A review of 5 months at the UPC

Daniel Berndt
Outline

- Inicial Plans
- Work on FreeLing
- Work on the Nominalisations Project
Inicial Plans

- Work on FreeLing
- Work on Nominalisations Project
- Acquisition of background knowledge due to papers and conferences
Work on FreeLing

Establish an option for transforming this

Mr_Brown  mr_brown  NNP  1
likes  like  VBZ  1
the  the  DT  1
Bank_of_Scotland  bank_of_scotland  NNP  1
Mr Brown likes the Bank of Scotland into this.
Creating a new English Lexicon

- The old lexicon was based on the Wall Street Journal Corpus
- Each word of that corpus was considered
Problems

- common words were missing (words like god, poem or thing)
- Typing errors were included (like in commmon)
- Automatic inflecting also caused errors (like funs, shakeablely, forgiven)
Solution

- discard the old lexicon and create a new one

Sources:

- **UKWAC** - tagged online based corpus consisting of 2,128,165,157 tokens

- **SPECIALIST Lexicon** - hand crafted lexicon for medic purposes containing 625,041 word forms and pos information

dijous 12 de febrer de 2009
Solution

- To put a word and all its inflections into the new lexicon these conditions must have been met:
  - the word and all its inflections have to occur at least 100 times in the UKWAC
  - the word must be in the SPECIALIST Lexicon
  - it is not a capitalised word (proper nouns should be detected via NER not via the lexicon)
Some Numbers

- Size of the old lexicon: 51,783 different words
- Size of the new lexicon: 71,142 different words
- New words: 43,486
- Discarded words: 24,127
What’s next?

* Including **Word Sense Disambiguation** with the SSI-Dijkstra algorithm as implemented by **Montse Cuadros**
What is my part of this project about?

For expressions like French wine there are four possible constructions:

- French wine, France's wine,
- France wine, wine of France

What are contextual and/or structural information that influence the choice of construction
Selection of the Nations

- We looked at the **BNC** for the four constructions for all nations that were listed with their adjectives on a wikipedia page

- **Results:**
  - The NN and N‘s N constructions are negligible in number
  - We chose only nations with a frequency between 25,000 and 1000
Distribution of the head nouns

How are the different head nouns distributed over the two constructions?

Results:

- persons and institutions tend to prefer AN constructions
- abstract concepts prefer NpN constructions
Filtering by Entropy

• to achieve better results it was necessary to exclude fixed constructions like
  • Indian Subcontinent, German reunification, French Franc

• We therefore used an entropy based method to exclude head nouns that are too nation specific
Filtering by Entropy

\[
\text{entropy}(w) = \sum_{n \in \text{Nat}} \Pr(w, n) \cdot \log(\Pr(w, n))
\]

\(w\) = head noun

Nat = the set of all considered nations

\(\Pr(w, n)\) = relative frequency of word \(w\) occurring with nation \(n\)
Check for the productivity of the constructions

- Check for each nation how many different AN and NpN constructions exist

- **Results:**
  - Considering bigger nations there is not such a big difference
  - smaller countries have more different NpN constructions
  - countries with complicated names and easier adjectives go more often with AN constructions

*diou 12 de febrer de 2009*
Check the prepositions in the NpN constructions

- list for every nation how often which preposition is used

Results:

- prepositions yielding a local relation are quite frequent (in, to, from) and of
Distribution of the two Constructions over Concepts

* the Top Concept Ontology project puts an ontology over the WordNet1.6 data
Distribution of the two Constructions over Concepts

• We check the concept for each head noun and look at the distribution over the two constructions

• **Hint one:** one sense always belongs to several concepts. The base concept and all concepts above

• **dog** for example belongs to: Animal, Living, Natural, Origin, 1stOrderEntity and Top

• **Hint two:** one word might have several senses
Distribution of the two Constructions over Concepts

- **WordNet** sorts senses of polysemous words according to their probable frequency

- We included an algorithm such that:
  - \( \text{rank}(s_2) = \text{rank}(s_1) + 1 \rightarrow \text{weight}(s_2) = \text{weight}(s_1) \cdot 0.5 \)
  - \( \sum_{s \in S} \text{weight}(s) = 1 \)
Distribution of the two Constructions over Concepts

\[
weight(s_i) = \frac{F^n \cdot (1-F)}{F^i \cdot (1-F^n)}
\]

\(s_i\) = sense with rank \(i\)

\(F\) = the multiplicator (2 in our case)

\(n\) = the number of senses for the word
Distribution of the two Constructions over Concepts

- We used two filters:
  - Minimal Entropy (Either 0.0, 1.0, 2.0 or 3.0)
  - Minimal Frequency of the sum of AN and NpN constructions of the head noun (Either 0, 25 or 100)
- Moreover we distinguished between token and type counts
TYPES VS. TOKENS
MINIMAL ENTROPY 2.0, MINIMAL FREQUENCY 25

dijous 12 de febrer de 2009
EFFECTS OF MINIMAL AMOUNT FILTERING
MINIMAL ENTROPY 2.0, TYPES

dijous 12 de febrer de 2009
EFFECTS OF ENTROPY FILTERING
MINIMAL FREQUENCY 25, TYPES
Distribution of the two Constructions over Concepts

Results:

- **1stOrderEntities** are more likely within AN constructions except for Part and Place.

- **2ndOrderEntities** are more likely to be part of NpN constructions.

- The more the data is filtered by entropy the clearer are this findings.
One strange Thing...

- If there was filtering by minimal frequency and no filtering by entropy the ratio was *always* very close to 1.0

- This caused us to create this graph:
Nominalisations

For exploring the behaviour of nominalisations contrasted to all nouns we did most of the data analysis once more for a restricted set of 45 nominalisations:

agreement, arrangement, application, arrival, attraction, construction, consultation, consumption, criticism, crossing, debate, decision, demand, development, distribution, education, evolution, examination, exploration, fear, growth, hope, interest, introduction, investigation, investment, management, movement, negotiation, production, reaction, reading, recognition, reconstruction, reduction, request, response, shift, shooting, statement, support, treatment, trip, understanding, visit
Distribution of the head nouns

- As expected most nominalisations tend to prefer NpN constructions
- The same accounts for the head noun variation
COMPARISON OF PREPOSITION DISTRIBUTION
PREPOSITION DISTRIBUTION: NATIONS VS. ALL NOUNS
Determiner Analysis

* check how often a AN and NpN construction occur with or with determiner

* If they occur with determiner, what kind of determiner is that?
Finally: Automatons

• Apply the **CSSR** (Causal State Splitting Reconstruction) method as implemented by Muntsa Padró to create two automatons for both constructions

• **Very important:** consider the right features!

• Still in progress
Future work:

* Until the end of my Internship:
  * Complete Documentation
  * Extract some more Corpus Data for Louise and Berit

* Until the end of my Bachelors Degree (probably):
  * I will dedicate myself to do some more analysis concerning the strange plot
Thank-you-Slide