SemEval-2007 Task 08: Metonymy Resolution at SemEval-2007

Katja Markert
School of Computing
University of Leeds, UK
markert@comp.leeds.ac.uk

Malvina Nissim
Dept. of Linguistics and Oriental Studies
University of Bologna, Italy
malvina.nissim@unibo.it

Abstract

We provide an overview of the metonymy resolution shared task organised within SemEval-2007. We describe the problem, the data provided to participants, and the evaluation measures we used to assess performance. We also give an overview of the systems that have taken part in the task, and discuss possible directions for future work.

1 Introduction

Both word sense disambiguation and named entity recognition have benefited enormously from shared task evaluations, for example in the Senseval, MUC and CoNLL frameworks. Similar campaigns have not been developed for the resolution of figurative language, such as metaphor, metonymy, idioms and irony. However, resolution of figurative language is an important complement to and extension of word sense disambiguation as it often deals with word senses that are not listed in the lexicon. For example, the meaning of *stopover* in the sentence *He saw teaching as a stopover on his way to bigger things* is a metaphorical sense of the sense "stopping place in a physical journey", with the literal sense listed in WordNet 2.0 but the metaphorical one not being listed.[1] The same holds for the metonymic reading of *rattlesnake* (for the animal's meat) in *Roast rattlesnake tastes like chicken*.[2] Again, the meat reading of *rattlesnake* is not listed in WordNet whereas the meat reading for *chicken* is.

As there is no common framework or corpus for figurative language resolution, previous computational works (Fass, 1997; Hobbs et al., 1993; Barnden et al., 2003, among others) carry out only small-scale evaluations. In recent years, there has been growing interest in metaphor and metonymy resolution that is either corpus-based or evaluated on larger datasets (Martin, 1994; Nissim and Markert, 2003; Mason, 2004; Peirsman, 2006; Birke and Sarkaar, 2006; Krishnakamuran and Zhu, 2007). Still, apart from (Nissim and Markert, 2003; Peirsman, 2006) who evaluate their work on the same dataset, results are hardly comparable as they all operate within different frameworks.

This situation motivated us to organise the first shared task for figurative language, concentrating on metonymy. In metonymy one expression is used to refer to the referent of a related one, like the use of an animal name for its meat. Similarly, in Ex. 1, *Vietnam*, the name of a location, refers to an event (a war) that happened there.

(1) Sex, drugs, and **Vietnam** have haunted Bill Clinton's campaign.

In Ex. 2 and 3, *BMW*, the name of a company, stands for its index on the stock market, or a vehicle manufactured by BMW, respectively.

(2) **BMW** slipped 4p to 31p

(3) His **BMW** went on to race at Le Mans

The importance of resolving metonymies has been shown for a variety of NLP tasks, such as ma-

[1] This example was taken from the Berkely Master Metaphor list (Lakoff and Johnson, 1980) .

[2] From now on, all examples in this paper are taken from the British National Corpus (BNC) (Burnard, 1995), but Ex. 23.

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 36–41,
Prague, June 2007. ©2007 Association for Computational Linguistics

chine translation (Kamei and Wakao, 1992), question answering (Stallard, 1993), anaphora resolution (Harabagiu, 1998; Markert and Hahn, 2002) and geographical information retrieval (Leveling and Hartrumpf, 2006).

Although metonymic readings are, like all figurative readings, potentially open ended and can be innovative, the regularity of usage for word groups helps in establishing a common evaluation framework. Many other location names, for instance, can be used in the same fashion as *Vietnam* in Ex. 1. Thus, given a semantic class (e.g. location), one can specify several regular metonymic patterns (e.g. place-for-event) that instances of the class are likely to undergo. In addition to literal readings, regular metonymic patterns and innovative metonymic readings, there can also be so-called mixed readings, similar to zeugma, where both a literal and a metonymic reading are evoked (Nunberg, 1995).

The metonymy task is a lexical sample task for English, consisting of two subtasks, one concentrating on the semantic class *location*, exemplified by country names, and another one concentrating on *organisation*, exemplified by company names. Participants had to automatically classify preselected country/company names as having a literal or non-literal meaning, given a four-sentence context. Additionally, participants could attempt finer-grained interpretations, further specifying readings into prespecified metonymic patterns (such as place-for-event) and recognising innovative readings.

2 Annotation Categories

We distinguish between literal, metonymic, and mixed readings for locations and organisations. In the case of a metonymic reading, we also specify the actual patterns. The annotation categories were motivated by prior linguistic research by ourselves (Markert and Nissim, 2006), and others (Fass, 1997; Lakoff and Johnson, 1980).

2.1 Locations

Literal readings for locations comprise *locative* (Ex. 4) and *political* entity interpretations (Ex. 5).

(4) coral coast of **Papua New Guinea**.

(5) **Britain**'s current account deficit.

Metonymic readings encompass four types:

- place-for-people a place stands for any persons/organisations associated with it. These can be governments (Ex. 6), affiliated organisations, incl. sports teams (Ex. 7), or the whole population (Ex. 8). Often, the referent is underspecified (Ex. 9).

(6) **America** did once try to ban alcohol.

(7) **England** lost in the semi-final.

(8) [...] the incarnation was to fulfil the promise to **Israel** and to reconcile the world with God.

(9) The G-24 group expressed readiness to provide **Albania** with food aid.

- place-for-event a location name stands for an event that happened in the location (see Ex. 1).

- place-for-product a place stands for a product manufactured in the place, as *Bordeaux* in Ex. 10.

(10) a smooth **Bordeaux** that was gutsy enough to cope with our food

- othermet a metonymy that does not fall into any of the prespecified patterns, as in Ex. 11, where *New Jersey* refers to typical local tunes.

(11) The thing about the record is the influences of the music. The bottom end is very New York/**New Jersey** and the top is very melodic.

When two predicates are involved, triggering a different reading each (Nunberg, 1995), the annotation category is **mixed**. In Ex. 12, both a literal and a place-for-people reading are involved.

(12) they arrived in **Nigeria**, hitherto a leading critic of [...]

2.2 Organisations

The **literal** reading for organisation names describes references to the organisation in general, where an organisation is seen as a legal entity, which consists of organisation members that speak with a collective voice, and which has a charter, statute or defined aims. Examples of literal readings include (among others) descriptions of the structure of an organisation (see Ex. 13), associations between organisations (see Ex. 14) or relations between organisations and products/services they offer (see Ex. 15).

(13) **NATO** countries

(14) **Sun** acquired that part of Eastman-Kodak Cos Unix subsidary

(15) **Intel**'s Indeo video compression hardware

Metonymic readings include six types:

- org-for-members an organisation stands for its members, such as a spokesperson or official (Ex. 16), or all its employees, as in Ex. 17.

(16) Last February **IBM** announced [. . .]

(17) It's customary to go to work in black or white suits. [. . .] **Woolworths** wear them

- org-for-event an organisation name is used to refer to an event associated with the organisation (e.g. a scandal or bankruptcy), as in Ex. 18.

(18) the resignation of Leon Brittan from Trade and Industry in the aftermath of **Westland**.

- org-for-product the name of a commercial organisation can refer to its products, as in Ex. 3.

- org-for-facility organisations can also stand for the facility that houses the organisation or one of its branches, as in the following example.

(19) The opening of a **McDonald's** is a major event

- org-for-index an organisation name can be used for an index that indicates its value (see Ex. 2).

- othermet a metonymy that does not fall into any of the prespecified patterns, as in Ex. 20, where *Barclays Bank* stands for an account at the bank.

(20) funds [. . .] had been paid into **Barclays Bank**.

Mixed readings exist for organisations as well. In Ex. 21, both an org-for-index and an org-for-members pattern are invoked.

(21) **Barclays** slipped 4p to 351p after confirming 3,000 more job losses.

2.3 Class-independent categories

Apart from class-specific metonymic readings, some patterns seem to apply across classes to all names. In the SemEval dataset, we annotated two of them.

object-for-name all names can be used as mere signifiers, instead of referring to an object or set of objects. In Ex. 22, both *Chevrolet* and *Ford* are used as strings, rather than referring to the companies.

(22) **Chevrolet** is feminine because of its sound (it's a longer word than **Ford**, has an open vowel at the end, connotes Frenchness).

object-for-representation a name can refer to a representation (such as a photo or painting) of the referent of its literal reading. In Ex. 23, *Malta* refers to a drawing of the island when pointing to a map.

(23) This is **Malta**

3 Data Collection and Annotation

We used the CIA Factbook[3] and the Fortune 500 list as sampling frames for country and company names respectively. All occurrences (including plural forms) of all names in the sampling frames were extracted in context from all texts of the BNC, Version 1.0. All samples extracted are coded in XML and contain up to four sentences: the sentence in which the country/company name occurs, two before, and one after. If the name occurs at the beginning or end of a text the samples may contain less than four sentences.

For both the location and the organisation subtask, two random subsets of the extracted samples were selected as training and test set, respectively. Before metonymy annotation, samples that were not understood by the annotators because of insufficient context were removed from the datsets. In addition, a sample was also removed if the name extracted was a homonym not in the desired semantic class (for example *Mr. Greenland* when annotating locations).[4]

For those names that do have the semantic class `location` or `organisation`, metonymy annotation was performed, using the categories described in Section 2. All training set annotation was carried out independently by both organisers. Annotation was highly reliable with a *kappa* (Carletta, 1996) of

[3] `https://www.cia.gov/cia/publications/factbook/index.html`

[4] Given that the task is not about standard Named Entity Recognition, we assume that the general semantic class of the name is already known.

Table 1: Reading distribution for locations

reading	train	test
literal	737	721
mixed	15	20
othermet	9	11
obj-for-name	0	4
obj-for-representation	0	0
place-for-people	161	141
place-for-event	3	10
place-for-product	0	1
total	925	908

Table 2: Reading distribution for organisations

reading	train	test
literal	690	520
mixed	59	60
othermet	14	8
obj-for-name	8	6
obj-for-representation	1	0
org-for-members	220	161
org-for-event	2	1
org-for-product	74	67
org-for-facility	15	16
org-for-index	7	3
total	1090	842

.88/.89 for locations/organisations.[5] As agreement was established, annotation of the test set was carried out by the first organiser. All cases which were not entirely straightforward were then independently checked by the second organiser. Samples whose readings could not be agreed on (after a reconciliation phase) were excluded from both training and test set. The reading distributions of training and test sets for both subtasks are shown in Tables 1 and 2.

In addition to a simple text format including only the metonymy annotation, we provided participants with several linguistic annotations of both training and testset. This included the original BNC tokenisation and part-of-speech tags as well as manually annotated dependency relations for each annotated name (e.g. *BMW subj-of-slip* for Ex. 2).

4 Submission and Evaluation

Teams were allowed to participate in the location or organisation task or both. We encouraged supervised, semi-supervised or unsupervised approaches.

Systems could be tailored to recognise metonymies at three different levels of granularity: *coarse*, *medium*, or *fine*, with an increasing number and specification of target classification categories, and thus difficulty. At the *coarse* level, only a distinction between literal and non-literal was asked for; *medium* asked for a distinction between literal, metonymic and mixed readings; *fine* needed a classification into literal readings, mixed readings, any of the class-dependent and class-independent metonymic patterns (Section 2) or an innovative metonymic reading (category `othermet`).

Systems were evaluated via accuracy (acc) and coverage (cov), allowing for partial submissions.

$$acc = \frac{\#\ correct\ predictions}{\#\ predictions} \quad cov = \frac{\#\ predictions}{\#\ samples}$$

For each target category c we also measured:

$$precision_c = \frac{\#\ correct\ assignments\ of\ c}{\#\ assignments\ of\ c}$$
$$recall_c = \frac{\#\ correct\ assignments\ of\ c}{\#\ dataset\ instances\ of\ c}$$
$$fscore_c = \frac{2 precision_c recall_c}{precision_c + recall_c}$$

A baseline, consisting of the assignment of the most frequent category (always literal), was used for each task and granularity level.

5 Systems and Results

We received five submissions (`FUH`, `GYDER`, `up13`, `UTD-HLT-CG`, `XRCE-M`). All tackled the location task; three (`GYDER`, `UTD-HLT-CG`, `XRCE-M`) also participated in the organisation task. All systems were full submissions (coverage of 1) and participated at all granularity levels.

5.1 Methods and Features

Out of five teams, four (`FUH`, `GYDER`, `up13`, `UTD-HLT-CG`) used supervised machine learning, including single (`FUH`,`GYDER`, `up13`) as well as multiple classifiers (`UTD-HLT-CG`). A range of learning paradigms was represented (including instance-based learning, maximum entropy, decision trees, etc.). One participant (`XRCE-M`) built a hybrid system, combining a symbolic, supervised approach based on deep parsing with an unsupervised distributional approach exploiting lexical information obtained from large corpora.

Systems `up13` and `FUH` used mostly shallow features extracted directly from the training data (including parts-of-speech, co-occurrences and collo-

[5] The training sets are part of the already available Mascara corpus for metonymy (Markert and Nissim, 2006). The test sets were newly created for SemEval.

cations). The other systems made also use of syntactic/grammatical features (syntactic roles, determination, morphology etc.). Two of them (GYDER and UTD-HLT-CG) exploited the manually annotated grammatical roles provided by the organisers.

All systems apart from up13 made use of external knowledge resources such as lexical databases for feature generalisation (WordNet, FrameNet, VerbNet, Levin verb classes) as well as other corpora (the Mascara corpus for additional training material, the BNC, and the Web).

5.2 Performance

Tables 3 and 4 report accuracy for all systems.[6] Table 5 provides a summary of the results with lowest, highest, and average accuracy and f-scores for each subtask and granularity level.[7]

The task seemed extremely difficult, with 2 of the 5 systems (up13, FUH) participating in the location task not beating the baseline. These two systems relied mainly on shallow features with limited or no use of external resources, thus suggesting that these features might only be of limited use for identifying metonymic shifts. The organisers themselves have come to similar conclusions in their own experiments (Markert and Nissim, 2002). The systems using syntactic/grammatical features (GYDER, UTD-HLT-CG, XRCE-M) could improve over the baseline whether using manual annotation or parsing. These systems also made heavy use of feature generalisation. Classification granularity had only a small effect on system performance.

Only few of the fine-grained categories could be distinguished with reasonable success (see the f-scores in Table 5). These include literal readings, and place-for-people, org-for-members, and org-for-product metonymies, which are the most frequent categories (see Tables 1 and 2). Rarer metonymic targets were either not assigned by the systems at all ("undef" in Table 5) or assigned wrongly (low f-scores). An exception is the object-for-name pattern, which XRCE-M and UTD-HLT-CG could distinguish with good success. Mixed readings also proved problematic since more than one pattern is involved, thus limiting the possibilities of learning from a single training instance. Only GYDER succeeded in correctly identifiying a variety of mixed readings in the organisation subtask. No systems could identify unconventional metonymies correctly. Such poor performance is due to the non-regularity of the reading by definition, so that approaches based on learning from similar examples alone cannot work too well.

Table 5: Overview of scores

	base	min	max	ave
LOCATION-coarse				
accuracy	0.794	0.754	0.852	0.815
literal-f		0.849	0.912	0.888
non-literal-f		0.344	0.576	0.472
LOCATION-medium				
accuracy	0.794	0.750	0.848	0.812
literal-f		0.849	0.912	0.889
metonymic-f		0.331	0.580	0.476
mixed-f		0.000	0.083	0.017
LOCATION-fine				
accuracy	0.794	0.741	0.844	0.801
literal-f		0.849	0.912	0.887
place-for-people-f		0.308	0.589	0.456
place-for-event-f		0.000	0.167	0.033
place-for-product-f		0.000	undef	0.000
obj-for-name-f		0.000	0.667	0.133
obj-for-rep-f		undef	undef	undef
othermet-f		0.000	undef	0.000
mixed-f		0.000	0.083	0.017
ORGANISATION-coarse				
accuracy	0.618	0.732	0.767	0.746
literal-f		0.800	0.825	0.810
non-literal-f		0.572	0.652	0.615
ORGANISATION-medium				
accuracy	0.618	0.711	0.733	0.718
literal-f		0.804	0.825	0.814
metonymic-f		0.553	0.604	0.577
mixed-f		0.000	0.308	0.163
ORGANISATION-fine				
accuracy	0.618	0.700	0.728	0.713
literal-f		0.808	0.826	0.817
org-for-members-f		0.568	0.630	0.608
org-for-event-f		0.000	undef	0.000
org-for-product-f		0.400	0.500	0.458
org-for-facility-f		0.000	0.222	0.141
org-for-index-f		0.000	undef	0.000
obj-for-name-f		0.250	0.800	0.592
obj-for-rep-f		undef	undef	undef
othermet-f		0.000	undef	0.000
mixed-f		0.000	0.343	0.135

[6] Due to space limitations we do not report precision, recall, and f-score per class and refer the reader to each system description provided within this volume.

[7] The value "undef" is used for cases where the system did not attempt any assignment for a given class, whereas the value "0" signals that assignments were done, but were not correct.

[8] Please note that results for the FUH system are slightly different than those presented in the FUH system description paper. This is due to a preprocessing problem in the FUH system that was fixed only after the run submission deadline.

Table 3: Accuracy scores for all systems for all the location tasks.[8]

task ↓ / system →	baseline	FUH	UTD-HLT-CG	XRCE-M	GYDER	up13
LOCATION-coarse	0.794	0.778	0.841	0.851	0.852	0.754
LOCATION-medium	0.794	0.772	0.840	0.848	0.848	0.750
LOCATION-fine	0.794	0.759	0.822	0.841	0.844	0.741

Table 4: Accuracy scores for all systems for all the organisation tasks

task ↓ / system →	baseline	UTD-HLT-CG	XRCE-M	GYDER
ORGANISATION-coarse	0.618	0.739	0.732	0.767
ORGANISATION-medium	0.618	0.711	0.711	0.733
ORGANISATION-fine	0.618	0.711	0.700	0.728

6 Concluding Remarks

There is a wide range of opportunities for future figurative language resolution tasks. In the SemEval corpus the reading distribution mirrored the actual distribution in the original corpus (BNC). Although realistic, this led to little training data for several phenomena. A future option, geared entirely towards system improvement, would be to use a stratified corpus, built with different acquisition strategies like active learning or specialised search procedures. There are also several options for expanding the scope of the task, for example to a wider range of semantic classes, from proper names to common nouns, and from lexical samples to an all-words task. In addition, our task currently covers only metonymies and could be extended to other kinds of figurative language.

Acknowledgements

We are very grateful to the BNC Consortium for letting us use and distribute samples from the British National Corpus, version 1.0.

References

J.A. Barnden, S.R. Glasbey, M.G. Lee, and A.M. Wallington. 2003. Domain-transcending mappings in a system for metaphorical reasoning. In *Proc. of EACL-2003*, 57-61.

J. Birke and A Sarkaar. 2006. A clustering approach for the nearly unsupervised recognition of nonliteral language. In *Proc. of EACL-2006*.

L. Burnard, 1995. *Users' Reference Guide, British National Corpus*. BNC Consortium, Oxford, England.

J. Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249-254.

D. Fass. 1997. *Processing Metaphor and Metonymy*. Ablex, Stanford, CA.

S. Harabagiu. 1998. Deriving metonymic coercions from WordNet. In *Workshop on the Usage of WordNet in Natural Language Processing Systems, COLING-ACL '98*, 142-148, Montreal, Canada.

J.R. Hobbs, M.E. Stickel, D.E. Appelt, and P. Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69-142.

S. Kamei and T. Wakao. 1992. Metonymy: Reassessment, survey of acceptability and its treatment in machine translation systems. In *Proc. of ACL-92*, 309-311.

S. Krishnakamuran and X. Zhu. 2007. Hunting elusive metaphors using lexical resources. In *NAACL 2007 Workshop on Computational Approaches to Figurative Language*.

G. Lakoff and M. Johnson. 1980. *Metaphors We Live By*. Chicago University Press, Chicago, Ill.

J. Leveling and S. Hartrumpf. 2006. On metonymy recognition for gir. In *Proceedings of GIR-2006: 3rd Workshop on Geographical Information Retrieval*.

K. Markert and U. Hahn. 2002. Understanding metonymies in discourse. *Artificial Intelligence*, 135(1/2):145–198.

K. Markert and M. Nissim. 2002. Metonymy resolution as a classification task. In *Proc. of EMNLP-2002*, 204-213.

K. Markert and M. Nissim. 2006. Metonymic proper names: A corpus-based account. In A. Stefanowitsch, editor, *Corpora in Cognitive Linguistics. Vol. 1: Metaphor and Metonymy*. Mouton de Gruyter, 2006.

J. Martin. 1994. Metabank: a knowledge base of metaphoric language conventions. *Computational Intelligence*, 10(2):134-149.

Z. Mason. 2004. Cormet: A computational corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23-44.

M. Nissim and K. Markert. 2003. Syntactic features and word similarity for supervised metonymy resolution. In *Proc. of ACL-2003*, 56-63.

G. Nunberg. 1995. Transfers of meaning. *Journal of Semantics*, 12:109-132.

Y Peirsman. 2006. Example-based metonymy recognition for proper nouns. In *Student Session of EACL 2006*.

D. Stallard. 1993. Two kinds of metonymy. In *Proc. of ACL-93*, 87-94.