FUH (FernUniversität in Hagen):
Metonymy Recognition Using
Different Kinds of Context for a Memory-Based Learner

Johannes Leveling
Intelligent Information and Communication Systems (IICS)
FernUniversität in Hagen (University of Hagen)
johannes.leveling@fernuni-hagen.de

Abstract

For the metonymy resolution task at SemEval-2007, the use of a memory-based learner to train classifiers for the identification of metonymic location names is investigated. Metonymy is resolved on different levels of granularity, differentiating between literal and non-literal readings on the coarse level; literal, metonymic, and mixed readings on the medium level; and a number of classes covering regular cases of metonymy on a fine level. Different kinds of context are employed to obtain different features: 1) a sequence of \( n_1 \) synset IDs representing subordination information for nouns and for verbs, 2) \( n_2 \) prepositions, articles, modal, and main verbs in the same sentence, and 3) properties of \( n_3 \) tokens in a context window to the left and to the right of the location name.

Different classifiers were trained on the Mascara data set to determine which values for the context sizes \( n_1 \), \( n_2 \), and \( n_3 \) yield the highest accuracy (\( n_1 = 4 \), \( n_2 = 3 \), and \( n_3 = 7 \), determined with the leave-one-out method). Results from these classifiers served as features for a combined classifier. In the training phase, the combined classifier achieved a considerably higher precision for the Mascara data. In the SemEval submission, an accuracy of 79.8% on the coarse, 79.5% on the medium, and 78.5% on the fine level is achieved (the baseline accuracy is 79.4%).

1 Introduction

Metonymy is typically defined as a figure of speech in which a speaker uses one entity to refer to another that is related to it (Lakoff and Johnson, 1980). The identification of metonymy becomes important for NLP tasks such as question answering (Stallard, 1993) or geographic information retrieval (Leveling and Hartumpf, 2006).

For regular cases of metonymy for locations and organizations, Markert and Nissim have proposed a set of metonymy classes. Annotating a subset of the BNC (British National Corpus), they extracted a set of metonymic proper nouns from two categories: country names (Markert and Nissim, 2002) and organization names (Nissim and Markert, 2003).

In the metonymy resolution task at SemEval-2007, the goal was to identify metonymic names in a subset of the BNC. The task consists of two subtasks for company and country names, which are further divided into classification on a coarse level (recognizing literal and non-literal readings), on a medium level (differentiating non-literal readings into mixed and metonymic readings), and on a fine level (identifying classes of regular metonymy, such as a name referring to the population, place-for-people). The task is described in more detail by Markert and Nissim (2007).

2 System Description

2.1 Tools and Resources

The following tools and resources are used for the metonymy classification:

- TiMBL 5.1 (Daelemans et al., 2004), a memory-based learner for classification is em-
ployed for training the classifiers (supervised
learning).

- Mascara 2.0 – Metonymy Annotation Scheme
  And Robust Analysis (Markert and Nissim, 2003; Nissim and Markert, 2003; Markert
  and Nissim, 2002) contains annotated data for
  metonymic names from a subset of the the
  BNC.

- WordNet 2.0 (Fellbaum, 1998) serves as a lin-
  guistic resource for assigning synset IDs and
  for looking up subordination information and
  frequency of readings.

- The TreeTagger (Schmid, 1994) is utilized for
  sentence boundary detection, lemmatization,
  and part-of-speech tagging. The English tag-
 ger was trained on the PENN treebank and uses
  the English morphological database from the
  XTAG project (Karp et al., 1992). The param-
  eter files were obtained from the web site.

2.2 Different Kinds of Context

Following the assumption that metonymic location
names can be identified from the context, there are
different kinds of context to consider. At most, the
context comprises a single sentence in this setup.
Three kinds of context were employed to extract fea-
tures for the memory-based learner TiMBL:

- $C_1$: Subordination (hyponymy) information for
  nouns and verbs from the left and right context
  of the possibly metonymic name.

- $C_2$: The sentence context for modal verbs, main
  verbs, prepositions, and articles.

- $C_3$: A context window of tokens left and right
  of the location name.

The trial data provided (a subset of the Mascara
data) contained 188 non-literal location names (of
925 samples total). For a supervised learning ap-
proach, this is too few data. Therefore, the full
Mascara data was converted to form training data
consisting of feature values for context $C_1$, $C_2$, and
$C_3$. The training data contained 509 metonymic an-
notations (of 2797 samples total). Some cases in
the Mascara corpus are filtered during processing,
including cases annotated as homonyms and cases
whose metonymy class could not be agreed upon.
The test data had a majority baseline of 82.8% accu-
rance for country names.

2.3 Features

The Mascara data was processed to extract the fol-
lowing features (no hand-annotated data from Mas-
cara was employed for feature values, i.e. no gram-
matical roles):

- For $C_1$ (WordNet context): From a context of
  $n_1$ verbs and nouns in the same sentence, their
distance to the location name is calculated. A
  sequence of eight feature values of WordNet
  synset IDs is obtained by iteratively looking up
  the most frequent reading for a lemma in Word-
  Net and determining its synset ID. Subordina-
  tion information between synsets is used to find
  a parent synset. This process is repeated until
  a top-level parent synset is reached. No actual
  word sense disambiguation is employed.

- For $C_2$ (sentence context): Sentence bound-
  aries, part-of-speech tags, and lemmatization
  are determined from the TreeTagger output.
  From a context window of $n_2$ tokens, lemma
  and distance are encoded as feature values for
  prepositions, articles, modal, and main verbs

- For $C_3$ (word context): From a context of $n_3$
tokens to the left and to the right, the distance
  between token and location name, three pre-
  fix characters, three suffix characters, part-of-
  speech tag, case information (U=upper case,
  L=lower case, N=numeric, O=other), and word
  length are used as feature values.

Table 1 and Table 2 show results for mem-
ory based learners trained with TiMBL. Perform-
ance measures were obtained with the leave-one-
out method. The classifiers were trained on fea-
tures for different context sizes ($n_i$ ranging from 2
to 7) to determine the setting for which the highest
accuracy is achieved (e.g. 1c, 2c, and 3c). In the
next step, classifiers with a combined context were
trained, selecting the setting with the highest accuracy for a single context for the combination (e.g. 4c, 5c, 6c, and 7c). As an additional experiment, a classifier was trained on classification results of the classifiers described above (combination of 1–7, e.g. 8c). It was expected that the combination of features from different kinds of context would increase performance, and that the combination of classifier results would increase performance.

### 3 Evaluation Results

Table 3 shows results for the official submission. Compared to results from the training phase on the Mascara data (tested with the leave-one-out method), performance is considerably lower. For this data, the combined classifier achieved a considerably higher precision (63.9% for non-literal readings; 57.3% for the fine class place-for-people and even 83.3% for the rare class place-for-event).

Performance may be affected by several reasons: A number of problems were encountered while processing the data. The TreeTagger automatically tokenizes its input and applies sentence boundary detection. In some cases, the sentence boundary detection did not work well, returning sentences of more than 170 words. Furthermore, the tagger output had to be aligned with the test data again, as multi-word names (e.g. New York) were split into different tokens. In addition, the tag set of the tagger differs somewhat from the official PENN tag set and includes additional tags for verbs.

In earlier experiments on metonymy classification on a German corpus (Leveling and Hartrumpf, 2006), the data was nearly evenly distributed between literal and metonymic readings. This seems to make a classification task easier because there is no hidden bias in the classifier (i.e. the baseline of always selecting the literal readings is about 50%).

Features are obtained by shallow NLP methods only, not making use of a parser or chunker. Thus, important syntactic or semantic information to decide on metonymy might be missing in the features. However, semantic features are more difficult to determine, because reliable automatic tools for semantic annotation are still missing. This is also indicated by the fact that the grammatical roles (comprising syntactic features) in Mascara data are hand-annotated.

However, some linguistic phenomena are already implicitly represented by shallower features from
Table 3: Results for the coarse (908 samples: 721 literal, 187 non-literal), medium (721 literal, 167 metonymic, 20 mixed), and fine classification (721 literal, 141 place-for-people, 10 place-for-event, 1 place-for-product, 4 object-for-name, 11 othermet, 20 mixed) of location names.

| class                  | P   | R   | F   |
|------------------------|-----|-----|-----|
| FUH.location.coarse (0.798 accuracy) |     |     |     |
| literal                | 0.812 | 0.971 | 0.884 |
| non-literal            | 0.543 | 0.134 | 0.214 |
| FUH.location.medium (0.795 accuracy) |     |     |     |
| literal                | 0.810 | 0.970 | 0.883 |
| metonymic              | 0.500 | 0.132 | 0.208 |
| mixed                  | 0.0   | 0.0  | 0.0  |
| FUH.location.fine (0.785 accuracy) |     |     |     |
| literal                | 0.808 | 0.965 | 0.880 |
| place-for-people       | 0.386 | 0.120 | 0.183 |

the surface level (given enough training instances). For instance, active/passive voice may be encoded by a combination of features for main verb/modal verbs. If only a small training corpus is available, overall performance will be higher when utilizing explicit syntactic or semantic features.

Finally, the data may be too sparse for a supervised memory-based learning approach. The identification of rare classes of metonymy (e.g. place-for-event) would greatly benefit from a larger corpus covering these classes.

4 Conclusion

Evaluation results on the training data were very promising, indicating a boost of precision by combining classification results. In the training phase, an accuracy of 83.7% was achieved on the coarse level, compared to the majority baseline accuracy of 81.8%. For the submission for the metonymy resolution task at SemEval-2007, accuracy is close to the majority baseline (79.4%) on the coarse (79.8%), medium (79.5%), and fine (78.5%) level.

In summary, using different context sizes for different kinds of context and combining results of different classifiers for metonymy resolution increases performance. The general approach would profit from combining results of more diverse classifiers, i.e. classifiers employing features extracted from the surface, syntactic, and semantic context of a location name.

Acknowledgments

The research described was in part funded by the DFG (Deutsche Forschungsgemeinschaft) in the project IRSAW (Intelligent Information Retrieval on the Basis of a Semantically Annotated Web).

References

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2004. TiMBL: Tilburg memory based learner, version 5.1. TR 04-02, ILK.

Christiane Fellbaum, editor. 1998. Wordnet. An Electronic Lexical Database. MIT Press, Cambridge, Massachusetts.

Daniel Karp, Yves Schabes, Martin Zaidel, and Dania Egedi. 1992. A freely available wide coverage morphological analyzer for English. In Proc. of COLING-92, pages 950–955, Morristown, NJ.

George Lakoff and Mark Johnson. 1980. Metaphors We Live By. Chicago University Press.

Johannes Leveling and Sven Hartrumpf. 2006. On metonymy recognition for GIR. In Proc. of GIR-2006, the 3rd Workshop on Geographical Information Retrieval (held at SIGIR 2006), Seattle, Washington.

Katja Markert and Malvina Nissim. 2002. Towards a corpus for annotated metonymies: The case of location names. In Proc. of LREC 2002, Las Palmas, Spain.

Katja Markert and Malvina Nissim. 2003. Corpus-based metonymy analysis. Metaphor and symbol, 18(3).

Katja Markert and Malvina Nissim. 2007. Task 08: Metonymy resolution at SemEval-07. In Proc. of SemEval 2007.

Malvina Nissim and Katja Markert. 2003. Syntactic features and word similarity for supervised metonymy resolution. In Proc. of ACL-2003, Sapporo, Japan.

Yves Peirsman. 2006. Example-based metonymy recognition for proper nouns. In Proc. of the Student Research Workshop of EACL-2006, pages 71–78, Trento, Italy.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In International Conference on New Methods in Language Processing, Manchester, UK.

David Stallard. 1993. Two kinds of metonymy. In Proc. of ACL-93, pages 87–94, Columbus, Ohio.