Recuperación de información textual
TEXT INFORMATION RETRIEVAL
2005; pp.
ISBN: 84-9750-525-5

Enfoques sintáctico y pseudo-sintáctico para la recuperación de información en español

Jesús Vilares
Carlos Gómez-Rodríguez
Miguel A. Alonso
UNIVERSIDADE DA CORUÑA

Resumen

La utilización de técnicas de procesamiento del lenguaje natural permite mejorar el rendimiento de aquellos sistemas de recuperación de información que trabajan sobre textos escritos en español. En este artículo nos centraremos en el nivel sintáctico, estudiando dos aproximaciones, una basada en la utilización de analizadores sintácticos superficiales, y otra que trata de explotar la relación difusa existente entre la proximidad de dos palabras y el hecho de que dichas palabras estén ligadas por algún tipo de relación sintáctica.

1. Introducción

A lo largo de las últimas décadas se han desarrollado diversos modelos y técnicas para realizar el emparejamiento entre consultas y documentos, tomando casi siempre como premisa que se trataba de textos escritos en inglés. El proceso de emparejamiento puede verse perturbado por múltiples factores, entre los cuales destacan la inherente ambigüedad de las lenguas naturales y la variación lingüística, tanto a nivel morfológico como sintáctico y semántico, ya que la forma de expresar conceptos utilizada en la consulta puede no corresponderse con la utilizada en todos o parte de los documentos.

En este contexto, el Grupo COLE de las universidades de Coruña y Vigo (http://www.grupocole.org) viene investigando en los últimos años sobre la posibilidad de aplicar técnicas de Procesamiento de Lenguaje Natural (NLP) para mejorar el rendimiento de los sistemas de Recuperación de Información (IR) que trabajan sobre textos escritos en español. En este artículo repasaremos brevemente parte del trabajo realizado, centrándonos principalmente en el tratamiento de la variación sintáctica.

2. Tratamiento de la variación morfológica

El primer paso en el procesamiento tanto de documentos como de consultas consiste en preprocesar el texto para, por una parte, segmentarlo adecuadamente, y por otra, realizar diversas transformaciones en el mismo y así facilitar el trabajo de las fases posteriores. Tales transformaciones comprenden el reconocimiento de números, fechas, abreviaturas, acrónimos, separación de contracciones, separación de pronombres clíticos, identificación de locuciones e identificación de nombres propios. Debemos señalar que la mayoría de sistemas de IR no suelen tratar tales fenómenos, lo cual en ocasiones redunda en normalizaciones erróneas que afectan negativamente el buen rendimiento del sistema.

Tras segmentar y preprocesar el texto, el siguiente paso consiste en etiquetarlo y lematizarlo [4]. Con ello conseguimos obtener los rasgos morfosintácticos más relevantes de cada palabra, información que será utilizada en el análisis sintáctico, al tiempo que logramos eliminar la variación flexiva, normalizando las distintas realizaciones de verbos, sustantivos y adjetivos, a una sola forma, su lema: es decir, el infinitivo en el caso de los verbos, y la formas masculina singular en el caso de sustantivos y adjetivos. Nos centramos en estas tres categorías de palabras, denominadas *palabras con contenido*, porque existe un acuerdo general respecto a que es en ellas en las que reside la semántica de un documento, aunque se podría discutir si ciertos adverbios merecen ser tenidos también en consideración.

Una vez eliminada la flexión del texto mediante la lematización, la variación morfológica queda reducida a la variación derivativa. Con objeto de eliminarla, la siguiente etapa consiste en agrupar las palabras derivables unas de otras mediante los mecanismos propios de la morfología derivativa. Cada una de estas agrupaciones recibe el nombre de *familia morfológica*. Ante la imposibilidad de crear a mano todas las familias, se optó por diseñar una herramienta automática que las generase a partir de las reglas más comunes de derivación, entre las cuales citamos la prefijación, la sufijación apreciativa, la sufijación no apreciativa, la parasíntesis y la derivación regresiva. Si bien nuestra solución comete errores, su tasa de fallo resultó ser lo suficientemente baja como para permitir su aplicación práctica en entornos de IR [8].

3. Procesamiento sintáctico

Con el fin de tratar la variación sintáctica se procede a realizar un análisis sintáctico superficial de los documentos y las consultas, mediante el cual se pretende extraer aquellos pares de palabras que se encuentran ligadas por medio de algún tipo de dependencia sintáctica. En particular, nos centramos en las dependencias que se establecen entre el núcleo de un sintagma nominal y el núcleo de sus modificadores, entre el núcleo del sujeto y el verbo, y entre el verbo y el núcleo de sus complementos. Esta tarea se podría realizar de una forma bastante precisa aplicando cualquiera de los potentes algoritmos de análisis sintáctico descritos en la literatura empleando una gramática de amplia cobertura del español. Desafortunadamente, el coste computacional de los algoritmos de análisis sintáctico tradicionales nos hace desistir de su utilización. Por si esto no fuese obstáculo suficiente, no se dispone de ninguna gramática de amplia cobertura del español, ni tampoco de un banco de árboles a partir del cual extraerla. Con estas limitaciones en mente, hemos desarrollado dos analizadores sintácticos superficiales para el español, uno basado en patrones y otro basado en cascadas de autómatas finitos.

3.1. Análisis sintáctico basado en patrones

La construcción de este analizador se realizó en dos fases. En una primera etapa se estudiaron las construcciones sintácticas a considerar, tomando como base la estructura de los sintagmas nominales del español. Para ello se levantaron los árboles correspondientes a cada una de las posibles formas de construir un sintagma nominal, teniendo en cuenta sus posibles complementos, y tratando de generar un árbol lo más bajo posible. Una vez creados los árboles, se procedió a aplicar sobre ellos las transformaciones sintácticas y morfosintácticas más frecuentes en español, dando lugar a nuevos árboles, cuyo alcance puede llegar a abarcar la oración completa en el caso de entrar en juego verbos derivados de los términos del árbol original.

En una segunda etapa los árboles resultantes fueron aplanados con el fin de obtener expresiones regulares, en base a las categorías gramaticales de los términos involucrados, que representasen de forma aproximada los sintagmas y frases generados por los árboles obtenidos en la etapa precedente. Finalmente, se estableció la correspondencia entre las palabras involucradas en las dependencias descritas por los árboles y las palabras de la expresión regular asociada.

3.2. Análisis sintáctico basado en cascadas de traductores finitos

El enfoque basado en patrones, si bien es robusto, es difícil de extender y presenta problemas en el tratamiento de dependencias distantes. Por ello optamos por desarrollar un nuevo analizador sintáctico superficial basado en cascadas. Este tipo de analizadores ha mostrado ser de utilidad en diversos ámbitos de aplicación del NLP, particularmente en el de la Extracción de Información. Su aplicación en IR, no tan estudiada, ha sido ensayada por Xerox [5] para el caso del inglés, demostrando su superioridad respecto a aproximaciones clásicas basadas en pares de palabras contiguas.

En nuestro sistema hemos utilizado una arquitectura basada en cuatro capas. Cada una de ellas ha sido implementada mediante traductores finitos, lo cual nos permite mantener una complejidad lineal respecto al tamaño del texto de entrada:

- **Capa 1**: sintagmas adverbiales y grupos verbales no perifrásticos en sus formas simples y compuestas, tanto en sus formas activas como pasivas.
- **Capa 2**: sintagmas adjetivales y grupos verbales perifrásticos
- **Capa 3**: sintagmas nominales.
- **Capa 4**: sintagmas preposicionales, diferenciando entre los precedidos por la preposición *por*, los precedidos por *de* y los restantes.

Una vez el análisis ha finalizado, tratamos de identificar las funciones sintácticas de los sintagmas reconocidos para, a continuación, identificar las siguientes dependencias: Sustantivo-Adjetivo, Sustantivo-Complemento nominal, Sujeto-Verbo predicativo, Sujeto-Atributo, Verbo predicativo activo-Objeto directo, Verbo predicativo pasivo-Agente, Verbo predicativo-Complemento circunstancial, Sujeto-Complemento circunstancial (sólo con verbos copulativos).

3.3. Indexación y emparejamiento

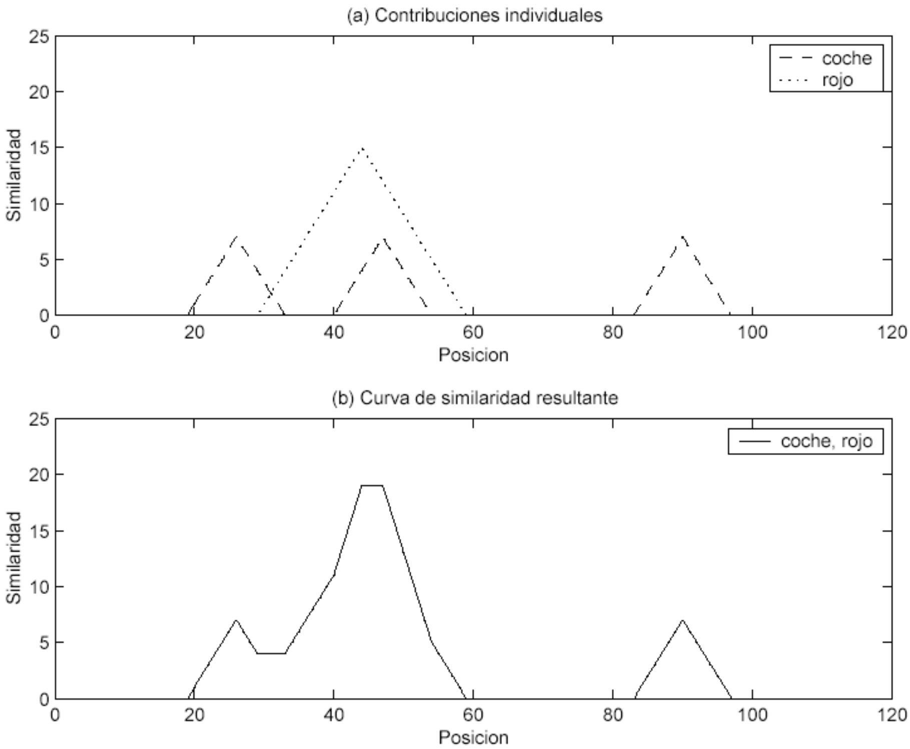
Finalmente las dependencias obtenidas tras el análisis se utilizan, conjuntamente con los lemas de las palabras individuales, para indexar los documentos. En lo que respecta al tratamiento de las consultas, hemos probado dos alternativas:

- Extraer directamente los lemas y las dependencias sintácticas presentes en la consulta.
- Extraer los lemas de las palabras con contenido, realizar una primera recuperación, y realizar un análisis de los 5 primeros documentos devueltos (mediante el método de realimentación de Rocchio) para identificar las dependencias sintácticas más relevantes. Estas dependencias son añadidas a los lemas de la consulta inicial para realizar la consulta final al sistema.

4. Un enfoque pseudo-sintáctico basado en distancias

Un enfoque alternativo a la utilización de analizadores sintácticos consiste en utilizar información pseudo-sintáctica basada en la distancia entre términos, utilizando como hipótesis de trabajo la existencia de una relación difusa entre la proximidad de dos términos y el hecho de que ambos estén ligados por algún tipo de relación sintáctica. Con ello abrimos una nueva vía de estudio que evita los problemas de los analizadores sintácticos, al no ser necesario el desarrollo de gramática o analizador alguno, y al integrar de modo consistente la información obtenida, tanto a nivel de la ocurrencia de los términos en sí, como de su proximidad, frecuentemente ligada a la existencia de una relación sintáctica entre los mismos.

En el modelo de recuperación imperante, el denominado *basado en documentos*, el usuario solicita del sistema los documentos relevantes a su necesidad de información. Frente a ello, el modelo *basado en localidad* va un paso más allá y busca las posiciones concretas del texto que pueden resultar relevantes al usuario, ya que considera la colección a indexar como una secuencia de palabras donde cada ocurrencia de un término de la consulta ejerce una influencia sobre los términos circundantes. Dichas influencias son aditivas, de forma que la contribución de diferentes ocurrencias de términos de la consulta pueden sumarse, dando lugar a una medida de similaridad, tal y como se muestra en la figura 1. Aquellas áreas del texto con mayor densidad de términos de la consulta, o con términos de mayor peso, darán lugar a picos en la curva de influencia resultante, señalando posiciones del texto potencialmente relevantes.



1. Posiciones del texto con ocurrencias de términos de la consulta y sus áreas de influencia (a), y curva de similaridad resultante (b).

La contribución a dicha similaridad por parte de una ocurrencia de un término de la consulta viene dada por una *función de contribución de similaridad* c_t definida en base a los siguientes factores [2]:

- La *forma* de la función, siendo la misma para todos los términos. En español, las que mejores resultados obtiene son las formas triangular y circular.
- La *altura* máxima de la función, que se da en la posición del término que ejerce la influencia.
- El *alcance* de la función, es decir, su radio de influencia.
- La *distancia* en palabras entre los dos términos considerados.

El planteamiento elegido a la hora de integrar la similaridad por distancias dentro de nuestro sistema de IR, fue el de la reordenación de resultados en una fase de postprocesado. Para ello hemos tomado como punto de partida la lista de documentos devuelta utilizando la indexación por lemas. A continuación, ese conjunto de documentos es procesado empleando el modelo basado en localidad. Debemos señalar que los parámetros de altura y alcance utilizados se calculan en base a los parámetros globales de la colección, y no en base a los parámetros locales al subconjunto de documentos devueltos, para así evitar los problemas derivados de la correlación que esto conllevaría.

El cálculo de relevancia de un documento se realiza en base a las medidas de similaridad asignadas a las ocurrencias de términos de la consulta que dicho documento contiene. Frente al algoritmo iterativo inicial [2], nuestra solución calcula la medida de relevancia $sim(D,Q)$ de un documento D respecto a una consulta Q como la suma de las medidas de similaridad asignadas a las ocurrencias de términos de la consulta que en él aparecen.

Los resultados iniciales [7] mostraron que la reordenación por distancias producía una caída general en el rendimiento del sistema. Dado que el número de documentos relevantes devueltos es el mismo, la caída en el rendimiento sólo podía ser debida a una peor ordenación de los resultados fruto de la aplicación del modelo basado en distancias. Es por ello que decidimos estudiar la variación en la distribución de documentos relevantes y no relevantes a los K documentos devueltos. Con este fin estudiamos los coeficientes de solapamiento de Lee [6] de los documentos relevantes y no relevantes para la ejecución inicial y la reordenada mediante distancias. Observamos que el factor de solapamiento entre documentos relevantes era mucho mayor que entre los documentos no relevantes, por lo que obedecía la *propiedad del solapamiento desigual* [6], un buen indicador de la efectividad de la fusión de ambas ejecuciones. Conforme a estas observaciones, decidimos abordar una nueva aproximación para la reordenación, esta vez basada en la fusión de datos, combinando los resultados obtenidos inicialmente mediante la indexación de lemas con los resultados obtenidos durante su reordenación con distancias: una vez fijado un valor K dado, los documentos son devueltos en el siguiente orden:

1. En primer lugar los documentos pertenecientes a la intersección de los K primeros documentos de ambas aproximaciones. Esto permite aumentar la precisión en los primeros documentos devueltos.
2. A continuación los documentos pertenecientes a los K primeros documentos de ambas aproximaciones que no estén en la intersección. Esto nos permite mejorar la ordenación de los documentos relevantes devueltos únicamente en las distancias sin perjudicar a los devueltos únicamente por la indexación de lemas.
3. Finalmente, los restantes documentos devueltos por la indexación con lemas.

Tras probar con diferentes valores de K, el mejor compromiso resultó ser $K = 50$, puesto que aunque con valores menores se obtenían picos de precisión en los primeros documentos devueltos, su comportamiento global era peor.

5. Resultados experimentales

La evaluación comparativa de las diferentes técnicas se ha realizado utilizando el motor de indexación SMART [1] con una configuración atn-ntc, que podemos considerar en cierta medida estándar, lo que facilita la reproducción y comparación de los experimentos con los realizados por otros investigadores. Las colecciones sobre las que se ha ensayado has sido obtenidas del Corpus Monolíngüe Español del CLEF (Cross-Language Evaluation Forum, http://www.clef-campaign.org) de la siguiente manera:

1. Resultados para *stemming* con realimentación(*stm*), un enfoque sintáctico con realimentación (*tsd*) y un enfoque pseudo-sintáctico (*dist*) con reordenación mediante fusión de datos con *K* = 50

colección técnica	CLEF 2001-02 · A			CLEF 2001-02 · B			CLEF 2003		
	stm	*syn*	*dist*	*stm*	*syn*	*dist*	*stm*	*syn*	*dist*
# consultas	46	46	46	45	45	45	47	47	47
# docs. rel.	3007	3007	3007	2513	2513	2513	2335	2335	2335
# docs. rel. rec.	2760	2775	**2779**	2395	**2419**	2406	**2227**	2222	2223
Pr.N.I.	0,5510	**0,5770**	0,5589	0,5281	**0,5629**	0,5497	0,5135	**0,5241**	0,5167
Pr. Doc.	0,5855	**0,6059**	0,5921	0,5754	**0,6321**	0,6305	**0,5917**	0,5852	0,5793
Pr.-R	0,5294	**0,5693**	0,5433	0,5002	**0,5207**	0,5188	0,4926	**0,4975**	0,4865
Pr. a 5	0,6913	0,7043	**0,7217**	**0,6978**	0,6933	0,6933	0,6340	0,6383	**0,6553**
Pr. a 10	0,6630	**0,7043**	**0,7065**	0,6022	**0,6422**	0,6400	0,5936	**0,6064**	0,5979
Pr. a 15	0,6174	**0,6522**	0,6449	0,5600	**0,6015**	0,6000	0,5461	0,5489	**0,5560**
Pr. a 20	0,5946	**0,6315**	0,6185	0,5256	**0,5689**	**0,5722**	0,5138	**0,5202**	0,5170
Pr. a 30	0,5449	**0,5739**	0,5652	0,4793	**0,5207**	0,5148	0,4660	**0,4716**	**0,4716**
Pr. a 100	0,3615	**0,3707**	0,3539	0,3027	**0,3249**	**0,3304**	**0,2885**	0,2847	0,2809
Pr. a 200	0,2373	**0,2404**	0,2380	0,2014	**0,2104**	0,2100	0,1791	**0,1821**	0,1800
Pr. a 500	0,1126	0,1120	**0,1126**	0,0987	**0,1018**	0,1012	**0,0874**	**0,0874**	**0,0874**
Pr. a 1000	0,0600	0,0603	**0,0604**	0,0532	**0,0538**	0,0535	**0,0474**	0,0473	0,0473

CLEF 2001-02 · A: colección para el entrenamiento y la estimación de parámetros, formada por 215.738 despachos de noticias realizados por EFE en 1994 y las consultas impares de las ediciones de 2001 y 2002 del CLEF.

CLEF 2001-02 · B: colección de evaluación formada por 215.738 despachos de noticias realizados por EFE en 1994 y las consultas pares de las ediciones de 2001 y 2002 del CLEF.

CLEF 2003: colección de evaluación formada por 454.045 despachos de noticias de EFE de 1994 y 1995 y las consultas de la edición de 2003 del CLEF.

Las consultas constan de tres campos: un breve *título*, una *descripción* de una frase y una *narración* más compleja que especifica los criterios que determinan qué documentos son relevantes. De acuerdo con [5], no hemos considerado aquellas consultas con menos de 6 documentos relevantes con el fin de eliminar ruido.

La tabla 1 muestra una comparación entre una técnica basada en palabras sin conocimiento lingüístico profundo (*stemming* con realimentación), utilizando para ello la herramienta de Snowball, la que mejor resultado obtiene para español según nuestros experimentos, disponible en http://snowball.tartarus.org; y dos técnicas de NLP, una basada en las dependencias sintácticas extraídas mediante el analizador de cascadas y otra pseudo-sintáctica basada en las distancias. Las filas muestran los datos correspondientes a las siguientes medidas: número de consultas, número de documentos relevantes, número de documentos relevantes devueltos, precisión media (no interpolada) para todos los documentos relevantes (media sobre consultas), precisión media por documento para todos los documentos relevantes (media sobre documentos), precisión-R, y precisión a los N documentos devueltos.

Como se puede observar, las técnicas basadas en NLP muestran mejores rendimientos que el stemming para casi todas las medidas. Los métodos sintáctico y pseudo-sintáctico muestran un rendimiento similar, con mejores resultados para el primero en lo que respecta a las medidas globales, y mejores para el segundo en lo que respecta a la precisión a 5 y 10 documentos.

6. Trabajo futuro

Con respecto al análisis sintáctico, intentaremos extraer dependencias más distantes utilizando el analizador basado en cascadas, al tiempo que nos planteamos las posibilidad de incorporar formas más elaboradas de análisis sintáctico robusto [9,10]. Finalmente, intentaremos experimentar nuestro enfoque en otros idiomas y en IR multilíngüe. Con respecto a la similaridad basada en distancias, estamos estudiando su conveniencia en aplicaciones tales como la búsqueda de respuestas. En ambos casos, una actividad que nos gustaría retomar es la utilización de relaciones de sinonimia ponderada [3].

Agradecimientos:

La investigación descrita en este artículo ha sido financiada en parte por el Ministerio de Educación y Ciencia y FEDER (TIN2004-07246-C03-01, TIN2004-07246-C03-02), y por la Xunta de Galicia (PGIDIT02SIN01E, PGIDIT05SIN044E, PGIDIT05PXIC10501PN).

Referencias

[1] C. Buckley. Implementation of the SMART information retrieval system. Technical report, Department of Computer Science, Cornell University, 1985. Sources avaliable in ftp://ftp.cs.cornell.edu/pub/smart.

[2] O. de Kretser and A. Moffat. Locality-based information retrieval. *Australian Computer Science Communications*, 21:177–188, 1999.

[3] S. Fernández, J. Graña, and A. Sobrino. Introducing FDSA (Fuzzy Dictionary of Synonyms and Antonyms): Applications on Information Retrieval and Stand-Alone Use. *Mathware & Soft Computing*, 10(2-3):57-70, 2003.

[4] J. Graña, F. M. Barcala, and M. A. Alonso. Compilation methods of minimal acyclic automata for large dictionaries. *Lecture Notes in Computer Science*, 2494:135–148. 2002.

[5] D. A. Hull, G. Grefenstette, B. M. Schulze, E. Gaussier, H. Schütze, and J. O. Pedersen. Xerox TREC-5 site report: Routing, filtering, NLP, and Spanish tracks. In *Pooc. of TREC-5*, 1996.

[6] J. H. Lee. Analysis of multiple evidence combination. In *Proc. SIGIR'97*, pages 267–276, Philadelphia, PA, USA, 1997. ACM.

[7] J. Vilares and M. A. Alonso. Dealing with syntactic variation through a locality-based approach. *Lecture Notes in Computer Science*, 3246:255–266. 2004.

[8] J. Vilares, D. Cabrero, and M. A. Alonso. Applying productive derivational morphology to term indexing of Spanish texts. *Lecture Notes in Computer Science*, 2004:336–348. 2001.

[9] M. Vilares, V. M. Darriba, J. Vilares, and F. J. Ribadas. A formal frame for robust parsing. *Theoretical Computer Science*, 328:171-186, 2004.

[10]M. Vilares, J. Otero and J. Graña. Regional Finite-State Error Repair. *Lecture Notes in Computer Science*, 3317:269-280, 2005.