Técnicas de Procesamiento del Lenguaje Natural en la Recuperación de Información

Pablo GAMALLO OTERO y Marcos GARCÍA GONZÁLEZ

Centro de Investigación sobre Tecnoloxías da Lingua (CITIUS)

Universidade de Santiago de Compostela

{pablo.gamallo, marcos.garcia.gonzalez}@usc.es

Telf: 881816426 / Fax: 881813602

RESUMEN

En este artículo se describe el efecto de la integración de varias técnicas basadas en el procesamiento del lenguaje natural en sistemas de recuperación de información. Se estudiarán, en concreto, métodos de lematización, anotación de categorías morfosintácticas, identificación de nombres propios compuestos y análisis en dependencias. Una evaluación a gran escala con colecciones de documentos en español nos permitirá verificar que la combinación de estas técnicas con otras menos sofisticadas, tales como tokenización y eliminación de palabras gramaticales, contribuye a una mejora significativa de la calidad de los sistemas de recuperación.

Palabras clave: Recuperación de información, procesamiento del lenguaje natural, análisis en dependencias.

Methods on Natural Language Processing for Information Retrieval

ABSTRACT

In this article, we describe the way in which different methods based on Natural Language Processing (NLP) can be integrated in Information Retrieval systems. More precisely, we will study NLP strategies such as lemmatization, PoS tagging, named entities recognition, and dependency-based parsing. A large scale evaluation on Spanish documents will be performed. This will allow us to verify whether these strategies combined with less complex NLP techniques (e.g., tokenization and stopwords removal) improve the quality of IR systems. The results reported at the end of the paper show that NLP-based strategies yield significant improvements.

Keywords: information retrieval, natural language processing, dependency-based parsing.

1. INTRODUCCIÓN

En la mayoría de los sistemas de Recuperación de Información (RI), la búsqueda de documentos relevantes a partir de una consulta depende casi exclusivamente de la presencia o ausencia de las mismas palabras (llamadas *términos* en el ámbito de RI) en los documentos de la colección objeto de la recuperación. A pesar de la enorme simplicidad lingüística de este enfoque, lo cierto es que las numerosas experiencias llevadas a cabo en los últimos 25 años demuestran su grande y sorprendente efectividad.

Ha habido numerosos intentos de mejora centradas, especialmente, en la búsqueda de nuevos modelos estadísticos. Los modelos bayesianos han intentado mejorar los modelos clásicos basados en la proximidad entre espacios vectoriales y, más recientemente, los modelos del lenguaje están intentado

mejorar los enfoques probabilísticos (como el bayesiano) más clásicos. Se han conseguido progresos, pero no de la magnitud esperada. Todos los modelos, clásicos y menos clásicos, siguen compitiendo sin que exista un claro vencedor.

Otras tentativas, bastante menos numerosas, se han centrado en la mejora de la representación lingüística del texto de documentos y consultas. Se han utilizado diferentes técnicas derivadas del Procesamiento del Lenguaje Natural (PLN), pero los resultados no han sido muy satisfactorios. Se han logrado mejoras en la calidad de los resultados obtenidos, pero a costa de un procesamiento lingüístico de los textos pesado y difícilmente adaptable a grandes colecciones de documentos, especialmente cuando lo que se pretende es recuperar páginas web en dominios en continuo crecimiento. Otro de los problemas de estas técnicas es la necesidad de encontrar herramientas de PLN costosas y sólo disponibles para algunas lenguas. Esta situación hace que los sistemas de RI basados en PLN sean demasiado dependientes de una lengua particular.

El objetivo de este artículo es comparar la efectividad de diferentes técnicas de representación lingüística para la recuperación de información, empleando métodos de PLN con diferentes grados de complejidad. Nuestras experiencias y evaluaciones se centran en colecciones de documentos en español. Los resultados obtenidos son similares a los de trabajos precedentes. En general, las técnicas de PLN permiten una ligera mejoría en la eficacia del sistema.

Nuestra principal contribución frente a experiencias anteriores está en el uso de herramientas de PLN eficientes (10 mil palabras por segundo) y robustas, capaces de analizar *gigabytes* de texto en un tiempo razonable, y que han sido elaboradas en código abierto. Esta última característica permite que la comunidad científica pueda ir adaptando periódicamente las herramientas a diferentes lenguas. En concreto, para realizar nuestras experiencias, hemos utilizado FreeLing (Atserias et al., 2006) y DepPattern (Gamallo et al., 2009), dos herramientas de análisis robustas y plurilingües en continua actualización. A raíz de los resultados que presentaremos en detalle más adelante, creemos que es importante seguir apostando por el procesamiento inteligente del lenguaje para mejorar los sistemas de

RI y, con ello, ayudar a gestionar el acceso a grandes colecciones de documentos.

Este artículo se organiza de la siguiente manera. En la sección 2 introduciremos algunas nociones básicas de RI y ofreceremos una visión general sobre el el estado-del-arte en el área. La sección 3 la dedicaremos a describir cómo fueron elaborados los sistemas de RI que, seguidamente, en la sección 4 serán evaluados y comentados. Finalizaremos esbozando las conclusiones finales del trabajo.

2. NOCIONES BÁSICAS DE RI Y ESTADO-DEL-ARTE

La esencia de la Recuperación de Información (RI) consiste en la búsqueda de documentos relevantes de una colección dada una consulta que expresa la necesidad de información del usuario. Los documentos devueltos por un sistema de RI son, en general, ordenados por grado de similitud o relevancia respecto a la consulta. Esta tarea de recuperación se lleva a cabo, por regla general, separando varios procesos:

- *Análisis y normalización*: selección de los términos que mejor representan el contenido de los documentos (y consultas) y transformación de los términos seleccionados con el objetivo de reducirlos a formas canónicas que faciliten las correspondencias posteriores en el proceso de búsqueda. Los términos pueden ser palabras, frases, n-gramas, u otras unidades.
- *Cálculo de pesos*: asignación a cada uno de los términos de un valor numérico (peso) que representa su importancia a la hora de representar el contenido de un documento.
- *Indexación*: creación de un índice que facilite el acceso a los documentos que contengan los términos que los representan.
- *Búsqueda*: proceso basado en el cálculo de correspondencias y semejanzas entre la representación de la consulta y la de cada documento. Para obtener representaciones compatibles y así permitir comparar consultas con documentos, el texto de cada consulta deberá ser analizado de la misma manera que el de los documentos.

Para diseñar un sistema de RI, existen muchas posibilidades y variantes en cuanto al modelo de recuperación utilizado (Vilares, 2005; Cacheda, 2008). Los modelos más comunes son los siguientes:

- *modelo booleano*, basado en la teoría de conjuntos,
- *modelo vectorial*, basado en el álgebra (Salton & Buckley, 1988),
- *modelos probabilísticos*, el más clásico de tipo bayesianao (Robertson & Spark-Jones, 1976) frente a los basados en modelos del lenguaje (Ponte & Croft, 1998).

Los sistemas más populares hasta la fecha se basan en el modelo vectorial con esquemas de pesos *tf-idf* (Baeza-Yates & Ribeiro, 1999), donde se penalizan los términos promiscuos que aparecen en la mayoría de los documentos de una colección. En los últimos años, numerosos estudios basados en los modelos vectorial y probabilístico han propuesto sofisticadas estrategias para mejorar los procesos de pesos, de indexación y de búsqueda. Hasta la fecha, ninguno de los modelos y estrategias más novedosos se han consolidado claramente sobre los modelos y estrategias más básicos, como por ejemplo el vectorial con pesos *tf-idf.*

Una alternativa interesante para mejorar el proceso de recuperación se centra en la expansión de las consultas. Varias estrategias han sido propuestas:

- Realimentación por relevancia (*relevance feedback method*), donde la consulta inicial se expande con términos extraídos de los documentos más relevantes recuperados a partir de la consulta inicial (Roccio,1971).
- Expansión con texto integral (*full-text expansion*), donde la expansión se realiza con textos enteros (y no sólo términos) que contienen las expresiones de la consulta inicial, recogidos tanto de textos relevantes como no relevantes (Strzalkowski et al., 1997).
- Expansión con sinónimos, donde los términos de la consulta inicial se amplían con sinónimos generalmente obtenidos de bases léxicas o tesáuricas, como WordNet. A diferencia de los dos

métodos anteriores, que han dado en general buenos resultados, la expansión con sinónimos no ha contribuido a mejorar los sistemas de RI.

Por último, también se han realizado estudios sobre la incidencia de métodos de análisis lingüístico y normalización en los sistemas de RI. Sin embargo, estos trabajos han sido más bien escasos. Por un lado, su escasez se debe al elevado coste computacional que requiere un análisis lingüístico de grandes colecciones textuales; por otro, a su baja tasa de mejora en el rendimiento del sistema. Los métodos más populares —pues conllevan un coste computacional bajo— se han centrado en técnicas de *stemming* (Hechavarría et al., Figuerola et al., 2004); es decir, en la reducción de variantes morfológicas de los términos a una forma léxica común (o raíz). Sin embargo, un problema crucial que todavía sigue vigente es el de encontrar un análisis lingüístico más profundo (sintáctico-semántico) que consiga una representación fiel de los documentos y las consultas. Algunos sistemas de RI integran estrategias de PLN para enriquecer la representación de los documentos con información lingüística. Entre las estrategias de análisis lingüístico más populares, además de la identificación de nombres propios compuestos y de la anotación morfosintáctica (*PoS tagging*), el principal foco de atención ha sido el análisis sintáctico (Strzalkowski et al., 1997; Flank,1998; Matsumura, 1999; Vilares et al., 2002; Brants, 2004; Koster, 2004; Benkö & Takona, 2005; Vilares et al., 2006). Así mismo, se están realizando estudios para mejorar, con dependencias sintácticas, sistemas probabilísticos basados en modelos del lenguaje (Gao et al., 2004; Straková & Pecina, 2010).

Como ya hemos mencionado en la introducción, nuestro objetivo será usar y evaluar un sistema de RI utilizando diferentes tipos de análisis. En particular usaremos y compararemos análisis básicos (tokenización y *stemming*) frente a análisis basados en técnicas más sofisticadas de PLN, en concreto, lematización, identificación de entidades con nombre, anotación morfosintáctica y análisis en dependencias.

3. ELABORACIÓN DE SISTEMAS RI CON DIFERENTES NIVELES DE INFORMACIÓN LINGÜÍSTICA

En esta sección, pasamos a describir sucintamente los sistemas de RI que han sido elaborados con el propósito de poder evaluar posteriormente (en la siguiente sección) la eficacia de la información lingüística que los integra.

Fueron elaborados 11 sistemas de RI con pequeñas diferencias en relación al tipo de información lingüística utilizada en el proceso de análisis y normalización. Todos los sistemas tienen en común los motores de indexación y búsqueda. Para ello, se utilizó la biblioteca Kinosearch (http://www.marvinhumphrey.com/kinosearch/), que es un módulo en Perl inspirado en las bibliotecas Lucene para Java (http://lucene.apache.org/). Kinosearch se caracteriza por implantar un modelo vectorial enriquecido con esquemas de pesos *tf-idf* y con un índice invertido. El índice invertido es la estructura de datos más popular en sistemas de recuperación.

3.1. PROCESOS LINGÜÍSTICOS INCORPORADOS

Como decíamos, las únicas diferencias entre los 11 sistemas se refieren al proceso de análisis. La mayoría de los sistemas que hemos elaborado incluye el proceso de *análisis básico* de Kinosearch. Llamamos análisis básico (*base*) a los procesos de tokenización y eliminación de *stopwords*. En el análisis básico, la estrategia para eliminar *stopwords* depende de una lista preexistente de palabras vacías. Por otro lado, todos los sistemas elaborados incluyen la normalización en minúsculas.

Los restantes procesos lingüísticos que fueron implantados en el módulo de análisis de los diferentes sistemas son los siguientes:

- Identificación de la raíz (*stem),* utilizando para ello la herramienta de Snowball, la que mejor resultado obtiene para el español según los experimentos realizados en (Vilares et al. 2005), y disponible en http://snowball.tartarus.org bajo licencia libre BSD.

- Lematización flexiva, anotación morfosintáctica e identificación de entidades con nombre, utilizando para los 3 tipos de análisis la herramienta de PLN FreeLing, disponible en http://www.lsi.upc.edu/~nlp/freeling/ bajo licencia libre GPL. Aquí, vamos a distinguir varios subprocesos:
    - lematización completa (*lemas*): identificación de todos los lemas, incluidos los nombres propios (simples y compuestos).
    - lematización parcial (*lemas'*):identificación de los lemas de las categorías léxicas conceptualmente más discriminantes, es decir, aquéllos etiquetados como verbos, nombres (incluidos los nombres propios) y adjetivos.
    - identificación de nombres propios (*NPs*). Se trata de un subconjunto de *lemas'* (que a su vez es un subconjunto de *lemas*). Incluye únicamente los lemas caracterizados como nombres propios, simples y compuestos. Los nombres propios denotan entidades con nombre, que son unidades conceptuales de naturaleza no composicional.
    - anotación morfosintáctica (*tags*): consiste en asignar una etiqueta (o PoS tag) a cada lema. Se trata por tanto de un desambiguador morfosintáctico.
- Extracción de dependencias sintácticas (*deps*), utilizando para ello el analizador DepPattern, disponible en http://gramatica.usc.es/pln/deppattern.html bajo licencia libre GPL. Se trata de un analizador robusto que devuelve un análisis parcial del texto. Puede aplicarse a 5 lenguas: español, gallego, portugués, francés e inglés. Hay que destacar que la gramática que está por detrás del analizador español contiene apenas unas 30 reglas y tiene, por tanto, una cobertura limitada.

3.2. COMBINACIÓN DE PROCESOS LINGÜÍSTICOS

Tomando en cuenta los diferentes procesos lingüísticos citados arriba (*base, stem, lemas, lemas',*

NPs, tags y *deps*), construimos 11 sistemas diferentes. Cada sistema puede contener en el módulo de análisis un único proceso lingüístico o una combinación de ellos (unidos por el símbolo "+"). Los sistemas se organizan, por lo tanto, en un árbol (c.f. Figura 1) a diferentes niveles de complejidad. Los más complejos, que se encuentran en la parte baja del árbol, integran un mayor número de procesos lingüísticos, mientras que los menos complejos sólo contienen uno de ellos y se sitúan en la parte superior.

Un sistema complejo es una combinación de procesos agrupados en sistemas más elementales. Por ejemplo, *base+deps* combina, por un lado, los procesos agrupados en *base*, es decir, tokenización y eliminación de *stopwords* y, por otro el análisis de dependencias de *deps*. Debemos reseñar que la figura 1 contiene en la parte superior tres sistemas entre corchetes (*stem*, *NPs* y *tags),* que no fueron implantados ni, por lo tanto, evaluados. Un sistema con sólo nombres propios en el índice tendría una cobertura demasiado baja. Por el contrario, un sistema con sólo PoS tags (sin lemas) recuperaría todos los documentos de la colección para cualquier consulta, lo que penalizaría drásticamente la precisión.

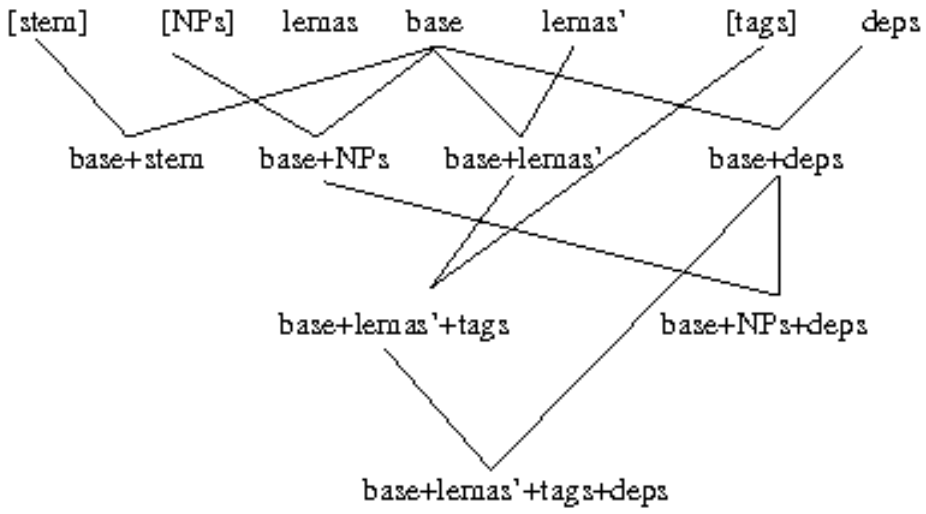

Figura 1. *Sistemas de RI organizados en función de los procesos de análisis lingüístico*

A continuación, pasamos a enumerar los 11 sistemas evaluados (sin corchetes en la figura), junto

con una pequeña descripción de sus propiedades lingüísticas:

- ***base*****,** utiliza el análisis básico de Kinosearch, es decir, tokenización y eliminación de *stopwords*.
- ***lemas****,* utiliza el lematizador (y por tanto tokenizador) de FreeLing para reducir la forma de todas las palabras a sus lema flexivo. Se incluye la identificación de nombres propios compuestos, pues estos se reducen a un único lema. No se eliminan las *stopwords*.
- ***lemas'****,* utiliza el lematizador de FreeLing y selecciona únicamente los lemas nominales, verbales y adjetivales. Entre los nominales, se incluyen los nombres propios simples y compuestos. Esta selección es, en realidad, un proceso de eliminación de *stopwords.* A diferencia de la eliminación efectuada en *base*, aquí no se recurre a una lista de palabras preexistente, sino al tipo de etiqueta morfosintáctica asociada al lema. En concreto, se eliminan todos los lemas con etiquetas consideradas no léxicas.
- **deps,** utiliza el analizador sintáctico DepPattern para identificar dependencias entre núcleos y modificadores. Los términos indexados y de consulta no son, por tanto, tokens, lemas o raíces léxicas, sino estructuras lingüísticas más complejas, concretamente relaciones sintácticas entre pares de lemas.
- ***base+stem*****,** incluye el análisis básico (tokenización y eliminación de *stopwords*), junto con el módulo de *stemming* de la herramienta Snowball. Los términos indexados y de consulta son las raíces generadas a partir de los tokens no eliminados.
- ***base+NPs*****,** incluye los tokens generados por el análisis básico junto con los nombres propios, simples y compuestos, identificados con FreeLing.
- ***base+lemas'*****,** incluye los tokens generados por el análisis básico junto con los lemas de las tres categorías léxicas: nombres, verbos y adjetivos. De esta manera, un token como “fusiones” no se reduce al lema “fusión”, sino que, por ser un nombre, va a mantenerse como token y como lema. No hay reducción de términos sino refuerzo.

- ***base+deps***, incluye los tokens generados por el análisis básico junto con las dependencias extraídas por DepPattern. Los términos indexados y de consulta son, por consiguiente, tanto tokens como relaciones sintácticas.
- ***base+lemas'+tags***, es una extensión de *base+lemas'*. Asigna a cada lema seleccionado como término la etiqueta morfosintáctica (PoS tag) correspondiente.
- ***base+NPs+deps*,** extiende *base+NPs* con la inclusión de nuevos términos representando dependencias sintácticas.
- ***base+lemas'+tags+deps***, extiende *base+lemas'+tags* con dependencias sintácticas. Es el análisis lingüístico más elaborado, ya que incluye más información lingüística y requiere más procesamiento textual que ningún otro. Se trata, por tanto, del análisis que genera el índice de términos más numeroso, englobando tres niveles de procesamiento: tokenización, anotación y análisis sintáctico.

4. EXPERIMENTOS Y EVALUACIÓN

4.1. DOCUMENTOS Y CONSULTAS

Hemos llevado a cabo un número importante de experimentos, con documentos y consultas en español, para evaluar la efectividad de los diferentes métodos de análisis. Las colecciones de documentos y las listas de consultas empleadas son las proporcionadas por TREC (Text Retrieval Conference: http://trec.nist.gov/). En concreto, hemos utilizado los recursos disponibles en los TRECs 3-5, orientados al español. Se trata de dos colecciones de textos: una contiene 250 M de noticias del diario mexicano El Norte, y la otra 300 M de teletipos de la Agencia France Presse que datan de 1994. Junto a los documentos, se les asocian tres listas de *tópicos* a partir de los cuales hemos generado las consultas. Cada tópico está formado, en general, por tres campos: un breve *título*, una *descripción* que se corresponde con una frase u oración explicativa, y una *narración* formada por un texto que

especifica los criterios de relevancia. Se generaron tres tipos de consultas:

- *Cortas*, creadas a partir del título exclusivamente
- *Medias*, creadas a partir del título y la descripción
- *Largas*, creadas a partir del tópico entero

Debemos mencionar que, para generar las consultas, no siempre hemos podido utilizar los tres campos de las listas de tópicos. En un caso no existía el campo *narración*, y en otro hemos desechado el campo *descripción* por contener únicamente oraciones en forma interrogativa, inaccesibles a nuestro analizador de dependencias cuya gramática todavía no incluye reglas que traten este tipo de oraciones.

Con el material suministrado por TREC, hemos podido llevar a cabo 7 experimentos diferentes. Para ello, hemos empleado dos colecciones de documentos (anotados A y B), tres listas de tópicos (anotadas t1, t2 y t3) y siete consultas generadas a partir de esos tópicos: tres cortas (C), dos medias (M) y dos largas (L). A continuación, nombramos y describimos de manera somera los 7 experimentos:

- **At1C**: artículos de El Norte, lista de tópicos 1 y generación de una consulta corta
- **At1M**: artículos de El Norte, lista de tópicos 1 y generación de una consulta media
- **At1L**: artículos de El Norte, lista de tópicos 1 y generación de una consulta larga
- **At2C**: artículos de El Norte, lista de tópicos 2 y generación de una consulta corta
- **At2M**: artículos de El Norte, lista de tópicos 2 y generación de una consulta media
- **Bt2C**: artículos de France Presse, lista de tópicos 3 y generación de una consulta corta
- **Bt2M**: artículos de France Presse, lista de tópicos 3 y generación de una consulta larga

4.2. MEDIDAS DE EVALUACIÓN

Las medidas de evaluación empleadas miden el rendimiento de un sistema en función de su capacidad para devolver, por un lado, *sólo* documentos relevantes (precisión) y, por otro, *todos* los documentos relevantes (cobertura) de una colección dado un conjunto de consultas. En nuestros

experimentos empleamos medidas que miden tanto la precisión como la relación entre precisión y cobertura. Las medidas utilizadas son las siguientes: precisión a los *n* documentos devueltos (donde *n*~= 5 y 10), precisión-R y precisión media de documento (*Mean Average Precision* – MAP). Los valores de estas medidas son obtenidos mediante la caja de herramientas llamada *trec_eval* (http://trec.nist.gov/trec_eval), de uso ampliamente difundido en este tipo de investigación.

La precisión a los *n* documentos devueltos se calcula contando el número de documentos relevantes devueltos entre los *n* primeros documentos y dividiendo ese número por *n*. Esta precisión refleja el rendimiento del sistema sobre los primeros documentos devueltos, por lo que es útil para medir su comportamiento sobre los resultados que un usuario estándar podría observar. Lo que le interesa a un usuario es que en la primera página mostrada por un servicio web de búsqueda se encuentren documentos relevantes. La principal desventaja de esta medida de precisión sobre los primeros documentos devueltos es que no toma en cuenta el número total de documentos relevantes para una consulta.

Una alternativa que alivia esta desventaja consiste en aplicar la *precisión-R*, que calcula la precisión de los *R* documentos devueltos, donde *R* es el número total de documentos relevantes de cada consulta. Aun así, la *precisión-R* sólo describe un punto sobre la curva precisión-cobertura. Otra medida que permite medir la cualidad del sistema a través de diferentes niveles de cobertura y, por tanto, cubrir un área extensa bajo la curva que relaciona precisión y cobertura, es la precisión media no interpolada para todos los documentos relevantes (*Mean Average Precision* – MAP). El valor es calculado como la media de precisiones obtenidas después la recuperación de cada documento relevante. Se trata, tal vez, de la medida más popular en el área de RI. Da una visión bastante completa y realista de la relación entre precisión y cobertura.

4.3. RESULTADOS

Las tablas 1, 2, 3 y 4 muestran los resultados obtenidos mediante las 4 medidas empleadas: Precisión-5, Precisión-10, Precisión-R y Precisión media no interpolada para todos los documentos relevantes (MAP). Cada tabla consta de 8 columnas correspondientes a los valores de los 7 experimentos definidos en la subsección 4.1, y una columna final (TOTAL) con la media de los valores obtenidos en los 7 experimentos. Las 11 líneas representan los sistemas de RI evaluados, cada uno de ellos caracterizado por una combinación de procesos lingüísticos en el módulo de análisis (ver subsección 3.2). En negrita, colocamos el valor más alto por cada experimento realizado, así como el conjunto de valores significativamente más elevados en la columna TOTAL.

	At1C	*At1M*	*At1L*	*At2S*	*At2M*	*Bt3C*	*At3L*	*TOTAL*
base	0.48	0.648	0.68	0.232	0.352	0.496	0.568	0.494
base+stem	0.496	0.632	0.664	0.224	0.344	0.504	0.568	0.490
lemas'	0.456	0.68	0.64	0.256	0.344	**0.536**	0.576	0.498
lemas	0.44	0.584	0.576	0.224	0.312	**0.536**	0.552	0.461
base+NPs	0.552	**0.712**	**0.752**	0.336	0.424	0.408	0.528	**0.530**
base+lemas'	**0.576**	0.632	0.616	0.312	0.472	0.464	0.608	**0.526**
base+lemas'+tags	0.544	0.632	0.648	0.288	0.392	0.512	0.584	0.514
deps	0.233	0.44	0.536	0.216	0.312	0.256	0.336	0.333
base+deps	0.504	0.64	0.688	0.312	0.408	0.504	**0.616**	**0.525**
base+NPs+deps	0.496	0.68	0.728	**0.344**	**0.456**	0.528	0.592	**0.546**
base+lemas'+tags+deps	0.552	0.656	0.68	0.312	0.424	0.488	**0.616**	**0.533**

Tabla 1. *Precisión-5 de los 11 sistemas evaluados*

	At1C	*At1M*	*At1L*	*At2S*	*At2M*	*Bt3C*	*At3L*	*TOTAL*
base	0.484	0.592	0.608	0.208	0.328	0.432	0.476	0.4469
base+stem	0.472	0.612	0.616	0.2	0.328	**0.464**	0.492	0.4549
lemas'	0.4	0.6	0.572	0.248	0.328	0.452	0.524	0.4463
lemas	0.4	0.576	0.576	0.236	0.3	0.456	0.492	0.4337
base+NPs	0.53	**0.644**	**0.692**	0.284	0.292	0.384	0.46	0.4694
base+lemas'	**0.552**	0.608	0.616	0.288	**0.416**	0.452	0.524	**0.4937**
base+lemas'+tags	**0.544**	0.592	0.616	0.26	0.388	0.456	0.516	**0.4817**
deps	0.2	0.452	0.468	0.2	0.264	0.24	0.28	0.3006
base+deps	0.452	0.64	0.652	0.288	0.388	0.42	0.524	**0.4806**
base+NPs+deps	0.504	0.608	0.652	**0.328**	0.364	0.432	**0.54**	**0.4897**
base+lemas'+tags+deps	0.54	0.6	0.636	0.28	0.4	0.46	0.524	**0.4914**

Tabla 2. *Precisión-10 de los 11 sistemas evaluados*

	At1C	*At1M*	*At1L*	*At2S*	*At2M*	*Bt3C*	*At3L*	*TOTAL*
base	0.224	0.228	0.249	0.111	0.132	0.246	0.28	0.21
base+stem	0.214	0.232	0.236	0.105	0.15	0.268	0.286	0.213
lemas'	0.189	0.219	0.221	0.128	0.136	**0.286**	0.31	0.213
lemas	0.183	0.197	0.196	0.116	0.123	0.271	0.282	0.195
base+NPs	0.212	0.229	0.246	0.114	0.097	0.258	0.293	0.207
base+lemas'	0.239	0.218	0.237	0.128	0.146	0.28	0.302	0.221
base+lemas'+tags	0.235	0.221	0.238	0.116	0.138	0.281	0.291	0.217
deps	0.092	0.197	0.209	0.08	0.114	0.115	0.15	0.137
base+deps	**0.241**	**0.293**	**0.311**	**0.138**	**0.18**	0.262	0.283	**0.244**
base+NPs+deps	0.195	0.225	0.241	0.133	0.114	0.29	**0.315**	0.216
base+lemas'+tags+deps	0.239	0.252	0.259	0.133	0.158	0.282	0.307	**0.233**

Tabla 3. *Precisión-R de los 11 sistemas evaluados*

	At1C	*At1M*	*At1L*	*At2S*	*At2M*	*Bt3C*	*At3L*	*TOTAL*
base	0.168	0.164	0.183	0.069	0.084	0.217	0.253	0.163
base+stem	0.16	0.163	0.172	0.066	0.083	0.23	0.26	0.162
lemas'	0.153	0.17	0.172	0.086	0.088	0.253	0.277	0.171
lemas	0.135	0.14	0.138	0.077	0.072	0.242	0.252	0.151
base+NPs	0.166	0.164	0.189	0.079	0.06	0.231	0.278	0.167
base+lemas'	0.185	0.163	0.178	0.086	0.093	0.25	0.271	0.175
base+lemas'+tags	0.182	0.16	0.173	0.079	0.083	0.244	0.266	0.170
deps	0.055	0.132	0.147	0.048	0.075	0.091	0.114	0.095
base+deps	0.173	**0.224**	**0.245**	0.087	**0.127**	0.219	0.273	**0.193**
base+NPs+deps	0.141	0.173	0.193	**0.092**	0.074	**0.254**	**0.317**	0.178
base+lemas'+tags+deps	**0.186**	0.187	0.203	0.09	0.102	0.246	0.282	**0.185**

Tabla 4. *Precisión media no interpolada de los 11 sistemas evaluados*

4.4. DISCUSIÓN DE LOS RESULTADOS

Los resultados obtenidos reflejan, en general, tendencias ya descritas en anteriores trabajos relacionados (Vilares, 2005; Strzalkowski, 1999). En concreto, resaltamos los siguientes aspectos:

– El mejor comportamiento global recae en la estrategia que mezcla tokens con dependencias, *base+deps,* como lo demuestra el valor alcanzado con la medida MAP, la que mejor permite conocer la eficacia global. Con esta medida, la mejora de *base+deps* con respecto al sistema

base(line) se sitúa en el 19%, lo que se puede considerar un avance significativo. En la tabla 5, mostramos el porcentaje de mejora de *base+deps* con respecto a *base* para las cuatro medidas calculadas:

	Prec-5	**Prec-10**	**Prec-R**	**MAP**
% mejora *base+deps*	+11	+11	+14	+19

Tabla 5. *Porcentaje de mejora de **base+deps** con respecto a **base***

– El uso de *stemming (base+stem)* no mejora ni empeora los resultados con respecto a *base*. Los resultados obtenidos con las cuatro medidas, tanto a partir de consultas cortas, medias o largas, son casi idénticos. Estos resultados para el español vienen a confirmar los clásicos trabajos de Harman para el inglés (1991) en los que concluyó que diferentes algoritmos de *stemming* no aumentan la efectividad en la recuperación.

– El uso de la lematización flexiva junto con la eliminación de todo lo que no son nombres, verbos y adjetivos (sistema *lemas'*) iguala los valores obtenidos por *base* y *base+stem*. En un interesante trabajo sobre el húngaro (Halácsy, 2006), lengua con una morfología muy rica, las técnicas de lematización basadas PLN superan claramente los métodos clásicos de *stemming*. Por otro lado, cabe destacar que la lista de términos indexables tras la lematización es la más pequeña de todos los sistemas que hemos evaluado. Es decir, es posible reducir significativamente los términos del índice y de la consulta sin perder efectividad. A diferencia de *lemas'*, los valores de precisión obtenidos por el sistema *lemas*, que incluye como términos los pertenecientes a todas las categorías morfosintácticas, son sensiblemente inferiores. Estos bajos valores son debidos, probablemente, a la presencia de las *stopwords* (aunque lematizadas).

- La inclusión de los nombres propios, simples y compuestos, en el sistema base (*base+NPs*) mejora sensiblemente la precisión para los 5 primeros documentos devueltos, pero tiene un comportamiento menos efectivo cuando se trata de medidas más globales, como Precisión-R y MAP.
- En general, el comportamiento de los sistemas que agrupan dos niveles de análisis lingüístico (es decir, *base+NPs, base+lemas'* y *base+deps*) tienen un comportamiento muy similar a aquellos que integran más de dos niveles: *base+lemas'+tags*, *base+NPs+deps*, y *base+lemas'+tags+deps*. De aquí se deduce que no parece necesario enriquecer excesivamente el análisis lingüístico. Dos niveles de análisis son suficientes, especialmente, si esos niveles son la tokenización y el análisis en dependencias.

- El nivel sintáctico en solitario (*deps*), sin el análisis en tokens, empobrece significativamente los resultados. Hay que resaltar, sin embargo, que el análisis sintáctico realizado se ha llevado a cabo con una pequeña gramática (+-30 reglas), de escasa cobertura, con un tratamiento básico de la coordinación y la subordinación.

- En cuanto al tamaño de las consultas, existe una mejora significativa cuando se emplean las de mayor tamaño, como así se describe en trabajos previos (Brants, 2004). También se puede observar, tal y como ya fue señalado en (Strzalkowski, 1999), que los sistemas con información sintáctica tienden a ser más efectivos cuanto mayor es el tamaño de las consultas.

Debemos puntualizar que todos los métodos de análisis propuestos son compatibles con técnicas de expansión de las consultas, tales como realimentación por relevancia, expansión con texto integral, o expansión con sinónimos.

5. CONCLUSIONES

Las técnicas más sofisticadas de PLN mejoran los sistemas de recuperación del español con respecto a técnicas básicas como la tokenización, simple eliminación de *stopwords* y el *stemming*. Sin embargo, estas mejoras, no demasiado significativas, se producen con un coste computacional algo elevado. Primero, es necesario procesar los documentos con lematizadores, etiquetadores y/o analizadores sintácticos antes de iniciar la indexación y la expansión de consultas. Además de este procesamiento lingüístico, los sistemas con varios niveles de análisis también cargan los índices y las consultas con más términos, haciendo con ello menos eficientes las tareas de indexación y de búsqueda de documentos relevantes.

Pese a todo, en los últimos años las herramientas de PLN han ido mejorando, no sólo en precisión, sino especialmente en eficiencia computacional y robustez. En cuanto a la eficiencia, las dos herramientas utilizadas en nuestro trabajo, FreeLing y DepPattern, alcanzan una velocidad de procesamiento en torno a 10 mil palabras por segundo. En lo que respecta la robustez, estas herramientas siempre devuelven un análisis, sea cual sea el ruido presente en los textos de entrada. Además, dado que se distribuyen en código abierto en ambos casos, la comunidad científica las está adaptando a diferentes lenguas, posibilitando su uso en contextos plurilingües. Por último, los algoritmos de indexación y almacenamiento de datos también han ido mejorando, lo que no penaliza excesivamente las estrategias que incorporan varios niveles de información lingüística en la caracterización de los términos indexables.

Por todo ello, nuestro trabajo contribuye a mostrar los beneficios, modestos pero beneficios al fin y al cabo, obtenidos mediante la inclusión del procesamiento lingüístico del lenguaje en los sistemas de recuperación de información.

REFERENCIAS

ATSERIAS, Jordi; Casas, Bernardino; Comelles, Elisabet; González, Meritxell; Padró, Lluís y Padró, Muntsa: "Freeling 1.3: Syntactic and semantic services in a open-source NLP library", *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*, ELRA.Genoa, Italy. Mayo, 2006.

BAEZA-YATES, Ricardo; Ribeiro Neto, Berthier: *Modern Information Retrieval,* Addison Wesley; 1999.

BENKÖ, Borbála Katalin; Katona, Tomás: "On the Efficient Indexing of Grammatical Parse Trees for Information Retrieval", *INISTA2005*, Istanbul, Turkey, 2005.

BRANTS, Thorstern: "Natural Language Processing in Information Retrieval", *14th Meeting of Computational Linguistics in the Netherlands*, 2004.

CACHEDA, Fidel: "Introducción a los modelos clásicos de Recuperación de Información", *Revista General de Información y Documentación,* 18, 2008, 365-374.

FIGUEROLA, Carlos; Zazo, Ángel; Rodríguez Vázquez de Aldana, Emilio; Alonso Berrocal, José Luis: "La Recuperación de Información en español y la normalización de términos", Inteligencia Artificial, Vol. VIII, Num. 22, 2004, 135-145.

FLANK, Sharon: "A layered approach to NLP-Based Information Retrieval", *17th international conference on Computational linguistics*, Montreal, Quebec, Canada, 1998, 397 – 403.

GAMALLO Pablo; González, Isaac: "Una gramática de dependencias basada en patrones de etiquetas", *Procesamiento del Lenguaje Natural*, 43, 2009, 315-324.

GAO, Jianfeng; Nie, Jian-Yun; Wu, Guangyuan; Cao, Guihong: "Dependence language model for information retrieval", 27th International ACM-SIGIR'04 conference on Research and Development in Information Retrieval, New York, USA, 2004, 170-177.

HALÁSCY, Péter: “Benefits of deep NLP-based lemmatization for information retrieval”, *Working Notes for the CLEF-2006 Workshop*, 2006

HECHAVARRÍA, Abdel; Pérez Suárez, Airel: “La lematización en el preprocesamiento de textos para RI. Evaluación de distintos algoritmos de lematización”, *IV Congreso de Reconocimiento de Patrones,* Cuba, Octubre, 2006.

HARMAN, D: “How effective is suffixing?” *Journal of the American Society of Information Science,* Vol. 42, No. 1, 1991, pp. 7-15.

KOSTER, Cornelis H.A.: “Head/Modifier Frames for Information Retrieval”, CICLing-2004, *Lecture Notes in Computer Science*, Vol. 2945, 2004, 420-432.

MATSUMURA, Atsushi; Takasu, Atsuhiro; Adachi, Jun: “Structured Index System at NTCIR1: Information Retrieval using Dependency Relationships between Words”, *First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition (NTCIR Workshop)*, 1999, 117-122.

PONTE, Jay; Croft, Bruce: “A language modeling approach to information retrieval”, *21st annual international ACM SIGIR conference on Research and development in information retrieval*, Melbourne, Australia, 1998, 275 – 281.

SALTON, Gerard; Buckley, Chris: “Term Weighting Approaches in Automatic Text Retrieval”, *Information Processing & Management*, 24(5), 1988, 513-523.

ROBERTSON, S.E.; Sparck-Jones, K.: “Relevance weighting of search terms”, *Journal of the American Society for Information Science*, Vol. 27, No. 3. 1976, 129-146.

ROCCIO, J.J.: “Relevance feedback in information retrieval”, En G. Salton (Ed.), T*he smart retrieval system* 313–323. Englewood Cliffs, NJ: Prentice-Hall, 1971.

STRAKOVÁ, Jana; Pecina, Pavel: “Czech Information Retrieval with Syntax-based Language Models”, S*eventh International conference on Language Resources and Evaluation (LREC 2010)*. Valletta, Malta, 2010.

STRZALOKWSKY, Tokek; Lin, Fang; Pérez-Carballo, José; Wang, Jin: "Building Effective Queries in Natural Language Information Retrieval", Fifth conference on Applied natural language processing, Washington, DC., 1997, 299-306.

VILARES, Jesús; Gómez Rodríquez, Carlos; Alonso, Miguel A.: "Enfoque sintáctico y pseudo-sintáctico para la recuperación de información en español", En: por José Angel Olivas, Alejandro Sobrino (eds.), *Recuperación de información textual*, 126-137, 2006.

VILARES, Jesús; Barcala, Mario; Alonso, Miguel A.: "Using syntactic dependency-pairs conflation to improve retrieval performance in Spanish", *Lecture Notes In Computer Science*; Vol. 2276, 2002.

VILARES, Jesús: *Aplicaciones del procesamiento del lenguaje natural en la Recuperación de Información en español*, Tesis doctoral presentada en la Universidade de A Coruña, el 20 de Mayo, 2005.