Preprocesamiento

Miguel A. Alonso Jorge Graña Jesús Vilares

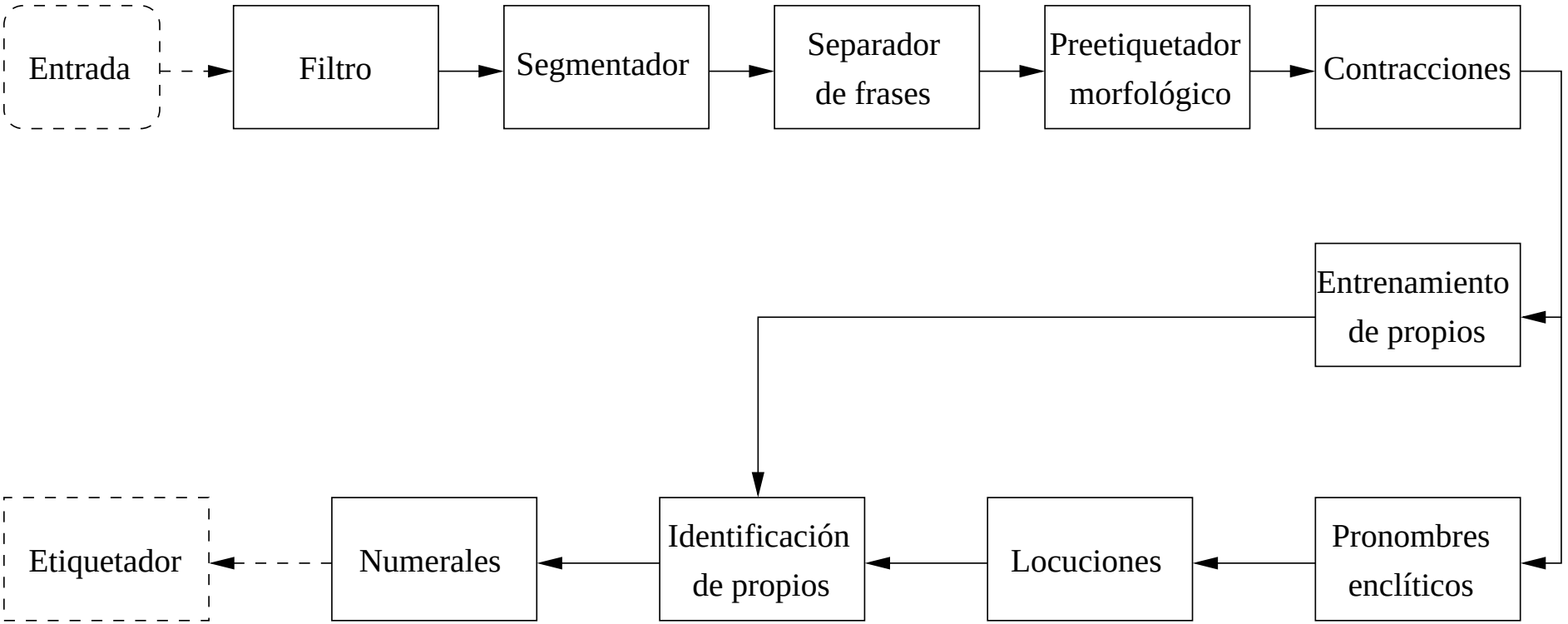
Departamento de Computación, Facultad de Informática, Universidade da Coruña

Índice

Introducción

- Una de las tareas previas más importantes en un sistema real de NLP es es la correcta segmentación y preprocesamiento de los textos
- Frecuentemente se obvia esta fase, lo que conduce a errores que afectarán al rendimiento del sistema entero
- La creciente disponibilidad de grandes corpus ha hecho que esta fase incremente su importancia, prestando especial atención a la robustez
- Problema de la dependencia con respecto a:
  - Idioma
  - Aplicación en la que se integrará el preprocesador, y que dictará las necesidades a cubrir
  - Ámbito de aplicación (por ejemplo, el preprocesamiento de textos literarios dista mucho del de textos científicos)

Esquema general



Filtro

- Procesos de conversión de otros formatos a texto plano
- Compactación de los separadores redundantes que existen en el texto (eliminar múltiples espacios, espacios a inicio de frase, etc.)

Segmentador

- Identificar y separar los **tokens** presentes en el texto, de manera que cada palabra ortográfica individual y cada signo de puntuación constituyan un **token** diferente.
- Tiene en consideración la existencia de:
  - abreviaturas: *etc.*
  - siglas: *CC.OO.*
  - números con decimales: *12,5*
  - fechas en formato numérico: *12/10/1492*
- Para ello se emplea un **diccionario de abreviaturas**, así como una serie de **patrones y reglas heurísticas** (expresiones regulares) para la detección de estos fenómenos.

Separador de frases: casuística

- Regla general: separar una frase ante un punto seguido de mayúscula
- Excepciones:
  - Abreviaturas a final de frase no seguidas por punto:
    "Traje queso, patatas, etc. Mi amigo trajo pollo"
  - Signos de interrogación y admiración finalizando la frase no seguidos por punto: *"¿Cuándo llegaste? No te vi entrar"*
  - Empleo de puntos suspensivos como fin de enunciado:
    "Dudé... Tenía miedo"
  - Omisión del punto fin de frase tras una sigla:
    "Trabaja en CC.OO. Da cursillos a parados"
- Excepciones de las excepciones:
  - Abreviaturas especiales como las empleadas en el tratamiento formal, direcciones postales, etc., que suelen ir acompañadas de mayúscula:
    Sr. Alonso y *avda. Fernández Latorre*
  - Abreviaturas en los nombres propios: *Miguel A. Alonso*
  - Puntos suspensivos para introducir matices de intriga o duda:
    "Me regaló... un coche"

Separador de frases: resolución

- Dos lexicones: uno de abreviaturas y otro de siglas
- Patrones y reglas heurísticas que permiten identificar cada caso y resolverlo adecuadamente
- La validez de las reglas reglas viene dada por el estilo y dominio de los documentos

Preetiquetador morfológico

- Etiquetar aquellos elementos cuya etiqueta se puede deducir a partir de la morfología de la palabra, sin que exista otra manera más fiable de hacerlo
- Patrones y reglas heurísticas
- Ejemplos:
  - los números y porcentajes se etiquetan como `Cifra`
  - Se asigna la etiqueta `Fecha` a las fechas en formatos diversos tales como *7/4/82*, *7 de abril de 1982* o *7 de abril*

Contracciones

- Desdoblar una contracción en sus diferentes componentes, etiquetando además cada uno de ellos
- Diccionario externo que especifica cómo se deben descomponer dichas contracciones
- Ejemplo: la salida correspondiente a la contracción `del` es:

```
de      [X     de]
+el     [DAMS  el]
```

Pronombres enclíticos

- Separar el verbo de sus pronombres, etiquetando correctamente cada una de las partes
- Problema importante en lenguas como el español y el gallego
- Se precisa:
  - Un diccionario que contiene el máximo número de formas verbales
  - Un diccionario con el máximo número de raíces verbales que pueden llevar pronombres enclíticos
  - Una lista con todas las combinaciones válidas de pronombre enclíticos
  - Una lista con todos los posibles pronombres enclíticos, junto con sus correspondientes etiquetas y lemas

Pronombres enclíticos: ejemplo

La descomposición de `cógeselo` es:

```
cóge    [V2SRM coger]
+se     [PY3P  le] [PY3P  se] [PY3S  le] [PY3S  se]
+lo     [PY3S  lo]
```

donde:

- *coge*: forma verbal de la segunda persona del singular del imperativo de `coger`
- *+se*: pronombre personal de la tercera persona, con cuatro pares etiqueta/lema posibles dependiendo de si se trata de una flexión de *le* o de *se*, o de una forma singular o plural
- *+lo*: pronombre personal de la tercera persona del singular.

Locuciones

- Concatenar los tokens que componen una locución y etiquetarlos como una unidad conjunta
- dos diccionarios de locuciones:
  - uno de locuciones que se sabe con seguridad que siempre son locuciones: *en vez de*
  - uno donde se encuentran las que pueden serlo o no: *sin embargo* puede ser una locución o la preposición *sin* y el sustantivo *embargo*.
- Una locución insegura conlleva una ambigüedad en la segmentación

Procesamiento de nombres propios

- Una de las tareas más complejas del preprocesamiento
- Imposible disponer de un diccionario con todos los posibles nombres de personas, lugares y entidades
- Dotar al sistema de la capacidad de aprender los nombres propios que aparecen en los documentos:
  1. Fase de entrenamiento: el sistema aprende los nombres propios no ambiguos contenidos en los documentos
  2. Fase de identificación de nombres propios

Entrenamiento de nombres propios

- Identificación de nuevos nombres propios situados en posiciones no ambiguas del texto: palabras ubicadas en posiciones en donde la utilización de mayúsculas indica sin ambigüedad que estamos ante un nombre propio
- Ejemplo: no se consideran las palabras en mayúscula ubicadas inmediatamente después de un punto
- Las palabras identificadas en esta fase dan lugar al diccionario de entrenamiento de nombres propios

Entrenamiento de nombres propios complejos

- Secuencias de palabras en mayúsculas interconectadas con nexos válidos (ciertas preposiciones y determinantes)
- Problema: no se puede determinar con seguridad si se trata de un único nombre propio o de una secuencia de nombres propios, por lo que deben considerarse todas las posibilidades
- Ejemplo: la secuencia *Consejo Superior de Cámaras de Comercio* genera los siguientes nombres propios válidos:

```
Consejo&Superior&de&Cámaras&de&Comercio
Consejo&Superior&de&Cámaras
Consejo&Superior
Superior&de&Cámaras&de&Comercio
Superior&de&Cámaras
Cámaras&de&Comercio
```

- Enfoque alternativo: identificación y extracción de nombres propios en base a reglas generadas automáticamente a partir de corpus etiquetados preexistentes

Identificación de nombres propios

- Entrada:
  - diccionario de entrenamiento de nombres propios
  - diccionario externo de nombres propios
- Salida:
  Etiqueta los nombres propios del texto, tanto simples como compuestos, y tanto en posiciones no ambiguas como ambiguas.

Proceso

- Posiciones no ambiguas:
  - Detecta el alcance del nombre propio (secuencias válidas que empiezan y terminan con una palabra en mayúscula)
  - Si el alcance total o una subsecuencia se encuentran en el diccionario externo, se etiqueta con la etiqueta correspondiente del diccionario
  - Si no existe una subsecuencia en el diccionario externo, se etiqueta como nombre propio pero sin especificar el género
- Posiciones ambiguas:
  - Detecta el alcance del nombre propio
  - Si el alcance o una subsecuencia se encuentran en el diccionario externo, se le asigna su correspondiente etiqueta
  - Si no existe ninguna subsecuencia en el diccionario externo, pero sí en el diccionario de entrenamiento de nombres propios, se etiqueta como nombre propio sin especificar el género
  - Si no existe ninguna subsecuencia en ninguno de los diccionarios, no asigna ninguna etiqueta.

Ejemplo

- Aparece el el texto nombre propio *Javier Pérez del Río* (nótese que *río* es un sustantivo común)
- Durante el entrenamiento sólo ha aparecido *Pérez del Río* en una posición no ambigua
- Además, *Javier* figura en el diccionario externo como nombre propio masculino singular
- Resultado: todo el nombre se etiqueta como nombre propio masculino singular.

Numerales

- Identificación de numerales compuestos mediante reglas heurísticas
- Ante la aparición de un numeral compuesto, se concatenan sus componentes de la misma manera que una locución, produciendo un único token
- Ejemplo, ante el numeral *mil doscientas* se genera:

  `mil&doscientas` `[DCFP mil&doscientas]` `[PCFP mil&doscientas]`

Ejemplo de descomposición de pronombres enclíticos

la descomposición de *ténselo* es ambigua:

- Forma verbal *tense* (de *tensar*) más el enclítico *lo*
- Forma verbal *ten* (de *tener*) más dos enclíticos, *se* y *lo*.

```
<alternativa>
<alternativa1>
ténse    [V2SRM   tensar]
+lo      [PY3S    lo]
</alternativa1>
<alternativa2>
tén      [V2SRM   tener]
+se      [PY3P    le] [PY3P    se] [PY3S    le] [PY3S    se]
+lo      [PY3S    lo]
</alternativa2>
</alternativa>
```

Ejemplo de locuciones inseguras

En gallego, la expresión *polo tanto* es muy ambigua:

- **Sustantivo+Adverbio**: *Coméche-lo polo tanto, que non quedaron nin os osos*
- **Preprosición+Artículo+Sustantivo**: *Gañaron o partido polo tanto da estrela*
- **Verbo+Pronombre+Adverbio**: *Pois agora polo tanto ti coma el*
- **Locución**: *Estou enfermo, polo tanto quédome na casa*

```
<alternativa>
<alternativa1>
polo     [Scms polo]
tanto
</alternativa1>
<alternativa2>
por      [P por]
+o       [Ddms o]
tanto
</alternativa2>
<alternativa3>
po       [Vpi2s0 pór] [Vpi2s0 poñer]
+o       [Raa3ms o]
tanto
</alternativa3>
<alternativa4>
por&+o&tanto
</alternativa4>
</alternativa>
```