Diploma de Estudios Avanzados

Víctor Josué Peinado Herencia
Licenciado en Lingüística por la UCM

Grupo de Procesamiento del Lenguaje Natural
Dept. de Lenguajes y Sistemas Informáticos
ETS. de Informática - UNED

14 de julio de 2004

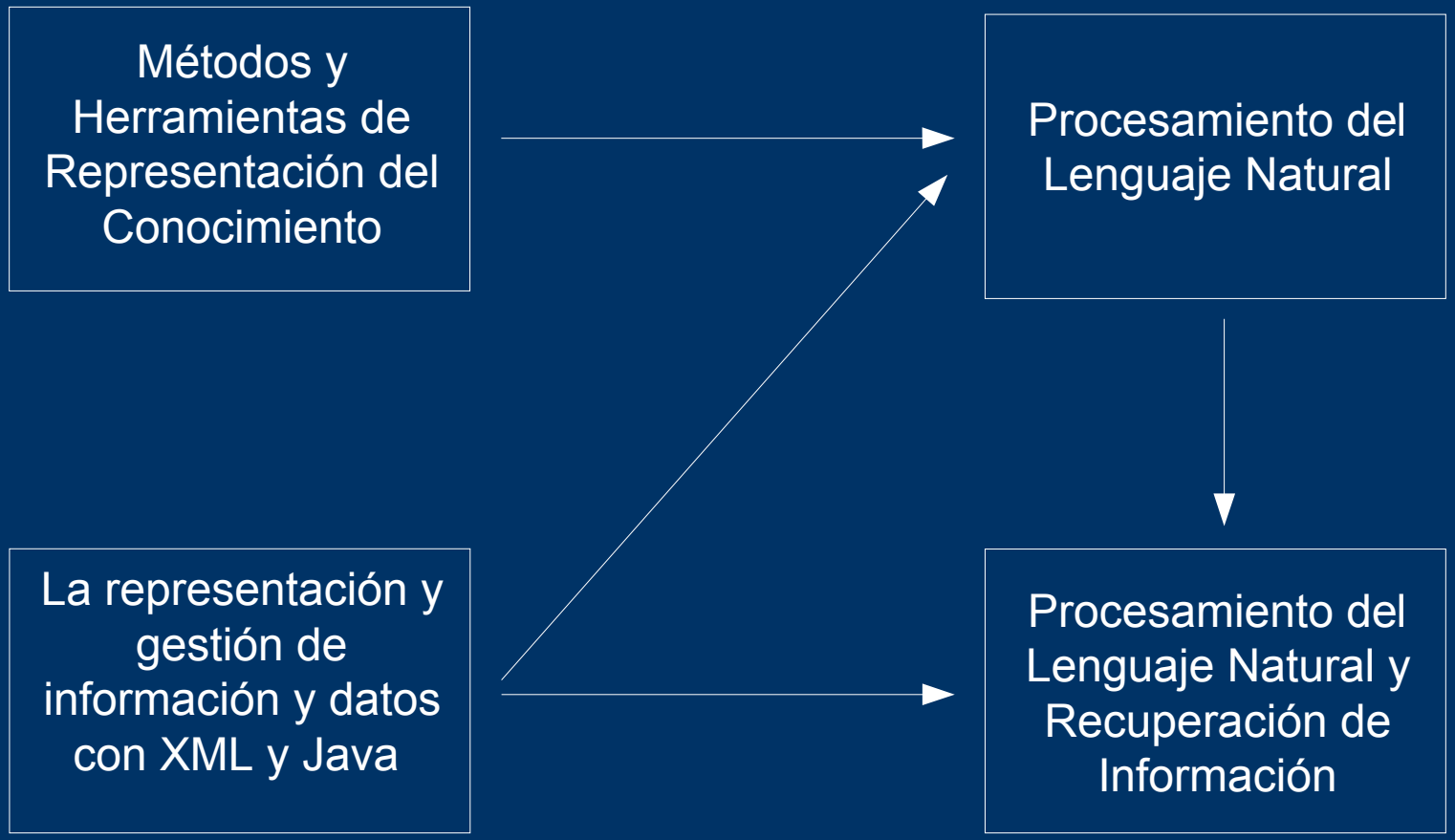
Índice

- Periodo de docencia
  - Formación previa
  - Cursos de doctorado realizados
- Periodo de investigación
  - Colaboración en proyectos y seminarios realizados
  - Trabajo de investigación
    - Antecedentes: sintagmas nominales en WTB e iCLEF
    - Objetivos
    - Fusión del diccionario
    - Experimentos de RI
    - Resultados
    - Conclusiones
- Publicaciones y Referencias

Periodo de docencia
Formación previa

- Licenciado en Lingüística en septiembre de 2002 por la UCM con la especialidad de Lingüística Computacional.
- Doctorando en el grupo de Procesamiento del Lenguaje Natural de la UNED.
- Becario predoctoral de la UNED desde abril de 2003.

Periodo de docencia
Cursos de doctorado realizados



Periodo de investigación
Participación en proyectos

- Elaboración de córpora y *testbeds*.
  - QA CLEF 2003: corpus de 450 preguntas y respuestas en cuatro lenguas (inglés, italiano, holandés, español).
  - Senseval3 (etiquetado semántico para el corpus de entrenamiento en español).
- Participación y propuestas
  - ImageCLEF 2004: RI Translingüe utilizando sintagmas nominales sobre la información contenida en pies de imagénes.
  - iCLEF 2004: búsqueda interactiva de respuestas sobre documentos completos y sobre párrafos con localización de entidades.

Periodo de investigación
Seminarios impartidos

- *Modelos de Lenguaje Estadísticos*
  - Miden probabilidad de aparición de secuencias de palabras.
  - *Toolkits* de PLN estadístico (CMU y SRI).
  - MLs en traducción automática (modelos de IBM).
  - MLs en recuperación de información.
- *Métricas de Evaluación Automática*
  - BLEU: Evaluación de sistemas de traducción automática (Papineni et al, 2002).
  - ROUGE: Evaluación de sistemas de resumen automático (Lin & Hovy, 2003).

Periodo de investigación
Estancias en el extranjero

- Sept – Dic de 2004: estancia en el *Institute of Information Sciences* de la *University of Southern California*, bajo la supervisión de Eduard Hovy.
- Tema: Análisis de Redes Sociales en grandes corporaciones a partir del contenido de mensajes de correo electrónico.

Recuperación de Información Translingüe

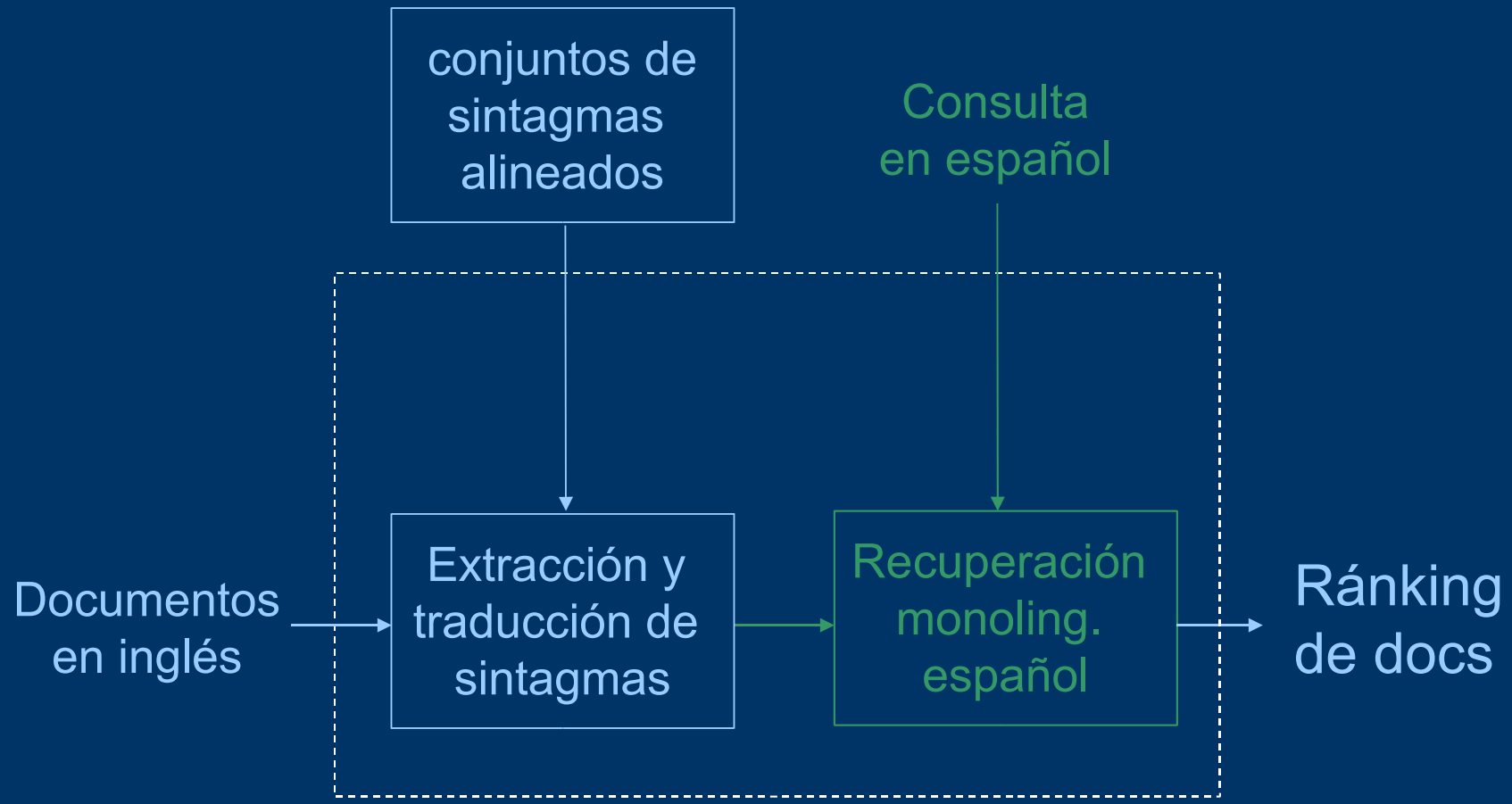
Necesidad información

consulta

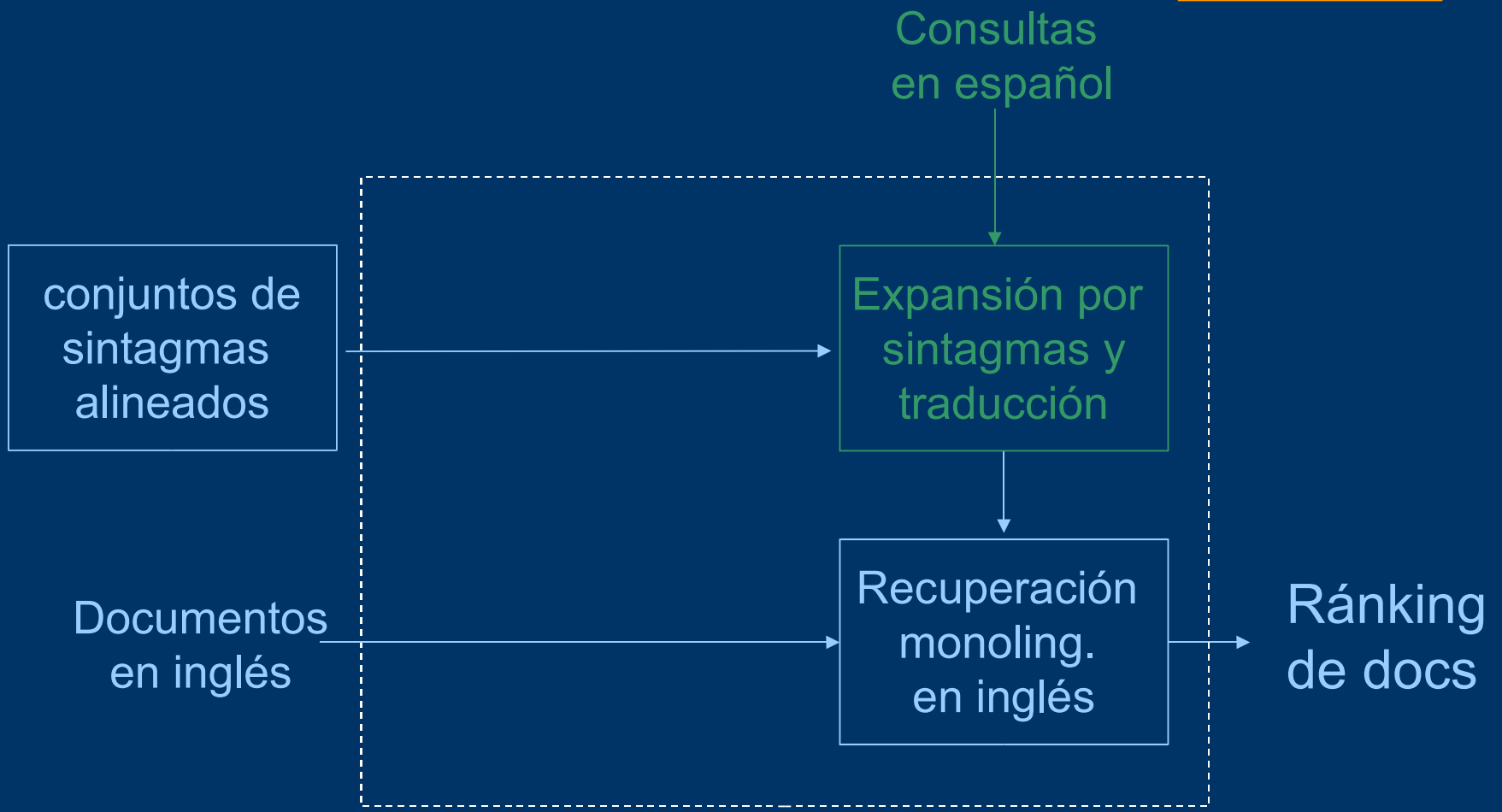
colección documentos

documentos

RI Translingüe
Traducción de documentos



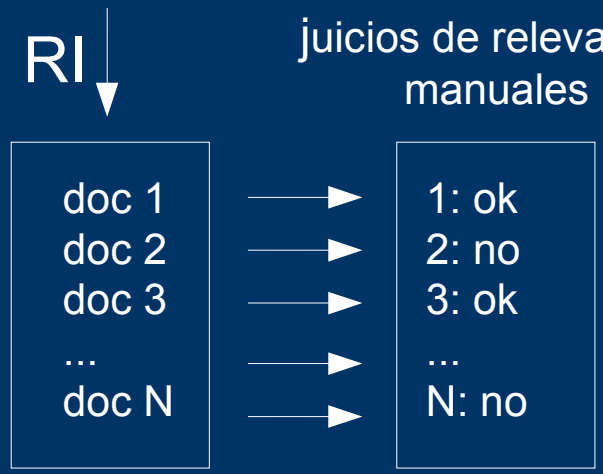
RI Translingüe
Traducción de consultas



RI Translingüe
Evaluación: Precisión y Cobertura

conjunto de prueba

- Consultas
- Colección de documentos



$$\text{PRECISIÓN} = \frac{\text{relevantes recuperados}}{\text{recuperados}}$$

$$\text{COBERTURA} = \frac{\text{relevantes recuperados}}{\text{relevantes}}$$

Periodo de investigación
Antecedentes: WTB e iCLEF

- Los sintagmas nominales son útiles en tareas de búsqueda de información interactivas multilingüe.
  - *WebSite Term Browser* (Peñas, 2002).
  - Sintagmas nominales en *Noodle* (López Ostenero, 2002).
  - iCLEF 2001-2004 (Oard & Gonzalo, 2001; Gonzalo & Oard, 2002; Oard & Gonzalo, 2003).
- Es posible utilizarlos como unidades de traducción intermedias entre la palabra y la oración.

¿Podríamos aprovecharlos en tareas de RI Translingüe automática?

Periodo de investigación
Trabajo de investigación

Objetivos

- Mejora y fusión de los recursos léxicos existentes.
- Comparación del uso de alineaciones de sintagmas nominales translingües en RI con aproximaciones de referencia.

Periodo de investigación
Trabajo – Fusión de diccionarios

- Fuentes disponibles: diccionarios bilingües VOX, ontologías basadas en WordNet (WN 1.5 y dos versiones de EWN) y fuentes FreeDict.
- Preprocesado: eliminación de erratas, inconsistencias y transformación de cada fuente en un formato estructurado intermedio.
- Fusión de las cuatro fuentes en dos diccionarios bilingües en XML (50 MB) con posibilidades de crecimiento futuro.

Periodo de investigación
Trabajo – Fusión de diccionarios

```
<entrada lema="spring">
  <trad lema="primavera" dict="fd ewn2 ewn vox">
    <fd />
    <ewn2 synset="09151839"/>
    <ewn synset="07062238 09151839"/>
    <vox tipo="simple" cat="n" uso="" dominio="" info="season" sentido="1"/>
  </trad>
  <trad lema="manantial" dict="ewn vox">
    <ewn synset="02535022 05380600 05380888 05727069"/>
    <vox tipo="simple" cat="n" uso="" dominio="" info="source" sentido="2"/>
  </trad>
  <trad lema="muelle" dict="fd ewn2 ewn vox">
    <fd />
    <ewn2 synset="03114639"/>
    <ewn synset="03114639 03260261"/>
    <vox tipo="simple" cat="n" uso="" dominio="" info="of_furniture" sentido="3"/>
  </trad>
  <trad lema="resorte" dict="vox">
    <vox tipo="simple" cat="n" uso="" dominio="" info="of_watch lock"
      sentido="4"/>
  </trad>
  ...
</entrada>
```

Periodo de investigación
Trabajo – Fusión de diccionarios

Aumentos de cobertura

COBERTURAS	VOX	diccionario fusionado	
español - inglés	16.346	57.648	+ 252%
inglés - español	15.137	85.052	+ 461%

Variaciones en precisión media en RI

EXPERIMENTO	<title> + <desc>	<title>
naive-sucio	.13	.14
naive	.19	.19
	+ 46%	+ 35%

Periodo de investigación
Trabajo – Condiciones del experimento

- Colección original en inglés etiquetada del *LA Times* de 1994 y su traducción en español con *Systran*.
- Motor de búsqueda `INQUERY` junto con las herramientas para lematizar, indexar córpora, realizar búsquedas.
- Diccionario bilingüe español-inglés creado a partir de distintos recursos.
- Conjuntos de sintagmas nominales alineados entre español e inglés (López Ostenero, 2002).

Periodo de investigación
Trabajo – Condiciones del experimento

- Herramientas para evaluar `trec_eval` y los juicios de relevancia del CLEF.
- 140 consultas CLEF en español e inglés correspondientes a los años 2000-02.

```
<top>

  <num> C054 </num>

  <EN-title> Final Four Results </EN-title>

  <EN-desc> Find documents giving the results of the European Basketball Final Four
  </EN-desc>

  <EN-narr> Relevant documents will give details on the results of at least one of the
  three matches (two semi-finals and one final) of the final phase of the European
  basketball championship. Documents written prior to the semi-finals that give the
  names of possible winners are not relevant </EN-narr>

</top>
```

Periodo de investigación
Trabajo – RI Trad. de Documentos

- Corpus traducido con Systran.
  - En general, traducciones comprensibles aunque poco naturales.
  - Alto coste de procesamiento
- Corpus traducido y resumido con sintagmas.
  - Sólo traduce los sintagmas nominales que es capaz de reconocer.
  - Tamaño y el coste de procesamiento menores.

EXPERIMENTO	Precisión media	variación
referencia monoling	.45	
systran_DT	.35	- 22%
phrases_DT	.24	- 46%

Periodo de investigación
Trabajo – RI Trad. de Consultas

- Experimentos combinando todos los recursos disponibles: diccionarios, sintagmas y TA.

EXPERIMENTO	<title> + <desc>	<title>
referencia monoling	.45	
phrases + pirkola	.33	.29
naive-sucio	.13	.14
naive	.19	.19
frec	.26	.25
pirkola	.31	.27
systran	.34	.27
phrases_multi + pirkola	.31	.31
phrases + pirkola + systran	.35	.30

estado del arte ◄

Periodo de investigación
Trabajo - Conclusiones

- Creación de un nuevo diccionario por fusión de recursos léxicos que mejora sustancialmente el rendimiento en tareas de Recuperación de Información Translingüe.
- El uso de sintagmas nominales, junto con traducción palabra por palabra en consultas estructuradas (Pirkola, 1998), mejora (aunque no de forma significativa) los resultados obtenidos con Traducción Automática, y es un mecanismo más universal y menos costoso.
- Los documentos resumidos y traducidos con sintagmas nominales no son suficientes para hacer RI translingüe automática, aunque hayan funcionado muy bien en experimentos interactivos.

Publicaciones

- Amigó, E., Gonzalo, J., Peinado, V., Peñas, A., Verdejo, F. An empirical study of Information Synthesis tasks. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL).* Barcelona, julio, 2004.
- Amigó, E., Gonzalo, J., Peinado, V., Peñas, A., Verdejo, F. Using syntactic information to extract relevant terms for multi-document summarization. *Proceedings of the 20th International Conference on Computational Linguistics (COLING). Ginebra, agosto, 2004.*
- Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F. and Rijke, M. Creating the DISEQuA Corpus: a test set for Multilingual Question Answering. *Evaluation of Cross-Language Information Systems,* LNCS 3237, Springer, 2004.

Publicaciones

- Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F. and Rijke, M. The multiple language Question Answering Track at CLEF 2003. *Evaluation of Cross-Language Information Systems,* LNCS 3237, Springer, 2004.
- Amigó, E., Gonzalo, J., Peinado, V., Peñas, A., Verdejo, F. PRISMA: un modelo interactivo de Síntesis de Información. Revista del *XX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN).* Barcelona, julio, 2004.
- Peinado, V., Artiles, J., López-Ostenero, F., Gonzalo, J., Verdejo, M. F. UNED@ImageCLEF 2004: using document structure and noun-phrase based query expansion for Cross-Language Image Caption Retrieval. (próximamente)

Publicaciones

- López-Ostenero, F., Gonzalo, J., Peinado, V., Verdejo, M. F. UNED@iCLEF 2004: document versus filtered paragraph retrieval for interactive Cross-Language Question Answering. (próximamente)

Referencias

- Gonzalo, J., Oard, D. The CLEF 2002 Interactive Track. *Evaluation of Cross-Language Information Systems,* LNCS 2406, Springer, 2002.
- Lin, C-Y., Hovy, E. Automatic Evaluation of Summaries Using N-gram Co-ocurrence Statistics. *Proceedings of the Human Technology Conference 2003 (HLT-NAACL-2003),* 2003.
- López Ostenero, F. *Un sistema interactivo para la búsqueda de información en idiomas desconocidos por el usuario*. Tesis Doctoral, Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, 2002.
- Oard, D., Gonzalo, J. The CLEF 2001 Interactive Track. *Evaluation of Cross-Language Information Systems,* LNCS 2785, Springer, 2001.
- Oard, D., Gonzalo, J. The CLEF 2003 Interactive Track. *Evaluation of Cross-Language Information Systems,* LNCS 3237, Springer, 2004.

Referencias

- Papineni, K., Roukos, S., Ward, T., Zhu, W. BLEU: a Method for Authomatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, julio, 2002.
- Peñas, A., *Website Term Browser: Un sistema interactivo y multilingüe de búsqueda textual basado en técnicas lingüísticas*. Tesis Doctoral, Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia, 2002.
- Pirkola, A., The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. *Proceedings of SIGIR'98, 21st ACM International Conference on Research and Development in Information Retrieval,* pp 55-63, 1998.

Diploma de Estudios Avanzados

Víctor Josué Peinado Herencia
Licenciado en Lingüística por la UCM

Grupo de Procesamiento del Lenguaje Natural
Dept. de Lenguajes y Sistemas Informáticos
ETS. de Informática - UNED

14 de julio de 2004