GYDER: maxent metonymy resolution

Richárd Farkas
University of Szeged
Department of Informatics
H-6720 Szeged, Árpád tér 2.
rfarkas@inf.u-szeged.hu

Eszter Simon
Budapest U. of Technology
Dept. of Cognitive Science
H-1111 Budapest, Stoczek u 2.
esimon@cogsci.bme.hu

György Szarvas
University of Szeged
Department of Informatics
H-6720 Szeged, Árpád tér 2.
szarvas@inf.u-szeged.hu

Dániel Varga
Budapest U. of Technology
MOKK Media Research
H-1111 Budapest, Stoczek u 2.
daniel@mokk.bme.hu

Abstract

Though the GYDER system has achieved the highest accuracy scores for the metonymy resolution shared task at SemEval-2007 in all six subtasks, we don't consider the results (72.80% accuracy for `org`, 84.36% for `loc`) particularly impressive, and argue that metonymy resolution needs more features.

1 Introduction

In linguistics *metonymy* means using one term, or one specific sense of a term, to refer to another, related term or sense. For example, in 'the pen is mightier than the sword' *pen* refers to writing, the force of ideas, while *sword* refers to military force. Named Entity Recognition (NER) is of key importance in numerous natural language processing applications ranging from information extraction to machine translation. Metonymic usage of named entities is frequent in natural language. On the basic NER categories `person`, `place`, `organisation` state-of-the-art systems generally perform in the mid to the high nineties. These systems typically do not distinguish between literal or metonymic usage of entity names, even though this would be helpful for most applications. Resolving metonymic usage of proper names would therefore directly benefit NER and indirectly all NLP tasks (such as anaphor resolution) that require NER.

Markert and Nissim (2002) outlined a corpus-based approach to proper name metonymy as a semantic classification problem that forms the basis of the 2007 SemEval metonymy resolution task. Instances like 'He was shocked by Vietnam' or 'Schengen boosted tourism' were assigned to broad categories like `place-for-event`, sometimes ignoring narrower distinctions, such as the fact that it wasn't the signing of the treaty at Schengen but rather its actual implementation (which didn't take place at Schengen) that boosted tourism. But the corpus makes clear that even with these (sometimes coarse) class distinctions, several metonymy types seem to appear extremely rarely in actual texts. The shared task focused on two broad named entity classes as metonymic sources, `location` and `org`, each having several target classes. For more details on the data sets, see the task description paper Markert and Nissim (2007).

Several categories (e.g. `place-for-event`, `organisation-for-index`) did not contain a sufficient number of examples for machine learning, and we decided early on to accept the fact that these categories will not be learned and to concentrate on those classes where learning seemed feasible. The shared task itself consisted of 3 subtasks of different granularity for both organisation and location names. The fine-grained evaluation aimed at distinguishing between all categories, while the medium-grained evaluation grouped different types of metonymic usage together and addressed literal / mixed / metonymic usage. The coarse-grained subtask was in fact a literal / nonliteral two-class classification task.

Though GYDER has obtained the highest accuracy for the metonymy shared task at SemEval-2007 in all six subtasks, we don't consider the results

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 161–164,
Prague, June 2007.

(72.80% accuracy for `org`, 84.36% for `loc`) particularly impressive. In Section 3 we describe the feature engineering lessons learned from working on the task. In Section 5 we offer some speculative remarks on what it would take to improve the results.

2 Learning

GYDER (the acronym was formed from the initials of the author' first names) is a maximum entropy learner. It uses Zhang Le's [1] maximum entropy toolkit, setting the Gaussian prior to 1. We used random 5-fold cross-validation to determine the usefulness of a particular feature. Due to the small number of instances and features, the learning algorithm always converged before 30 iterations, so the cross-validation process took only seconds.

We also tested the classic C4.5 decision tree learning algorithm Quinlan (1993), but our early experiments showed that the maximum entropy learner was consistently superior to the decision tree classifier for this task, yielding about 2-5% higher accuracy scores on average on both tasks (on the training set, using cross-validation).

3 Feature Engineering

We tested several features describing orthographic, syntactic, or semantic characteristics of the Possibly Metonymic Words (PMWs). Here we follow Nissim and Markert (2005), who reported three classes of features to be the most relevant for metonymy resolution: the grammatical annotations provided for the corpus examples by the task organizers, the determiner, and the grammatical number of the PMW. We also report on some features that didn't work.

3.1 Grammatical annotations

We used the grammatical annotations provided for each PMW in several ways. First, we used as a feature the type of the grammatical relation and the word form of the related word. (If there was more than one related word, each became a feature.) To overcome data sparseness, it is useful to generalize from individual headwords Markert and Nissim (2003). We used three different methods to achieve this:

First, we used Levin's (1993) verb classification index to generalize the headwords of the most relevant grammatical relations (subject and object). The added feature was simply the class assigned to the verb by Levin.

We also used WordNet (Fellbaum 1998) to generalize headwords. First we gathered the hypernym path from WordNet for each headword's sense#1 in the train corpus. Based on these paths we collected synsets whose tree frequently indicated metonymic sense. We indicated with a feature if the headword in question was in one of such collected subtrees.

Third, we have manually built a very small verb classification 'Trigger' table for specific cases. E.g. *announce, say, declare* all trigger the same feature. This table is the only resource in our final system that was manually built by us, so we note that on the test corpus, disabling this 'Trigger' feature does not alter `org` accuracy, and decreases `loc` accuracy by 0.44%.

3.2 Determiners

Following Nissim and Markert (2005), we distinguished between definite, indefinite, demonstrative, possessive, wh and other determiners. We also marked if the PMW was sentence-initial, and thus necessarily determinerless. This feature was useful for the resolution of organisation PMWs so we used it only for the `org` tasks. It was not straightforward, however, to assign determiners to the PMWs without proper syntactic analysis. After some experiments, we linked the nearest determiner and the PMW together if we found only adjectives (or nothing) between them.

3.3 Number

This feature was particularly useful to separate metonymies of the `org-for-product` class. We assumed that only PMWs ending with letter *s* might be in plural form, and for them we compared the web search result numbers obtained by the Google API. We ran two queries for each PMWs, one for the full name, and one for the name without its last character. If we observed a significant increase in the number of hits returned by Google for the shorter phrase, we set this feature for plural.

[1] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

3.4 PMW word form

We included the surface form of the PMW as a feature, but only for the `org` domain. Cross-validation on the training corpus showed that the use of this feature causes an 1.5% accuracy improvement for organisations, and a slight degradation for locations. The improvement perfectly generalized to the test corpora. Some company names are indeed more likely to be used in a metonymic way, so we believe that this feature does more than just exploiting some specificity of the shared task corpora. We note that the ranking of our system would have been unaffected even if we didn't use this feature.

3.5 Unsuccessful features

Here we discuss those features where cross-validation didn't show improvements (and thus were not included in the submitted system).

Trigger words were automatically collected lists of word forms and phrases that more frequently appeared near metonymic PMWs.

Expert triggers were similar trigger words or phrases, but suggested by a linguist expert to be potentially indicative for metonymic usage. We experimented with sample-level, sentence-level and vicinity trigger phrases.

Named entity labels given by a state-of-the-art named entity recognizer (Szarvas et al. 2006).

POS tags around PMWs.

Ortographical features such as capitalisation and and other surface characteristics for the PMW and nearby words.

Individual tokens of the potentially metonymic phrase.

Main category of Levin's hierarchical classification.

Inflectional category of the verb nearest to the PMW in the sentence.

4 Results

Table 1. shows the accuracy scores of our submitted system on fine classification granularity. As a baseline, we also evalute the system without the WordNet, Levin, Trigger and PMW word form features. This baseline system is quite similar to the one described by Nissim and Markert (2005). We also publish the majority baseline scores.

run	majority	baseline	submitted
org train 5-fold	63.30	77.51	80.92
org test	61.76	70.55	**72.80**
loc train 5-fold	79.68	85.58	88.36
loc test	79.41	83.59	**84.36**

Table 1: Accuracy of the submitted system

We could not exploit the hierarchical structure of the fine-grained tag set, and ended up treating it as totally unstructured even for the mixed class, unlike Nissim and Markert, who apply complicated heuristics to exploit the special semantics of this class.

For the coarse and medium subtasks of the `loc` domain, we simply coarsened the fine-grained results. For the coarse and medium subtasks of the `org` domain, we coarsened the train corpus to medium coarseness before training. This idea was based on observations on training data, but was proven to be unjustified: it slightly decreased the system's accuracy on the medium subtask.

	coarse	medium	fine
location	85.24	84.80	84.36
organisation	76.72	73.28	72.80

Table 2: Accuracy of the GYDER system for each domain / granularity

In general, the coarser grained evaluation did not show a significantly higher accuracy (see Table 2.), proving that the main difficulty is to distinguish between literal and metonymic usage, rather than separating metonymy classes from each other (since different classes represent significantly different usage / context). Because of this, data sparseness remained a problem for coarse-grained classification as well.

Per-class results of the submitted system for both domains are shown on Table 3. Note that our system never predicted `loc` values from the four small classes `place-for-event and product`, `object-for-name` and `other` as these had only 26 instances altogether. Since we never had significant results for the `mixed` category, in effect the `loc` task ended up a binary classification task between `literal` and `place-for-people`.

loc class	#	prec	rec	f
literal	721	86.83	95.98	91.17
place-for-people	141	68.22	51.77	58.87
mixed	20	25.00	5.00	8.33
othermet	11	-	0.0	-
place-for-event	10	-	0.0	-
object-for-name	4	-	0.0	-
place-for-product	1	-	0.0	-
org class	#	prec	rec	f
literal	520	75.76	90.77	82.59
org-for-members	161	65.99	60.25	62.99
org-for-product	67	82.76	35.82	50.00
mixed	60	43.59	28.33	34.34
org-for-facility	16	100.0	12.50	22.22
othermet	8	-	0.0	-
object-for-name	6	50.00	16.67	25.00
org-for-index	3	-	0.0	-
org-for-event	1	-	0.0	-

Table 3: Per-class accuracies for both domains

While in the `org` set the system also ignores the smallest categories `othermet`, `org-for-index` and `event` (a total of 11 instances), the six major categories `literal`, `org-for-members`, `org-for-product`, `org-for-facility`, `object-for-name`, `mixed` all receive meaningful hypotheses.

5 Conclusions, Further Directions

The features we eventually selected performed well enough to actually achieve the best scores in all six subtasks of the shared task, and we think they are useful in general. But it is worth emphasizing that many of these features are based on the grammatical annotation provided by the task organizers, and as such, would require a better dependency parser than we currently have at our disposal to create a fully automatic system.

That said, there is clearly a great deal of merit to provide this level of annotation, and we would like to speculate what would happen if even more detailed annotation, not just grammatical, but also semantical, were provided manually. We hypothesize that the metonymy task would break down into the task of identifying several journalistic cliches such as "location for sports team", "capital city for government", and so on, which are not yet always distinguished by the depth of the annotation.

It would be a true challenge to create a data set of non-cliche metonymy cases, or a corpus large enough to represent rare metonymy types and challenging non-cliche metonymies better.

We feel that at least regarding the corpus used for the shared task, the potential of the grammatical annotation for PMWs was more or less well exploited. Future systems should exploit more semantic knowledge, or the power of a larger data set, or preferably both.

Acknowledgement

We wish to thank András Kornai for help and encouragement, and the anonymous reviewers for valuable comments.

References

Christiane Fellbaum ed. 1998. WordNet: An Electronic Lexical Database. MIT Press.

Beth Levin. 1993. English Verb Classes and Alternations. A Preliminary Investigation. The University of Chicago Press.

Katja Markert and Malvina Nissim. 2002. Metonymy resolution as a classification task. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Philadelphia, USA.

Katja Markert and Malvina Nissim. 2003. Syntactic Features and Word Similarity for Supervised Metonymy Resolution. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL2003)*. Sapporo, Japan.

Malvina Nissim and Katja Markert. 2005. Learning to buy a Renault and talk to BMW: A supervised approach to conventional metonymy. *International Workshop on Computational Semantics (IWCS2005)*. Tilburg, Netherlands.

Katja Markert and Malvina Nissim. 2007. SemEval-2007 Task 08: Metonymy Resolution at SemEval-2007. In Proceedings of SemEval-2007.

Ross Quinlan. 1993. C4.5: Programs for machine learning. Morgan Kaufmann.

György Szarvas, Richárd Farkas and András Kocsor. 2006. Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. *Proceedings of Discovery Science 2006, DS2006, LNAI 4265 pp. 267-278*. Springer-Verlag.