Trabajo basado en Corpus

ITALICA
Universidad de Sevilla
José A. Troyano

Índice

- **Herramientas y recursos**
- Identificación de palabras
- Análisis morfológico
- Identificación de frases
- Esquemas de marcado
- Etiquetado gramatical

Herramientas y recursos

Corpus

- Linguistic Data Consortium (LDC)
  `http://www.ldc.upenn.edu`
- European Language Resources Association (ELRA)
  `http://www.icp.grenet.fr/ELRA/`
- International Computer Archive of Modern English (ICAME)
  `http://nora.hd.uib.no/icame.html`
- Oxford Text Archive (OTA)
  `http://ota.ahds.ac.uk/`
- Child Language Data Exchange System (CHILDES)
  `http://childes.psy.cmu.edu`

Herramientas y recursos

Etiquetadores gramaticales

- Basado en modelos ocultos de Markov
  http://www.english.bham.ac.uk/staff/oliver/software/tagger/
  http://www.coli.uni-sb.de/~thorsten/tnt/
- Basado en árboles de decisión
  http://www.cis.upenn.edu/~adwait/statnlp.html
  http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
- Basado el modelo de transformación de Brill
  http://www.cs.jhu.edu/~brill/
- Basado en el modelo de máxima entropía
  http://www.cis.upenn.edu/~adwait/statnlp.html
- Para el español:
  http://elvira.lllf.uam.es/~fernando/projects/esT.html

Herramientas y recursos

Software y lenguajes de desarrollo

- Editores planos
- Lenguajes de programación:
  - C/C++ más bibliotecas ER
  - Perl
  - Python
  - Java

Herramientas y recursos

Estrategias de programación

- Codificación de palabras
  - entrada (texto) ⇒ proceso(número) ⇒ salida (texto)
  - tablas *hash*
- Contar palabras
  - estructuras de datos (en memoria)
  - “tokenizar”, ordenar y contar (en ficheros)

Índice

- Herramientas y recursos
- **Identificación de palabras**
- Análisis morfológico
- Identificación de frases
- Esquemas de marcado
- Etiquetado gramatical

Identificación de palabras

Mayúsculas y minúsculas

- Normalización (pasar a minúsculas)
- Identificar nombres propios:
  - – Lista de nombres (problema: *Richard Brown* y *brown paint*)
  - – Sólo pasar a minúsculas las palabras del principio de una frase (problema: identificar el final de una frase)
  - – Otro problema: uso de las mayúsculas para resaltar palabras dentro de una frase: *Object Oriented Specification*

Identificación de palabras

Tokenización

Tokenizar: Identificar elementos básicos del lenguaje (palabras, números, símbolos de puntuación)

Palabra gráfica: Una cadena de caracteres alfanuméricos con espacios a cada lado; los únicos símbolos de puntuación que puede incluir son guiones o apóstrofes.

Excepciones: *$22.50*, *Micro$oft*, *C++*, *:-)*

Mejor pista: Espacios en blanco, pero hay varios problemas que hacen que no sea tan simple.

Identificación de palabras

Puntos

Además de por blancos, una palabra puede estar limitada por comas, puntos, puntos y comas, interrogaciones, admiraciones, paréntesis, ...

Abreviaturas: Suponen un problema porque el punto forma parte de la palabra, por ejemplo *Wash.* (además en este caso se puede tratar del nombre común *wash*).

Haplología: Síncopa de dos sílabas iguales o semejantes (cejunto por cejijunto). Se da un fenómeno similar cuando una abreviatura aparece al final de una frase, por ejemplo *etc.*

Identificación de palabras

Apóstrofes

¿Son *we'll* o *isn't* una palabra o dos?

En inglés este problema se acentúa por:

- Sufijos iguales para distintos verbos *she's* puede ser *she is* o *she has*
- El posesivo se forma de la misma manera

Identificación de palabras

Guiones

¿Deben contar las secuencias de caracteres separadas por un guión como una palabra o como dos?

Origen de los guiones:

- Tipográfico: Para justificar el texto al final de una línea (este tipo puede provocar haplología).
- Palabras gráficas: No tiene sentido tratarlas como palabras independientes, por ejemplo *e-mail*.
- Agrupaciones de palabras: Expresiones o incluso frases completas:
  - the alumnium-export ban
  - a final "take-it-or-leave-it" offer
  - he's 26-year-old

Identificación de palabras

Guiones (II)

Otros problemas:

• Inconsistencia: En un mismo texto puede haber distintas formas para una misma expresión. Por ejemplo *data-base*, *database* y *data base* (hilo de noticias de *Dow Jones*).

• Anotaciones: Se suelen utilizar guiones —como estos— para insertar notas o comentarios adicionales.

¿Qué hacer en general?

• Separar siempre que se pueda

• Separar y mantener información que recuerde que el original tenía guiones

Identificación de palabras

Espacios en blanco

Lenguajes sin espacios en blanco:

• Lenguas asiáticas: Chino, Japonés

• Alemán: En los nombres compuestos, por ejemplo

Lebensversicherungsgesellschaftsangestellter

"empleado de una empresa de seguros de vida"

Espacios que no suponen una división de palabras:

• Nombres compuestos: *New York-New Haven railroad*

• Verbos preposicionales: *make up*, *make it up*

• Expresiones hechas: *in spite of*, *in order to*

Identificación de palabras

Informaciones especiales

Distintas formas de escribir un número de teléfono:

Número	País	Número	País
0171 378 0674	UK	+45 43 48 60 60	Dinamarca
(44.171) 830 1007	UK	95-51-279648	Pakistán
+44 (0) 1225 753678	UK	+411/284 3797	Suiza
01256 468551	UK	(94-1) 866854	Sri Lanka
(202) 522-22300	USA	+49 69 136-2 98 05	Alemania
1-925-225-3000	USA	+34 96 387 9350	España
212. 995.5402	USA	91 538 21 06	España
33 1 34 43 32 26	Francia	900 900 900	España
++31-20-5200161	Holanda	95 455 2767-2753	España

Identificación de palabras

Corpus hablados

- Más contracciones
- Variantes de pronunciación
- Fragmentos de frases
- Ausencia de mayúsculas, símbolos de puntuación
- Expresiones de relleno como *er*, *um*

> *Also I [cough] not convinced that the, at least the kind of people that I work with, I'm not convinced that that's really, uh, doing much for the progr-, for the, uh, drug problem.*

Índice

- Herramientas y recursos
- Identificación de palabras
- **Análisis morfológico**
- Identificación de frases
- Esquemas de marcado
- Etiquetado gramatical

Análisis morfológico

Stemming y Lemmatization

Stemmig: Búsqueda del “tallo” de la palabra, habitualmente consiste en un simple algoritmo heurístico que elimina o sustituye partes de la palabra original.

Lemmatization: Búsqueda del lexema del que deriva la palabra, puede implicar un proceso de desambiguación.

Se ha demostrado empíricamente que este tipo de procesos no mejora sensiblemente los sistemas clásicos de recuperación de información. Sí son eficaces en aplicaciones más complejas.

Análisis morfológico

Problemas con el análisis morfológico

Pérdida de información: No es lo mismo recuperar documentos con la consulta *business* que con la consulta *busy*.

Fragmentación: El análisis morfológico puede separar una palabra en varias. En muchas ocasiones puede resultar beneficioso justamente lo contrario.

Dependencia del lenguaje: Hay lenguajes "poco morfológicos" como el inglés y otros con muchas posibilidades de inflexión y derivación. Por ejemplo el bantú:

akabimúha $\Rightarrow$ *a-ka-bi-mú-ha* $\Rightarrow$ *1ªsing - pasado- 3ªplu-3ªsing-dar*

$\Rightarrow$ *Yo se los di a él*

En estos lenguajes el análisis morfológico es más necesario, si no, el lexicón sería inmanejable.

Índice

Identificación de frases

Los límites de una frase

Problemas:

Puntos: El 90% de los puntos señala el final de una frase.

Otros símbolos de puntuación: Por ejemplo los dos puntos pueden dar paso a una frase o no.

Frases anidadas: Por ejemplo

You remind me, she remarked, of your mother.

Soluciones:

Árboles de clasificación estadísticos: Utilizan el caso y la longitud de las palabras que rodean a los símbolos de puntuación y sus probabilidades de ocurrencia a priori.

Otros métodos: Redes neuronales y análisis morfológico (98%), maximizar la entropía (99%).

Identificación de frases

Algoritmo de decisión heurístico

1. Marcar como límites provisionales las ocurrencias de . ? ! (y probablemente ; : —)
2. Mover la frontera hacia la derecha si se encuentra un nuevo símbolo de puntuación.
3. Ignorar el punto si:
   - Está precedido por una abreviatura conocida o una palabra que no aparece normalmente como final de una frase.
4. Ignorar el fin de una interrogación o admiración si:
   - Está seguida de una letra minúscula o un nombre conocido.
5. Cualquier otro límite provisional se considerará definitivo.

Identificación de frases

Tamaño de las frases

Longitud	Número	Porcentaje	Acumulado
1-5	1317	3.13	3.13
6-10	3215	7.64	10.77
11-15	5906	14.03	24.80
16-20	7206	17.12	41.92
21-25	7350	17.46	59.38
26-30	6281	14.92	74.30
31-35	4740	11.26	85.56
36-40	2826	6.71	92.26
41-45	1606	3.82	96.10
46-50	858	2.04	98.14
51-100	780	1.85	99.99
101+	6	0.01	100.0

Moda: 23

Parsing: Coste polinomial

Índice

- Herramientas y recursos
- Identificación de palabras
- Análisis morfológico
- Identificación de frases
- **Esquemas de marcado**
- Etiquetado gramatical

Esquemas de marcado

Markup (etiquetado)

Información adicional al texto (estructural)

COCOA: Información de cabecera como autor, fecha o título.

SGML: Standard Generalized Markup Language. Combina texto libre con marcas (sujetas a una gramática muy simple).

HTML: Instancia de SGML, especializada para páginas webs.

XML: Versión simplificada de SGML, también orientada a la web pero más flexible y potente que HTML. Separa los aspectos de representación de los datos.

Esquemas de marcado

SGML. Entornos anidados y referencias

Entornos anidados: Gramática independiente del contexto con símbolos de inicio y fin para cada entorno.

```
<p><s> And then he left. </s> <s> He did
not say another word.</s></p>
```

Referencias: Permiten especificar caracteres que no están el el conjunto de caracteres estándar.

```
&#x43; is the less than symbol

cami&oacute;n

This chapter was written on &docdate;.
```

Esquemas de marcado

XML. Un lenguaje de marcas adaptable

Las marcas no están prefijadas:

```
<?xml version="1.0" ?>
<!--My first XML document -->
  <BankAccount>
     <Number>1234</Number>
     <Name>Darshan Singh</Name>
     <Type>Checking</Type>
     <OpenDate>11/04/1974</OpenDate>
     <Balance>25382.20</Balance>
  </BankAccount>
```

Definición de Tipo de Documento (DTD): Permite describir las nuevas marcas y las relaciones entre ellas.

Transformación (XSLT): Permite describir transformaciones de documentos XML a otros documentos (XML, HTML, ...).

Índice

Etiquetado gramatical

Conjuntos de etiquetas

Reflejan información del “Part-Of-Speech”

Brown tag set: Usado en el corpus de la Universidad de Brown (años 60).

CLAWS1-5: Serie de cinco conjuntos desarrollado por la Universidad de Lancaster, usados para etiquetar los corpora *LOB* y el *British National Corpus* (años 80-90).

Penn Treebank tag set: Es uno de los más usados, es una versión simplificada del *Brown tag set* (años 90).

Suelen ser bastante dependientes del lenguaje

Etiquetado gramatical

Tamaño de los conjuntos de etiquetas

Conjunto	Número de etiquetas
Brown	87(179)
Penn	45
CLAWS1	132
CLAWS2	166
CLAWS5 (c5)	62

Varían en precisión y en orientación

Etiquetado gramatical

Ejemplo de etiquetado

Frase	c5	Brown	Penn
she	PNP	PPS	PRP
was	VBD	BEDZ	VBD
told	VVN	VBN	VBN
that	CJT	CS	IN
the	AT0	AT	DT
journey	NN1	NN	NN
might	VM0	MD	MD
kill	VVI	VB	VB
her	PNP	PPO	PRP
.	PUN	.	.

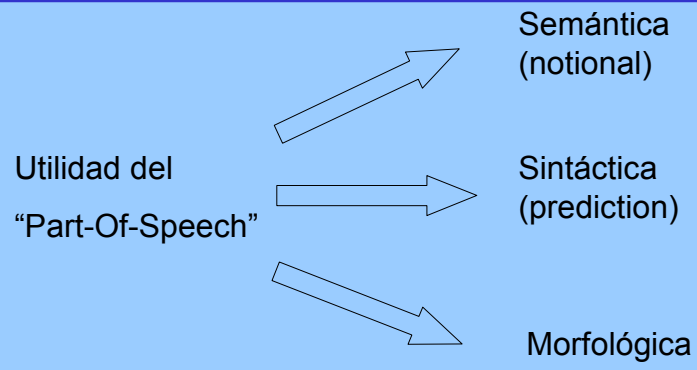
Etiquetado gramatical

Comparativa (p.e. pronombres)

Categorías (clásicas)	Ejemplos	c5	Penn	Brown
pronombre, nominal, indefinido	none, one	PNI	NN	PN
pronombre, personal, sujeto	you, we	PNP	PRP	PPSS
pronombre, personal, sujeto, 3SG	she, he, it	PNP	PRP	PPS
pronombre, personal, objeto	you, them, me	PNP	PRP	PPO
pronombre, reflexivo	herself, myself	PNX	PRP	PPL
pronombre, reflexivo, plural	themselves	PNX	PRP	PPLS
pronombre, interrogativo, sujeto	who, whoever	PNQ	WP	WPS
pronombre, interrogativo, objeto	who, whoever	PNQ	WP	WPO
pronombre, existencial	there	EX0	EX	EX

Etiquetado gramatical

Diseño de un conjunto de etiquetas



La elección del conjunto de etiquetas dependerá de las necesidades de nuestra aplicación.

Reconocimiento de entidades, desambiguadores:
Producen un etiquetado distinto (diferente semántica).