El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines

Autores: Mari Vallez (Universitat Pompeu Fabra) y Rafael Pedraza-Jimenez (Universitat Pompeu Fabra)

Citación recomendada: Mari Vallez y Rafael Pedraza-Jimenez. *El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines* [en linea]. "Hipertext.net", núm. 5, 2007. <http://www.hipertext.net>

Sumario

1. Introducción

El " Procesamiento del Lenguage Natural " (NLP) es una disciplina con una larga trayectoria. Nace en la década de 1960, como un subárea de la Inteligencia Artificial y la Lingüística, con el objeto de estudiar los problemas derivados de la generación y comprensión automática del lenguaje natural.

En sus orígenes, sus métodos tuvieron gran aceptación y éxito, no obstante, cuando sus aplicaciones fueron llevadas a la práctica, en entornos no controlados y con vocabularios genéricos, empezaron a surgir multitud de dificultades. Entre ellas, pueden mencionarse por ejemplo los problemas de polisemia y sinonimia.

En los últimos años, las aportaciones que se han hecho desde este dominio han mejorado sustancialmente, permitiendo el procesamiento de ingentes cantidades de información en formato texto con un grado de eficacia aceptable. Muestra de ello es la aplicación de estas técnicas como una componente esencial en los motores de búsqueda web, en las herramientas de traducción automática, o en la generación automática de resúmenes.

Este artículo tiene por objeto hacer una revisión de las principales características de las técnicas de procesamiento del lenguaje natural , centrándose en su aplicación a la recuperación de información y áreas afines [Strzalkowski, 1999]. Concretamente, en el segundo apartado se estudiaran los diferentes fenómenos que dificultan el procesamiento automático del lenguaje natural; en el apartado tercero se describen las principales metodologías del NLP aplicadas en la recuperación de información; en el capítulo cuarto se enuncian diversos campos de investigación relacionados con la recuperación de información y el procesamiento del lenguaje natural; a continuación se presentan las conclusiones de este trabajo y, finalmente a modo de anexo (Anexo 1), mencionamos algunas de las peculiaridades del NLP en el caso concreto del castellano.

2. Problemática del procesamiento del lenguaje natural: la variación y la ambigüedad lingüísticas

El lenguaje natural, entendido como la herramienta que utilizan las personas para expresarse, posee propiedades que merman la efectividad de los sistemas de recuperación de información textual. Estas propiedades son la variación y la ambigüedad lingüística. Cuando hablamos de la variación lingüística nos referimos a la posibilidad de utilizar diferentes palabras o expresiones para comunicar una misma idea. En cambio, la ambigüedad lingüística se produce cuando una palabra o frase permite más de una interpretación.

Ambos fenómenos inciden en el proceso de recuperación de información aunque de forma distinta. La variación lingüística provoca el silencio documental, es decir la omisión de documentos relevantes para cubrir la necesidad de información, ya que no se han utilizado los mismos términos que aparecen en el documento. En cambio, la ambigüedad implica el ruido documental, es decir la inclusión de documentos que no son significativos, ya que

se recuperan también documentos que utilizan el término pero con significado diferente al requerido. Estas dos características dificultan considerablemente el tratamiento automatizado del lenguaje. A continuación se muestran algunos ejemplos que ilustran la repercusión de estos fenómenos en el proceso de recuperación de información:

A nivel morfológico una misma palabra puede adoptar diferentes roles morfo-sintácticos en función del contexto en el que aparece, ocasionando problemas de ambigüedad (ejemplo 1).

Ejemplo 1. Deja la comida que sobre sobre la mesa de la cocina, dijo llevando el sobre en la mano.

La palabra sobre es ambigua morfológicamente ya que puede ser un sustantivo masculino singular, una preposición, y también la primera o tercera persona del presente de subjuntivo del verbo sobrar.

A nivel sintáctico , centrado en el estudio de las relaciones establecidas entre las palabras para formar unidades superiores, sintagmas y frases, se produce ambigüedad a consecuencia de la posibilidad de asociar a una frase más de una estructura sintáctica. Por otro lado, esta variación supone la posibilidad de expresar lo mismo pero cambiando el orden de la estructura sintáctica de la frase. (ejemplo 2).

Ejemplo 2. María vio a un niño con un telescopio en la ventana.

La interpretación de la dependencia de los dos sintagmas preposicionales, con un telescopio y en la ventana, otorga diferentes significados a la frase: (i) María vio a un niño que estaba en la ventana y que tenía un telescopio, (ii) María estaba en la ventana, desde donde vio a un niño que tenía un telescopio, y (iii) María estaba en la ventana, desde donde miraba con un telescopio, y vio a un niño.

A nivel semántico , donde se estudia el significado de una palabra y el de una frase a partir de los significados de cada una de las palabras que la componen. La ambigüedad se produce porque una palabra puede tener uno o varios sentidos, es el caso conocido como polisemia (ejemplo 3).

Ejemplo 3. Luís dejó el periódico en el banco.

El término banco puede tener dos significados en esta frase, (i) entidad bancaria y (ii) asiento. La interpretación de esa frase va más allá del análisis de los componentes que forman la frase, se realiza a partir del contexto en que es formulada.

Y también hay que tener en cuenta la variación léxica que hace referencia a la posibilidad de utilizar términos distintos a la hora de representar un mismo significado, es decir el fenómeno conocido como sinonimia (ejemplo 4):

Ejemplo 4: Coche / Vehículo / Automóvil.

A nivel pragmático , basado en la relación del lenguaje con el contexto en que es utilizado, en muchos casos no puede realizarse una interpretación literal y automatizada de los términos utilizados. En determinadas circunstancias, el sentido de las palabras que forman una frase tiene que interpretarse a un nivel superior recurriendo al contexto en que es formulada la frase. (ejemplo 5).

Ejemplo 5. Se moría de risa.

En esta frase no puede interpretarse literalmente el verbo morirse si no que debe entenderse en un sentido figurado.

Otra cuestión de gran importancia es la ambigüedad provocada por la anáfora, es decir, por la presencia en la oración de pronombres y adverbios que hacen referencia a algo mencionado con anterioridad (ejemplo 6).

Ejemplo 6. Ella le dijo que los pusiera debajo

La interpretación de esta frase tiene diferentes incógnitas ocasionadas por la utilización de pronombres y adverbio: ¿quién habló?, ¿a quién?, ¿qué pusiera qué?, ¿debajo de dónde?. Por tanto, para otorgar un significado a esta frase debe recurrirse nuevamente al contexto en que es formulada.

Con todos los ejemplos expuestos queda patente la complejidad del lenguaje y que su tratamiento automático no resulta fácil ni obvio.

3. El procesamiento del lenguaje natural en la recuperación de información textual

Como el lector habrá deducido, la complejidad asociada al lenguaje natural cobra especial relevancia cuando necesitamos recuperar información textual [Baeza-Yates, 1999] que satisfaga la necesidad de información de un usuario. Es por ello, que en el área de Recuperación de Información Textual las técnicas de NLP son muy utilizadas [Allan, 2000], tanto para facilitar la descripción del contenido de los documentos, como para representar la consulta formulada por el usuario, y ello, con el objetivo de comparar ambas descripciones y presentar al usuario aquellos documentos que satisfagan en mayor grado su necesidad de información [Baeza-Yates, 2004].

Dicho de otro modo, un sistema de recuperación de información textual lleva a cabo las siguientes tareas para responder a las consultas de un usuario (imagen 1):

1. Indexación de la colección de documentos: en esta fase, mediante la aplicación de técnicas de NLP, se genera un índice que contiene las descripciones de los documentos. Normalmente, cada documento es descrito mediante el conjunto de términos que, hipotéticamente, mejor representa su contenido.
2. Cuando un usuario formula una consulta el sistema la analiza, y si es necesario la transforma, con el fin de representar la necesidad de información del usuario del mismo modo que el contenido de los documentos.
3. El sistema compara la descripción de cada documento con la descripción de la consulta, y presenta al usuario aquellos documentos cuyas descripciones más se asemejan a la descripción de su consulta.
4. Los resultados suelen ser mostrados en función de su relevancia, es decir, ordenados en función del grado de similitud entre las descripciones de los documentos y de la consulta.

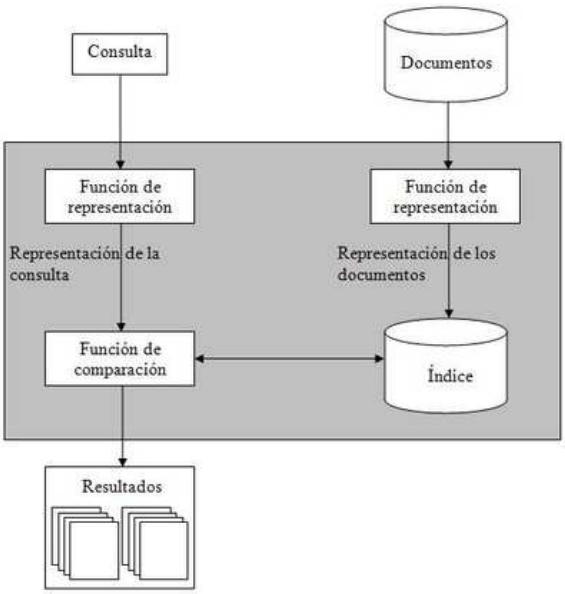

Imagen 1: Arquitectura de un sistema de recuperación de información.

De momento no existen técnicas de NLP que permitan extraer de forma inequívoca el significado de un documento o una consulta. De hecho, la comunidad científica está dividida en cuanto a los procedimientos a seguir para alcanzar tal objetivo. A continuación, detallamos el funcionamiento y las peculiaridades de las dos principales aproximaciones para el procesamiento del lenguaje natural: de un lado, la aproximación estadística, de otro, el enfoque lingüístico. Ambas propuestas difieren considerablemente, aunque en la práctica los sistemas para el procesamiento del lenguaje natural suelen utilizar una aproximación mixta, combinando técnicas propias de ambos enfoques.

3.1. Procesamiento estadístico del lenguaje natural

El procesamiento estadístico del lenguaje natural [Manning, 1999] representa el modelo clásico de los sistemas de recuperación de información, y se caracteriza porque cada documento está descrito por un conjunto de palabras clave denominadas términos índice.

Este enfoque es muy simple, y se basa en lo que se ha denominado como "bolsa de palabras" (o "bag of words"). En esta aproximación, todas las palabras de un documento se tratan como términos índices para ese documento. Además se asigna un peso a cada término en función de su importancia, determinada normalmente por su frecuencia de aparición en el documento. De este modo, no se toma en consideración el orden, la estructura, el significado, etc. de las palabras.

Estos modelos se limitan, por tanto, a emparejar las palabras en los documentos con las palabras en las consultas. Su simplicidad y eficacia los han convertido hoy en los modelos más utilizados en los sistemas de recuperación de información textual.

En este modelo el procesamiento de los documentos consta de las siguientes etapas:

1. Preprocesado de los documentos: consiste fundamentalmente en preparar los documentos para su parametrización, eliminando aquellos elementos que se consideran superfluos.
2. Parametrización: es una etapa de complejidad mínima una vez se han identificado los términos relevantes.

Consiste en realizar una cuantificación de las características (es decir, de los términos) de los documentos.

Vamos a ilustrar su funcionamiento mediante el uso del primer párrafo de este mismo documento, suponiendo que éste está etiquetado en XML. Así, el documento sobre el que aplicaríamos las técnicas de preprocesado y parametrización sería el siguiente:

```
<document document_ID="000127" source=http://www.hipertext.net>
  <title>
El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines
  </title>
  <body>
1. Introducción
El "Procesado del Lenguaje Natural" (PLN) es una disciplina con una larga trayectoria. Nace en la
década de 1960, como un subárea de la Inteligencia Artificial y la Lingüística, con el objeto de estudiar
los problemas derivados de la generación y comprensión automática del lenguaje natural.
...
  </body>
</document>
```

El preprocesado de los documentos consta de tres fases básicas:

1. Eliminación de los elementos del documento que no son objeto de indexación (o stripping), como podrían ser ciertas etiquetas o cabeceras de los documentos (ejemplo 5).

```
El Procesamiento del Lenguaje Natural en la Recuperación de Información Textual y áreas afines
1. Introducción
El "Procesado del Lenguaje Natural" (PLN) es una disciplina con una larga trayectoria. Nace en
la década de 1960, como un subárea de la Inteligencia Artificial y la Lingüística, con el objeto de
estudiar los problemas derivados de la generación y comprensión automática del lenguaje natural.
...
```

Ejemplo 5. Documento sin cabeceras ni etiquetas

2. Normalización de textos, que consiste en homogeneizar todo el texto de la colección de documentos sobre la que se trabajará, y que afecta por ejemplo a la consideración de los términos en mayúscula o minúscula; el control de determinados parámetros como cantidades numéricas o fechas; el control de abreviaturas y acrónimos, eliminación de palabras vacías mediante la aplicación de listas de palabras función (preposiciones, artículos, etc.), la identificación de N-Gramas (los términos compuestos, subrayados en el ejemplo), etc. (Ejemplo 6).

```
procesamiento lenguaje natural recuperacion de informacion textual areas afines
StringNumber introduccion
procesado lenguaje natural PLN disciplina larga trayectoria nace decada StringNumber subarea
inteligencia artificial linguistica objeto estudiar problemas derivados generacion comprension automatica
lenguaje natural
...
```

Ejemplo 6. Documento normalizado

3. Lematización de los términos, que es una parte del procesamiento lingüístico que trata de determinar el lema de cada palabra que aparece en un texto. Su objetivo es reducir una palabra a su raíz, de modo que las palabras clave de una consulta o documento se representen por sus raíces en lugar de por las palabras originales. El lema de una palabra comprende su forma básica más sus formas declinadas. Por ejemplo, "informa" podría ser el lema de "información", "informaciones", e "informar". El proceso de lematización (ejemplo 7) se lleva a cabo utilizando algoritmos de radicación (o stemming), que permiten representar de un mismo modo las distintas variantes de un término, a la vez que reducen el tamaño del vocabulario y mejoran, en consecuencia, la capacidad de almacenamiento de los sistemas y el tiempo de procesamiento de los documentos. No obstante, estos algoritmos presentan el inconveniente de no agrupar en ocasiones palabras que deberían estarlo, y viceversa, mostrar como iguales palabras que realmente son distintas.

procesa lenguaje natural recuperación de información text area afin
StringNumber introdu
procesa lenguaje natural PLN disciplina larga trayecto nace decada StringNumber subarea inteligencia artificial linguist objeto estudi problema deriva genera compren automatica lenguaje natural
…

Ejemplo 7. Documento con términos lematizados.

En cuanto a la parametrización de los documentos, consiste en asignar un peso a cada uno de los términos relevantes asociados a un documento. El peso de un término se calcula normalmente en función de su frecuencia de aparición en el documento, e indica la importancia de dicho término como descriptor del contenido de ese documento(ejemplo 8).

afin	1	linguist	1
area	1	nace	1
automatica	1	objeto	1
compren	1	PLN	1
decada	1	problema	1
deriva	1	procesa	2
disciplina	1	recuperación de inforción	1
estudi	1	StringNumber	--
genera	1	subarea	1
inteligencia artificial	1	text	1
introd	1	trayecto	1
larga	1	[...]	
lenguaje natural	3		

Ejemplo 8. Fragmento de un documento parametrizado (nótese que las frecuencias asociadas a cada término cambiarían a medida que se avanzara en la cuantificación de los restantes términos del documento).

Uno de los métodos más utilizados para estimar la importancia de un término es el conocido sistema TF.IDF (Term Frecuency, Inverse Document Frecuency). Está pensado para calcular la importancia de un término en función de su frecuencia de aparición en un documento, pero supeditado a su frecuencia de aparición total en el conjunto de documentos de la colección. Es decir, el hecho de que un término aparezca muchas veces en un documento es indicativo de que ese término es representativo del contenido del mismo, pero siempre y cuando este término no aparezca con una frecuencia muy alta en todos los documentos. De ser así, no tendría ningún valor discriminatorio (por ejemplo, en una base de datos de recetas no tendría ningún sentido representar el contenido de un documento por la palabra alimento , por más veces que ésta aparezca).

Por último, y aunque se han mencionado de pasada, es necesario describir dos técnicas muy utilizadas en el procesamiento estadístico del lenguaje natural, a saber:

a) La detección de N-Gramas: consiste en la identificación de aquellas palabras que suelen aparecer juntas (palabras compuestas, nombres propios, etc.) con el fin de tratarlas como una sola unidad conceptual. Suele hacerse estimando la probabilidad de que dos palabras que aparecen con cierta frecuencia juntas constituyan realmente un solo término (compuesto). Estas técnicas tratan de identificar términos compuestos tales como "accomodation service" o "European Union".

b) Listas de palabras vacías o palabras función (stopwords lists): una lista de palabras vacías es un listado de términos (preposiciones, determinantes, pronombres, etc.) considerados de escaso valor semántico, que cuando se identifican en un documento se eliminan, sin considerarse términos índices para la colección de textos a analizar. La supresión de todos estos términos evita los problemas de ruido documental y supone un considerable ahorro de recursos, ya que aunque se trata de un número relativamente reducido de elementos tienen una elevada tasa de frecuencia en los documentos.

3.2. Procesamiento lingüístico del lenguaje natural

Esta aproximación se basa en la aplicación de diferentes técnicas y reglas que codifican de forma explicita el conocimiento lingüístico [Sanderson, 2000]. Los documentos son analizados a partir de los diferentes niveles lingüísticos, citados ya anteriormente, por herramientas lingüísticas que incorporan al texto las anotaciones propias de cada nivel. A continuación se muestran los diferentes pasos a realizar para llevar a cabo un análisis lingüístico de los documentos aunque ello no implica que se apliquen en todos los sistemas.

El análisis morfológico es ejecutado por los etiquetadores (taggers) que asignan a cada palabra su categoría gramatical a partir de los rasgos morfológicos identificados.

Después de identificar y analizar las palabras que forman un texto, el siguiente paso consiste en ver como éstas se relacionan y combinan entre sí para formar unidades superiores, los sintagmas y las frases. Por tanto, se trata de realizar el análisis sintáctico del texto. En este punto se aplican gramáticas (parsers) que son formalismos descriptivos del lenguaje que tienen por objetivo fijar la estructura sintáctica del texto. Las técnicas empleadas para aplicar y construir las gramáticas son muy variadas y dependen del objetivo con el que se realiza el análisis sintáctico. En el caso de la recuperación de la información acostumbra a aplicarse un análisis superficial, donde

se identifican únicamente las estructuras más significativas: frases nominales, sintagmas verbales y preposicionales, entidades, etc. Este nivel de análisis suele utilizarse para optimizar recursos y no ralentizar el tiempo de respuesta de los sistemas.

A partir de la estructura sintáctica del texto, el siguiente objetivo es obtener el significado de las frases que lo componen. Se trata de conseguir la representación semántica de las frases, a partir de los elementos que la forman.

Una de las herramientas más utilizadas en el procesamiento semántico es la base de datos lexicográfica WordNet. Se trata de un léxico semántico anotado en diferentes lenguas, formado por grupos de sinónimos llamados synsets de los que se facilitan definiciones cortas y se almacenan las distintas relaciones semánticas entre estos grupos de sinónimos.

Imagen 2: Ejemplo de información semántica facilitada por WordNet. http://wordnet.princeton.edu/perl/webwn

4. Campos de investigación relacionados

Existen diferentes campos de investigación relacionados con la recuperación de información y el procesamiento del lenguaje natural que enfocan el problema desde otra perspectiva, pero cuyo objetivo final es facilitar el acceso a la información.

La extracción de información consiste en extraer las entidades, los eventos y relaciones existentes entre los elementos de un texto o de un conjunto de textos. Es una forma de acceder eficientemente a documentos grandes, pues extrae partes del documento que muestran el contenido de éste. La información generada puede utilizarse para bases de conocimiento u ontologías.

La generación de resúmenes se basa en condensar la información más relevante de un texto. Las técnicas utilizadas varían según la tasa de compresión, la finalidad del resumen, el género del texto, el idioma (o idiomas) de los textos de partida, entre otros factores.

La búsqueda de respuesta tiene como objetivo dar una respuesta concreta a la pregunta formulada por el usuario. Las necesidades de información han de estar muy definidas: fechas, lugares, etc. En este caso el procesamiento del lenguaje natural trata de identificar el tipo de respuesta a facilitar (mediante la desambiguación de la pregunta, el análisis de las restricciones fijadas, y el uso de técnicas para la extracción de información). Estos sistemas son considerados como uno de los potenciales sucesores de los actuales sistemas de recuperación de información. START natural language system es un ejemplo de estos sistemas.

La recuperación de información multilingüe consiste en la posibilidad de recuperar información aunque la pregunta y/o los documentos estén en diferentes idiomas. Son utilizados traductores automáticos de los documentos y/o de las preguntas, o mecanismos interlingua para crear interpretaciones de los documentos. Estos sistemas suponen un gran reto, pues combinan dos aspectos claves en el actual contexto de la web, la recuperación de información y el tratamiento de información multilingüe.

Para acabar, hay que citar las técnicas automáticas de clasificación de textos , consistentes en la asignación automática de un conjunto de documentos a categorías de clasificación predefinidas. La correcta descripción de las características de los documentos (normalmente mediante el uso de técnicas estadísticas - preprocesado y parametrización) determinará en gran medida la calidad de los agrupamientos/categorizaciones propuestos por estas técnicas.

5. Conclusiones

Con el objeto de dar a conocer el estado actual del Procesamiento del Lenguaje Natural se han definido, de forma

muy concisa, los principales conceptos y técnicas asociados a esta disciplina, que además se han ilustrado con sencillos ejemplos para facilitar su comprensión al lector.

Así mismo, se ha comprobado que, pese a su madurez, el NLP es una disciplina viva y en pleno desarrollo, con multitud de retos que superar fruto de la ambigüedad subyacente al lenguaje natural.

Hemos prestado especial atención a la diferenciación entre los métodos estadístico y lingüístico para el procesamiento del lenguaje natural. Pese a que las comunidades científicas defensoras de ambas aproximaciones suelen aparecer enfrentadas, las aplicaciones de NLP suelen hacer un uso combinado de las técnicas procedentes de ambos enfoques. Nuestra experiencia en esta disciplina nos hace pensar que no es posible afirmar que el uso de una u otra aproximación, e incluso del enfoque mixto, sea más o menos apropiado.

En relación con la recuperación de información, las técnicas de procesamiento estadístico son las más extendidas en las aplicaciones comerciales. No obstante, y en nuestra opinión, el comportamiento y eficacia de las distintas técnicas de NLP variará en función de la naturaleza de la tarea que tratemos de resolver, del tipo de documentos a analizar, y del coste computacional que podamos asumir.

De todo lo dicho, se deduce la necesidad de continuar trabajando con el fin de dilucidar nuevas técnicas o enfoques que contribuyan a superar las deficiencias de las existentes. Sólo así podremos alcanzar la quimera que hoy es la comprensión automática del lenguaje natural.

Por último, y a modo de anexo (Anexo 1), se han descrito algunas de las peculiaridades del procesamiento del castellano, y se han mencionado las principales iniciativas desarrolladas para su tratamiento.

6. Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio de Educación y Ciencia (España), como parte del proyecto HUM2004-03162/FILO.

7. Referencias

[Allan, 1995] J. Allan [et al.]. (1995). Recents experiments with INQUERY , en: HARMAN, D.K. de: The Fourth Text Retrieval Conference, NIST SP 500-236, Gaithersburg, Maryland.

[Allan, 2000] Allan, J. (2000). NLP for IR - Natural Language Processing for Information Retrieval http://citeseer.ist.psu.edu/308641.html [consultado 26-2-2007].

[Baeza-Yates, 1999] Baeza-Yates, R. and Ribeiro-Neto, Berthier. (1999). Modern information retrieval. Addison-Wesley Longman.

[Baeza-Yates, 2004] Baeza-Yates, R. (2004). Challenges in the Interaction of Information Retrieval and Natural Language Processing. in Proc. 5 th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2004), Seoul , Corea. Lecture Notes in Computer Science vol. 2945, pages 445-456, Springer.

[Carmona, 1998] J. Carmona [et al.]. (1998). An environment for morphosyntactic processing of unrestricted spanish text. In: LREC 98: Procedings of the First International Conference on Language Resources and Evaluation, Granada, España.

[Figuerola, 2000] C. G. Figuerola. (2000). La investigación sobre recuperación de información en español. En: C.Gonzalo García y V. García Yedra, editores, Documentación, Terminología y Traducción , pages 73-82. Síntesis, Madrid.

[Figuerola, 2004] C. G. Figuerola [et al.]. (2004). La recuperación de información en español y la normalización de términos , en: Revista Iberoamericana de Inteligencia Artificial, vol VIII, nº 22, pp. 135-145.

[Manning, 1999] Manning, C. D. and Schütze, H. (1999). Foundations of statistical natural language processing. MIT Press. Cambridge, MA: May, 680 p.

[Rodríguez, 1996] S. Rodríguez y J. Carretero. (1996). A formal approach to spanish morphology: the COES tools. En: XII Congreso de la SENLP, Sevilla, pp. 118-126.

http://www.datsi.fi.upm.es/~coes/

[Sanderson, 2000] Sanderson, M. (2000). Retrieving with good sense , In: Information Retrieval, 2, 49-69.

[Santana, 1997] O. Santana [et al.]. (1997). Flexionador y lematizador automático de formas verbales , en: Lingüística Española Actual, XIX(2), pp. 229-282.

http://protos.dis.ulpgc.es/

[Santana, 1999] O. Santana [et al.]. (1999). Flexionador y lematizador de formas nominales , en: Lingüística

Española Actual, XXI(2), pp. 253-297.

http://protos.dis.ulpgc.es/

[Strzalkowski, 1999] Strzalkowski, T. (1999). Natural Language Information Retrieval. Netherlands: Kluwer Academic Publishers.

[Vilares Ferro, 2005] Vilares Ferro, J. (2005). Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español.http://coleweb.dc.fi.udc.es/cole/library/ps/Vil2005a.pdf [consultado 11-3-07]

8. Anexo1. Peculiaridades del procesamiento del lenguaje natural en castellano

Para terminar señalaremos algunas de las características más importantes del procesado del lenguaje natural en castellano:

1. Listas de palabras vacías [Figuerola, 2000]: la construcción de estas herramientas en castellano representa un problema importante, principalmente por la ausencia de colecciones y estudios estadísticos de la lengua castellana que aconsejen o no su uso. Además, la generación de dichas listas varía en función de si son utilizadas en el procesamiento de información de carácter general o específico. Si nuestro corpus es independiente de dominio, lo más apropiado será utilizar una lista de palabras vacías formada principalmente por: determinantes, pronombres, adverbios, preposiciones y conjunciones. Si por el contrario la información a analizar depende de algún dominio, esta lista deberá ser modificada y/o ampliada por un experto del dominio en cuestión. Mencionar también que hay investigadores que han señalado la idoneidad de utilizar expresiones compuestas como elementos de las listas de palabras vacías. Concretamente [Allan, 1995] recomienda utilizar esta breve lista de "frases vacías": indicaciones de, cuáles son, cómo van, información sobre.

2. Técnicas de radicación (o stemming): la mayoría de las técnicas de recuperación de información utilizan el recuento de las frecuencias de los términos que aparecen en los documentos y las consultas. Esto implica la necesidad de normalizar dichos términos para que los recuentos puedan efectuarse de manera adecuada, tomando en consideración aquellos términos que derivan de un mismo lema o raíz. Existen varios lematizadores y analizadores morfológicos para el castellano. Entre ellos cabe destacar: las herramientas COES [Rodríguez, 1996], puestas a disposición del público por sus autores bajo licencia GNU; el analizador morfosintáctico MACO+ [Carmona, 1998], o los lematizadores FLAMON [Santana, 1997] / FLAVER [Santana, 1999].

Por último, señalar que la experimentación con este tipo de algoritmos en corpus en castellano ha demostrado que la normalización de los términos mediante las técnicas de radicación produce mejoras. Son sorprendentes los resultados obtenidos por el algoritmo S-stemmer. Este algoritmo es muy simple, y básicamente lo que hace es reducir las formas plurales al singular. En su versión original (para el inglés), este algoritmo sólo elimina las -s finales de cada palabra. En el caso del castellano, este algoritmo puede enriquecerse teniendo en cuenta que los plurales de sustantivos y adjetivos terminados en consonante se consiguen con el sufijo -es. Eliminar directamente las terminaciones en -es puede producir inconsistencia en el caso de las palabras que directamente terminan en -e en su forma singular, por lo que también es adecuado eliminar las -e finales. También se ha demostrado que produce mejoras la eliminación de las -a y -o finales, sorteando así los problemas de género.

La mayor ventaja de este algoritmo es su simplicidad, sin embargo, su inconveniente es que el S-stemmer es incapaz de distinguir sustantivos y adjetivos de otras categorías gramaticales y se aplica indiscriminadamente a todas las palabras; tampoco contempla plurales irregulares. En compensación, al tratar todas las palabras de la misma forma, no introduce ruido adicional.

Última actualització 15-03-2011