Preprocessing and tokenization

Miguel A. Alonso

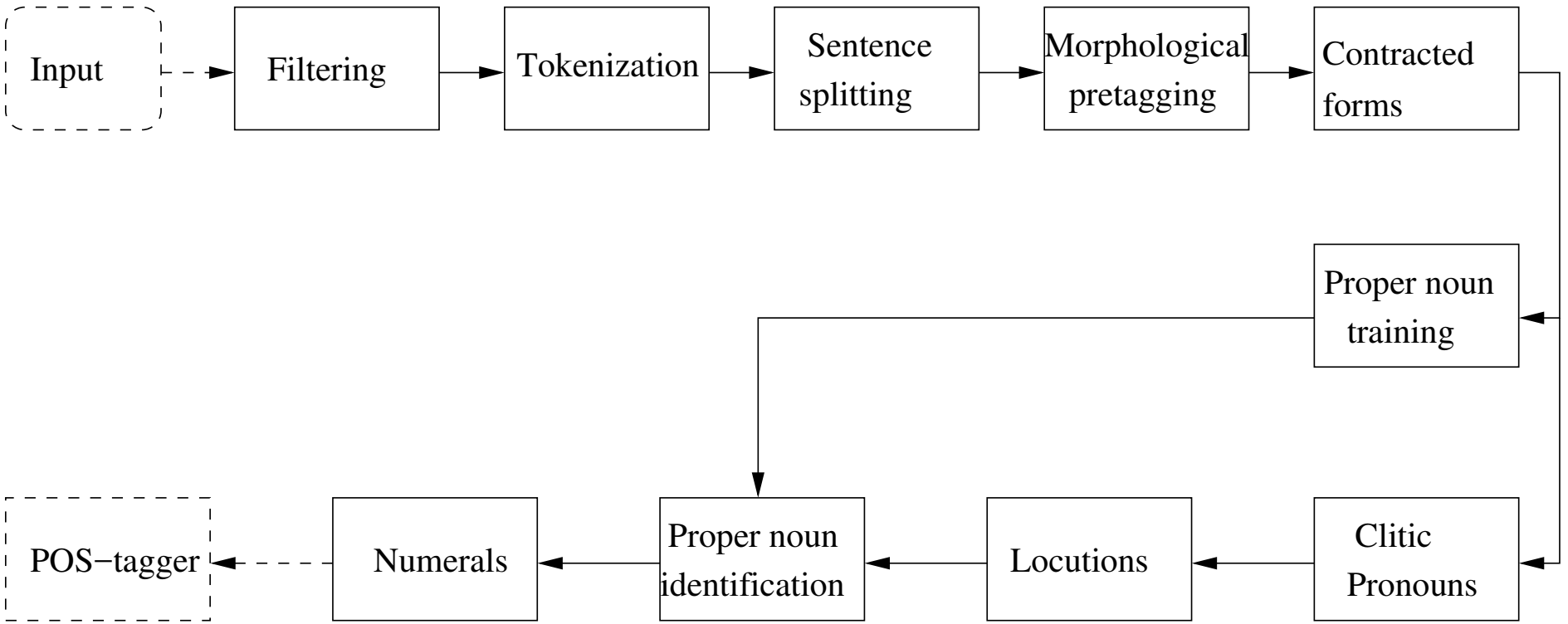
Departamento de Computación, Facultad de Informática, Universidade da Coruña

Contents

Introduction

- One of the most important previous tasks in a real NLP system NLP is segmentation and preprocessing of texts
- Frequently, this phase is ignored, which leads to errors that will affect the performance of the entire system
- The increasing application of NLP systems to real texts increses the importance of these tasks. We must pay special attention to robustness
- Theses tasks are strongly dependent on:
  - Language
  - Application
  - Domain

General scheme



Filtering

- Conversion from other formats (HTML, PDF, MS Word, ...) to plain text
- Treatment of redundant separators existing in text (deleting multiple spaces; spaces at the beginning of sentence; etc.)

Tokenization

- Identify and separate the **tokens** present in the text, so that each individual word and each punctuation constitute a different **token**.
- Consider:
  - abbreviations: *etc.*
  - acronyms: *ACL, SIGIR*
  - decimal numbers: *12.5*
  - dates in numeric format: *12/10/1492*
- For this, a **dictionary of abbreviations** is used, as well as a series of **heuristic patterns and rules** (regular expressions) for the detection of these phenomena.

Sentence splitting

- Separate sentences at each period followed by a capital letter
- Exceptions:
  - Abbreviations at the end of a sentence not followed by a period: *"I brought cheese, potatoes, etc. My friend brought chicken"*
  - Question and admiration marks ending the sentence not followed by a period: *"When did you arrive? I did not see you comming in"*
  - Use of ellipses as the end of the sentence: *"I hesitated... I was afraid"*
  - Omitting the end of sentence after an acronym: *"She works at X.Y.Z. She earns a lot of money"*
- Exceptions to exceptions:
  - Special abbreviations such as those used in formal treatment, postal addresses, etc., which are usually accompanied by a capital letter: *Mr. Alonso* y *Avda. Fernández Latorre*
  - Abbreviations in proper names: *Miguel A. Alonso*
  - Ellipsis to introduce nuances of intrigue or hesitation: *"She gave me... a gift"*

Sentence splitting

- Two lexicons: one of abbreviations and one of acronyms
- Heuristic patterns and rules that allow each case to be identified and solve it properly
- The reliability of the rules is dependent on the style and domain of documents

Morphological pre-tagging

- To tag those elements whose tag can be deduced from the morphology of the word, without another more reliable way of doing it
- Heuristic patterns and rules
- Examples:
  - Numbers and percentages are labeled as `Number`
  - The label `Date` is assigned to dates in formats such as *7/4/82*, *7 April 1982* or *April 7, 1982*

Contracted forms

- Unfold a contraction into its components, tagging each of them
- An external lexicon specifies how we should be break down those contractions
- Example: the output corresponding to the Spanish contraction `del` is:

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

Clitic pronouns

- Separate the verb from its pronouns, tagging each of the parts
- A major problem in languages such as Spanish and Galician
- It is required:
  - A lexicon with as many verb forms as possible.
  - A lexicon with verbal roots that can carry enclitic pronouns
  - A list of valid clitic pronoun combinations
  - A list of all possible clitic pronouns, along with their corresponding tags and lemmas

Clitic pronouns: an example in Spanish

The word `cógeselo` is decomposed into:

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

where:

- *coge*: verbal form of the second person of the singular of the imperative of `coger` (to take)
- *+se*: personal pronoun of the third person, with four possible tag/lemma pairs depending on whether it is from *le* or *se*, or in a singular or plural form (it/them)
- *+lo*: personal pronoun of the third person singular (to him)

Locutions (adverbial phrases, ...)

- Concatenate the tokens that make up a locution and tag them as a joint unit
- Two lexicons of locutions:
  - one of locutions that are known for sure that they are always locutions: *en vez de* (instead of)
  - a lexicon with dubious locutions: *sin embargo* (however) can be a single adverb or the preposition *sin* (without) and the noun *embargo* (seizure)
- A dubious locution leads to ambiguous segmentation

Proper nouns

- One of the most complex preprocessing tasks
- It is unfeasible to have a lexicon with all possible names of people, places and entities
- We must provide the system with the ability to learn proper names that appear in documents:
  1. Training phase: the system learns the unambiguous proper names contained in the documents
  2. Identification phase

Training proper nouns

- Identification of new proper names located in unambiguous positions of the text: words located in positions where the use of capital letters unambiguously indicates that it is a proper name (in particular: uppercase words located immediately after a period are NOT considered unambiguous)
- Words identified in this phase result in the learned lexicon of proper names

Training complex proper nouns

- Uppercase word sequences interconnected with valid links (certain prepositions and determinants)
- Issue: it cannot be determined with certainty whether we are in the presence of a unique proper name or in the presence of a sequence of proper names, so all possibilities must be considered
- Example: the sequence *High Council of Chambers of Commerce* generates the following valid proper names:

```
High&Council&of&Chambers&of&Commerce
High&Council&de&Chambers
High&Council
Council&of&Chambers&of&Commerce
Council&of&Chambers
Chambers&of&Commerce
```

- Alternative approach: identification and extraction of proper names based on machine learning or rules automatically generated from pre-existing labeled corpora (information extraction approach)

Identification of proper nouns

- Input:
  - learned lexicon of proper nouns
  - external lexicon of proper nouns
- Output:
  proper names of the text, both simple and complex, and both in unambiguous and ambiguous positions.

Process

- Non-ambiguous positions:
  - Detects the scope of the proper name (valid sequences that start and end with a capitalized word)
  - If such scope or a subsequence of it is found in the external lexicon, it is tagged with the corresponding lexicon tag
  - If there is no subsequence in the external dictionary, it is tagged as its own name without specifying the gender
- Ambiguous positions:
  - Detects the scope of the proper name
  - If such a scope or a subsequence of it is in the external lexicon, it is assigned the corresponding tag
  - If there is no such subsequence in the external lexicon, but it is found in the learned lexicon of proper names, it is tagged as its own name without specifying the gender
  - If there is no subsequence in any of the lexicons, no tag is assigned.

Example in Spanish

- The proper noun *Javier Pérez del Río* appears in the text (Note that *río* (river) is also a common noun)
- In the training phase, only *Pérez del Río* have appear in a non-ambiguous position
- *Javier* is found in the external lexicon as a proper noun, masculine singular
- Result: the whole name is tagged as a proper noun, masculine singular

Numerals

- Identification of compound numeral using heuristic rules
- Before the appearance of a compound numeral, its components are concatenated in the same way as a phrase, producing a single token
- For example, the numeral *two hundred and fifty*:

  `two&hundred&and&fifty` `[DCFP two&hundred&and&fifty]`

Example of segmenting clitic pronouns

The segmentation of *Ténselo* is ambiguous:

- Verbal form *tense* (to pull) and the pronoun *lo* (it)
- Verbal form *ten* (to have) and two pronouns *se* (it) and *lo* (to him)

```
<alternative>
<alternative1>
ténse    [V2SRM   tensar]
+lo      [PY3S    lo]
</alternative1>
<alternative2>
tén      [V2SRM   tener]
+se      [PY3P    le] [PY3P    se] [PY3S    le] [PY3S    se]
+lo      [PY3S    lo]
</alternative2>
</alternative>
```

Example of dubious locutions

In Galician, the expression *polo tanto* is highly ambiguous:

- **Noun+Adverb** (chicken so much): *Coméche-lo polo tanto, que non quedaron nin os osos*
- **Preposition+Determiner+Noun** (by the goal): *Gañaron o partido polo tanto da estrela*
- **Verb+Pronoun+Adverb** (put it both): *Pois agora polo tanto ti coma el*
- **Adverbial phrase** (thus): *Estou enfermo, polo tanto quédome na casa*

```
<alternative>
<alternative1>
polo     [Scms polo]
tanto
</alternative1>
<alternative2>
por      [P por]
+o       [Ddms o]
tanto
</alternative2>
<alternative3>
po       [Vpi2s0 pór] [Vpi2s0 poñer]
+o       [Raa3ms o]
tanto
</alternative3>
<alternative4>
por&+o&tanto
</alternative4>
</alternative>
```

End