Final project
Project Text Analysis

Mart Busger op Vollenbroek
Olivier Louwaars

Table of contents

- Assignment
- Our approach
- Challenges
- Results
- Evaluation

Assignment

- NER
- Wikification
- Just like previous assignments…

Our approach (1)

- Input: development.set
- Train NER tagger
- Extract first columns in development.set
- Add NER tags

Our approach (2)

- Wikification
- Extract context from sentences
- Using wikipedia API to find link
- Compare with WordNet synsets
- Add best link to column

Challenges

- Multi-word entities:
    - Reverse list -> add previous tag
- Ambiguous words
    - Lesk
- Getting the right wiki using machine learning

Results (1)

- Good results:
  - g3.11 64 71 1013 Baghdad NNP CIT http://en.wikipedia.org/wiki/Baghdad,1
  - g12.12 173 181 2011 al-Qaida CD ORG http://en.wikipedia.org/wiki/Al-Qaeda,1
  - g10.2 44 62 1010 Bosnia-Herzegovina NN COU http://en.wikipedia.org/wiki/Bosnia_and_Herzegovina,1
  - g13.4 188 194 2001 Mullah NNP PER http://en.wikipedia.org/wiki/Mohammed_Omar,1
  - g13.4 195 199 2002 Omar NNP PER http://en.wikipedia.org/wiki/Mohammed_Omar,1

Results (2)

- Bad results:
  - g4.5 18 26 1004 National NNP ORG http://en.wikipedia.org/wiki/National,1
  - g4.5 27 37 1005 Geographic NNP ORG http://en.wikipedia.org/wiki/Geographic,1
  - g4.5 38 45 1006 Society NN ORG http://en.wikipedia.org/wiki/Society,1
  - g12.7 35 43 1007 southern CD
  - g12.7 44 55 1008 Afghanistan VBP COU http://en.wikipedia.org/wiki/Afghanistan,1
  - g12.7 56 64 1009 district CD COU http://en.wikipedia.org/wiki/district,1

Evaluation

- Measures.py
- ……..

Questions?