

EXTRACCIÓN AUTOMÁTICA DE DICIONARIOS E IDENTIFICACIÓN DE COMPUESTOS. PRIMERA APROXIMACIÓN

Carla Parra Escartín
15/12/2009

0. PLANTEAMIENTO PRELIMINAR

Extracción automática de diccionarios a partir de textos traducidos de/al alemán

- Inducción de diccionarios
- Alineación palabra-palabra
- Segmentación de compuestos
- ¿Es posible extrapolar los resultados a otros idiomas germánicos?

2

1. INDUCCIÓN DE DICIONARIOS

- Aplicaciones en PLN:
 - Traducción automática
 - Traducción asistida
 - Permite inventariar las distintas traducciones de un mismo término y calcular frecuencias.
 - Cross-language information retrieval
 - Enseñanza de idiomas por ordenador
 - Desambiguación semántica
- Inicios → corpus bilingües alineados
- Nuevas corrientes → corpus comparables
 - Fung, 2000; Chiao and Zweigenbaum, 2002; Daille and Morin, 2005; Robitaille et al., 2006; Morin et al., 2007; Otero, 2008; Saralegi et al., 2008

3

1. INDUCCIÓN DE DICIONARIOS

A partir de corpus comparables: 2 estrategias

- a) Basada en el contexto → un término y su traducción aparecen en contextos léxicos similares
 - Daille and Morin, 2008; Fung, 2000; Chiao and Zweigenbaum, 2002; Daille and Morin, 2005; Robitaille et al., 2006; Morin et al., 2007; Daille and Morin, 2008; Saralegi et al., 2008.
- b) Basada en la sintaxis → se adquieren los contextos sintácticos, lo que aumenta la información léxica empleada para calcular la similitud léxica
 - Tanaka, 2002; Otero, 2007; Otero, 2008; Yu and Tsujii, 2009

4

1. INDUCCIÓN DE DICIONARIOS

A partir de corpus bilingües

1. Alineación frase-frase del corpus (si no lo está aún)
2. Alineación palabra-palabra del corpus
 - *** Gale 1991: correspondencia vs. alineación!
3. Generación del diccionario y validación
 - Algoritmo maximización de la expectativa (algoritmo EM)
 - Algoritmo voraz

5

1. INDUCCIÓN DE DICIONARIOS

Algoritmo maximización de la expectativa

Encontrar estimadores de máxima verosimilitud de parámetros en modelos probabilísticos que dependen de variables no observables.

1. Se computa la expectativa (E) de verosimilitud mediante la inclusión de variables ocultas como si fueran observables.
2. Maximización (M): se computan expectativas de máxima verosimilitud de los parámetros mediante la maximización de la verosimilitud esperada del paso E.
3. Los parámetros resultantes de M se emplean para el siguiente paso E, etc.

6

1. INDUCCIÓN DE DICCIONARIOS

Algoritmo voraz

Elegir la opción óptima en cada paso local con la esperanza de llegar a una solución general óptima

1. Elección de una similitud léxica S entre las palabras en $L1$ y $L2$.
→ Frecuencia de coaparición de palabras en las regiones correspondientes de un corpus de textos paralelos.
2. Cálculo de las puntuaciones de asociación $S(v,w)$ para un conjunto de parejas de palabras: $(v,w) \in (L1 \times L2)$
3. Clasificación de las parejas de palabras por orden descendiente
4. Selección de un umbral t . Las parejas de palabras que sobrepasen el umbral pasan a formar parte del diccionario.

7

1. INDUCCIÓN DE DICCIONARIOS

A partir de corpus anotado semánticamente (Dagan et al. 1993)

Identificación de la traducción correcta a partir de la anotación semántica en la LM

→ Mapeo de relaciones semánticas

8

2. ALINEACIÓN DE PALABRAS

Modelos estadísticos 1-5 de Brown et al. (1993)

- Basados en el par de lenguas inglés-francés
- Implementados en Giza ++
- Basados en el algoritmo EM para calcular la probabilidad condicional $\Pr(f|e)$ de que f sea la traducción de e : $t(f, e)$.

9

2. ALINEACIÓN DE PALABRAS

- **Modelos 1 y 2:** basados en la longitud de un string.
 - **M1:** orden de las palabras en e y f no afecta a $\Pr(f|e)$.
→ Todas las conexiones para cada posición en francés tienen la misma probabilidad.
 - **M2:** $\Pr(f|e)$ depende del orden de las palabras en e y f .
→ La probabilidad de conexión depende de las posiciones que conecta así como de la longitud de los dos strings.

10

2. ALINEACIÓN DE PALABRAS

- **Modelos 3, 4, y 5:** desarrollan el string en francés determinando, para cada palabra del string en inglés:

1. El número de palabras del string en francés que se conectarán a él.
2. La identidad de dichas palabras en francés
3. La posición real que dichas palabras ocuparán en el string en francés
 - a) **M3:** la probabilidad de una conexión depende de las posiciones que conecta y de las longitudes de los strings en inglés y francés.
 - b) **M4:** La probabilidad depende además de las identidades de las palabras en francés e inglés conectadas y de las posiciones de cualquier otra palabra en francés que esté conectada con la misma palabra en inglés.
 - c) **M5:** mejora más eficiente del M4.

11

3. PARTICULARIDADES DEL ALEMÁN (Y OTRAS LENGUAS GERMÁNICAS)

- Orden sintáctico problemático para la alineación de palabras:
 - Verbos separables
 - División del sintagma verbal
 - Verbos a final de frase en las subordinadas...
- Frecuencia muy alta de compuestos
 - Implica equivalencia $1 - n$ en la alineación
 - Compuestos lexicalizados vs. compuestos "nuevos"
 - Dificultad para inducción de diccionarios hacia el alemán!
 - Mayor frecuencia de aparición en textos especializados
 - Los compuestos son un problema en más lenguas: (inglés), holandés, noruego, finlandés, danés, sueco...

12

3. PARTICULARIDADES DEL ALEMÁN (Y OTRAS LENGUAS GERMÁNICAS)

- Compuestos: diversos análisis y casos posibles:

Prüfsystem → Prüfen (v) + System (n)
 Testing system → Testing (n) + system (n)
 Sistema de pruebas → SN con PP

- Bushaltestelle



- Handbremsvorrichtung



13

3. PARTICULARIDADES DEL ALEMÁN (Y OTRAS LENGUAS GERMÁNICAS)

Mecanismos de composición nominal en alemán:

- a) Composición de 2 ó más nombres

- No núcleo = complemento → traducción a español con PP (de...): *Busfahrer, Programmentwicklung, Datenschutz, Problemlösung*
- No núcleo = modificador → estructura en LM no predecible: *Landhaus, Fabrikarbeiter, Nordseeöl, Metallindustrie, Hauptaufgabe, Grundfähigkeit, Endprodukt, Schlüsselwort, Mitgliedstaat, Preisleistungsverhältnis*

14

3. PARTICULARIDADES DEL ALEMÁN (Y OTRAS LENGUAS GERMÁNICAS)

Mecanismos de composición nominal en alemán:

- b) Composición de raíz verbal y nombre

- El elemento no nuclear siempre es modificador
- Rol temático del marco argumental: *Schimmkran* (grúa flotante) → Kran = TEMA

- c) Composición de adjetivo y nombre

- El adjetivo siempre tiene la función de modificador del verbo: *Gesamtausgabe* (edición completa), *Höchstgeschwindigkeit* (velocidad máxima), *Zentraleinheit* (unidad central), *Privatbereich* (sector privado).

15

3. PARTICULARIDADES DEL ALEMÁN (Y OTRAS LENGUAS GERMÁNICAS)

Mecanismos de composición verbal en alemán:

- a) Composición de nombre + verbo (*teilnehmen – asistir, participar*)
- b) Composición de adjetivo + verbo (*trockenlegen – desecar, drenar*)
- c) Composición de verbo + verbo (*gefriertrocknen – liofilizar*)
- d) Composición de adverbio + verbo (*herunterschauen – mirar hacia abajo*)

→ Los neologismos no son tan frecuentes como en el caso de los sustantivos

16

3. PARTICULARIDADES DEL ALEMÁN (Y OTRAS LENGUAS GERMÁNICAS)

Mecanismos de composición adjetival en alemán:

- a) Composición de nombre + adjetivo (*wasserdicht – impermeable*)
- b) Composición de adjetivo + adjetivo (*frühreif – precoz, prematuro*)
- c) Composición de verbo + adjetivo (*fahrbereit – en estado de marcha*)
- d) Composición de adverbio + adjetivo (*andersfarbig – de otro color*)
- e) Composición de pronombre + adjetivo (*selbsterhell – autocrático*)

→ Los neologismos no son tan frecuentes como en el caso de los sustantivos

17

3. PARTICULARIDADES DEL ALEMÁN (Y OTRAS LENGUAS GERMÁNICAS)

Mecanismos de composición adverbial en alemán:

- a) Composición de adverbio + adverbio (*dorthin – ahí/allí*)
- b) Composición de adverbio + preposición (*hierbei*)
- c) Composición de preposición + adverbio (*nebenher*)
- d) Composición de preposición + preposición (*mitunter*)
- e) Composición de preposición + pronombre (*überdies*)
- f) Composición de nombre + adverbio (*flussabwärts*)
- g) Composición de adjetivo + adverbio (*schlechthin*)

→ No suelen aparecer en diccionarios y pueden tener diversas interpretaciones en función del contexto!!!

18

4. SEGMENTACIÓN DE COMPUESTOS

- Necesaria para inducción de diccionarios???
 - Podría ser un *by-product* final
- Útil para:
 - Correctores ortográficos y gramaticales
 - Information Retrieval
 - Segmentación de palabras en distintas líneas
 - Reconocimiento del habla
 - Traducción automática
 - Generación de textos

19

4. SEGMENTACIÓN DE COMPUESTOS

Metodologías de segmentación de compuestos

- a) Consultas recursivas a diccionarios
- b) División de palabras en n-gramas de caracteres que no aparecen en palabras no compuestas.
- c) Corpus paralelos: división a partir de cognados en la LM (Brown 2002)
- d) Enfoque lingüístico → análisis
- e) *¿Podemos crear un modelo estadístico a partir de diccionarios bilingües extraídos de corpus alineados?*

20

5. NEXT STEPS...

1. *State of the art* de las metodologías de inducción de diccionarios.
2. *State of the art* de los alineadores estadísticos y análisis de errores en el caso de lenguas germánicas (e.g. alemán).
3. Estudio de la tipología de los compuestos y su posible segmentación/generación.
4. Creación de un corpus de pruebas a partir de memorias de traducción de textos de DG Enterprise de la UE.
5. Primeras pruebas y evaluación de resultados.

21