Text Mining Wikipedia to extract historical facts

João Valente – ei05029@fe.up.pt
João Gradim – ei05030@fe.up.pt

Seminário de Sistemas Inteligentes, Interacção e Multimédia

January 7th, 2010

The Problem

- Information is available;
  - But is not easily searchable: requires manual searching through entire articles to find a specific piece of information
- A relational, cross-referenced database would provide easy access to specific information
  - Queries could be performed using natural language

Objectives

- To provide an easily queryable database of historical events of major importance
- To allow users to use natural language to perform queries
- To be able to cross-reference historical events and link figures, places, etc...

Motivation

"Those who don't know history are destined to repeat it."

Edmund Burke (1729-1797)

Approach

- Text classification
  - Naive Bayes classifier
- Natural Language Processing
  - The Stanford Parser

Approach

- Ruby language
- Extract events from Wikipedia
  - Parse HTML, extract date and description
  - Classify event
  - Extract features
- Use processed events to build a web app

Text Classification

- Naive Bayes Classifier
  - Ruby implementation
- 7 categories
  - 50 training examples per category, total of 350 examples

Text Classification - Results

- With a test set of 145 elements:

	Correctly classified as	Incorrectly classified as
Accidents	19	4
Crime	14	5
Cultural	12	2
Economy	5	16
Politics	29	1
Science	12	7
War	10	7
Total	103	42

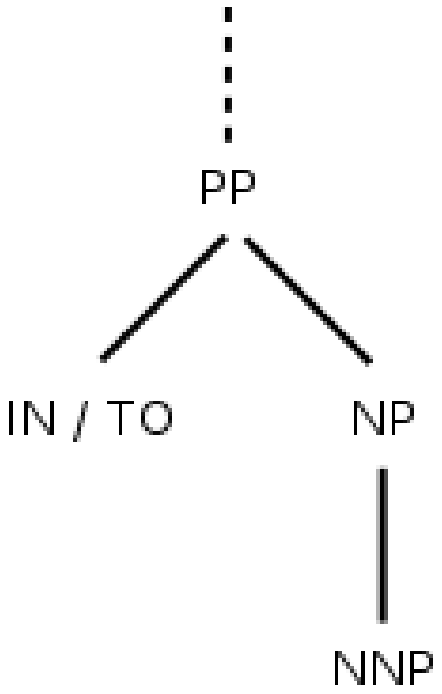
Accuracy: 71%

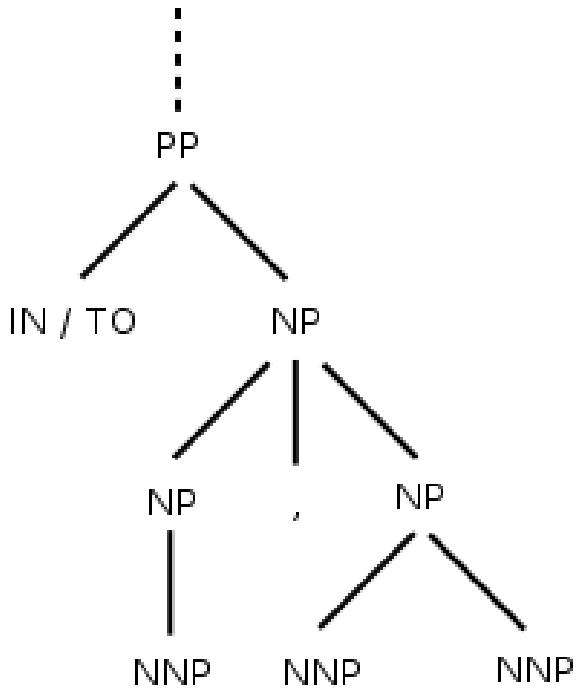
Feature Extraction

- Stanford Parser
- 3 different parse trees
  - Able to extract locations and people involved

Feature Extraction

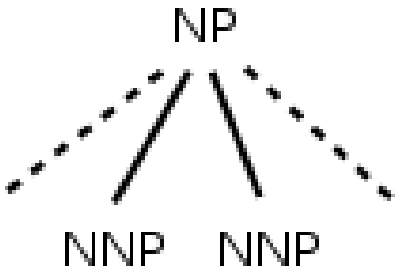
Location extraction





Feature Extraction

People extraction



Feature Extraction - Results

- With a test set of 145 elements:

	Location Extraction	People Extraction
Accidents	19	4
Crime	14	5
Total	103	42

Accuracy:
People Extraction – 72%
Location Extraction – 66%

Conclusions

- Natural language processing can effectively enrichen plain data with significant meaning
- A larger and less biased training set could improve the classifications results
- A more detailed analysis of the phrasal structures could improve the results of feature extraction