Procesamiento del Lenguaje Natural

Comparación de algoritmos de lematización

Jorge Mansilla Sierra
Magíster Informática Médica
Dr. Alejandro Mauro
20 / Diciembre / 2014

Agenda

- Procesamiento del Lenguaje Natural (NLP)
- Aplicaciones
- Lematización
- Algoritmos de lematización
- Trabajos futuros

NLP

El Procesamiento del Lenguaje Natural es una disciplina de la Inteligencia Artificial que se ocupa de la formulación e investigación de mecanismos computacionales para la comunicación entre personas y maquinas mediante el uso de Lenguajes Naturales.

Es el reconocimiento y utilización de la información expresada en lenguaje humano, a través del sistemas informáticos.

NLP

- Filosofía
- Matemáticas
- Psicología
- **Lingüística**
- **Ingeniería informática**

NLP

Lingüística

- Existen 4 niveles del lenguaje
  - Morfológico
  - Semántico
  - Sintáctico
  - Pragmático

NLP

Aplicaciones

- **Minería de datos**
  - Patrones ocultos y relaciones de datos estructurados
  - Como se encuentras estructurados, se realiza una limpieza y normalización de datos, generando numerosas relaciones entre los datos de las bases de datos
  - Usa técnicas IR, EI y corpus procesados con técnicas lingüísticas

NLP

Aplicaciones

- **Traducción automática**
  - Transformar el texto de un idioma a otro
  - Consta de tres pasos
    1. Texto original transformado a una representación intermedia
    2. Modificaciones morfológicas de acuerdo al lenguaje de destino
    3. Transformación al lenguaje de destino
  - La evaluación no es trivial, personas capacitadas
  - Se realizan estadísticas y comparaciones contra corpus
  - Un área muy difícil aún, debido a lo radical que pueden llegar a ser diferentes idiomas

NLP

Aplicaciones

- **Sistemas de búsqueda de respuestas**
  - Diseñados para tomar una pregunta en LN y proporcionar una respuesta
  - Motores de búsqueda
- **Generación de resúmenes automáticos**
  - A nivel de grupo y de documento
  - Extractivo y abstractivo

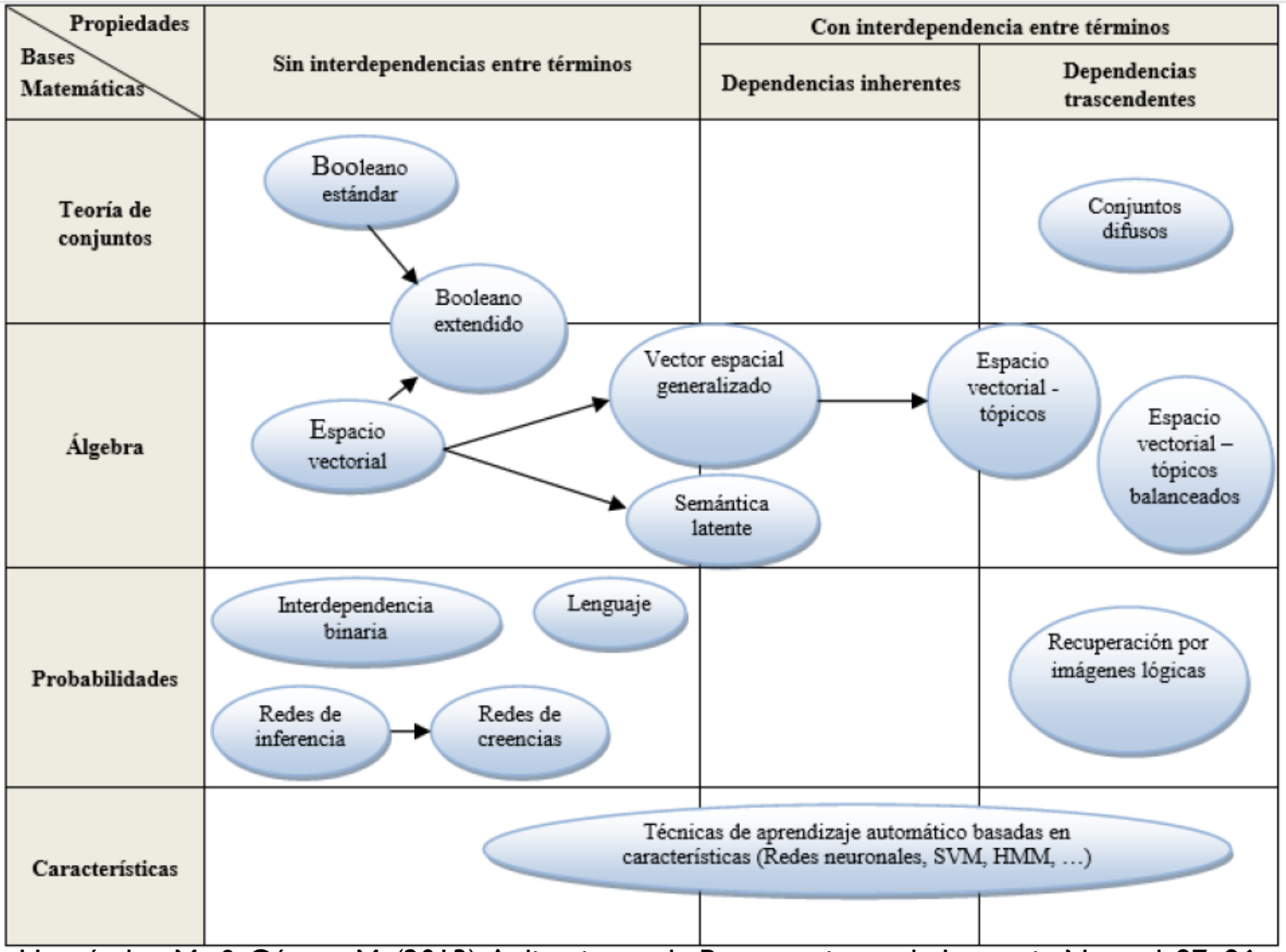
NLP

Aplicaciones

- **Recuperación de Información (RI)**
  - Material no estructurado (documentos)
  - Material semiestructurado (paginas web)
  - Material estructurado (Bases de datos)
- La RI transforma texto en representaciones adecuadas basada en modelos específicos entendibles por las maquinas

NLP

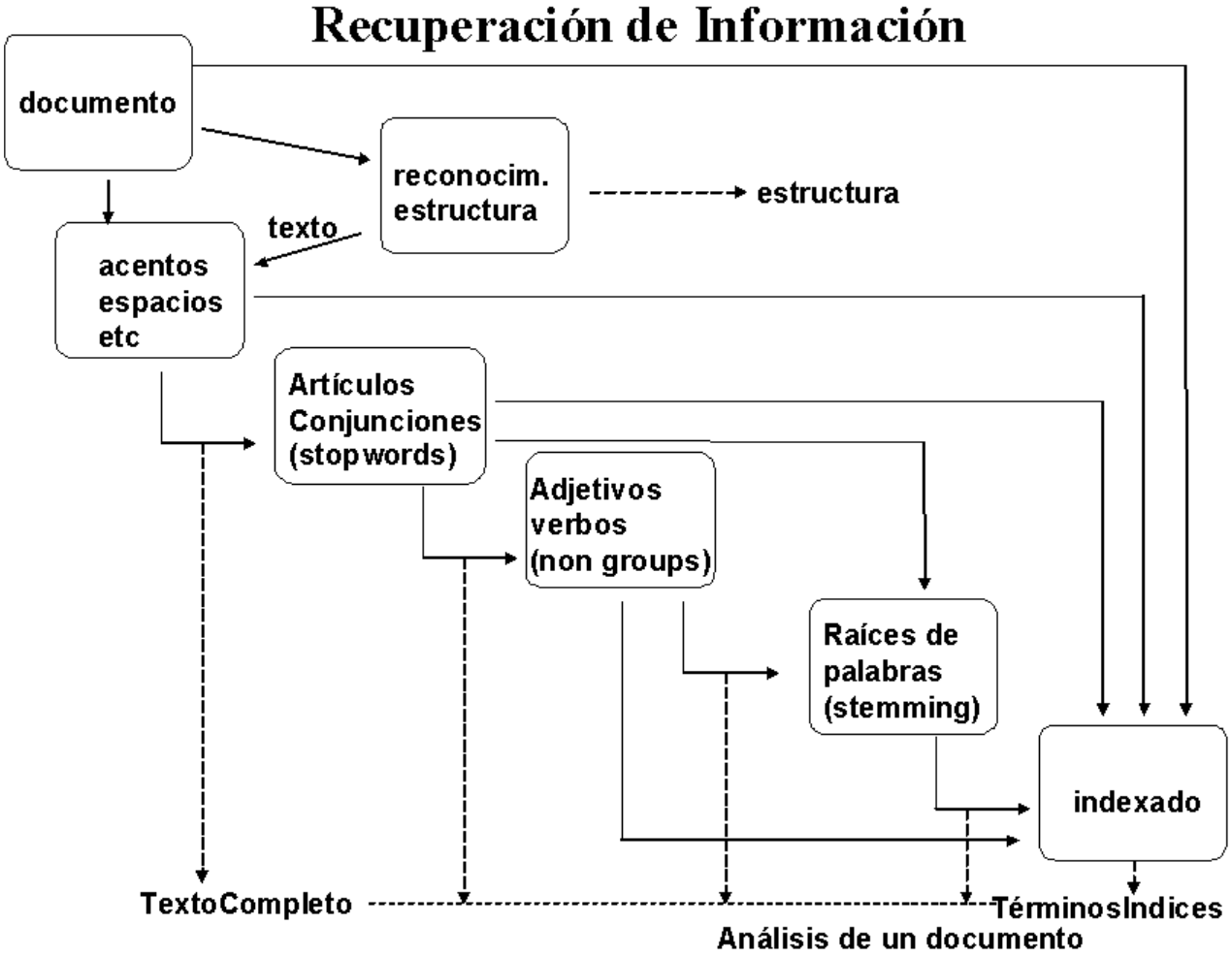
Aplicaciones



Hernández, M., & Gómez, M. (2013). Aplicaciones de Procesamiento de Lenguaje Natural, 87–96.

NLP

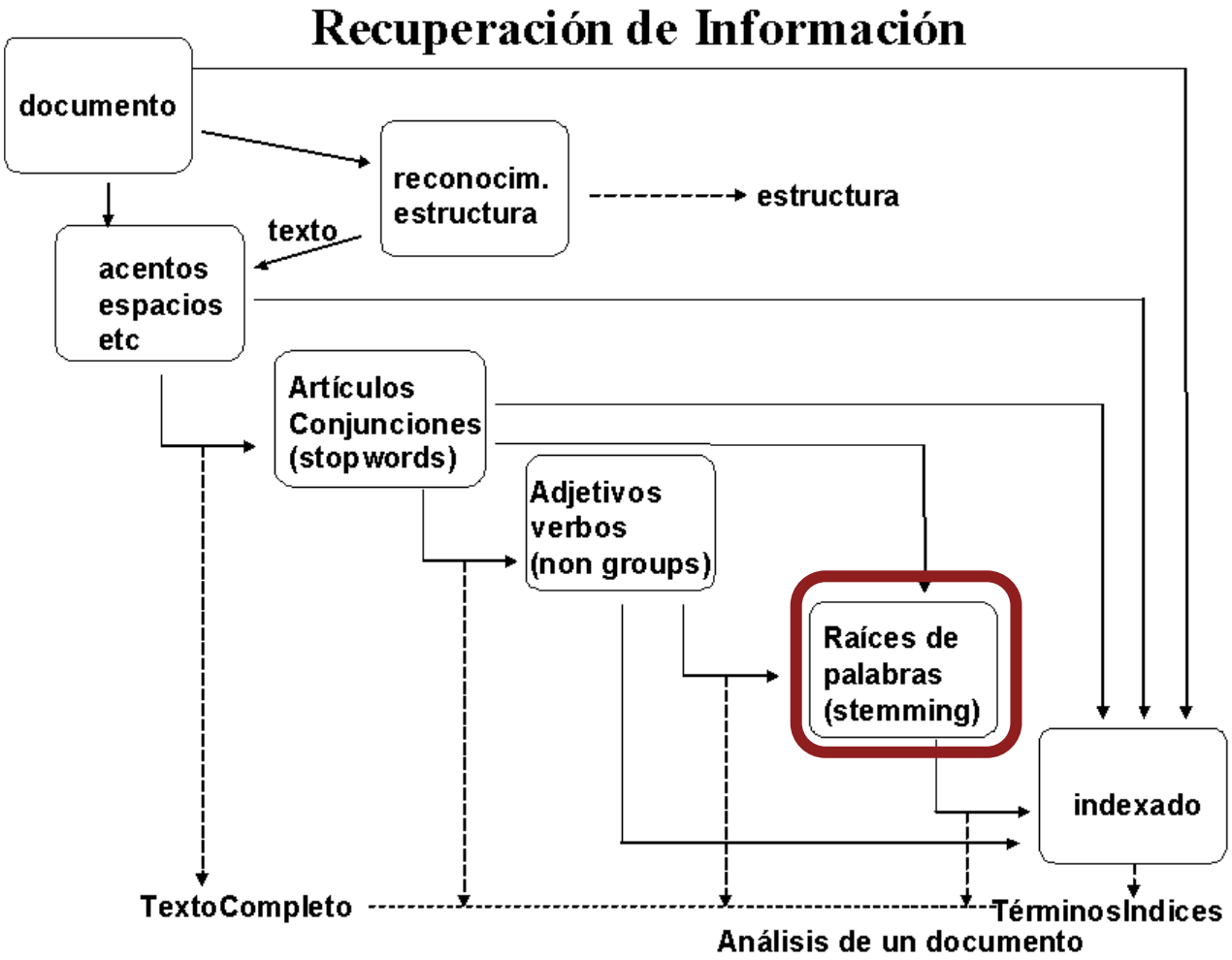
Aplicaciones



http://ict.udlap.mx/people/carlos/is346/admon08.html

NLP

Aplicaciones



http://ict.udlap.mx/people/carlos/is346/admon08.html

R.I. -> Lematización

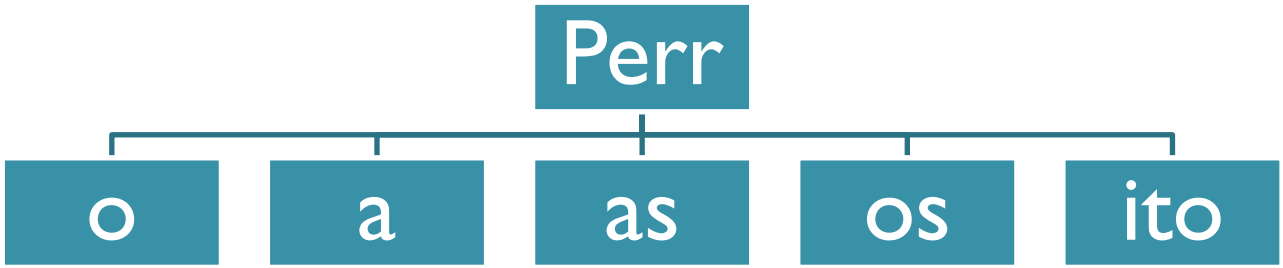
- El gran inconveniente de los sistemas tradicionales de R.I, es que no tomaban en cuenta las variantes morfológicas de una palabra, es por esto que surge el interés de crear algoritmos que reduzcan el numero de variantes morfológicas de las palabras, como lo hacen los lematizadores

Lematización

- *Lematización* es un proceso de eliminación automática de partes no esenciales de las palabras (sufijos, prefijos) para reducirlas a su parte original (lema).
- Es uno de los procesos fundamentales en el NLP.
- Es una técnica en la recuperación de datos de en los sistemas de información

Lematización

- Reduce variantes morfológicas de las formas de una palabra a raíces comunes, lema o lexemas y no tiene que tener significado

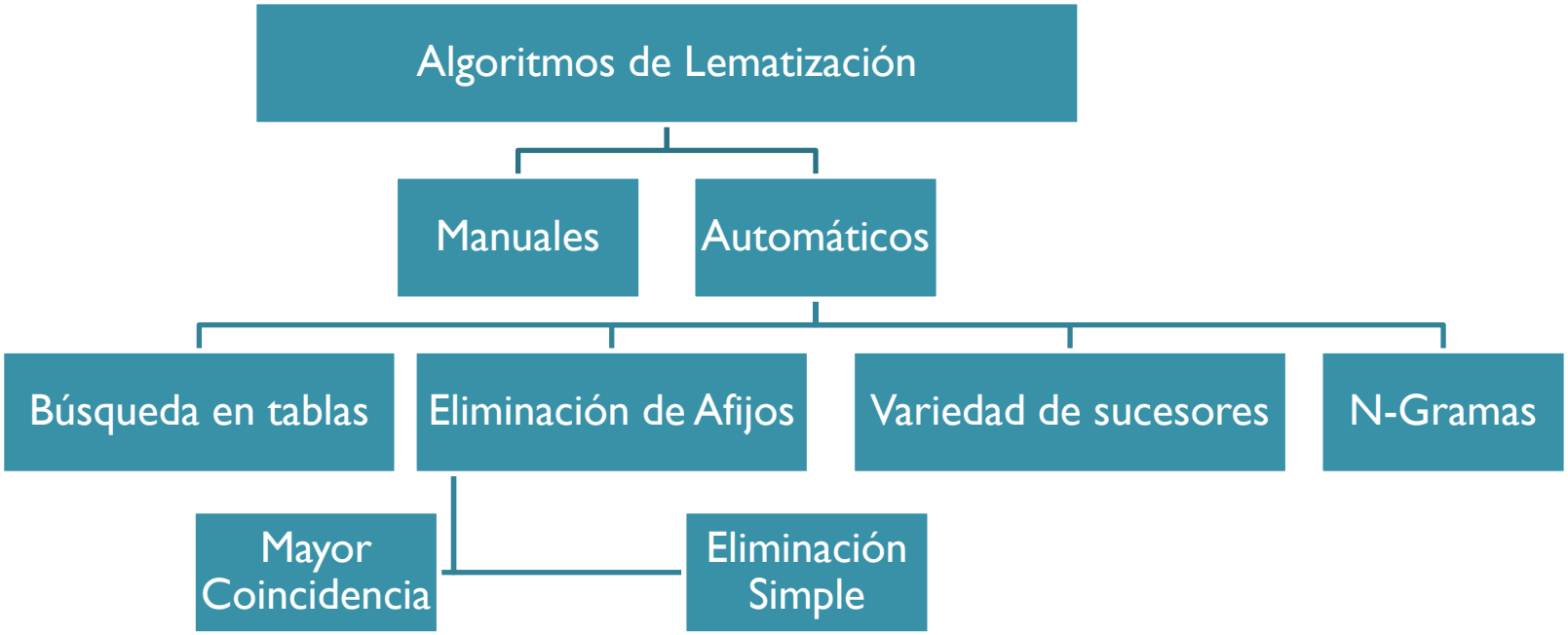


- La lematización también ayuda a reducir el tamaño de los índices, cerca de un 50%

Lematización

Algoritmos

- Existen diferentes algoritmos para realizar este proceso



Lematización

Algoritmos – Búsqueda en tabla

- Almacenamiento en tablas de los términos índices y sus lexemas
- Los términos de las consultas y de los índices podrán ser lematizados a través de la búsqueda.

Término	Lema
Tomar**la**	Tomar
Tomar**las**	Tomar
Tomar**lo**	Tomar
Tomar**los**	Tomar
Tomar**le**	Tomar
Tomar**les**	Tomar

Lematización

Algoritmos – Búsqueda en tabla

- Ventajas
  - Utilizando un arbol-b estas búsquedas serán muy rápidas
  - Es sencillo
- Desventajas
  - Se debe crear la tabla la primera vez
  - Difícil para palabras de un dominio especifico
  - Puede llegar a generar una alta cantidad de datos
  - No existen reglas para este proceso

Lematización

Algoritmos – Eliminación de afijos

- Consiste en la eliminación de ***sufijos*** y/o prefijos

Si una palabra termina en "**ies**" pero no en "**eies**" ni "**aies**", entonces: "**ies**" → "y"
Si una palabra termina en "**s**" pero no en "**us**" o "**ss**", entonces: "**s**" → "NULL" (se elimina)

- No es heurístico
- Se aplican reglas para obtener una forma común de una palabra
- Utiliza reglas gramaticales inversas
- Algoritmo de Porter

Lematización

Algoritmos – Eliminación de afijos

- Ventajas
  - Utiliza pocas reglas
  - Obtención de nuevos lemas
- Desventajas
  - Se debe implementar para cada idioma
  - Crear reglas
  - Reglas exhaustivas

Lematización

Algoritmos – Variedad de sucesores

- Se basan en lingüística estructural, determinado limites de las palabras y los morfemas
- Agrupa palabras con la misma raíz o lema, eliminando sufijos.
- Complejidad mayor a la eliminación de afijos
- La variedad de sucesores de una cadena es el numero de caracteres diferentes que siguen a esta cadena utilizando un corpus

Lematización

Algoritmos – Variedad de sucesores

	1	2	3	4	5	6	7	8	9
readable	r	e	a	d	a	b	l	e	

able	a	b	l	e					
ape	a	p	e						
beatable	b	e	a	t	a	b	l	e	
fixable	f	i	x	a	b	l	e		
read	r	e	a	d					
readable	r	e	a	d	a	b	l	e	
reading	r	e	a	d	i	n	g		
reads	r	e	a	d	s				
red	r	e	d						
rope	r	o	p	e					
ripe	r	i	p	e					

Prefijo	Variedad de sucesores	Letras
r	3	e-i-o
re	2	a-d
rea	1	d
read	3	a-i-s
reada	1	b
readab	1	l
readabl	1	e
readable	1	bco

Lematización

Algoritmos – Variedad de sucesores

- Luego de calcular la variedad se sucesores, se puede segmentar la palabra según 4 criterios
  - Valor de corte
    - Valor de corte muy pequeño o muy grande
  - Picos y valles
    - El corte se realiza cuando la variedad de sucesores excede a la del carácter que lo precede y la del que sigue.

Lematización

Algoritmos – Variedad de sucesores

- Palabra completa
  - Corte después de un segmento si es una palabra completa
- Método de la entropía
  - Método estadístico, que calcula distribuciones de las variedades de sucesores

Lematización

Algoritmos – N-Gramas

- Esta basado en el método bigramas
- Es heurístico
- No produce lematización
- Calcula medidas de asociación
- Se utiliza el coeficiente de Sorensen-Dice

$$^{2*C}/_{A+B} \rightarrow {}^{2*3}/_{4+5} = 0{,}6$$

cenar	CE		EN		NA		AR	
escena	ES		SC		CE		EN	NA

Lematización

Algoritmos – N-Gramas

- Se usa un umbral de 0,6 para hacer agrupaciones
- Ventajas
  - Agrupa palabras similares
  - Sencillo y fácil
- Desventajas
  - No es propiamente un lematizador
  - Over-Stemming (Escenario – Cena)
  - Under-Stemming (Maquina – Maquinaria)

Research paper

Migrating existing clinical content from ICD-9 to SNOMED

Prakash M Nadkarni,[1] Jonathan A Darer[2]

► Additional data are published online only. To view these files please visit the journal online (http://jamia.bmj.com)

[1]Yale University School of Medicine, USA
[2]Geisinger Health Systems, Danville, Pennsylvania, USA

Correspondence to
Dr Prakash M Nadkarni, Yale Center for Medical Informatics, 300 George St, New Haven, CT 06511, USA;
Prakash.Nadkarni@yale.edu

Received 15 September 2009
Accepted 30 May 2010

ABSTRACT

Objective To identify challenges in mapping internal International Classification of Disease, 9th edition, Clinical Modification (ICD-9-CM) encoded legacy data to Systematic Nomenclature of Medicine (SNOMED), using SNOMED-prescribed compositional approaches where appropriate, and to explore the mapping coverage provided by the US National Library of Medicine (NLM)'s SNOMED clinical core subset.

Design This study selected ICD-CM codes that occurred at least 100 times in the organization's problem list or diagnosis data in 2008. After eliminating codes whose exact mappings were already available in UMLS, the remainder were mapped manually with software assistance.

Results Of the 2194 codes, 784 (35.7%) required manual mapping. 435 of these represented concept types documented in SNOMED as deprecated: these included the qualifying phrases such as 'not elsewhere classified'. A third of the codes were composite, requiring multiple SNOMED code to map. Representing 45 composite concepts required introducing disjunction ('or') or set-difference ('without') operators, which are not currently defined in SNOMED. Only 47% of the concepts required for composition were present in the clinical core subset. Search of SNOMED for the correct concepts often required extensive application of knowledge of both English and medical synonymy.

Conclusion Strategies to deal with legacy ICD data

ones and failure to support multi-hierarchy.[2] ICD-10,[5] ICD-9's successor, remedies only the obsolescence problem.

Importantly, ICD lacks sufficient granularity to capture nuances of the clinical encounter that impact therapy/prognosis[6]; this impacts subsequent data analysis for purposes such as outcomes research.[7 8] By contrast, the Systematic Nomenclature of Medicine (SNOMED clinical terms; CT)[9] has been shown in numerous studies[10–15] to be significantly superior to any other single-source biomedical terminology for encounter encoding. SNOMED CT was originally developed by the College of American Pathologists and is now managed by the International Health Terminology Standards Development Organization (IHTSDO), an organization of organizations that includes the US National Library of Medicine (NLM).

Given, however, that most organizations have large volumes of existing (legacy) ICD-encoded data, several efforts focus on mapping between ICD and SNOMED CT, with the hope of eventually substituting the former with the latter. However, several challenges influence SNOMED CT's ultimate deployment; this paper explores some of these issues.

1. We explore issues in representing our legacy ICD-9 coded data with SNOMED building blocks, using a SNOMED CT-prescribed compo-

Trabajos futuros

- Teniendo el diccionario de terminologías "SNOMED-CT" en español, utilizar uno de estos algoritmos mostrados, para realizar la lematización de los términos.
- Esto tendría un impacto en los momentos que el usuario esta ingresando texto libre y el sistema ofrece términos como sugerencias a los que el usuarios va ingresando.

Referencias

- Hernández, M., & Gómez, M. (2013). Aplicaciones de Procesamiento de Lenguaje Natural, 87–96.
- Morales, L. P., Natural, P. D. L., Language, U. M., & Processing, N. L. (2010). Sistemas de Acceso Inteligente a la Información Biomédica : una revisión, (1), 7–15.
- Peinado R., J. (2003). Lematización para palabras médicas complejas: Implementación de un algoritmo en LISP.
- Benavides Cañón, P. A., & Correa, R. Procesamiento del lenguaje natural en la recuperación de información.
- Díaz Gómez, R. (2001). *Estudio de la incidencia del conocimiento lingüístico* en los sistemas de recuperación de la información para el español.
- Sosa, Eduardo, Procesamiento del lenguaje natural (2007): revisión del estado actual, bases teóricas y aplicaciones (Parte I y II), Revista internacional de Información y Comunicación
- Especial I+S: SNOMED CT Como teminología estándar en la historia clínica. (2010), 80, 9–13, 14–22. (2010), *80*, 9–13, 14–22.
- Hohendahl, A. T., & Zelasco, J. F. (2006). Lematizador y Flexionador con Estimador Idiomático , usando algoritmos eficientes para idiomas muy flexivos como el español. *CACIC*, 1–12.