ITEM: Un Motor de Búsqueda Multilingüe Basado en Indexación Semántica

Felisa Verdejo, Julio Gonzalo, David Fernández Anselmo Peñas, Fernando López

Depto. de Lenguajes y Sistemas Informáticos UNED, Spain

{felisa, julio, david, anselmo, flopez@ieec.uned.es}

Resumen. En este artículo se presenta el motor de búsqueda multilingüe ITEM. Este sistema realiza un procesamiento léxico completo (análisis morfológico, asignación de categoría y desambiguación semántica) sobre documentos y consultas, para proporcionar índices independientes del idioma en los procesos de consulta y recuperación. Los términos de indexación son los registros del Índice interLingual de la base de datos léxica EuroWordNet/ITEM, que contiene relaciones semánticas entre palabras de diez idiomas de la Comunidad Europea (el motor de búsqueda soporta en estos momentos Español, Inglés y Catalán.) Esta aplicación proporciona una forma de comparar, en contexto, el comportamiento de distintas estrategias de Procesamiento de Lenguaje Natural en Recuperación de Información Multilingüe y, en particular, distintas formas de desambiguar los términos de la consulta para realizar indexación conceptual o realizar una expansión a otros idiomas.

1 Introducción

En este artículo se presenta el motor de búsqueda multilingüe ITEM. El proyecto ITEM (`http://sensei.ieec.uned.es/item/principal.htm`), financiado por la CICyT, comenzó en 1996 y finalizó en 1999, y tenía dos objetivos principales: 1) integrar y desarrollar diversos recursos lingüísticos y herramientas de Procesamiento de Lenguaje Natural (PLN) para el castellano, el inglés, el catalán y el euskara, y 2) demostrar la aplicación de esos recursos y herramientas en un sistema de recuperación de información multilingüe.

El motor de búsqueda multilingüe es uno de los resultados del proyecto ITEM, y se puede consultar en `http://terral.ieec.uned.es/clir`. El motor de búsqueda permite consultar y recuperar documentos en tres idiomas (castellano, catalán e inglés); el usuario puede seleccionar entre distintas opciones de procesado, así como refinar los resultados del procesamiento léxico.

En su estado actual, el motor de búsqueda integra bases de datos léxicas y módulos de PLN (analizadores morfológicos, lematizadores, etiquetadores de categoría gramatical y desambiguadores) para el castellano, catalán e inglés [9], [10], [11]. También se espera poder incorporar el euskara en un futuro próximo.

Los recursos lingüísticos desarrollados en el proyecto -e integrados en el motor de búsqueda- tienen una relación muy estrecha con el proyecto EuroWordnet [13], e incluyen una base de datos léxica con relaciones semánticas entre palabras del inglés, castellano, catalán y euskara que sigue de cerca el diseño de EuroWordnet, que llamaremos en adelante base de datos EWN/ITEM [2], [5]. La característica principal de la red semántica EWN/ITEM es un índice interlingua mediante el cual se conectan todos los wordnets monolingües. Ese índice permite encontrar conceptos equivalentes entre cualquier par de idiomas de la base de datos. El índice interlingua es el superconjunto de todos los conceptos que aparecen en todos los idiomas.

En el motor de búsqueda, los documentos en la colección textual son procesados por completo para obtener la información léxica que permite la indexación conceptual de cada documento en términos del índice Interlingua EWN/ITEM. La colección usada en la interfaz web cubre unos 10000 artículos de la sección internacional de los periódicos Washington Post (en inglés), El País (en español) y El Periódico (en catalán), desde Abril de 1998 hasta Mayo de 1999. Los recursos y herramientas, sin embargo, no están adaptados a ningún dominio en particular.

El objetivo de esta aplicación es proporcionar una forma de comparar en contexto el comportamiento de distintas estrategias de PLN para la recuperación de información multilingüe (CLIR) y, en particular, distintas estrategias de desambiguación semántica (WSD) para la indexación conceptual y la expansión de la consulta en otros idiomas.

Las dos secciones siguientes describen dos sistemas distintos de realizar recuperación multilingüe utilizando la base de datos EWN/ITEM: El primero consiste en traducir la consulta del idioma original a los idiomas de la base documental, utilizando los enlaces del índice interlingua. La segunda consiste en proyectar tanto las consultas como los documentos en el índice Interlingua como un espacio de indexación independiente del idioma. La sección 4 enumera brevemente las herramientas léxicas integradas en el motor de búsqueda. La sección 5 describe la interfaz web para el motor de búsqueda y, finalmente, la última sección presenta algunas conclusiones y futuras mejoras al motor de búsqueda.

2 Traducción de la Consulta vía EuroWordnet

El motor de búsqueda ITEM implementa dos enfoques alternativos para la recuperación de información multilingüe. El primero consiste en traducir la consulta de su idioma original a los otros dos idiomas de búsqueda posibles, y en realizar a continuación tres procesos de búsqueda monolingüe con el motor de búsqueda standard INQUERY [4].

Este enfoque se aproxima a la recuperación multilingüe basada en diccionarios, donde cada palabra original se sustituye por las traducciones obtenidas en un diccionario bilingüe, después de ciertos filtros estadísticos (en especial para traducir expresiones multipalabra). Sin embargo, el uso de la red semántica

EWN/ITEM ofrece cierto número de ventajas sobre un conjunto de diccionarios bilingües que cubran todas las direcciones posibles de traducción.

Los wordnets de español, catalán e inglés juegan el papel de seis diccionarios bilingües. La ventaja de disponer de un índice interlingua crece rápidamente con el número de idiomas contemplados, y el número potencial de idiomas para el motor de búsqueda es actualmente 10 (inglés, español, catalán, euskara y el resto de idiomas de EWN: holandés, italiano, francés, alemán, estonio y checo).

La desambiguación semántica se puede realizar explícitamente en un nivel independiente del idioma (la representación del índice interlingua). La desambiguación proporciona los registros del ILI adecuados, y los registros del ILI están ligados a los conjuntos de palabras sinónimas en cada idioma contemplado.

Las relaciones semánticas en la base de datos léxica EWN/ITEM permiten una expansión controlada con términos semánticamente relacionados: hipónimos, merónimos, etc.

Las relaciones de hiperonimia/hiponimia en el índice interlingua permiten obtener traducciones aproximadas para los términos de la consulta que no tienen equivalentes en el (o los) idiomas objetivo. Por ejemplo, ``governor's race'' no tiene equivalente en España, y por tanto no hay ningún término equivalente en español. Sin embargo, ``governor's race'' se puede ligar a ``elecciones'' a través de ``elections'', que es el hiperónimo directo de ``governor's race''. Otro ejemplo es ``grand jury'', que no tiene equivalente en español pero puede tener como traducción aproximada ``jurado'', como un equivalente en español para el concepto ``jury'', que es hiperónimo directo de ``grand jury''. El proceso de traducción de la consulta se ilustra en la sección 5.

3 Indexación Conceptual

Traducir la consulta es el sistema más popular de realizar búsquedas multilingües, ya que el coste computacional asociado es mucho menor que el de traducir o procesar los documentos.

Sin embargo, la disponibilidad de la base de datos EWN/ITEM y su índice interlingua permite explorar una alternativa atractiva a la traducción de consultas: la utilización de los registros del índice interlingua para indexar tanto las consultas como los documentos, acercándose a la comparación de conceptos más que a la comparación de palabras clave. Las ventajas más apreciables frente a la traducción de consultas son:

- La comparación entre documentos y consultas se hace a un nivel conceptual, evitando los problemas de polisemia de las palabras como términos de indexación, y permitiendo identificar términos sinónimos como el mismo término de indexación.
- La comparación se hace en un espacio independiente del idioma, simplificando el problema de fusionar resultados de varias búsquedas monolingües en varias colecciones textuales distintas. Todos los textos

pueden ser indexados con los mismos términos de indexación, sin importar el idioma original en que han sido escritos.

- En un sistema interactivo de recuperación multilingüe, el refinamiento de la consulta se puede hacer, en gran parte, a un nivel conceptual, evitando la necesidad de realizar ese refinamiento para cada idioma contemplado en la base documental.

4 Procesamiento Léxico

Para realizar cualquiera de los dos enfoques de recuperación multilingüe descritos, es necesario realizar un procesamiento léxico profundo asociado a la base de datos EWN/ITEM.

En el motor de búsqueda ITEM, los documentos (en la aproximación de indexación conceptual) y las consultas (en los dos enfoques) son procesados por una cascada de analizadores léxicos, en la que sólo los lematizadores y los etiquetadores de categoría gramatical son dependientes del idioma:

Lematización y etiquetación de categoría. El español y el catalán son procesados con el analizador morfológico MACO+ y el etiquetador RELAX [9], [11]. El inglés se procesa con una versión del Brill tagger [3] y con el lematizador de WordNet 1.5 [8].

Detección de expresiones multipalabra. La detección de expresiones multipalabra es, en este enfoque, una tarea independiente del idioma que considera sólo las expresiones incluidas en la base de datos léxica EWN/ITEM.

Desambiguación semántica. Todos los nombres en el documento (o en la consulta) se desambiguan, asignando probabilidades para cada sentido posible de cada nombre. Actualmente estamos usando una implementación muy eficiente de un algoritmo no supervisado inspirado en [1] que sólo usa información jerárquica y medidas de distancia conceptual para realizar la desambiguación. El hecho de utilizar un algoritmo no supervisado fue tomado como una restricción necesaria en nuestro sistema, ya que no existen -que sepamos- corpora anotados semánticamente para el español ni el catalán. Nuestro sistema de desambiguación semántica se comporta (para la colección de prueba Semcor) mucho peor que la heurística del sentido más frecuente al escoger el sentido adecuado, pero sin embargo su distribución de probabilidad parece comportarse ligeramente mejor que escoger, simplemente, el sentido más frecuente en una tarea de recuperación de información [12]. En el estado actual del motor de búsqueda, se ofrecen dos opciones adicionales para desambiguar términos (tanto en documentos como en consultas): una opción ``First sense" que siempre toma el primer sentido de la base de datos léxica EWN/ITEM, y una opción ``All senses" que toma todos los posibles sentidos de cada nombre como índices igualmente válidos. Es necesario hacer notar, sin embargo, que en los wordnets español y catalán el primer sentido no es necesariamente el más frecuente.

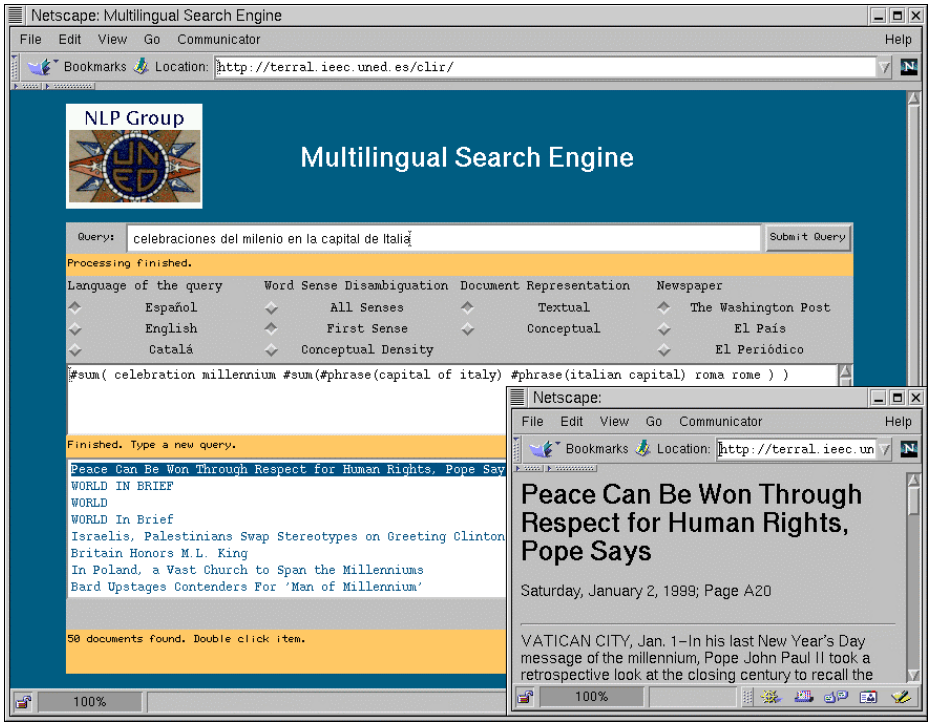

Figura 1. Interfaz del motor de búsqueda ITEM

Merece la pena mencionar que la base de datos EWN/ITEM no ha sido enriquecida ni adaptada al dominio concreto que se usa en la demo de Internet (en particular, la sección internacional de varios periódicos). Aunque este hecho perjudica claramente determinados tipos de consultas, nuestra intención era precisamente medir lo que se puede conseguir, y lo que no se puede esperar, de una base de datos léxica de gran tamaño en una aplicación de este tipo. Una evaluación del coste de adaptar -semiautomáticamente- la base de datos léxica a dominios particulares de búsqueda será realizada en próximas extensiones al proyecto.

5 La Interfaz de Búsqueda

La evaluación off-line de nuestro enfoque de recuperación multilingüe, en términos de colecciones de prueba y medidas de precisión/cobertura se ha presentado en otros artículos [6], [7], [12]. Sin embargo, las medidas de precisión y cobertura, por si solas, tienden a ocultar tanto los beneficios como los problemas asociados al procesamiento léxico, dependiendo del tipo de consulta.

El motor de búsqueda ITEM, por contraste, proporciona una experiencia directa con el uso de una red semántica multilingüe, así como con las peculiaridades de usar un procesamiento léxico exhaustivo. La interfaz web al motor de búsqueda

permite al usuario ajustar parámetros relacionados con el procesamiento de lenguaje natural de la consulta (y los documentos), refinar los resultados del procesamiento de la consulta y comparar resultados con distintos parámetros. En la figura 1 se puede ver el aspecto de la interfaz web. La caja de texto superior de la interfaz se usa para realizar la consulta. Las opciones de procesamiento se seleccionan en los botones que están inmediatamente debajo. Las consisten en:

Idioma de la consulta. Las posibilidades son inglés, español o catalán.

Idioma de los documentos. Depende del periódico que se selecciona: español para ``El País", inglés para el ``Washington Post" y catalán para ``El Periódico".

Representación de los documentos. Las opciones son ``textual" o ``conceptual". En la opción textual, la consulta se traslada al idioma de la colección a través del índice interlingua, y entonces se realiza una búsqueda estándar sobre la base documental original. En la opción conceptual, el procesado de la consulta se detiene al nivel de representación conceptual, y se compara con la representación conceptual de documentos en términos del índice interlingua.

Desambiguación conceptual. El usuario puede elegir entre considerar todos los posibles sentidos de cada palabra (``All Senses"), tomar siempre el sentido más frecuente (``First sense") o usar el algoritmo de desambiguación descrito más arriba (``Conceptual Density"): Si la representación del documento seleccionado es ``textual", el criterio de desambiguación seleccionado sólo afecta a la traducción de la consulta, restringiendo el número de conceptos que se traducen al idioma objetivo. Si la representación de documentos elegida es ``conceptual", entonces la elección del sistema de desambiguación sólo afecta a cómo se indexan los documentos. La distribución de probabilidad obtenida por el programa de desambiguación sobre los conceptos candidatos para cada nombre se usa de esta manera: en la representación de documentos, los sentidos que tienen al menos un 80% del valor del sentido más probable se toman como índices válidos, y el resto se descarta. Para las consultas (en las que el contexto es, usualmente, demasiado pequeño para una desambiguación fiable), se conservan todos los sentidos en la consulta expandida, pero pesándolos de acuerdo con sus probabilidades.

Detección de expresiones multipalabra. Cuando se activa, las expresiones multipalabra en documentos y consultas se toman como unidades de indexación únicas. En un futuro próximo, dispondremos de un tratamiento más sofisticado de las expresiones multipalabra en la que se distinguirán distintos tipos de ellas. Un sintagma se considerará unidad de indexación sólo en el caso de compuestos exocéntricos, como ``fisher cat", en los que el significado de los componentes no está relacionado con el significado de la expresión completa. Otras expresiones, como ``abstract art", se tratarán combinando los significados de las palabras que las componen. Una vez que se procesa la consulta, el sistema proporciona:

- En la caja de texto bajo los botones de selección, se muestra la consulta expandida con los términos de búsqueda en el idioma objetivo (para recuperación por texto, como en la figura) o los registros del índice interlingua (para recuperación por conceptos). El usuario puede refinar la consulta procesada añadiendo / eliminando términos (o conceptos) en esta

caja, y realizando una búsqueda directa (sin más procesamiento léxico) con esta consulta refinada en el idioma objetivo (o en la representación conceptual).

- En la caja inferior, una lista ordenada de documentos relevantes en el periódico seleccionado. El usuario puede pulsar en el título para ver el texto completo.

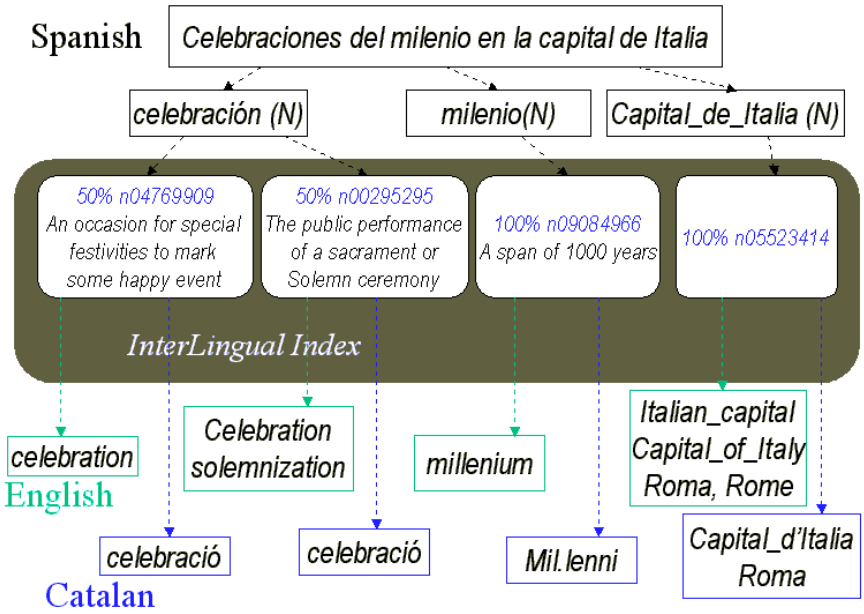

Figura 2. Ejemplo de procesamiento léxico de una consulta**.**

La consulta procesada (que se pasa al motor de búsqueda estándar INQUERY) se obtiene a partir del procesamiento léxico (tal y como se describe en la sección 4) y algún post-procesado para codificar la información léxica de acuerdo con la sintaxis de consultas de INQUERY (usando, vbg., los operadores de expresión multipalabra, #phrase y sinonimia, #syn). Por ejemplo, la consulta en español:

celebraciones del milenio en la capital de Italia

produce, tras el procesamiento léxico, el resultado de la figura 2. Los pasos principales de procesamiento son: 1) identificación de expresiones multipalabra, lemas y categoría gramatical adecuada, 2) representación en términos del índice interlingua, como probabilidades asignadas al algoritmo de desambiguación, y 3) expansión en los idiomas objetivo. La información en los pasos 2 y 3 se usa para construir la consulta final de acuerdo con las opciones seleccionadas por el usuario. Por ejemplo, cuando la representación documental es ``textual" y la opción de

desambiguación es ``All senses'', el resultado es:

#sum(#sum(celebration #sum(celebration solemnization)) millennium

#sum(#phrase(capital of italy) #phrase(italian capital) roma rome))

Cuando la opción de desambiguación es ``First Sense'', el resultado sería

#sum(celebration millennium #sum(#phrase(capital of italy)

#phrase(italian capital) roma rome))

y cuando la opción de desambiguación es ``Conceptual density'', se usan los pesos en la construcción de la consulta:

#sum(#wsum(100 50 celebration 50 #sum(celebration solemnization).

#wsum(100 100 millennium) #wsum(100 100 #sum(#phrase(capital ofitaly)

#phrase(italian capital) roma rome)))

Si la representación de los documentos es ``Conceptual'' y la estrategia de desambiguación es, por ejemplo, ``First Sense'', la consulta se convierte en:

#sum(n04769909 n09084966 n05523414)

donde, por ejemplo, n05523414 representa el registro del índice interlingua:

```
n05523414
English: Rome, Roma, Italian capital, capital of Italy
Spanish: capital de Italia, Roma
Catalan: capital d'Itàlia, Roma
    => hypernym: n05483778
          English: national capital
          Spanish: capital de nació
          Catalan: capital de nación
```

El usuario puede entonces refinar la consulta, ya sea reformulando la consulta original o, lo que es más interesante, añadiendo o eliminando términos directamente de la consulta expandida. Por ejemplo, el usuario puede escoger una expansión ``All senses'' en el idioma objetivo, y después eliminar manualmente aquellas traducciones que no son apropiadas, y entonces consultar directamente a la base textual con el resultado. En los próximos meses esperamos ofrecer también una descripción gráfica de los conceptos involucrados, así como sugerencias de conceptos / términos relacionados.

6 Conclusiones y Trabajo Futuro

Cierto número de experimentos en relación con el uso de conceptos en recuperación de información textual [6], [7], [12], junto con la experiencia directa utilizando la interfaz de búsqueda, nos permiten extraer unas primeras conclusiones sobre la calidad de los recursos y herramientas empleados, y sobre la utilidad de estos

recursos en recuperación de información multilingüe. Para traducción/expansión de consultas, las bases de datos de EuroWordnet y EWN/ITEM ofrecen características interesantes por comparación con los diccionarios bilingües. Las relaciones semánticas en el Índice Interlingua permite encontrar traducciones aproximadas cuando no se puede encontrar una equivalencia directa (o no existe en el idioma objetivo), y permite sugerir otros términos relacionados semánticamente. Sin embargo, como en el caso de los diccionarios electrónicos, es necesario adaptar el sistema al dominio para obtener traducciones adecuadas a términos y significados propios del dominio, especialmente con expresiones multipalabra. La indexación conceptual es una opción atractiva, a priori, para realizar recuperación multilingüe. Pero deben resolverse dos retos principales :

- Los distintos sentidos, para una palabra dada, que se consideran en la base de datos léxica deberían reflejar diferencias de uso en contexto. En caso contrario, esas distinciones sólo perturban el proceso de recuperación de información. Este requisito significa, en la práctica, que debemos encontrar formas de agrupar los sentidos de EWN/ITEM, buscando un grano menos fino y más apropiado para los propósitos de Recuperación de Información. Nuestra experiencia con el motor de búsqueda ITEM nos confirma que la granularidad apropiada para distinguir sentidos depende de la aplicación, y para Recuperación de Información es crucial tener la granularidad adecuada.
- La desambiguación semántica es todavía un tema de investigación abierto especialmente cuando la tarea es realizar una anotación semántica exhaustiva de los nombres y verbos en una colección de textos en tres idiomas distintos. Nuestro algoritmo satisface los requisitos de cobertura, ya que es no supervisado e independiente del idioma (para idiomas que tenga una base de datos léxica tipo WordNet). Sin embargo, no es suficientemente preciso, al igual que el resto de algoritmos no supervisados que conocemos. Sin embargo, parece que la distribución de probabilidades que produce nuestro sistema funciona mejor que la heurística del ``primer sentido" en Recuperación de Información, a pesar de detectar el sentido más adecuado peor que esa heurística.

Creemos que la forma óptima de beneficiarse del software de Ingeniería Lingüística en Recuperación de Información multilingüe es de forma integrada con interfaces interactivas de búsqueda, capaces de sugerir términos y conceptos, y de guiar al usuario para obtener una combinación óptima de términos para sus necesidades de información. Una ventaja de tomar un enfoque de recuperación basada en conceptos en recuperación interactiva es que la selección de los conceptos adecuados se hace sólo una vez para todos los idiomas objetivo, mientras que la selección de traducciones adecuadas debe hacerse una vez por cada idioma objetivo. La búsqueda de información multilingüe es uno de los retos para la investigación en PLN en la llamada ``sociedad de la información", y una de las razones por las que las comunidades científicas de Recuperación de Información y PLN se están aproximando la una a la otra. El motor de búsqueda ITEM es una contribución que permite realizar tests cuantitativos y cualitativos sobre el impacto de las bases de

datos léxicas y las herramientas de procesamiento de lenguaje natural en los sistemas de recuperación de información mono y multilingües.

7 Agradecimientos

Este trabajo ha sido financiado por la Comisión Interministerial de Ciencia y Tecnología (CICyT), proyecto ITEM (TIC96-1243-C03-01), y también parcialmente por la Comisión Europea, proyecto EuroWordnet (Language Engineering #4003).

Referencias

1. Eneko Agirre and German Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of COLING'96*.
2. L. Benítez, S. Cervell, G. Escudero, M. López, G. Rigau, and M. Taulé. 1998. Methods and tools for building the catalan wordnet. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98).*
3. E. Brill. 1992. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*.
4. J. Callan, B. Croft, and S. Harding. 1992. The INQUERY retrieval system. In *Proceedings of the 3rd Int. Conference on Database and Expert Systems applications.*
5. X. Farreres, G. Rigau, and H. Rodríguez. 1998. Using wordnet for building wordnets. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*.
6. J. Gonzalo, A. Peñas, and F. Verdejo. 1999a. Lexical ambiguity and information retrieval revisited. In *Proceedings of EMNLP/VLC'99 Conference*.
7. J. Gonzalo, F. Verdejo, and I. Chugur. 1999b. Using EuroWordNet in a concept-based approach to cross-language text retrieval. *Applied Artificial Intelligence*, 13(7):647--678
8. G. Miller, C. Beckwith, D. Fellbaum, D. Gross, and K. Miller. 1990. Five papers on Wordnet, CSL report 43. Technical report, Cognitive Science Laboratory, Princeton University.
9. L. Marquez and L. Padró. 1997. A flexible POS tagger using an automatically acquired language model. *Proceedings of ACL/EACL'97*.
10. G. Rigau, J. Atserias, and E. Agirre. 1997. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proceedings of ACL/EACL'97*.
11. H. Rodríguez, M. Taulé, and J. Turmo. 1998. An environment for morphosyntactic processing of unrestricted spanish text. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*.
12. P. Vossen, W. Peters, and J. Gonzalo. 1999. Towards a universal index of meaning. In *Proceedings of SIGLEX'99*.
13. P. Vossen. 1998. EuroWordNet: a multilingual database with lexica semantic networks. Kluwer Academic Publishers.