MorphoNet: Exploring the Use of Community Structure for Unsupervised Morpheme Analysis

Delphine Bernhard

Ubiquitous Knowledge Processing Lab, Darmstadt, Germany
LIMSI-CNRS, Orsay, France

Morpho Challenge 2009 – September 30, 2009

Main features of the method

- Algorithm relying on a *network representation* of morphological relations between words
- Goal: investigate the use of *community structure* for morphology induction
- Networks with community structure contain groups of nodes with dense interconnections
- In our case, communities correspond to families of morphologically related words
- Related to work on networks in other areas of NLP, e.g. word clustering [Matsuo et al., 2006], word sense disambiguation [Mihalcea, 2005] or keyword extraction [Mihalcea and Tarau, 2004]

Overview of the method

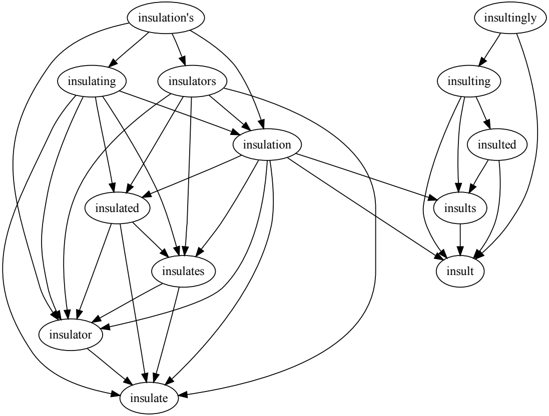
1. Acquisition of morphological transformation rules
2. Construction of a lexical network
3. Identification of word families using community structure
4. Acquisition of morpheme analyses

Step 1:
Acquisition of morphological transformation rules

- Morphological transformation rules make it possible to transform one word into another by performing substring substitutions
  `^(.+)ly$` → `\1`
  totally→ *total*
- These rules are acquired using a subset *L* of the wordlist *W* provided for each language (we used 10,000 words)
- Graphically similar words in *L* are first identified using a *gestalt* approach to fuzzy pattern matching based on the Ratcliff-Obershelp algorithm
- Rules are then obtained by comparing these graphically similar words
  democratic – undemocratic : `^un(.+)$` → `\1`

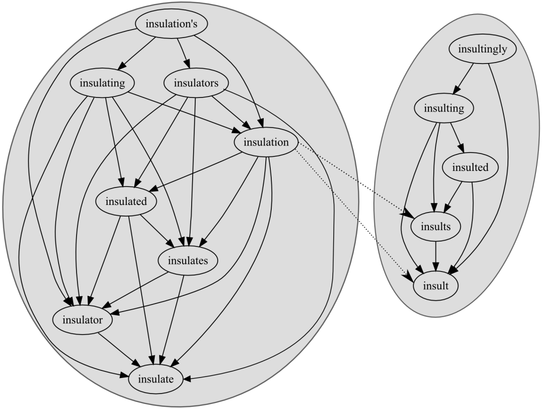
Step 2: Construction of a lexical network

- Rules are used to build a lexical network represented as a graph
- Nodes in the graph represent words from the input word list W
- Two words w_1 and w_2 are connected by an edge if there exists a transformation rule R such that $\mathrm{R}(w_1) = w_2$.



Step 3: Identification of word families

- Communities are detected in the lexical network, using a clustering algorithm
- Communities correspond to groups of tightly-knit nodes characterised by a high intra-group edge density and a lower inter-group density
- Use of the clustering algorithm proposed by Newman [Newman, 2004] to identify communities which correspond to word families



Step 4: Acquisition of morpheme analyses

- Identification of a representative word for each word family (shortest word)
- The full morpheme analysis for a word form *w* consists of its family representative and a string representation of the transformation rules that apply to *w*

Example

- Word family {*insulted*;*insulting*;*insult*;*insults*;*insultingly*}
- Family representative: *insult*
- Complete analyses:

`insultingly`	`insult _ly _ingly`
`insulting`	`insult _ing`
`insulted`	`insult _ed`
`insults`	`insult _s`
`insult`	`insult`

Conclusions and future work

- Promising results obtained at Morpho Challenge 2009
- Future improvements:
    - Increase recall by providing a better method for the acquisition of transformation rules
    - Weight edges in the network
    - Devise a more elaborate method for obtaining complete morpheme analyses
    - Address compounding

Thank you!

Questions?

Delphine.Bernhard@googlemail.com

References

Matsuo, Y., Sakaki, T., Uchiyama, K., and Ishizuka, M. (2006).
Graph-based Word Clustering using a Web Search Engine.
In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 542–550.

Mihalcea, R. (2005).
Unsupervised Large-Vocabulary Word Sense Disambiguation with Graph-based Algorithms for Sequence Data Labeling.
In Proceedings of the HLT/EMNLP 2005 Conference, pages 411–418.

Mihalcea, R. and Tarau, P. (2004).
TextRank: Bringing Order into Texts.
In Lin, D. and Wu, D., editors, Proceedings of EMNLP 2004, pages 404–411.

Newman, M. E. J. (2004).
Fast algorithm for detecting community structure in networks.
Physical Review E, 69.