Desambiguación semántica

con Procesamiento de Lenguaje Natural

Luis Briones Montecinos
Profesor Guía: Alejandro Mauro
Seminario Bibliográfico – Magister en Informática Medica

20 Diciembre 2014

Agenda

- Motivación
- Introducción
- Desambiguación semántica dentro del PLN
- Estado del Arte Desambiguación semántica
- Desafíos de la Desambiguación semántica
- Conclusiones y Propuesta futura
- Bibliografía

Motivación

Business Intelligence

Data Mining

Informes CDA

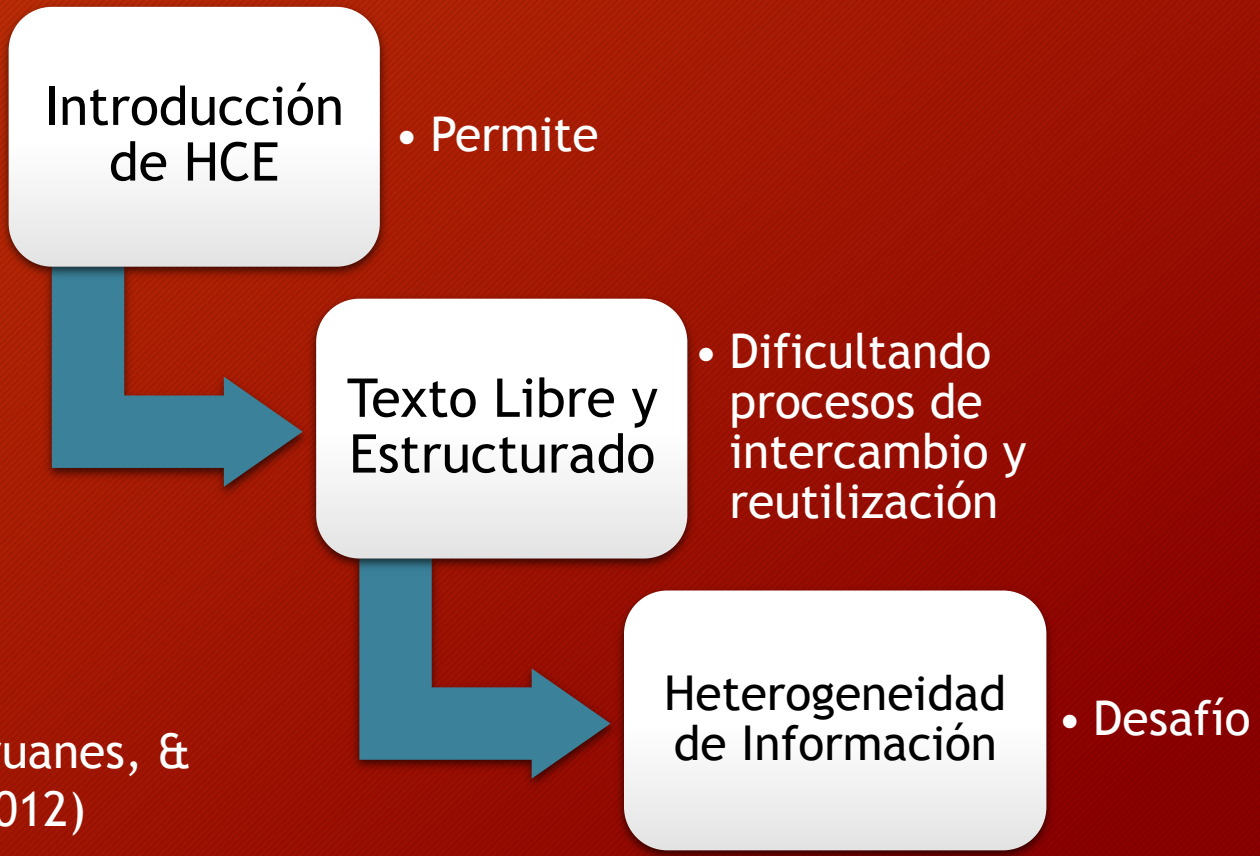
Sistemas de Soporte a la Decisión

Auto Codificación

Para automatizar procesos se requiere **PLN** y por ende: **Desambiguación Semántica**

Introducción

- Procesamiento de Lenguaje Natural (PLN)

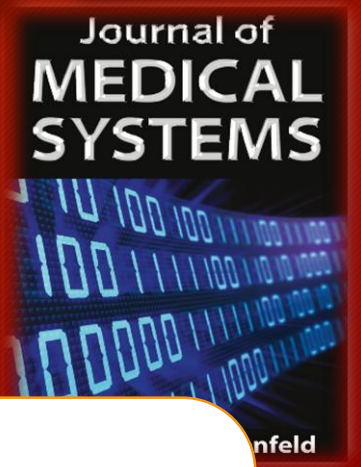


(Romá-Ferri, Cruanes, & Lloret Pastor, 2012)

Introducción

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Automated Mapping of Clinical Terms into SNOMED-CT. An Application to Codify Procedures in Pathology



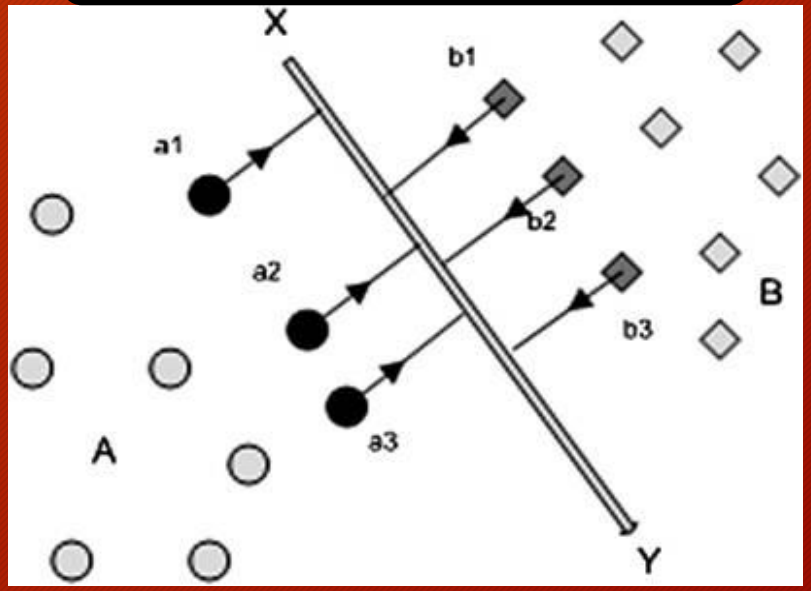
La terminología clínica es considerada la tecnología clave para la captura de datos clínicos de manera precisa y estandarizada, lo cual es fundamental para el intercambio de información entre diferentes aplicaciones, registros médicos y sistemas de soporte a la decisión, etc.

(Allones, Martinez, & Taboada, 2014)

Introducción

- Procesamiento de Lenguaje Natural (PLN) El enfoque mas utilizado y con mejores resultados son los métodos basados en aprendizaje de máquina

Support Vector Machines



Supervisados

- Cada elemento de los datos de entrenamiento se etiqueta con la respuesta correcta

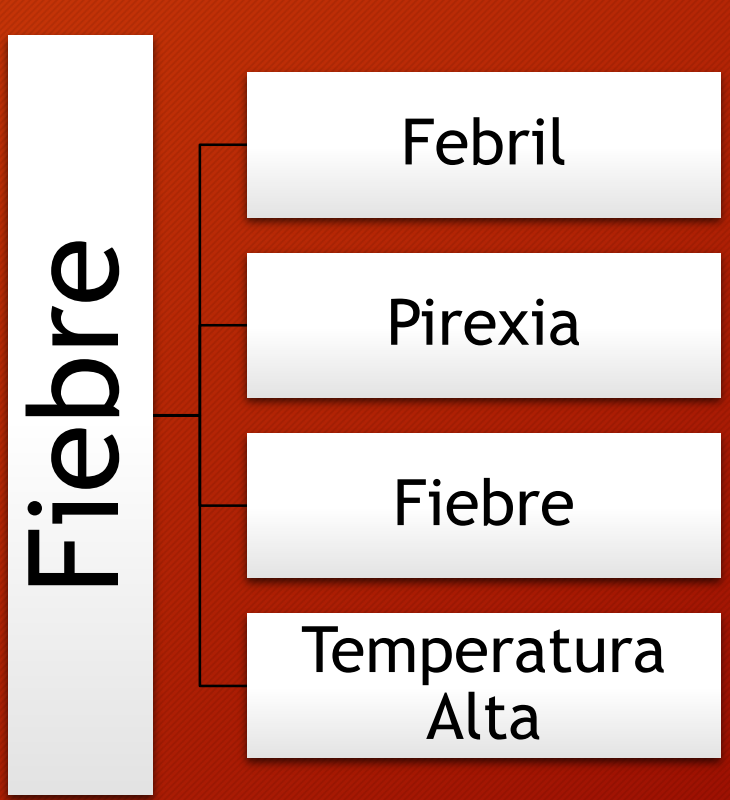
No Supervisados

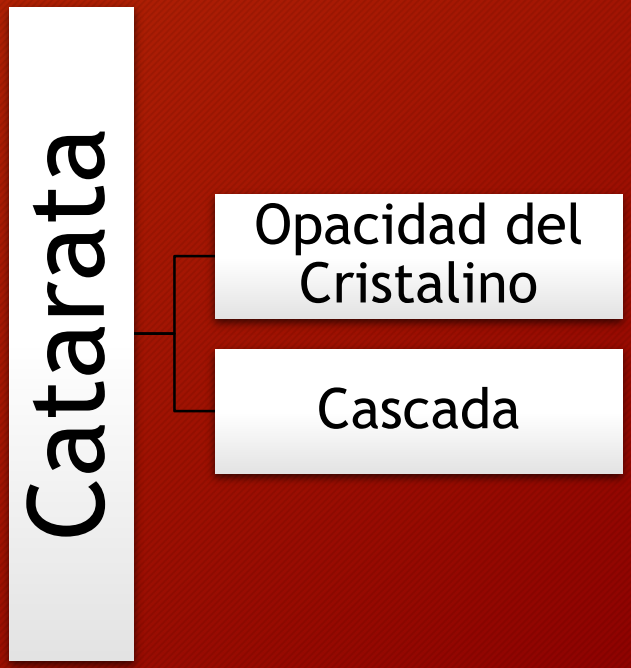
- Trata de reconocer patrones de forma automática

(Nadkarni, Ohno-Machado, & Chapman, 2011)

Desambiguación semántica dentro del PLN

- Que es la desambiguación semántica?





SNOMED-CT reconoce 3.741 homonimias en español

(Clase Representación del Conocimiento en Salud – Alenjandro Mauro – 2014)

Desambiguación semántica dentro del PLN

- Podemos identificar dos aproximaciones metodológicas para afrontar la desambiguación semántica (Marco, 2004):

Basados en Conocimientos	Basados en Corpus
• Diccionarios • Tesauros • Ontologías	• Etiquetados • Gran cantidad de ejemplos

- La identificación automática del sentido correcto de una palabra ambigua incrementa el rendimiento de aplicaciones clínicas y biomédicas tales como la codificación médica de diagnósticos, que se están transformando en tareas esenciales dentro del trabajo clínico cotidiano.

(McInnes & Stevenson, 2014).

Estado del Arte
Desambiguación semántica

Research and applications

JAMIA

Applying active learning to supervised word sense disambiguation in MEDLINE

Yukun Chen,[1] Hongxin Cao,[2] Qiaozhu Mei,[3,4]

Table 3 Accuracy of active learners and the passive learner across 197 ambiguous words when different numbers of training samples were used

Number of training samples	LC	Margin	Entropy	Random
10	0.819	0.819	0.820	0.751
20	0.887	0.887	0.886	0.844
30	0.915	0.914	0.915	0.884
40	0.927	0.928	0.927	0.903
50	0.932	0.932	0.932	0.914
60	0.937	0.937	0.937	0.922
70	0.939	0.940	0.940	0.927
80	0.941	0.942	0.941	0.931
90	0.942	0.942	0.942	0.934

LC, Least Confidence.

(Chen, Cao, Mei, Zheng, & Xu, 2013).

Estado del Arte
Desambiguación semántica

ELSEVIER

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Supervised word sense disambiguation using semantic diffusion kernel

Tinghua Wang [a,b,*], Junyang Rao [b], Qi Hu [c]

MFS: Most Frequent Sense
BoW: Bag-of-Words model
LSI: LSI-based – SVM
SDK: Semantic Diffusion Kernel

(Wang, Rao, & Hu, 2014)

Table 7
Classification results (micro-F_1) of four methods.

Data set	Micro-F_1 (%)			
	MFS	BoW	LSI	SDK
Interest	52.87	86.00	86.13	**87.32**
Line	53.48	82.94	82.98	**84.09**
Hard	79.74	83.70	83.62	**84.84**
Serve	41.43	86.29	86.73	**87.00**
W-T-L	4-0-0	4-0-0	3-1-0	–

Estado del Arte
Desambiguación semántica

Para el siguiente fragmento de un informe:

«El paciente ingresó con fiebre y signs de deshidratacion. Llega a planta con REG. ACR : sin arritmias.»

El sistema nos devolvería los siguientes resultados:

Correcciones:

signs → signos;

deshidratacion → deshidratación.

Acrónimos:

REG → «Regular estado general»;

ACR → «Auscultación cardiorespiratoria».

Conceptos:

Hallazgo - Fiebre (*afirmado*);

Hallazgo - Deshidratación (*afirmado*);

Condición - Estado general regular (*afirmado*);

Procedimiento - Auscultación cardiorespiratoria (*afirmado*);

Hallazgo - Arritmias (*negado*).

Estado del Arte
Desambiguación semántica

Procesador automático de informes médicos

Facultad de Informática de la Universidad Complutense

Departamento de Ingeniería del Software e Inteligencia Artificial

- Mejoras en la identificación de conceptos
- Mejoras en la fase de detección de la negación
- Mejoras en la fase de detección y desambiguación de acrónimos en español

- Fase de detección de especulaciones

(Barahona, 2012)

Desafíos de la Desambiguación semántica

- Como indica (Gonzalo, Verdejo, & Chugar, 2003) las causas de la no utilización de desambiguación semántica en muchas tareas finales como traducción automática, recuperación de información, búsqueda de respuestas, etc. son:
  - La desambiguación semántica muchas veces es más difícil que la propia tarea a la que quiere ayudar.
  - Los sistemas no supervisados aún obtienen un pobre rendimiento.
  - Los supervisados apenas tienen recursos con los que trabajar en idioma español.

(Gonzalo, Verdejo, & Chugar, 2003)

Conclusiones y Propuesta futura

- La tarea de desambiguación semántica en el contexto medico es crucial, dado lo delicado de la información. Por otra parte, cuando todos los esfuerzos van en el sentido de informatizar procesos, debemos solventar necesidades de automatizar y mantener la interoperabilidad semántica.

- Es necesario apoyar el procesamiento de texto libre ingresado por los clínicos, donde se requieren herramientas de PLN, desde la extracción de información de HCE, hasta la codificación automática.

- Los métodos de desambiguación basados en aprendizaje supervisado son los que obtienen mejores resultados en términos de precisión, pero deja un desafío para idioma español al requerir bases de conocimiento bien estructurada y conjuntos de datos de entrenamiento de alta calidad.

- Revisar otros puntos clave dentro del PLN en el contexto medico, con el fin de aportar una mejora significativa en la codificación automática con SNOMED-CT.

Bibliografía

- Allones, J. L., Martinez, D., & Taboada, M. (2014). Automated mapping of clinical terms into SNOMED-CT. An application to codify procedures in pathology. *Journal of Medical Systems*, *38*(10), 134.
- Barahona, E. B. (2012). Procesador automático de informes médicos. Retrieved from http://eprints.ucm.es/16070/
- Chen, Y., Cao, H., Mei, Q., Zheng, K., & Xu, H. (2013). Applying active learning to supervised word sense disambiguation in MEDLINE. *Journal of the American Medical Informatics Association : JAMIA*, *20*(5), 1001–6.
- Gonzalo, J., Verdejo, F., & Chugar, I. (2003). The Web as a Resource for WSD. *1st MEANING Workshop, Spain*.
- Marco, A. M. (2004). Universidad Politécnica de Valencia Desambiguación en procesamiento del lenguaje natural mediante técnicas de aprendizaje automático.
- McInnes, B. T., & Stevenson, M. (2014). Determining the difficulty of Word Sense Disambiguation. *Journal of Biomedical Informatics*, *47*, 83–90. doi:10.1016/j.jbi.2013.09.009
- Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association : JAMIA*, *18*(5), 544–51.
- Romá-Ferri, M. T., Cruanes, J., & Lloret Pastor, E. (2012). Analisis del uso de metodos de similitud lexica con conocimiento semantico superficial para mapear la informacion de enfermeria en español. *Procesamiento Del Lenguaje Natural*, 75–82.
- Wang, T., Rao, J., & Hu, Q. (2014). Supervised word sense disambiguation using semantic diffusion kernel. *Engineering Applications of Artificial Intelligence*, *27*, 167–174. doi:10.1016/j.engappai.2013.08.007