado por el *Multi-Lingual Theory and Technology Group* (MLTT). Una de las aplicaciones más relevantes de la herramienta diseñado por Xerox es la reducción de variantes léxicas en los sistemas de RI. Este analizador se ha aplicado a los idiomas inglés, danés, alemán, francés, italiano, portugués, o español, entre otras lenguas. Otra herramienta basada en métodos de estado-finito es el analizador morfológico para la lengua inglesa *ENGTWOL* (Voutilainen, 1995). Para el idioma español contamos con la herramienta *COES* (Rodríguez y Carretero, 1996), o el analizador *MACO* (Carmona *et alii*, 1998).

La evaluación de los métodos de unificación de variantes de término, en RI se realiza habitualmente a partir de dos medidas:

• *Medidas externas* relacionadas con la eficacia en RI. Dentro de esta aproximación, distintos experimentos llegan a resultados muy diversos y no siempre positivos, concluyendo que, en muchos casos, la eficacia de las técnicas de stemming y lematización está en función de la complejidad del lenguaje (Hull, 1996; Popoviv & Willett, 1992). En relación con el español, algunos experimentos comparables presentados en *Text REtrieval Conferences TREC* (Buckley *et alii*, 1994; Allan *et alii*, 1996; Hull *et alii*, 1996) y en *Cross-Language Evaluation Forum CLEF* (Vilares *et alii*, 2003) han puesto de manifiesto que los resultados en esta lengua son muy parecidos, o levemente mejores, a los obtenidos en inglés, el idioma habitual de las evaluaciones.

• *Medidas internas*, relacionadas con la corrección con la que los sistemas agrupan los términos en formas normalizadas. Dentro de esta aproximación se han realizado estudios sobre los fallos del *stemming* como el *understemming*, *overstemming*, y *misstemming* (Xu y Croft, 1988; Paice, 1996). En relación a la lematización, se han realizado pocas investigaciones sobre los errores producidos en los procesos de unificación (Jacquemin y Tzoukermann, 1999).

Este trabajo se encuadra en la segunda aproximación expuesta arriba, dada la carencia de estudios empíricos que evalúen la exactitud y corrección con la que los lematizadores unifican las variantes léxicas.

OBJETIVOS

Los objetivos de este trabajo son:

• Proponer la aplicación de métodos de estado-finito para realizar los procesos de unificación de términos en los sistemas de RI.

• Implementar una aplicación informática basada en tecnología de estado-finito para la construcción de las herramientas de análisis léxico.

• Evaluar las deficiencias y errores de los analizadores desarrollados con esta tecnología.

APROXIMACIÓN A LA TECNOLOGÍA DE ESTADO-FINITO

La *Teoría de los Lenguajes Formales* se dirige a aquellas expresiones que pueden ser descritos de forma muy precisa, como son los lenguajes de programación. Los lenguajes naturales no son lenguajes formales, y, por tanto, no hay un límite claramente definido entre una sentencia correcta de otra que no lo es. Sin embargo, se pueden adoptar algunas aproximaciones formales a ciertos fenómenos del lenguaje natural susceptibles de una codificación similar a la realizada en los lenguajes de programación. Estas descripciones formales se utilizan por los lingüistas computacionales para expresar teorías sobre aspectos específicos de los lenguajes naturales, tales como el análisis morfológico.

En un principio, Johnson (1972) fue el primero en observar que determinadas reglas fonológicas y morfológicas se podrían representar por mecanismos de estado-finito, denominando a su formalismo '*two level model*'. La idea del modelo de dos-niveles fue clave para el progreso del formalismo computacional sobre la morfología propuesto por Koskenniemi (1983). El modelo de Koskenniemi model estableció una correspondencia entre la forma canónica, o forma léxica, y la forma superficial de las palabras. Esta relación la representó usando Transductores de Estado-Finito, *Finite-State Transducers* (FST).

De forma sintetizada, un transductor es un sistema de representación computacional que comprende un conjunto de estados y una función de transición, que define el cambio de estado. La función de transición se etiqueta con un par de símbolo que constituyen el alfabeto del *input* y el alfabeto de *output*. Este mecanismo se puede representar en la forma de un diagrama o gráfico de estado-finito. El transductor tomaría cadenas en el *input* y las relacionaría con cadenas en el *output*. Formalmente un FST se define como una tupla de cinco elementos (Roche y Schabes, 1995) que se expresa de la forma siguiente:

$$FST = (\Sigma, Q, i, F, E)$$

donde

Σ = *alfabeto de input y output*
Q = *número de estados*
i = *estado inicial*, $i \in Q$
F = *estado final*, $F \in Q$
E = *número de relaciones de transición*, $E \subseteq Q \times \Sigma \cup \{\varepsilon\} \times \Sigma^* \times \Sigma$

En la figura 1 se muestra la representación gráfica de transductor cuyos arcos están etiquetado con pares de símbolos que constituyen el alfabeto de *input* y *output*. Por ejemplo, "*a*" denota el símbolo superior y "*b*" el símbolo inferior. Este transductor podría establecer una relación entre: el lenguaje superior y el inferior. Así, este mecanismo podría reconocer la cadena representada por "*ac*" y la podría transformar en la cadena "*bd*". La equiparación es bidireccional, y una cadena de un lenguaje se podría corresponder por una, o más, cadenas de otro lenguaje. Las transducciones son posibles si la cadena en la parte del *input* lleva al transductor a un estado final. Hopcroft y Ullman (1979), Mohri (1996) y Roche y Schabes (1997) proporcionan una explicación más completar de este tipo de sistemas de estado-finito.

La aplicación del formalismo de estado-finito a la unificación de términos parte básicamente de que se puede establecer una relación de equivalencia entre las distintas formas superficiales y la forma normalizada, o lema, a la que se le puede añadir una etiqueta de la categoría gramatical correspondiente, o etiqueta *POS*. Esta correspondencia se puede implementar computacionalmente por medio de transductores (Karttunen *et alii*, 1992). Una analizador de *dos-niveles* o *lematizador* desarrollado con tecnología de estado-finito se encargaría de equiparar formas variantes léxicas, a formas unificadas, tal y como se representa en la figura 2.

FIGURA 1
Transductor de Estado-Finito

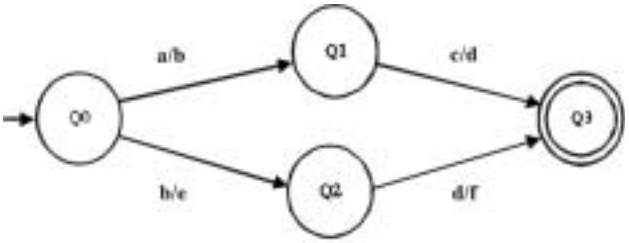
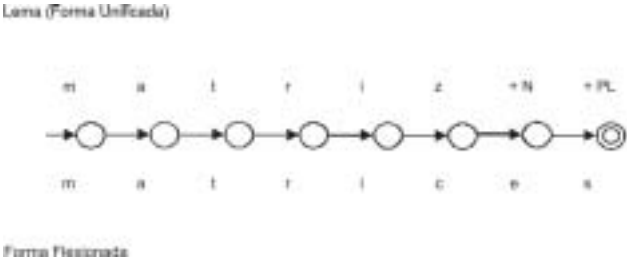


FIGURA 2
Relación entre una forma variante y una forma unificada (adaptado de Karttunen *et alii*, 1992)



CONSTRUCCIÓN DE DICCIONARIOS ELECTRÓNICOS

El planteamiento general para la identificación y agrupación de variantes flexionales que hemos adoptado parte del establecimiento de una *relación de equivalencia* entre el lema y la descripción flexional. Para manipular formalmente esta relación utilizamos tecnología de estado finito. El procedimiento que se ha seguido para la construcción de los recursos de análisis léxico es la implementación de una herramienta lingüística basada en Transductores de Estado-Finito (Silberztein, 1999). El desarrollo de los analizadores léxicos nos permitirá

distinguir las formas flexivas y las irregularidades que se producen en los distintos tipos de flexión y asignar las distintas categorías *POS* a las unidades lingüísticas. En el proceso de flexión se va a tomar como base una unidad genérica denominada *lema*. Por ejemplo, el lema de la palabra {*usuario*} estaría compuesto por el conjunto {*usuario, usuarios, usuaria, usuarias*} formado por todas las cadenas con la misma raíz y la misma categoría general de **N** (*Nombre*).

En relación con lo anterior, el análisis flexional se rige por reglas, que aportan los mecanismos para poder relacionar los distintos elementos en el contexto de la oración. Frente a esto, otro tipo de análisis, como es el derivacional, no se somete tan claramente a la regularidad de las reglas y muchas palabras compuestas, o derivadas de otras, llegan a transformarse totalmente y se alejan de la palabra origen. La variabilidad en la derivación, y el hecho de que en las palabras derivadas los afijos formen parte de la *raíz*, hace que sea muy difícil fijar cualquier tipo de regularidad en su representación, por esta razón este análisis queda excluido del presente trabajo. De la misma forma tampoco se van a tratar los pronombres enclíticos, que harían necesario el desarrollo de herramientas específicas que fueran capaces de incorporar estas partículas como parte de la conjugación verbal.

En general, el desarrollo de las herramientas de análisis léxicos ha estado guiado por la representación de la flexión de las palabras en sistemas cerrados o *paradigmas flexivos* (Matthews, 1974). Por otra parte, la estructura del paradigma hace referencia al número de categorías que puedan aparecer en el interior de los paradigmas. Dependiendo de la variación de ese número se puede hablar de estructuras simples, cuando intervenga sólo una dimensión o categoría, y estructuras complejas, cuando intervengan varias dimensiones o categorías.

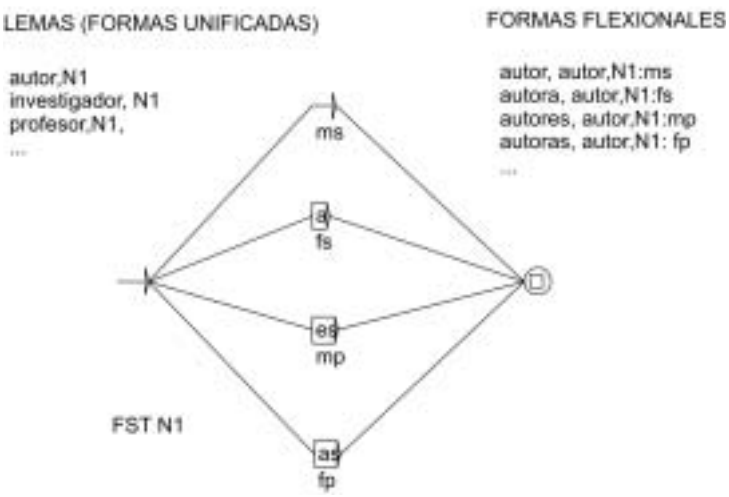
La información sobre la morfología flexiva se va a representar en transductores gráfico, por medio de una interfaz. Este recurso nos va a permitir la construcción de tres tipos de diccionarios representados internamente en transductores, que son los que vamos a utilizar para la unificación de términos: *Diccionario de Formas Unificadas*, o lemas (DELAS), *Diccionario Expandido de Formas Flexionadas* (DELAF) y *Diccionario de Formas Compuestas, Nombres Propios, Acrónimos y Abreviaturas* (DELACF).

Básicamente, el procedimiento seguido se sintetiza como sigue: hemos creado manualmente un diccionario (DELAS), que contiene una entrada para cada lema con su etiqueta *POS* correspondiente y un código numérico. Todos los lemas, que pertenecen al mismo paradigma flexivo, se relacionarán con el mismo transductor gráfico. Las entradas del diccionario DELAS se han seleccionado según la oposición binaria 'término marcado/término no marcado'. Por ejemplo, dentro de la categoría **N** (*Nombre*) y **A** (*Adjetivo*) se selecciona el término no marcado masculino/singular, y dentro de categoría **V** (*Verbo*) se elige el infinitivo. Las entradas del diccionario de lemas (DELAS) serían del siguiente tipo:

autor, **N1**
experimental, **A2**
documental, **A2**
documental, **N2**
éste, **PRODE1**
iniciar, **V1**
investigador, **N1**
solucionar, **V1**
alguno, **CUANT3**

Cada entrada del diccionario DELAS está vinculada a un transductor gráfico que describe el trayecto que el analizador morfológico debe seguir para producir todas las formas flexionadas de un término. Los nombres de los transductores gráficos se corresponderán con las etiquetas *POS* de los lemas y con los códigos numéricos, tal y como se muestra en la figura 3. Utilizando este procedimiento se han elaborado un total de 205 transductores gráficos que incluyen variantes flexivas, términos especializados, nombres propios, abreviaturas y acrónimos.

FIGURA 3
Transductor gráfico



Una vez compilados los gráficos de estado-finito en transductores se proyectan sobre las formas canónicas produciendo de forma automática el *Diccionario Expandido de Formas Flexionadas* (DELAF). Las entradas de este diccionario contienen: formas flexionadas, formas canónicas, etiquetas de categoría gramatical, e información flexiva, como masculino singular (**:ms**), o masculino plural (**:mp**). El tipo de entradas de DELAF sería el siguiente:

autor, *autor*. **N1: ms**
autora, *autor*. **N1 : fs**
autores, *autor*. **N1: mp**
autoras, *autor*. **N1: fp**
experimental, *experimental*. **A2 : ms : fs**
experimentales, *experimental*. **A2 : mp : fp**
documental, *documental*. **A2 : ms : fs**
documentales, *documental*. **A2 : mp : fp**
documental, *documental*. **N2 : ms**
documentales, *documental*. **N2 : mp**

Por último, el diccionario DELACF contiene términos compuestos, cada entrada de este diccionario está asociada a su correspondiente forma canónica y a su etiqueta de categoría léxica. Entradas de este tipo serían las siguientes:

a causa de que, *a causa de que*. **CONJS**
a disposición de, *a disposición de*. **PREP**
al igual, *al igual*. **ADV**
Univ. Granada, *Universidad de Granada*. **N : fs**
Univ. Carlos III, *Universidad Carlos III*. **N : fs**
Univ. Salamanca, *Universidad de Salamanca*. **N : fs**

EVALUACIÓN

Corpus de Verificación

Una vez desarrolladas las herramientas de análisis, el paso siguiente fue evaluar la corrección de los índices generados por los este tipo de analizadores. Para ello, aplicamos las herramientas desarrolladas a un corpus de verificación, obtenido de registros de la base de datos ISOC-Biblioteconomía y Documentación distribuida por el *Consejo Superior de Investigaciones Científicas* (CSIC). La base de datos ISOC reúne fundamentalmente revistas y actas de congreso sobre literatura científica española en el área temática de Documentación Científica.

Los registros obtenidos constituyen el Corpus de Verificación (CV), transformado para su posterior procesamiento en un fichero de texto en formato ASCII. La composición léxica del CV incluye 27.800 formas simples, o tokens, distribuidas del modo siguiente:

- 18.200 unidades léxicas;
- 4.082 dígitos (integrados por las cifras);
- 5.508 delimitadores (integrados por los distintos separadores, tales como puntos, comas o guiones).

Antes de aplicar los analizadores léxicos es necesario efectuar una etapa de pre-procesamiento consistente en someter el texto de entrada a una serie de transformaciones tales como:

- Identificar los elementos que reflejen la estructura lógica del texto (párrafos, oraciones, signos de puntuación, o delimitadores). Para la adecuada fragmentación de estas unidades lógicas de análisis hemos diseñado un transductor gráfico que inserta marcas de delimitación, {*S*}.
- Reconocer las formas compuestas no-ambiguas, locuciones o expresiones fijas, por medio de la aplicación del diccionario DELACF.
- Eliminar la ambigüedad de las formas contractas, integradas por aquellos términos que no se pueden adscribir a una sola categoría porque formalmente equivalen a dos categorías sucesivas, como ocurre con las formas contractas '*al*' ('*a el*') y '*del*' ('*de el*'). Para la separación de tales formas hemos diseñado distintos transductores gráficos.

Aplicación de los diccionarios electrónicos

Una vez pre-procesadas las sentencias del corpus, el paso siguiente fue analizar el texto en formas canónicas lematizadas, para ello hemos aplicado las bases de información léxicas, o diccionarios electrónicos, cuyas entradas se distribuyen de la forma siguiente:

- 4.500 entradas en el *Diccionario de Formas Unificadas*, o lemas (DELAS), integrado por lemas simples.
- 60.500 entradas en el *Diccionario Expandido de Formas Flexionadas* (DELAF), integrado por formas flexionadas, lemas a los que se asocia, etiqueta de clase distribucional a la que pertenece y las propiedades de número y género en nombres, o tiempo, modo, persona y número en los verbos.
- 1.200 entradas en el *Diccionario de Formas Compuestas, Nombres Propios, Acrónimos y Abreviaturas* (DELACF), integrado por lemas compuestos formados por más de una forma simple.

El resultado de la aplicación de las herramientas de análisis léxico se puede presentar de tres modos distintos en el etiquetado lineal:

• Análisis de las unidades léxicas en {*lemas*}

• Etiquetado de las unidades en {*lemas+etiquetas POS*}

• Etiquetado de las unidades léxicas en {*formas flexivas+etiquetas POS*}

Es necesario indicar que los diccionarios que hemos construido no tienen un tamaño muy elevado, porque han estado orientados a los datos de un dominio específico, con el propósito de poder tratar determinas expresiones propias de la ciencia de la información y documentación. Esta limitación, adoptada por los objetivos prácticos de este trabajo, no afecta a los resultados y, sin embargo, sí evita que muchas palabras no puedan ser analizadas por no estar incluidas en los diccionarios electrónicos.

Un extracto del etiquetado de las unidades léxicas del CV en {*lemas+etiquetas POS*} se muestra en la tabla 1.

TABLA 1
Etiquetado y reducción de las unidades léxicas del corpus a su correspondiente radical

{S}{número,.N}. Registro:{S} 408834
{S}{autor,.N}:{S} Moya Anegón, Félix {de,.PREP};{S}Moscoso, Purificación;{S}Olmeda, Carlos;{S}Ortiz Repiso, Virginia;{S}Herrero, Víctor;{S}Guerrero, Vicente
{S}{título,.N}:{S} {NeuroISOC,.N}:{S} {un,.DET} modelo {de,.PREP} {red,.N} {neuronal,.A} para la {representación,.N} {de,.PREP} {el,.DET} {conocimiento,.N}
{S}{lugar,.N} {de,.PREP} trabajo:{S} {Universidad de Granada,.N}, {España,.N}.
{S}{descriptor,.N}:{S} Bases {de,.PREP} {dato,.N};{S}{producción,.N} científica; {S}{representación,.N} {de,.PREP} {el,.DET} {conocimiento,.N};{S}{recuperación,.N} {de,.PREP} la {información,.N}
{S}Resumen:{S} El {propósito,.N} {de,.PREP} esta {ponencia,.N} {ser,.V} {presentar,.V} {un,.DET} modelo {de,.PREP} {red,.N} {neuronal,.A} que {se,.PRO} {haber,.V} desarrollado {con el fin de,.PREP} {representar,.V} {el,.DET} {conocimiento,.N} expresado {a través de,.PREP} la {producción,.N} científica {en,.PREP} {el,.DET} {campo,.N} {de,.PREP} las {ciencia,.N} {social,.A} {y,.CONJC} las {humanidad,.N}.{S} Dicho modelo {se,.PRO} {haber,.V} aplicado {a,.PREP} {el,.DET} {caso,.N} concreto {de,.PREP} la base {de,.PREP} {dato,.N} {ISOC,.N}, {producido,.PA} {y,.CONJC} {distribuido,.PA} {por,.PREP} {el,.DET} {Consejo Superior de Investigaciones Científicas,.N}.{S} Esta {aplicación,.N} forma parte {de,.PREP} {un,.DET} proyecto {de,.PREP} {investigación,.N} {cuyo,.ARE} objetivo {principal,.A} {ser,.V} {el,.DET} {desarrollar,.V} {de,.PREP} una {interfaz,.N} {de,.PREP} {realidad,.N} {virtual,.A}.

Parámetros de evaluación

El método de evaluación que vamos a utilizar es una adaptación de la métrica clásica habitualmente empleada en los sistemas de RI y que se utilizan para otras aplicaciones en las que intervienen las técnicas de procesamiento de lenguaje natural, como son las medidas de *cobertura* y *precisión* (Pereira, 1997). Aquí, el parámetro de cobertura se redefine como la proporción de variantes correctas lematizadas de entre el total de variantes posibles. El parámetro de precisión se redefine como la proporción de variantes correctas lematizadas de entre el total de variantes lematizadas. Las dos medidas se calcularían con las siguientes ecuaciones:

$$Cobertura\ \ (C) = \frac{Número\ de\ Formas\ Léxicas\ Lematizadas\ Correctamente}{Número\ Total\ de\ Formas\ Léxicas\ Posibles}$$

$$\mathbf{Pr}ecisión\ \ (P) = \frac{Número\ de\ Formas\ Léxicas\ Lematizadas\ Correctamente}{Número\ Total\ de\ Formas\ Léxicas\ Lematizadas}$$

Incorporando esta métrica de evaluación, nuestro objetivo ahora es medir el grado de corrección con el que las herramientas de análisis generan índices normalizados, con este objetivo necesitamos adquirir los siguientes datos:

• El total de las variantes léxicas reconocidas y agrupadas en lemas. Para obtener estos datos decidimos aplicar los analizadores léxicos en el modo {*lemas*}, en el que cada palabra del corpus se agrupa, o relaciona, con su lema correspondiente.

• El total de las variantes léxicas posibles reconocidas y agrupadas en lemas. Para obtener los datos del total de las variantes posibles hemos decidido aplicar los analizadores léxicos en el modo {*formas flexivas+etiquetas POS*}, en el que las palabras del corpus se identifican con las variantes flexionadas correspondientes y se les asigna la categoría *POS* a la que pertenecen.

• Las variantes léxicas correctas reconocidas y agrupadas en lemas. Para obtener los datos de las variantes correctas hemos optado por aplicar los analizadores léxicos en el modo {*lemas+etiquetas POS*}, en el que cada palabra del corpus se agrupa, o relaciona, a su correspondiente lema y a su categoría *POS*.

RESULTADOS Y DISCUSIÓN

Según los resultados mostrados en la tabla 2, la cobertura de los analizadores léxicos alcanza el 74%, esto significa que si el total de las variantes posibles susceptibles de

agruparse en lemas es de 2.216, con los analizadores basados en técnicas de estado-finito se han conseguido reducir correctamente las variantes a 1.632. El hecho de que estas herramientas consigan reducir las variantes en un 26.4%, reduciendo en esta proporción las entradas a los índices de los sistemas de RI, se puede considerar un resultado adecuado. Estos resultados no se pueden mejorar, es decir, no podemos aumentar el porcentaje obtenido porque no se trata de errores de los analizadores sino de restricciones propias de estas herramientas, diseñadas para reducir las variantes simplemente al lema. A pesar de lo anterior, como hemos mencionado, se puede considerar que este resultado es bastante satisfactorio.

La tasa de precisión de los analizadores léxicos muestra un pequeño porcentaje de errores del 3.4%, producidos en casos en los que dos o más variantes se corresponden con un único lema, cuando en realidad se trataría de variantes distintas porque cada una tiene asignada una etiqueta *POS* distinta, como ocurre con los términos {*científico, científico.* ***A1 : ms***} y {*científicos, científico.* ***N1 : ms***}. Según esto, la primera deficiencia de los analizadores desarrollados es que existen variantes léxicas que se lematizan erróneamente, provocando que determinadas variantes se vinculen al mismo lema.

Sin embargo, aunque la ratio de cobertura y precisión sea muy favorable, el gran problema de este tipo de analizadores léxicos es el *infra-análisis*. Cuando hemos evaluado los datos obtenidos lo hemos hecho siempre sobre variantes léxicas que se lemantizan, esto es, hemos evaluado los datos a partir de las variantes que se pueden agrupar a un lema. En la aplicación de los analizadores léxicos es fundamental tener en cuenta que sólo se lematizan las variantes que se corresponden con un lema, es decir cuando una mismas variante se puede agrupar a lemas distintos, como {*modelo, modelar.* ***V1 : P1s***} y {*modelo, modelo.* ***N4 : ms***}, el analizador no lematiza. En los casos de ambigüedad en los que las variantes se puedan agrupar a distintos lemas, los analizadores no son capaces de unificar las variantes. El número de variantes léxicas sin lematizar es de 365, a las que se sumarían las formas desconocidas, integradas por errores ortográficos, o términos en otras lenguas. Todo esto nos lleva a constatar que la principal deficiencia de los analizadores desarrollados es el *infra-análisis*.

CONCLUSIONES

Los resultados de la evaluación sobre las limitaciones de los analizadores léxicos basados en tecnología de estado-finito para la unificación de términos nos han llevado a las siguientes conclusiones. Primera, los analizadores léxicos consiguen reducir las variantes de términos a su radical o formas normalizada en un 26.4%, y este resultado se puede considerar satisfactorio. Segundo, los analizadores léxicos unifican las variantes con una alta precisión. Tercero, los analizadores desarrollados con técnicas de estado-finito tienen una limitación sólo agrupan las variantes que se puedan vincular a una sola forma normalizada, y en caso de ambigüedad no lematizan. En consecuencia, la principal deficiencia de los analizadores léxicos desarrollados con tecnología de estado-finito es el *infra-análisis*, y este obstáculo no se puede superar en el etiquetado lineal porque se trata de una limitación inherente de este procedimiento, los analizadores léxicos no son capaces de reducir los términos que se pueden agrupar a distintos lemas.

TABLA 2
Medidas de cobertura y precisión

Formas léxicas lematizadas	1689
Formas léxicas lematizadas correctamente	1632
Formas léxicas lematizadas incorrectamente	57
Formas léxicas posibles	2216
Cobertura	**0.74**
Precisión	**0.97**

Artigo submetido em 24/01/2006 e aceito em 07/03/2007.

REFERENCIAS

1. ADAMSON, G. W.; BOREHAM, J. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval*, v. 10, n. 1, p. 253-260, 1974.

2. ALLAN, J. et al. Inquery at TREC-5. In: TEXT RETRIEVAL CONFERENCE, TREC-5, 5., 1995, Gaithersburg. *Proceedings*... Gaithersburg, Maryland: National Institute of Standards and Technology, 1996. p. 119-132.

3. BUCKLEY, C. et al. Automatic query expansion using SMART: TREC 3. In: TEXT RETRIEVAL CONFERENCE, TREC-3, 3., 1994, Gaithersburg. *Proceedings*... Gaithersburg, Maryland: National Institute of Standards and Technology, 1994. p. 69-80.

4. CARMONA, J. et al. An environment for morphosyntactic processing of Spanish unrestricted text. In: CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, LREC'98, 1., 1998, Granada. *Proceedings*... [S.l.: s.n.], 1998.

5. DAWSON, J. L. Suffix removal for word conflation. *Bulletin of the Association for Literary & Linguistic Computing*, v. 2, n. 3, p. 33-46, 1974.

6. FRAKES, W. B. Stemming algorithms. In: FRAKES, W. B.; BAEZA-YATES, R. (Ed.). *Information retrieval:* data structures and algorithms. Englewood Cliffs: Prentice-Hall, 1992.

7. GALVEZ, C.; MOYA-ANEGON, F.; SOLANA, V. H. Term conflation methods in information retrieval: non-linguistic and linguistic approaches. *Journal of Documentation*, v. 61, n. 4, p. 520-547, 2005.

8. HARMAN, D. K. How effective is suffixing?. *Journal of the American Society for Information Science*, v. 47, n. 1, p. 70-84, 1991.

9. HOPCROFT, J. E.; ULLMAN, J. D. *Introduction to automata theory, languages, and computation*. [S.l.]: Addison-Wesley, 1979.

10. HULL, D. A. Stemming algorithms: a case study for detailed evaluation. *Journal of the American Society for Information Science*, v. 47, n. 1, p. 70-84, 1996.

11. _____ et al. Xerox TREC-5 site report: routing filtering, NLP and Spanish tracks. In: TEXT RETRIEVAL CONFERENCE, TREC-5, 5., 1995, Gaithersburg. *Proceedings*... Gaithersburg, Maryland: National Institute of Standards and Technology, 1996.

12. JACQUEMIN, C.; TZOUKERMANN, E. NLP for term variant extraction: synergy between morphology, lexicon, and syntax. In: STRZALKOWSKI, T. (Ed). *Natural language information retrieval*. Dordrecht: Kluwer Academic Publishers, 1999.

13. JONES, K. Sparck; TAIT, J. I. Automatic search term variant generation. *Journal of Documentation*, v. 40, n. 1, p. 50-66, 1984.

14. JOHNSON, C. D. *Formal aspects of phonological description*. La Haya: Mouton, 1972.

15. KARTTUNEN, L. Constructing lexical transducers. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 15., 1994, Kyoto. *Proceedings*... Kyoto: Coling 94, 1994.

16. _____; KAPLAN, R. M.; ZAENEN, A. Two-level morphology with composition. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS (COLING'92), 15., 1992, Nantes, France. *Proceedings*... Nantes, France: [s.n.], 1992.

17. KOSKENNIEMI, K. *Two-level morphology:* a general computational model for word-form recognition and production. University of Helsinki: Department of General Linguistics, 1983.

18. LENNON, M. et al. An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science*, v. 3, n. 4, p. 177-183, 1981.

19. LOVINS, J. B. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, v. 11, p. 22-31, 1968.

20. MATTHEWS, P. H. *Morphology:* an introduction to the theory of word-structure. Cambridge: Cambridge University Press, 1974.

21. MOHRI, M. On some applications of finite-state automata theory to natural language processing. *Journal of Natural Language Engineering*, v. 2, n. 1, p. 61-80, 1996.

22. PAICE, C. D. Another stemmer. *ACM SIGIR Forum*, v. 24, n. 3, p. 56-61, 1990.

23. _____. A method for evaluation of stemming algorithms based on error counting. *Journal of the American Society for Information Science*, v. 47, n. 8, p. 632-649, 1996.

24. PEREIRA, F. Sentence modeling and parsing. In: COLE, R. A. et al. *Survey of the state of the art in human language technology*. Cambridge, MA: Cambridge University Press, 1997. p. 130-140.

25. POPOVIC, M.; WILLET, P. The effectiveness of stemming for natural-language access to slovene textual data. *Journal of the American Society for Information Science*, v. 43, n. 5, p. 384-90, 1992.

26. PORTER, M. F. An algorithm for suffix stripping. *Program*, v. 14, p. 130-137, 1980.

27. ROBERTSON, A. M.; WILLETT, P. Applications of n-grams in textual information systems. *Journal of Documentation*, v. 54, n. 1, p. 48-69, 1998.

28. ROCHE, E.; SCHABES, Y. Deterministic part-of-speech tagging with finite state transducers. *Computational Linguistics*, v. 21, n. 2, p. 227-253, 1995.

29. RODRIGUEZ, S.; CARRETERO, J. A formal approach to spanish morphology: the COES tools. In: CONGRESO DE LA SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL (SEPLN), 12., 1996, Sevilla. *Anales*... Sevilla: SEPLN, 1996. p. 118-126.

30. SAVOY, J. Stemming of french words based on grammatical categories. *Journal of the American Society for Information Science*, v. 44, n. 1, p. 1-9, 1993.

31. SILBERZTEIN, M. Text indexation with INTEX. *Computers and the Humanities*, v. 33, n. 3, p. 265-80, 1999.

32. SNOWBALL web site. Disponível em: <http://snowball.tartarus.org>. Acesso em: 18 jun. 2006.

33. VILARES, J. et al. Experiments at CLEF 2002 spanish monolingual track. In: ADVANCES in cross-language information retrieval. Berlin: Springer-Verlag, 2003. p. 265-271.

34. VOUTILAINEN, A. Morphological disambiguation. In: KARLSSON, F.; VOUTILAINEN, A.; HEIKKILA, J. (Ed.). *Constraint grammar:* a language-independent system for parsing unrestricted text. New York: Mouton de Gruyter, 1995. p. 165-284.

35. XU, J.; CROFT, B. Corpus-based stemming using co-occurrence of word variants. *ACM Transactions on Information Systems*, v. 16, n. 1, p. 61-81, 1998.