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 Hartrumpf, 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-

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 153–156,
Prague, June 2007. ©2007 Association for Computational Linguistics

ployed for training the classifiers (supervised learning).[1]

- 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 linguistic 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 tagger was trained on the PENN treebank and uses the English morphological database from the XTAG project (Karp et al., 1992). The parameter files were obtained from the web site.[2]

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 features 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 approach, 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 annotations (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% accuracy for country names.

2.3 Features

The Mascara data was processed to extract the following features (no hand-annotated data from Mascara was employed for feature values, i.e. no grammatical 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 WordNet and determining its synset ID. Subordination 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 boundaries, 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 prefix 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 memory based learners trained with TiMBL. Performance measures were obtained with the leave-one-out method. The classifiers were trained on features for different context sizes (n_i ranging from 2 to 7) to determine the setting for which the highest accuracy is achieved (e.g. 1_c, 2_c, and 3_c). In the next step, classifiers with a combined context were

[1] Peirsman (2006) also employs TiMBL for metonymy resolution, but trains a single classifier.

[2] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

Table 1: Results for training the classifiers on the coarse location name classes (2797 instances, 509 *non-literal*, leave-one-out) for the Mascara data (P = precision, R = recall, F = F-score).

ID	n_1,n_2,n_3	coarse class	P	R	F
1_c	4,0,0	literal	0.850	0.893	0.871
1_c	4,0,0	non-literal	0.377	0.289	0.327
2_c	0,3,0	literal	0.848	0.874	0.860
2_c	0,3,0	non-literal	0.342	0.295	0.317
3_c	0,0,7	literal	0.880	0.889	0.885
3_c	0,0,7	non-literal	0.478	0.455	0.467
4_c	4,3,0	literal	0.848	0.892	0.896
4_c	4,3,0	non-literal	0.368	0.282	0.320
5_c	4,0,7	literal	0.860	0.913	0.885
5_c	4,0,7	non-literal	0.459	0.332	0.385
6_c	0,3,7	literal	0.875	0.905	0.889
6_c	0,3,7	non-literal	0.496	0.420	0.455
7_c	4,3,7	literal	0.860	0.918	0.888
7_c	4,3,7	non-literal	0.473	0.332	0.390
8_c	res. of 1_c–7_c	literal	0.852	0.968	0.907
8_c	res. of 1_c–7_c	non-literal	0.639	0.248	0.357

trained, selecting the setting with the highest accuracy for a single context for the combination (e.g. 4_c, 5_c, 6_c, and 7_c). As an additional experiment, a classifier was trained on classification results of the classifiers described above (combination of 1–7, e.g. 8_c). 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.

Table 2: Excerpt from results for training the classifiers on the fine location name classes (2797 instances, leave-one-out) for the Mascara data.

ID	n_1,n_2,n_3	fine class	P	R	F
1_f	4,0,0	literal	0.851	0.895	0.873
1_f	4,0,0	pl.-for-p.	0.366	0.280	0.318
1_f	4,0,0	pl.-for-e.	0.370	0.270	0.312
2_f	0,3,0	literal	0.848	0.876	0.862
2_f	0,3,0	pl.-for-p.	0.332	0.276	0.301
2_f	0,3,0	pl.-for-e.	0.222	0.270	0.244
3_f	0,0,7	literal	0.878	0.892	0.885
3_f	0,0,7	pl.-for-p.	0.463	0.424	0.442
3_f	0,0,7	pl.-for-e.	0.279	0.324	0.300
4_f	4,3,0	literal	0.851	0.899	0.875
4_f	4,3,0	pl.-for-p.	0.358	0.269	0.307
4_f	4,3,0	pl.-for-e.	0.435	0.270	0.333
5_f	4,0,7	literal	0.861	0.914	0.887
5_f	4,0,7	pl.-for-p.	0.452	0.322	0.377
5_f	4,0,7	pl.-for-e.	0.550	0.297	0.386
6_f	0,3,7	literal	0.871	0.906	0.888
6_f	0,3,7	pl.-for-p.	0.468	0.383	0.422
6_f	0,3,7	pl.-for-e.	0.400	0.324	0.358
7_f	4,3,7	literal	0.861	0.918	0.889
7_f	4,3,7	pl.-for-p.	0.459	0.323	0.378
7_f	4,3,7	pl.-for-e.	0.500	0.297	0.373
8_f	res. of 1_f–7_f	literal	0.854	0.963	0.905
8_f	res. of 1_f–7_f	pl.-for-p.	0.573	0.262	0.360
8_f	res. of 1_f–7_f	pl.-for-e.	0.833	0.270	0.408

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.