Análisis semántico del lenguaje natural para **expresiones geotemporales**

JIDEE 2008

V Jornadas Técnicas de la IDE de España
JIDEE 2008

Universidad Politécnica de Madrid

AGENDA

Análisis semántico para expresiones geotemporales

- Introducción
- Conceptos y trabajos relacionados
- Minería de datos geotemporal
- Analizador semántico
- Conclusión
- Aportes

Referencias

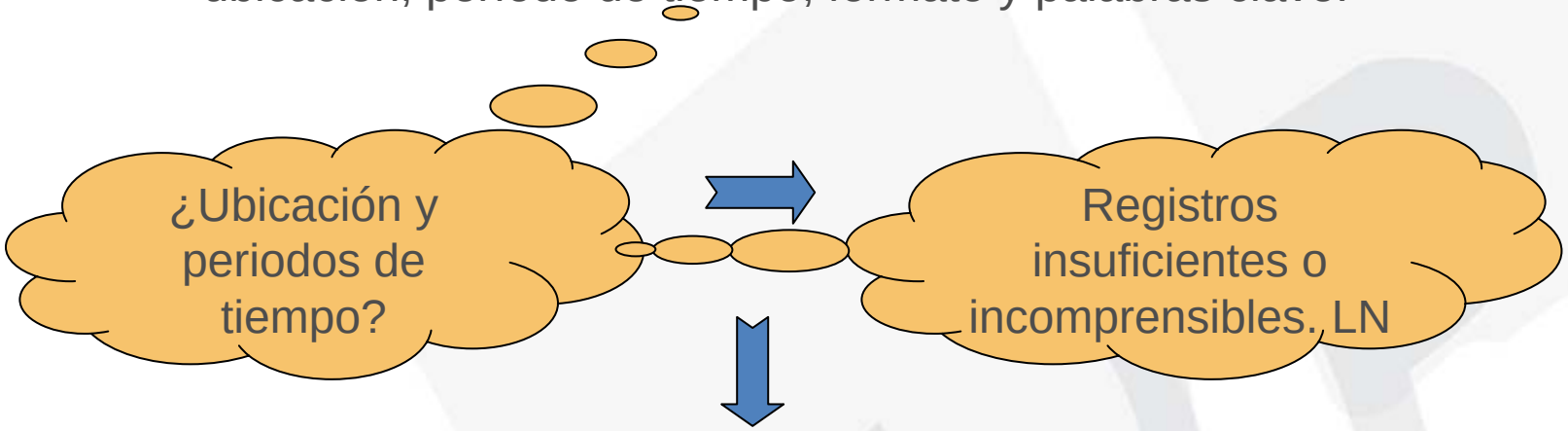
INTRODUCCIÓN

Análisis semántico del lenguaje natural para expresiones geotemporales

Universidad Politécnica de Madrid

Introducción -1-

En el contexto de bibliotecas y cartotecas digitales, los recursos son generalmente descritos en registros de metadatos que definen su objeto, ubicación, período de tiempo, formato y palabras clave.



Se presentan técnicas para la extracción de información geotemporal de colecciones de texto.

El objetivo es partir de referencias textuales geotemporales descritas por humanos, identificar las entidades geográficas y temporales y expresarlas en un lenguaje comprensible y procesable por un sistema informático.

wsiabato@acm.org
Mercator Research Group

Introducción -2-

El análisis semántico y la correcta interpretación por parte de los sistemas informáticos del lenguaje natural, es un tema de creciente estudio en la actualidad que se remonta a décadas de arduo trabajo.

En este contexto, el análisis de las expresiones geotemporales debe ser analizado de forma independiente.

Contextos, periodos, fechas puntuales, hechos de referencia.

Enseñar al sistema informático a interpretar este tipo de expresiones.

Implementa componentes como WordNet, GeoNames y elementos teóricos como el Algebra de Allen [1]

eXtensible Language Markup

wsiabato@acm.org
Mercator Research Group

CONCEPTOS

Análisis semántico del lenguaje natural para expresiones geotemporales

Universidad Politécnica de Madrid

Conceptos -1-

Múltiples estudios [3] [6] [15] muestran como los componentes temporal y geográfico desempeñan un importante rol al filtrar, agrupar y dar prioridad a recursos de información.

Tres procesos involucrados:

(i) Identificación y análisis del texto,

(ii) la búsqueda de elementos y referencias geográficas y temporales y su precisa ubicación sobre la superficie y el espacio, y

(iii) la combinación de estas referencias en recopilaciones semánticas significativas.

wsiabato@acm.org
Mercator Research Group

Conceptos -2-

Aunque el problema presentado ha sido tratado con relativo éxito en comunidades como la *Natural Language Processing* [17] y la *Geographical Information Retrieval* [15] [16], el asunto discutido en este artículo difiere del NER –*Named Entity Recognition*– convencional en múltiples elementos.

Diferencias:

1. Las entidades analizadas son más precisas.
2. Múltiples idiomas.
3. Reconocimiento de las entidades no implica un significado en sí mismo.
4. Técnicas y heurísticas para un tiempo de proceso aceptable.
5. Unidades individuales o como parte de un contexto semántico específico.

wsiabato@acm.org
Mercator Research Group

Conceptos -3-

Base tecnológica.

1. Fuentes léxicas (*Gazetteer)*
2. Operaciones de procesamiento secundarias en las que se incluyen por lo menos una semilla, un diccionario especializado y un conjunto de reglas de extracción.

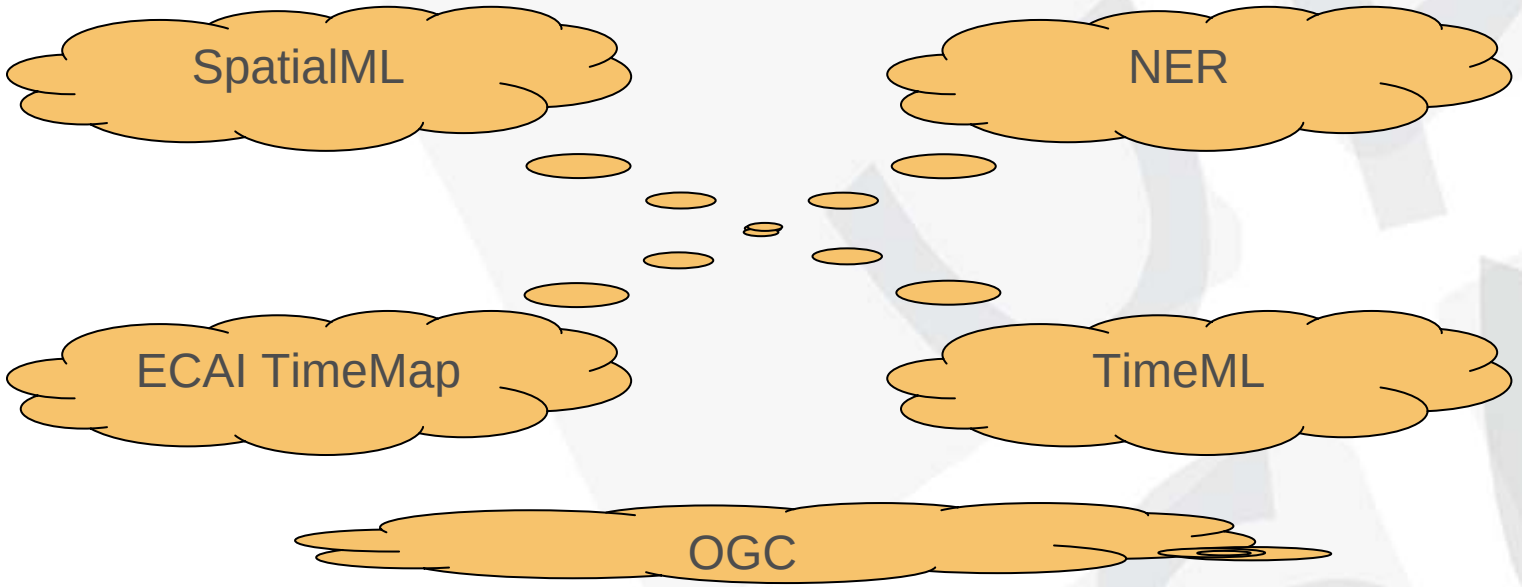
Las reglas para el reconocimiento de entidades son el núcleo del sistema, es posible por ejemplo combinar los nombres de elementos en el diccionario con características como las mayúsculas y textos contiguos o circundantes.

wsiabato@acm.org
Mercator Research Group

Conceptos -4-

Base tecnológica.

La investigación con analizadores semánticos geográficos (*geo-parsers*) está iniciando.



wsiabato@acm.org
Mercator Research Group

MINERÍA DE DATOS

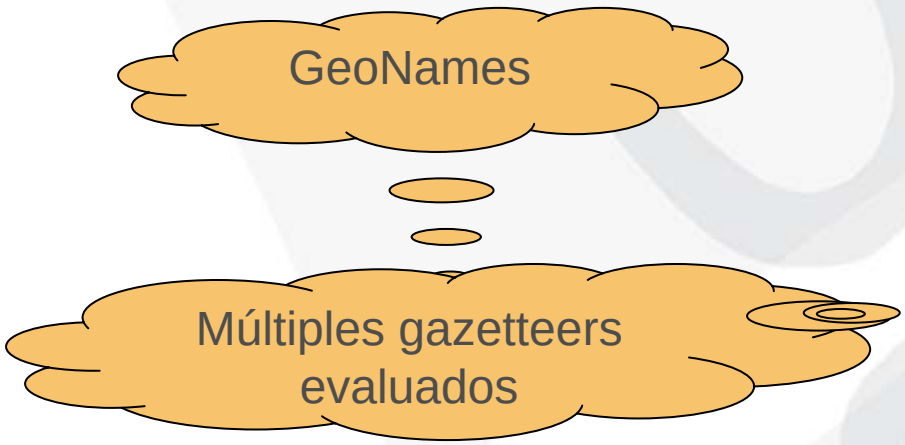
Análisis semántico del lenguaje natural para expresiones geotemporales

Universidad Politécnica de Madrid

Minería de datos -1-

Gazetteer geotemporal.

Contar con un *gazetteer* fiable, robusto, con información completa y detallada, con un registro de nombres de lugares, periodos históricos y la descripción de sus propiedades (ej. tipos de lugar, coordenadas, intervalos temporales, jerarquías, nombres alternativos, asociaciones semánticas) es un factor determinante en el desarrollo y éxito del trabajo presentado..

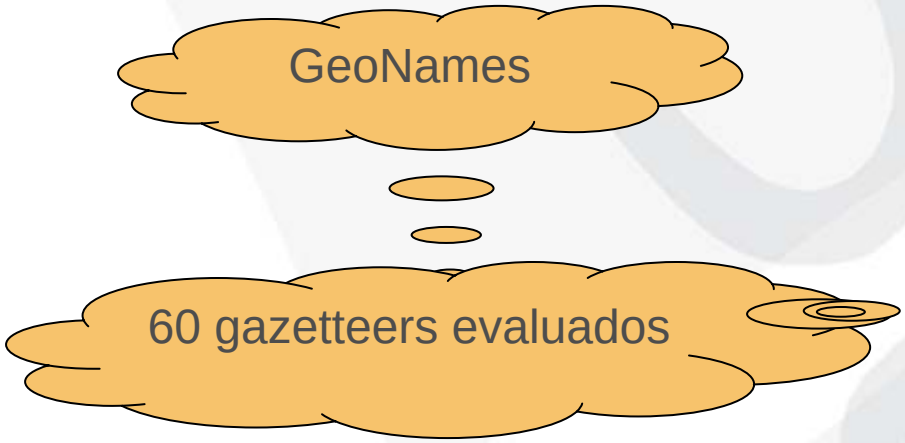


wsiabato@acm.org
Mercator Research Group

Minería de datos -2-

Gazetteer geotemporal.

Contar con un *gazetteer* fiable, robusto, con información completa y detallada, con un registro de nombres de lugares, periodos históricos y la descripción de sus propiedades (ej. tipos de lugar, coordenadas, intervalos temporales, jerarquías, nombres alternativos, asociaciones semánticas) es un factor determinante en el desarrollo y éxito del trabajo presentado..



DynCoopNet.

wsiabato@acm.org
Mercator Research Group

Minería de datos -3-

Extracción de información geográfica.

1. Reconocer de las referencias de lugares,
2. Desambiguar las referencias encontradas, y
3. Asignar un entorno geográfico a las mismas.

Método NER basado en semillas complementado con búsquedas en *gazetteers.*

$s \in [0,1]$

$s \leq 0.5$

wsiabato@acm.org
Mercator Research Group

Minería de datos -4-

Extracción de información geográfica.

1. Reconocer de las referencias de lugares,
2. Desambiguar las referencias encontradas y
3. Asignar un entorno geográfico a las mismas.

Para desambiguar las referencias encontradas, se requiere del uso del *gazetteer* y un conjunto de heurísticas.

wsiabato@acm.org
Mercator Research Group

Minería de datos -5-

Extracción de información geográfica.

1. Reconocer de las referencias de lugares,
2. Desambiguar las referencias encontradas y
3. Asignar un entorno geográfico a las mismas.

Esta tarea es realizada a través de una técnica similar a la propuesta en el proyecto *Web a Where* [2]. Jerarquías del gazetteer.
Pesos heredados.

wsiabato@acm.org
Mercator Research Group

Minería de datos -6-

TimeMap. Pesos

Extracción de información temporal.

Tipo de relación	*Ejemplo*	*Relación de limites**
x before y **y after x**	xxxx yyyy	$x^+ < y^-$
x meets y **y met-by x**	xxxx yyyy	$x^+ = y^-$
x overlaps y **y overlap-by x**	xxxx yyyy	$x^- < y^- < x^+$; $x^+ < y^+$
x during y **y includes x**	xxxx yyyyyyyyyy	$x^- > y^-$; $x^+ < y^+$
x starts y **y started by x**	xxxx yyyyyyyyyy	$x^- = y^-$; $x^+ < y^+$
x finishes y **y finished by x**	xxxx yyyyyyyyyy	$x^+ = y^+$; $x^- > y^-$
x equals y	xxxx yyyy	$x^- = y^-$; $x^+ = y^+$

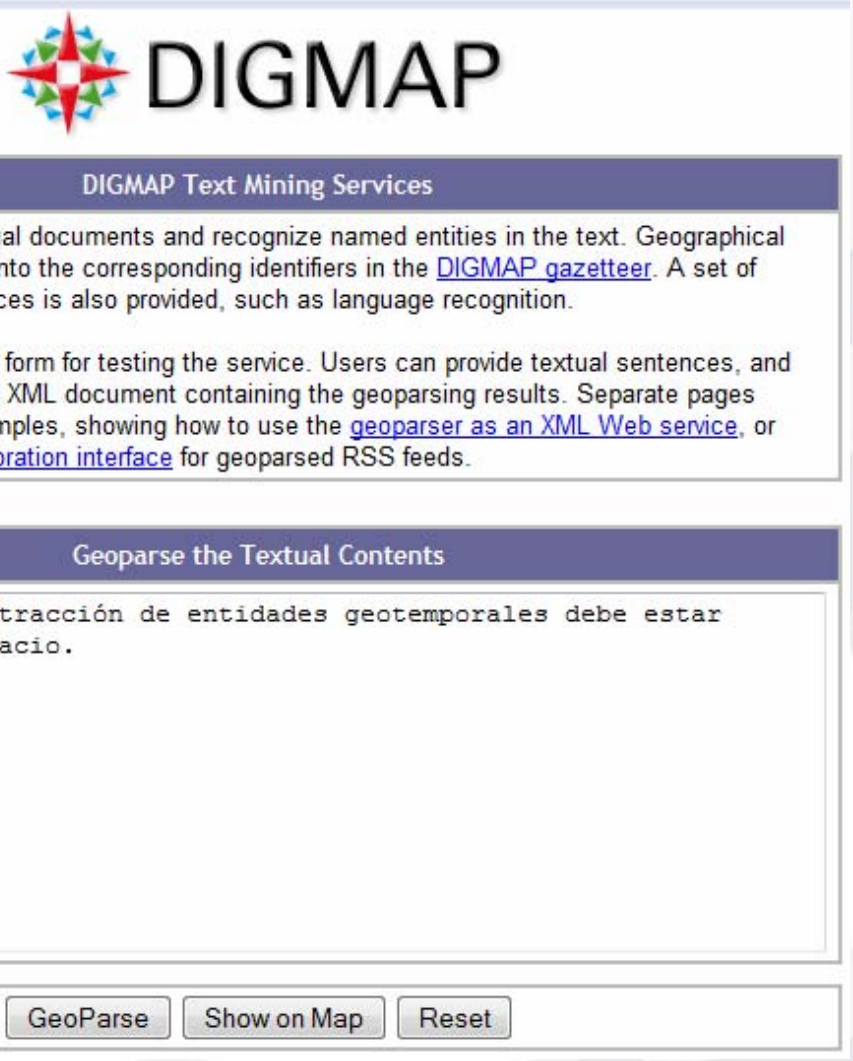
wsiabato@acm.org
Mercator Research Group

ANALIZADOR

Análisis semántico del lenguaje natural para expresiones geotemporales

Universidad Politécnica de Madrid

Analizador -1-



wsiabato@acm.org
Mercator Research Group

Analizador -2-

<GeoparseResult xsi:schemaLocation="http://www.opengis.net/gp ../gp/GetFeatureRequest.xsd http://www.opengis.net/wfs ../wfs/GetFeatureRequest.xsd">

<EntryCollection>

……

<TermName>Antonio Telo</TermName>

<Ocurrence>

<Range start="33" end="45"/>

</Ocurrence>

<Label>PERSON</Label>

</Person>

<PlaceName entryID="1">

…

<TermName>Portugal</TermName>

<Ocurrence>

<Range start="99" end="107"/>

</Ocurrence>

<Label>LOCATION</Label>

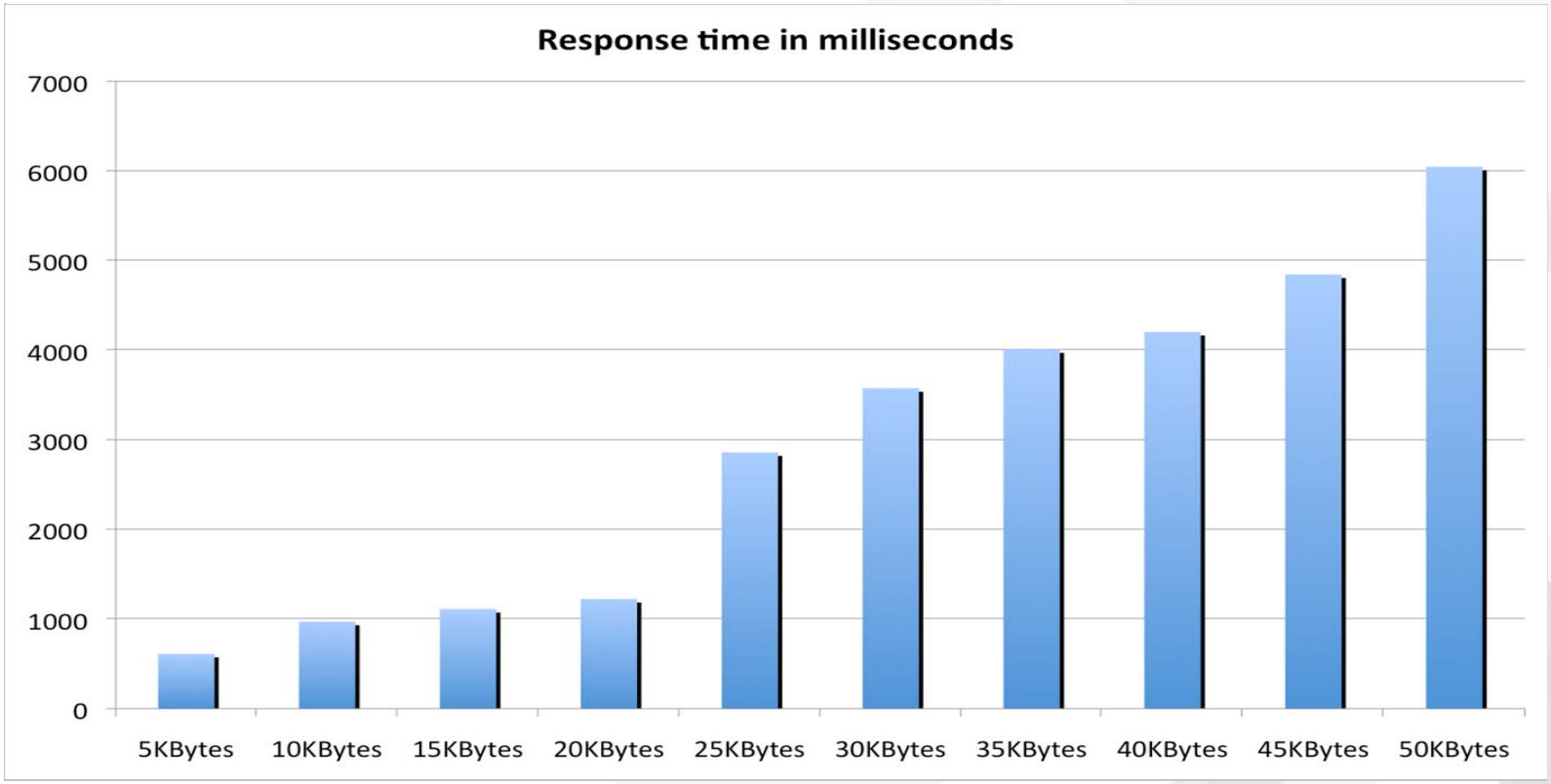
</PlaceName>

<DateTime entryID="2">

<TermName>20th century</TermName>

wsiabato@acm.org
Mercator Research Group

Analizador -3-



wsiabato@acm.org
Mercator Research Group

CONCLUSIÓN

Análisis semántico del lenguaje natural para expresiones geotemporales

Universidad Politécnica de Madrid

Conclusiones -1-

Los **registros de metadatos en bibliotecas y cartotecas digitales** describen con frecuencia recursos que **relacionan algún lugar en un momento específico**. Debido a ello, las colecciones pueden ser **organizadas** de acuerdo a criterios que involucren su distribución **espacial y/o temporal**.

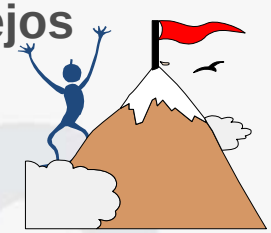
Aunque la idea parece simple, los recursos **están caracterizados a menudo por descripciones textuales** y tanto los nombres de lugares como los períodos de tiempo presentan una **alta ambigüedad**; por ejemplo, ¿cuál es el significado detrás de la Guerra del golfo o Periodo de la revolución?.

La ambigüedad en este tipo de descripciones y en las entidades que representan los lugares y periodos de tiempo **debe ser resuelta a fin de obtener una plena comprensión** del contexto geotemporal involucrado.

wsiabato@acm.org
Mercator Research Group

Conclusiones -2-

Este artículo muestra como sencillas **técnicas de extracción** (o relativamente sencillas) **son capaces de proporcionar resultados con calidad suficiente para ser utilizadas en procesos complejos y que requieren de una alta precisión en la identificación de entidades geotemporales**.



Se evidencia también como **se hace necesario** involucrar múltiples elementos externos para una correcta interpretación de las entidades identificadas y poder **desambiguarlas con un grado de precisión aceptable**.

FUNCIONA!!!!!.

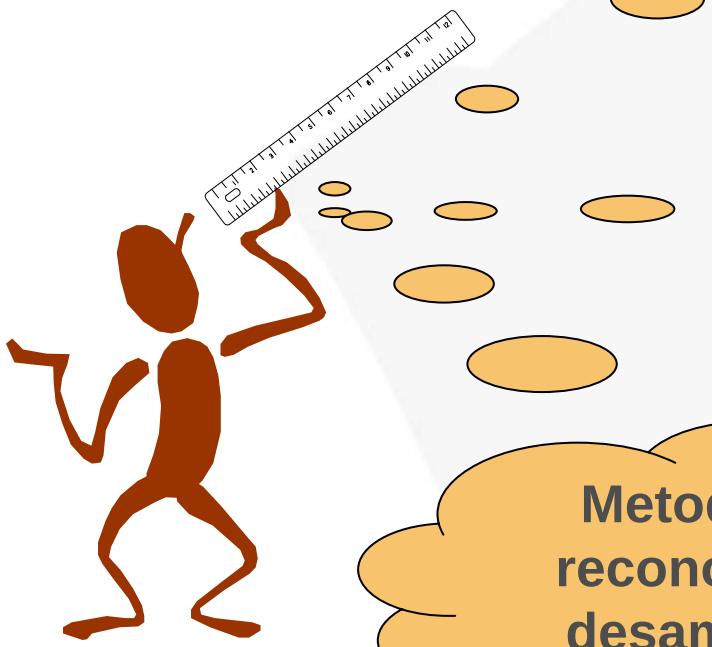
APORTES

Análisis semántico del lenguaje natural para expresiones geotemporales

Universidad Politécnica de Madrid

Aportes

En conclusión:



Servicio Web

Gazetteer

Metodología de reconocimiento y desambiguación

wsiabato@acm.org
Mercator Research Group

REFERENCIAS

Análisis semántico del lenguaje natural para expresiones geotemporales

Universidad Politécnica de Madrid

Referencias -1-

[1] Allen, J.F. 1991. “Temporal reasoning and planning”. In Reasoning about plans. Edited by J.F. Allen, H.A. Kautz, R.N. Pelavin and J.D. Tenenberg. Morgan Kaufmann, San Francisco - USA. pp. 1-67.

[2] Amitay E., Har'El N., Sivan R. and Soffer A. 2004 Web-a-where: geotagging Web content. Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval

[3] Bates M. J. and Wilde D. N. 1993 An analysis of search terminology used by humanities scholars: the Getty Online Searching Project Report Number 1. Library Quarterly, 63(1)

[4] Boguraev B. and Ando R. K. 2005 TimeML-compliant text analysis for temporal reasoning. Proceedings of the 19th International Joint Conference on Artificial Intelligence

[5] Buckland M. and Lancaster L. 2004 Combining Place, Time, and Topic : The Electronic Cultural Atlas Initiative. D-Lib Magazine, 10(5)

[6] Chen Y., Di Fabbrizio G., Gibbon D., Jana R., Jora S., Renger B. and Wei B. 2007 GeoTracker: Geospatial and temporal RSS navigation. Proceedings of the 16th World Wide Web conference

[7] Chinchor N. 1998 Proceedings of the 7th Message Understanding Conference

[8] Clough P. and Sanderson M. 2004 A proposal for comparative evaluation of automatic annotation for geo-referenced documents. Proceedings of the 1st Workshop on Geographic Information Retrieval

[9] Cohen W. and Sarawagi S. 2004 Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods. Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining

[10] Drakengren T. and Jonsson P. "Eight maximal tractable subclasses of Allen's algebra with metric time" Journal of Artificial Intelligence Research vol. 7, pp. 25-45, 1997.

[11] Gale W., Church K. and Yarowsky D. 1992 One sense per discourse. Proceedings of the 4th DARPA Speech and Natural Language Workshop

Referencias -1-

[12] Garbin E. and Mani I. 2005 Disambiguating toponyms in news. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing

[13] Harpring P. 1997 The limits of the world: Theoretical and practical issues in the construction of the Getty Thesaurus of Geographic Names. Proceedings of the 4th International Conference on Hypermedia and Interactivity in Museums, Archives and Museum Informatics

[14] Hill L. and Zheng Q. 1999. Indirect geospatial referencing through place names in the digital library: Alexandria Digital Library experience with developing and implementing gazetteers. Proceedings of the American Society for Information Science Annual Meeting

[15] Jones C., Abdelmoty A., Finch D., Fu G. and Vaid S. 2004 The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. Proceedings of the 3rd International Conference on Geographic Information Science

[16] Jones C. and Purves R. 2006 GIR'05: The 2005 ACM workshop on Geographical Information Retrieval, ACM SIGIR Forum, 40(1)

[17] Kornai A. 2003 Proceedings of the HLT-NAACL 2003 workshop on Analysis of Geographic References

[18] Lansing J. 2001 Geoparser service draft candidate implementation specification. OGC Discussion Paper 01-035

[19] Leidner J. 2007 Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names, Ph.D. thesis, School of Informatics, University of Edinburgh, Scotland, UK

[20] Li H., Srihari K. R., Niu C. and Li W. 2002 Location normalization for information extraction. Proceedings of the 19th Conference on Computational Linguistics

[21] Malouf R. 2002 Markov models for language-independent named entity recognition. In Proceedings of the 6th Conference on Natural Language Learning

wsiabato@acm.org
Mercator Research Group

Referencias -1-

[22] Manguinhas, H., Martins, B., Siabato, W. & Borbinha, J. 2008 "The DIGMAP Geo-Temporal Web Gazetteer Service" Proceedings of the 3th International Workshop Digital Approaches to Cartographic Heritage.

[23] Manov D., Kiryakov A., Popov B., Bontcheva K., Maynard D. and Cunningham H. 2003 Experiments with geographic knowledge for information extraction. Proceedings of the HTL/NAACL-03 Workshop on Analysis of Geographic References

[24] Mikheev A., Moens M. and Grover C. 1999 Named entity recognition without gazetteers. Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics

[25] Petras V., Larson R. R. and Buckland M. 2006 Time period directories: a metadata infrastructure for placing events in temporal and geographic context. Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries

[26] Pustejovsky J., Castano J., Ingria R., Saurı R., Gaizauskas R., Setzer A., Katz G., and Radev D. 2003 TimeML: Robust specification of event and temporal expressions in text. Proceedings of the AAAI Spring Symposium on New Directions in Question-Answering

[27] Sang E. T. K. and De Meulder F. 2003 Introduction to the CoNLL-2003 shared task: Language-Independent Named Entity Recognition. Proceedings of the 7th Conference on Natural Language Learning

GRACIAS!!!

Análisis semántico del lenguaje natural para expresiones geotemporales

Universidad Politécnica de Madrid