Using textual association measures and minimum edit distance to discover morphological relations

Marco Baroni
Austrian Research Institute for Artificial Inteligence
marco@oefai.at

Johannes Matiasek
Austrian Research Institute for Artificial Inteligence
john@oefai.at

Harald Trost
University of Vienna
harald@ai.univie.ac.at

Collocations Workshop
July 23, 2002

Introduction

- Computational models learning morphology from data are of great practical and theoretical interest.
- Practical interest: fast, economic development of morphological analyzers, stemmers, linguistic analysis tools...
- Theoretical interest: model human language acquisition.
- Recent "boom" of work on unsupervised, knowledge poor morphological discovery: ACL2002 workshop, Goldsmith 2001, Snover and Brent 2001, Baroni 2000, Schone and Jurafsky 2000, Yarowsky and Wicentowski 2000, Gaussier 1999, Déjean 1998, Jacquemin 1997...

Introduction (*continued*)

- Our model: unsupervised, knowledge-free extraction of *morphologically related pairs* from a raw corpus.
- The basic idea:
  - Look for pairs that are similar in both orthographic and semantic terms: probably, they are also morphologically related (Cf. *blue/glue*, *blue/green* vs. *green/greenish*).
  - Orthographic similarity: measured by *minimum edit distance*.
  - Semantic similarity: measured by several AMs.

Introduction (*continued*)

- New features of our approach:
  - Algorithm does not rely on the notions of morpheme and morpheme concatenation, nor on global distributional properties of substrings (e.g., affix frequency).
  - Thus, at least in principle, well-suited for discovery of non-concatenative and/or "sparse" morphological processes.
  - Semantic similarity measured using AMs.
- Here, we focus on latter aspect.

Orthographic similarity with minimum edit distance

- *Minimum edit distance*: Minimum number of editing operations (insert, delete, substitute) needed to transform one string into the other (Jurafsky and Martin 2000: section 5.6).
- Normalized by dividing by length of longer string.
- *Orthographic similarity score*: 1 - normalized minimum edit distance.
- We assign a cost of 1 to all editing operations, independently of source and target segments.
- Example: orthographic similarity score of *black*/*lack* = $1 - (1/5) = .8$.

Semantic similarity with AMs

- Basic intuition: related words will tend to co-occur more often than what we would predict based on their unigram frequencies.
- Several related AMs are based on comparison of empirical frequency of co-occurrence of two words and expected frequency of co-occurrence under assumption of independence: The larger the discrepancy, the more likely it is that the two words are *not* independent.
- See Evert's collocation webpage for review: http://www.collocations.de/EK/am-html/index.html

Semantic similarity with AMs (*continued*)

- What counts as a "co-occurrence" of two words?
  - Co-occurrence of adjacent directed n-grams: finds collocations;
  - Co-occurrence of long distance, non-adjacent, non-directed bigrams: finds *semantically related words* (Brown et al. 1990).
- We are interested in the second notion of co-occurrence.

Semantic similarity with AMs (*continued*)

- We collect bigrams that have the following properties:
  - *Long distance* (maximum distance: 498 words);
  - *Non-adjacency* (minimum distance: 3 words);
  - *Non-directionality* ($a...b == b...a$).

Semantic similarity with AMs (*continued*)

- We experiment with three AMs:
  - *Co-occurrence frequency*;
  - *Mutual information* (Church and Hanks 1989);
  - *Log-likelihood ratio* (Dunning 1993).

The pairing and ranking algorithm

- Construct *potential content word* list (potential content word: alphabetic string with fq $<$ .01% of total corpus token count).
- Compute relevant AM for all pairs of words from potential content word list that co-occur (in the sense defined above) three times or more.
- Compute orthographic similarity for same set of pairs.
- Discard pairs with orthographic similarity below .5.
- Rank remaining pairs based on weighted sum of their orthographic similarity and AM scores (simple interpolation model based on alignment of order of magnitude of maximum scores).

Empirical evaluation

- Input corpora:
  - German APA Newswire corpus: $28M$ words;
  - English Brown corpus: $1.2M$ words.
- German words longer than 9 characters not evaluated (to make results easier to interpret).

A quick look at the results

- Most frequent patterns found by algorithm do correspond to common German/English affixation patterns.
- Algorithm found many instances of discontinuous German processes (e.g., pluralization via vowel fronting and suffixation of *-e*: *Tiefflug/Tiefflüge*).
- Algorithm found many instances of German past participle circumfixation: *bilden/gebildet*, *fordern/gefordert*, *sendent/gesendet*...
- See our ACL02 morpho-learning workshop paper (http://www.ai.univie.ac.at/˜marco/).

Comparing AM performance

- Reference sets: Lists of all pairs of potential content words from input corpus that are morphologically related according to XEROX analyzer for relevant language
- XEROX analyzers provide limited derivational analysis for German, no derivational analysis for English (about 800 mostly derivationally related pairs added by hand to English reference set).
- Precision = (number of pairs in reference set) / (total number of pairs) at different cutoff points in the output ranked lists.
- Results using three different AMs and keeping everything else unchanged.

German precision

# of pairs	fq precision	mi precision	ll precision
500	95%	96%	96%
1000	94%	96%	96%
1500	94%	97%	95%
2000	93%	97%	93%
3000	92%	96%	92%
4000	91%	95%	92%
5000	90%	93%	90%
10000	71%	75%	73%
67032	16%	16%	16%

English precision

# of pairs	fq precision	mi precision	ll precision
500	98%	99%	97%
1000	95%	96%	96%
1500	91%	93%	92%
2000	83%	87%	87%
3000	72%	73%	73%
4000	60%	61%	62%
5000	50%	51%	52%
9043	30%	30%	30%

Overlap among measures

Number of shared items among top 100 pairs:

- German:
  - Co-occurrence frequency and log-likelihood ratio: 89.
  - Co-occurrence frequency and mutual information: 0.
  - Log-likelihood ratio and mutual information: 0.
- English:
  - Co-occurrence frequency and log-likelihood ratio: 79.
  - Co-occurrence frequency and mutual information: 7.
  - Log-likelihood ratio and mutual information: 7.

Overlap among measures (*continued*)

Average unigram frequency for top 100 pairs (counting most fq word of each pair):

- German:
  - Co-occurrence frequency: 1953 (s.d. 964).
  - Log-likelihood ratio: 1901 (s.d 986).
  - Mutual information: 34 (s.d. 26).
- English:
  - Co-occurrence frequency: 76 (s.d. 27).
  - Log-likelihood ratio: 69 (s.d. 28).
  - Mutual information: 22 (s.d. 18).

Conclusion

- Overall, encouraging results.
- Output of algorithm can be used to construct stemmers or morphological analyzers, or as a data exploration tool.
- Can we do better than frequency??? (Krenn and Evert 2001)
- Perhaps, with mutual information.
- Co-occurrence frequency and mutual information look at different sections of the frequency spectrum: could be used in combination.