Text Mining Wikipedia to extract historical facts

João Valente and João Gradim

Faculdade de Engenharia da Universidade do Porto,
Rua do Dr. Roberto Frias, s/n, Porto, Portugal

Abstract. The abstract should summarize the contents of the paper using at least 70 and at most 150 words. It will be set in 9-point font size and be inset 1.0 cm from the right and left margins. There will be two blank lines before and after the Abstract. ...

1 Natural Language Processing

1.1 People Extraction

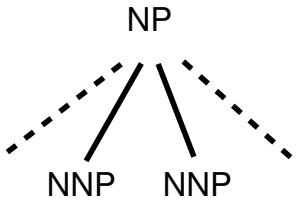
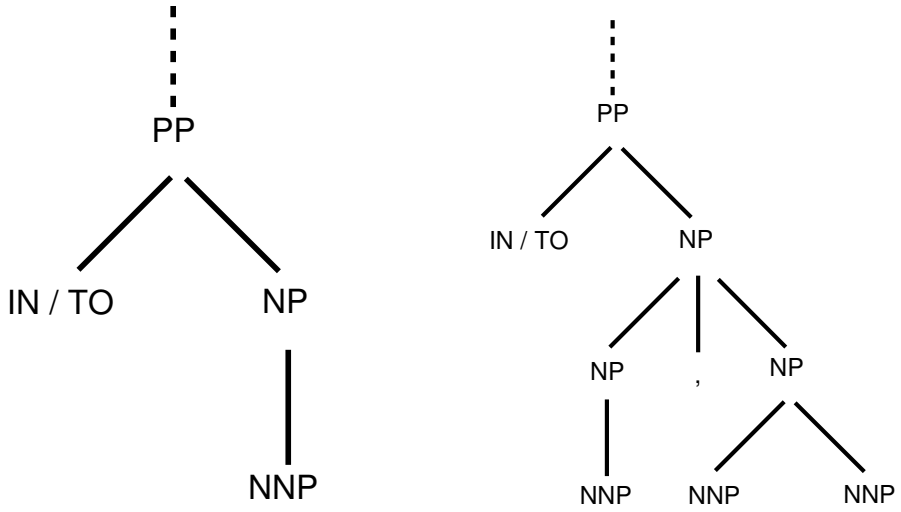


Fig. 1: Subtree relating to proper nouns, from where names of people involved can be extracted

1.2 Local extraction



(a) Local extraction pattern 1 (b) Local extraction pattern 2

Fig. 2: Subtrees for event location extraction

This kind of structure is present on most events parsed from Wikipedia and allows the retrieval of the phisical location of the occurence, mainly the country or a city in a determined country.

2 Main implementation problems

Wikipedia is an open knowledge base, relying mainly on user generated content. As such, it's difficult to ensure a proper and constant textual structure for information. Although the pages for the most recent centuries (aprox. 18th century) have a well defined an constant HTML strutcture that allows for reliable information retrieval, there are many years that don't follow this structure, leading to specific parsing cases.

This openness lead to another problem: an HTML structure not suited for easy parsing. A series of workarounds had to be implemented to successfuly extract useful information.

References

[1990] Santorini, B. : Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision, 2nd Printing). (July 1990)

[1995] Bies, A., Ferguson, M., Katz, K., MacIntyre, R. : Bracketing Guidelines for Treebank II Penn Treebank Project (June 1995)

Subject Index