Final project
Project Text Analysis

Mart Busger op Vollenbroek
Olivier Louwaars

Table of contents

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
- Data irregularities

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 (1)

- Measures.py

- Precision, Recall, F-score

Evaluation (2)

Confusion Matrix for NERS, Row= Reference, Column= Tagged

	-	ANI	CIT	COU	ENT	NAT	ORG	PER	SPO	http://en.wikipedia.org/wiki/Khost_Province
-	<.>	.	.	4	.	.	.	.	.	.
ANI	.	<5>	.	.	.	.	.	.	.	.
CIT	.	.	<105>	20	.	1	4	9	.	.
COU	.	.	24	<298>	.	2	5	19	.	.
ENT	.	.	.	4	<4>	.	21	4	.	.
NAT	.	.	5	16	.	<1>	2	5	.	.
ORG	.	.	10	35	.	.	<192>	10	.	.
PER	.	.	.	19	.	.	5	<294>	.	.
SPO	.	.	.	.	.	.	3	.	<.>	.
http://en.wikipedia.org/wiki/Khost_Province	.	.	1	.	.	.	.	.	.	<.>

Evaluation (3)

- Line 12345

- g13.0 323 328 4014 Khost NNP http://en.wikipedia.org/wiki/Khost_Province

```
Confusion Matrix for NERS, Row= Reference, Column= Tagged
     |          A   C   C   E   N   O   P   S |
     |          N   I   O   N   A   R   E   P |
     |    -     I   T   U   T   T   G   R   O |
-----+-----------------------------------------+
   - |   <.>    .   .   4   .   .   .   .   . |
ANI  |     .  <5>  .   .   .   .   .   .   . |
CIT  |     .    .<105> 20   .   1   4   9   . |
COU  |     .    .  25<298>  .   2   5  19   . |
ENT  |     .    .   .   4  <4>  .  21   4   . |
NAT  |     .    .   5  16   .  <1>  2   5   . |
ORG  |     .    .  10  35   .   .<192> 10   . |
PER  |     .    .   .  19   .   .   5<294>  . |
SPO  |     .    .   .   .   .   .   3   .  <.>|
-----+-----------------------------------------+
```

Evaluation (4)

- NER
- NER True Positives: 899
- NER False Positives: 220
- NER False Negatives: 220

- Precision: 0.8026785714285715
- Recall: 0.8026785714285715
- F-score: 0.8026785714285715

Evaluation (5)

- Wikis:
- 0.4371482176360225 of wikis 100% correctly added
- Wiki True Positives: 466
- Wiki False Positives: 599
- False Negatives: 2

- Precision: 0,4375586854
- Recall: 0,9957264957
- F-score: 0,6079601739

Questions?