Searching for linguistic phenomena in literary digital libraries

Felipe Sánchez-Martínez, Mikel L. Forcada, Rafael C. Carrasco

Transducens Group, Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain
{fsanchez,mlf,carrasco}@dlsi.ua.es

1. Introduction

Goal

- To search for specific linguistic phenomena, allowing to explore in a richer way the cultural heritage in current digital libraries

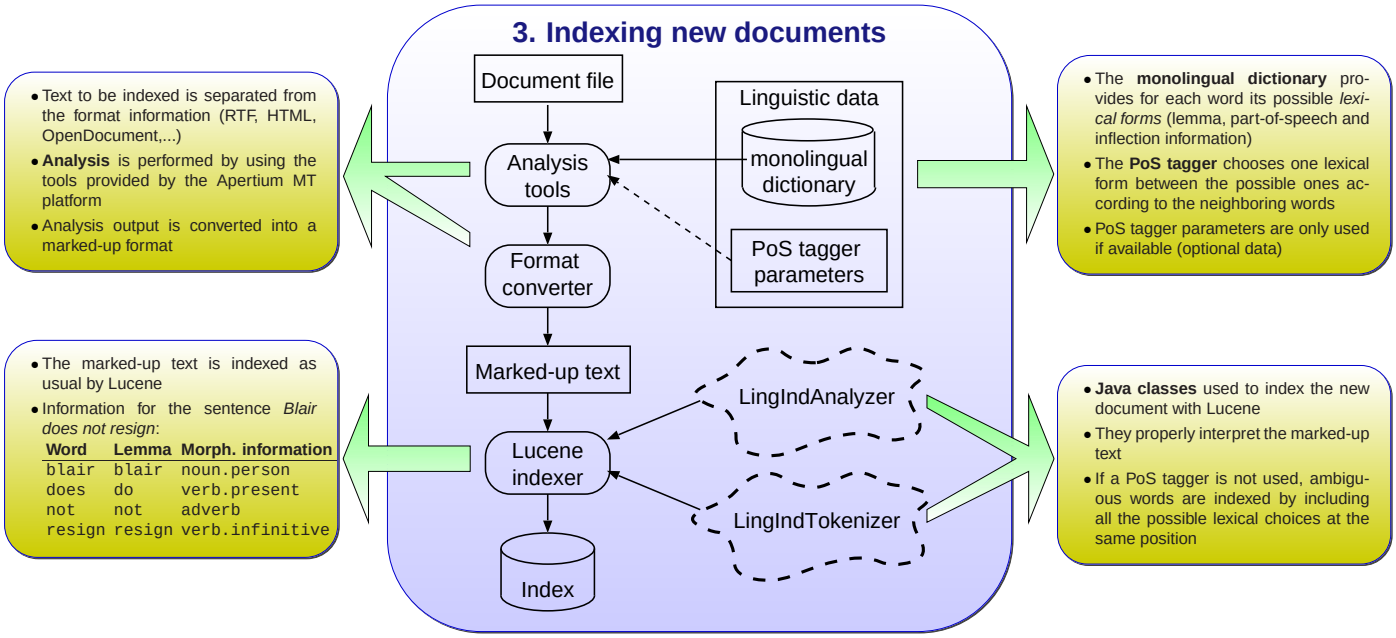
How?

- By using existing linguistic resources to allow the Java Lucene text search engine (`http://lucene.apache.org`) to use morphological information to index and search
  - Morphological attributes can be used to specify query terms

2. Linguistic resources

- Morphological dictionaries and part-of-speech (PoS) taggers developed for the open-source machine translation (MT) platform Apertium (`http://www.apertium.org`)
- Several languages available: Spanish, Catalan, Occitan, Galician, Portuguese, Romanian, French, English, ...
- Free download from `http://sf.net/projects/apertium`
- Apertium linguistic data is coded using XML-based formats
  - Allows for interoperability, and for easy data transformation and maintenance

3. Indexing new documents



- Text to be indexed is separated from the format information (RTF, HTML, OpenDocument,...)
- **Analysis** is performed by using the tools provided by the Apertium MT platform
- Analysis output is converted into a marked-up format

- The **monolingual dictionary** provides for each word its possible *lexical forms* (lemma, part-of-speech and inflection information)
- The **PoS tagger** chooses one lexical form between the possible ones according to the neighboring words
- PoS tagger parameters are only used if available (optional data)

- The marked-up text is indexed as usual by Lucene
- Information for the sentence *Blair does not resign*:

Word	Lemma	Morph. information
blair	blair	noun.person
does	do	verb.present
not	not	adverb
resign	resign	verb.infinitive

- **Java classes** used to index the new document with Lucene
- They properly interpret the marked-up text
- If a PoS tagger is not used, ambiguous words are indexed by including all the possible lexical choices at the same position

4. Searching for documents

- Queries are written in the language accepted by Lucene's query parser
- Prefixes to differentiate traditional and morphological terms:
  `lem#` for lemmas (e.g., `lem#do`)
  `tag#` for PoS tags and morphological information (e.g., `tag#verb.infinitive`)

Searches tested over Spanish texts

`"lem#dignar tag#prep"` searches for any form of the verb *dignar* followed by a preposition

> ... se han **dignado a** recibirnos ...
> ... una sede **digna de** mención ... *

`"lem#deber de tag#verb.infinitive"` searches for any form of the Spanish verb *deber* followed by the preposition *de* and by a verb in infinitive mood

- this query does not disambiguate between the verb *deber* and the noun *deber*

> ... la ONU **debe de reaccionar** urgentemente ...
> ... no **debiera de estar** representado ...

5. Example results

Spanish texts come from the news agency EFE (`http://ww.efe.com`)

Using a PoS tagger

	Precision	Recall	F-measure
`"lem#dignar tag#prep"`	60%	100%	75%
`"lem#deber de tag#verb.infinitive"`	82%	100%	90%

Using no PoS tagger

	Precision	Recall	F-measure
`"lem#dignar tag#prep"`	2%	100%	5%
`"lem#deber de tag#verb.infinitive"`	82%	100%	90%

- Large difference between the results achieved for the two queries
  - First query achieves better results when a PoS tagger is used to build the index
  - Second query achieves the same results in both cases

Possible explanation: Longer queries virtually disambiguate the text at search time

6. Discussion

- The use of morphological attributes to define query terms makes it possible to search for specific linguistic phenomena
  - Ease the access and study of the cultural heritage found in current digital libraries
- We plan to integrate this tool with the *Biblioteca Virtual Miguel de Cervantes* (`http://www.cervantesvirtual.com`)
- This tool has been released as open-source and can be freely downloaded from `http://sf.net/projects/apertium/`, package name `apertium-morph`

Further reading

Armentano-Oller, C. et al. (2006). "Open-source Portuguese-Spanish machine translation". In *Lecture Notes in Computer Science 3960 (Computational Processing of the Portuguese Language)*, p. 50–59, Rio de Janeiro, Brazil. (`http://www.dlsi.ua.es/~fsanchez/pub/pdf/armentano06.pdf`)

Presented at *the ECDL 2008 Workshop on Information Access to Cultural Heritage*, September 18, 2008, Århus, Denmark.