Lang Resources & Evaluation (2009) 43:123–138
DOI 10.1007/s10579-009-9087-y

Data and models for metonymy resolution

Katja Markert · Malvina Nissim

Published online: 26 February 2009
© Springer Science+Business Media B.V. 2009

Abstract We describe the first shared task for figurative language resolution, which was organised within SemEval-2007 and focused on metonymy. The paper motivates the linguistic principles of data sampling and annotation and shows the task's feasibility via human agreement. The five participating systems mainly used supervised approaches exploiting a variety of features, of which grammatical relations proved to be the most useful. We compare the systems' performance to automatic baselines as well as to a manually simulated approach based on selectional restriction violations, showing some limitations of this more traditional approach to metonymy recognition. The main problem supervised systems encountered is data sparseness, since metonymies in general tend to occur more rarely than literal uses. Also, within metonymies, the reading distribution is skewed towards a few frequent metonymy types. Future task developments should focus on addressing this issue.

Keywords Metonymy · Selectional restrictions · Shared task evaluation

1 Introduction

Both word sense disambiguation (WSD) and named entity recognition (NER) have benefited enormously from shared task evaluations, for example in the Senseval, MUC and CoNLL frameworks. Figurative language, such as metaphor, metonymy, idioms

K. Markert (✉)
School of Computing, University of Leeds, Woodhouse Lane, LS2 9JT Leeds, UK
e-mail: markert@comp.leeds.ac.uk

M. Nissim
Department of Linguistics and Oriental Studies, University of Bologna, via Zamboni 33,
40126 Bologna, Italy
e-mail: malvina.nissim@unibo.it

and irony, has not received a comparable amount of attention. However, resolution of figurative language is an important complement to and extension of WSD.

Sense distinctions in dictionaries do include conventionalized figurative readings, such as the metonymic meat reading and the metaphoric coward reading for the word *chicken*, both of which are listed in WordNet (Fellbaum 1998), for example. These patterns (using an animal for its meat and an animal for metaphoric descriptions of a person) are also present in other dictionary entries for animals, such as *lamb, pig* and *shark*. Thus, (fine-grained) WSD deals implicitly with the detection of some figurative readings. However, dictionaries normally do not link literal and figurative senses for semantic classes (such as animals) systematically, therefore WSD misses out on generalisation via usage patterns. In addition, figurative language resolution has to deal with word senses that are not listed in the lexicon. For example, the meaning of *stopover* in *He saw teaching as a stopover on his way to bigger things* is a metaphorical use of the sense "stopping place in a physical journey", with the literal sense listed in WordNet but the metaphorical one being absent.[1] Similarly, the metonymic meat reading of *rattlesnake* in *Roast rattlesnake tastes like chicken* is not in WordNet.[2] Named entities, which we focus on in this paper, are also often used figuratively, but not normally listed in dictionaries.

Most traditional computational approaches to figurative language resolution carried out only small-scale evaluations (Pustejovsky 1995; Fass 1997; Hobbs et al. 1993; Briscoe and Copestake 1999; Barnden et al. 2003). In recent years, there has been growing interest in figurative language resolution that is corpus-based or evaluated on larger datasets (Martin 1994; Lapata and Lascarides 2003; Nissim and Markert 2003; Mason 2004; Peirsman 2006; Birke and Sarkaar 2006; Krishnakamuran and Zhu 2007). Still, apart from Nissim and Markert (2003) and Peirsman (2006), who evaluated their work on the same data, results are not comparable.

This situation motivated us to organize the first shared task for figurative language resolution, focusing 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 shares that are traded on the stock market, or a vehicle manufactured by BMW, respectively.

(2) **BMW** slipped 4p to 31p
(3) She arrived in a big **BMW** of the type the East End villains drive.

Resolving metonymies is important for a variety of NLP tasks, such as machine translation (Kamei and Wakao 1992), question answering (Stallard 1993), anaphora resolution (Harabagiu 1998; Markert and Hahn 2002) and geographical IR (Leveling and Hartrumpf 2006).

[1] The example is from the Berkeley Master Metaphor list (http://cogsci.berkeley.edu/lakoff/).

[2] This and all following examples in this paper are from the British National Corpus (BNC) (Burnard 1995). An exception is Ex. 22.

The SemEval-2007 task set-up is described in Sect. 2 and its underlying principles, advantages and disadvantages are discussed. In Sect. 3 we explore simple automatic baselines and discuss evaluation measures, and in Sects. 4 and 5 we focus on the five participating systems and their strengths and weaknesses. In Sect. 6 we describe previous work on metonymy resolution, which was based mostly on selectional restriction violations, and simulate how well such an algorithm would do on our dataset. Finally, we draw conclusions on the current performance level of metonymy resolution systems and discuss the possibilities for future developments.

2 The SemEval-2007 shared task for metonymy resolution

The task was organized as a lexical sample task for English. We profited from the well-established observation that although metonymic readings are potentially open-ended and can be innovative, there exist usage regularities for semantic word classes (Lakoff and Johnson 1980). Many other location names, for instance, can be used in the same way as *Vietnam* in Ex. 1. Thus, given a semantic class (e.g. location), one can specify regular metonymic patterns (e.g. place-for-event) that class instances are likely to undergo. We focused on the classes *location* and *organisation*, exemplified by country and company names, respectively. Participants had to automatically classify preselected country/company names into literal and non-literal, given a four-sentence context. Additionally, they could attempt finer-grained interpretations, such as recognizing prespecified metonymic patterns and innovative readings. Training and test data was produced using the framework of Markert and Nissim (2006), summarised below.

2.1 Annotation framework

We distinguish between literal, metonymic, and mixed readings. In the case of a metonymic reading, we also specify the actual patterns.

2.1.1 Locations

Literal readings comprise *locative* (Ex. 4) and *political* senses (Ex. 5).

(4) coral coast of **Papua New Guinea**.
(5) The **Socialist Republic of Vietnam** was proclaimed in 1976.

Metonymic readings encompass four metonymic patterns:

place-for-people a place stands for any persons/organisations associated with it. These can be governments (Ex. 6), affiliated organisations, including 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) […] a perfect own goal which gave **Wales** a fortunate draw.

(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 jug of new **Bordaux**

othermet a metonymy that does not fall into any of the prespecified patterns. In Ex. 11, *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 trigger a different reading each (Nunberg 1995), the annotation category is **mixed**. In Ex. 12, both a literal (triggered by *in*) and a place-for-people reading (triggered by *a leading critic*) are involved.

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

2.1.2 Organisations

The **literal** reading of organisations describes references to the organisation as a legal entity that has members and a charter or defined aims. Examples include descriptions of the organisation's structure (Ex. 13) or relations between organisations and their products/services (Ex. 14).

(13) **NATO** countries
(14) **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. 15), or all its employees, as in Ex. 16.[3]

(15) **IBM** argued that the market should be analysed as a whole
(16) 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 such as a scandal (Ex. 17).

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

[3] Org-for-members metonymies referring to a spokesperson are quite commonplace so that it is tempting to see them as literal readings. We follow here previous linguistic research (Fass 1997; Lakoff and Johnson 1980) that see these as metonymies.

org-for-product a company name can refer to its products (Ex. 3).
org-for-facility organisations can also stand for the facility that houses the organisation or one of its branches, as in Ex. 18.

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

org-for-index an organisation name is used to indicate its value, such as by their shares on the stock market (see Ex. 2).
othermet a metonymy that does not fit any prespecified pattern. In Ex. 19, *Barclays Bank* stands for an account at the bank.

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

Mixed readings exist for organisations as well. In Ex. 20, both an org-for-index and an org-for-members pattern are invoked.

(20) **Barclays** slipped 4p […] after confirming 3,000 more job losses.

2.1.3 Class-independent categories

Some metonymic patterns can apply across classes to all names:

object-for-name all names can be used as mere signifiers or strings. Thus, in Ex. 21, both *Chevrolet* and *Ford* are used as strings, rather than referring to the companies.

(21) **Chevrolet** is feminine because of its sound (it's a longer word than **Ford**, has an open vowel at the end […]

object-for-representation a name can refer to a representation (such as a photo) of the referent of its literal reading. In Ex. 22, *Malta* refers to a drawing of the island when pointing to a map.

(22) This is **Malta**

2.2 Data collection, annotation, and distribution

We used the CIA Factbook[4] and the Fortune 500 list as sampling frames for country and company names, respectively. All occurrences (including plurals) of all names in the sampling frames were extracted in context from all texts of the BNC 1.0. All samples contain up to four sentences: the sentence with the country/company name, 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 task, two random subsets of the extracted samples were selected as training and test set. Before metonymy annotation, we removed samples that were not understood by the annotators because of insufficient context. A sample was also removed if the extracted name was a homonym not in the desired semantic class (for example, *Mr. Greenland* when annotating locations).

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

Table 1 Reading distributions

Reading	LOCATIONS		ORGANISATIONS	
	Train	Test	Train	Test
literal	737	721	690	520
mixed	15	20	59	60
othermet	9	11	14	8
obj-for-name	0	4	8	6
obj-for-rep	0	0	1	0
place-for-people	161	141	–	–
place-for-event	3	10	–	–
place-for-product	0	1	–	–
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	925	908	1,090	842

On all remaining cases metonymy annotation was performed, using the categories in Sect. 2.1. All training set annotation was carried out independently by both authors and proved highly reliable, with a percentage agreement of 0.94/0.95 and a *Kappa* (Carletta 1996) of 0.88/0.89 for locations/organisations (Markert and Nissim 2006). As agreement was established, test set annotation was carried out by the first author. Difficult cases were then independently checked by the second author. Samples whose readings could not be agreed on after a reconciliation phase were excluded from training and test sets. The reading distributions are shown in Table 1. We kept rare classes as target categories as they are regular sense extensions described in the linguistic literature and are clearly separate senses (for example, org-for-event).

The datasets also included the original BNC header information, tokenisation and part-of-speech tags for each sample. We also provided manually annotated head-modifier relations for each annotated name in training and test sets. Thus, Ex. 2 is annotated as *subj-of-slip*. Syntactic relations had proved useful for metonymy recognition, and we wanted all teams to be able to use them, while abstracting away from parser errors. We refer the reader to Nissim and Markert (2003) for a study on syntactic relations for metonymy recognition and on the influence of automatic parsing. The relations with examples and their distribution in the data are reported in Table 2. The upper part of Table 2 contains relations where the name is a modifier (such as of an adjective (Ex. 23) or in an apposition (Ex. 24)) and the lower part where it is a head (with modifications such as a genitive (Ex. 25) or a noun premodifier (Ex. 26), among others).

(23) […] the **IBM** compatible PC
(24) in their own countries—Italy, Germany and **France**—they are stars
(25) Germany’s **Lufthansa**

Table 2 Distribution of dependency relations for all datasets with reference to examples

Relation	Locations				Organisations				Ex.
	Train	Non-lit	Test	Non-lit	Train	Non-lit	Test	Non-lit	
subj	100	72	100	71	374	249	291	217	Ex. 2
obj	29	13	43	19	56	20	33	18	Ex. 9
subjpassive	9	5	5	2	7	2	12	7	Ex. 5
iobj	1	0	5	5	7	0	3	2	Ex. 7
adj	0	0	0	0	1	0	0	0	Ex. 23
app	35	14	59	16	46	22	48	28	Ex. 24
premod	95	13	91	10	198	33	163	23	Ex. 13
gen	93	20	72	11	146	16	125	14	Ex. 14
pred	9	2	8	2	11	6	7	2	Ex. 11
pp	529	60	518	63	277	80	230	73	Ex. 17
hasgen	1	0	0	0	6	2	1	0	Ex. 25
hasadj	28	6	30	6	36	23	24	22	Ex. 3
haspp	0	0	5	3	6	6	11	7	Ex. 3
hasapp	5	0	5	0	19	7	15	9	Ex. 12
haspremod	0	0	0	0	3	3	1	1	Ex. 26
None	37	4	22	0	39	11	10	2	
All	971	209	963	209	1232	490	974	440	

(26) including SCO board member **Microsoft Corp**

Instances were annotated with all relations in cases of coordination or cases such as Ex. 3, which is a modifier of a pp (*pp*), head of a pp (*haspp*), and has an adjective (*hasadj*). Thus, the number of relations is higher than the number of instances (see Table 2). Names without any relations (for example, in simple headlines) are marked as *none*.

2.3 Task analysis

Being the first task for figurative language, we adopted several simplifications, which had an impact on representativeness and feasibility.

The task was set up as a lexical sample instead of an all-words task. This follows the example of task development for WSD in the Senseval competitions and is a reasonable starting point for a newly evaluated phenomenon. The class-based sampling method still allows for the inclusion of a relatively wide range of word types, going beyond standard WSD lexical sample tasks.

Location and organisation names were chosen as their metonymic interpretation is a natural extension to standard NER. In addition, they are frequent, cover many different word types and undergo a wide variety of metonymic patterns. Metonymic usage of named entities also influences their syntactic and morphological behaviour such as pluralisation (*BMWs*) and determination (*a BMW*), making its detection potentially relevant for parsing applications. Our annotation scheme covered the full range of location/organisation metonymies, although we restricted sampling to country/company names. This is exemplified in including place-for-product metonymies which rarely apply to countries but more frequently to regions and cities. Sampling different location/organisation names would yield a different a priori frequency distribution of readings; however, the trigger environments for the readings would remain similar. For example, "<organisation> argues" triggers an org-for-members metonymy independent of organisation type.

We randomly extracted samples from a large, representative corpus as the most unbiased selection procedure possible. Thus, systems had to cope with facts that are common place in language such as a large skew in distribution (for example, 80% of locations being literal), lack of training data for some categories (for example, there were no object-for-name training instances for locations although examples in the annotation manual were given) and some samples with spelling or grammatical errors. It also did not bias the competition too much towards supervised systems which can profit from balanced training data. On the negative side, the training data alone does not provide a wide range of examples for some target classes and is not geared towards algorithm optimization. However, participants were free to use manually or automatically acquired additional data.

We assumed that the semantic class of the name is already known, i.e. that metonymy resolution can *follow* standard NER. This assumption is only reasonable if there is no pressing need to combine NER and metonymy resolution, i.e. if a priori NER performs equally well on literally and metonymically used named entities. We ran the GATE NE recognizer (Cunningham et al. 2002) on our datasets and computed how many of our annotated names were not detected by GATE as a location or organisation in the first place. There was no significant difference in GATE's error rate for metonymic and literal named entities, suggesting that a pipeline approach should indeed be feasible.

3 Evaluation and baselines

Teams could participate in the location or organisation task or both and recognise metonymies at three different levels of granularity: *coarse*, *medium*, or *fine*, with an increasing number and specification of target 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 or an innovative metonymic reading (category othermet). Systems were evaluated via accuracy (acc), i.e percentage of correct assignments, as well as precision, recall and f-score for each target category. All comparisons were conducted with a McNemar test with a significance level of 5%.

We use three baselines for system comparison. Their accuracy measures on the test sets are summarised in Table 3.

Table 3 Accuracy scores for all baselines and participating systems

task ↓ / system →	MFS	GRAMM	SUBJ	FUH	UTD	XRCE-M	GYDER	up13
LOC-coarse	0.794	0.833	0.834	0.778	0.841	0.851	0.852	0.754
LOC-medium	0.794	0.821	0.824	0.772	0.840	0.848	0.848	0.750
LOC-fine	0.794	0.817	0.819	0.759	0.822	0.841	0.844	0.741
ORG-coarse	0.618	0.748	0.736	–	0.739	0.732	0.767	–
ORG-medium	0.618	0.699	0.702	–	0.711	0.711	0.733	–
ORG-fine	0.618	0.688	0.688	–	0.711	0.700	0.728	–

The supervised baseline MFS assigns the most frequent sense in the training data ("literal") to all test instances, resulting in an accuracy of 79.4% for the location and 61.8% for the organisation test set.

The unsupervised baseline SUBJ assumes that subjects often play an active role and are therefore more likely to be metonymic for our semantic classes. Thus, it assigns a non-literal reading to all subjects, and literal otherwise. For medium and fine-grained evaluation we predict metonymic or place-for-people/org-for-members, respectively.

The supervised baseline GRAMM assigns each test instance the reading that was most frequent for its grammatical role in the training set (see Table 2). As an example, for organisations for the coarse-grained categories, only the roles of subjects, pred, hasadj, haspp and haspremod trigger a non-literal reading. If an instance has two relations which give conflicting information, a non-literal reading (or mixed for non-coarse) is assigned for both the SUBJ and GRAMM baseline.

SUBJ and GRAMM significantly outperform MFS on all tasks and granularity levels. However, they are mostly useful for the recognition of non-literal readings (coarse-grained set-up), instead of interpretation (see Table 3).

4 Participating systems

Five teams took part in the task: FUH (University of Hagen, Germany), GYDER (Universities of Budapest and Szeged, Hungary), up13 (University of Paris 13), UTD (University of Texas at Dallas) and XRCE-M (Xerox, Grenoble). All tackled the location task, and three—GYDER, UTD, XRCE-M—also the organisation task. All systems participated at all granularity levels. We refer you to Agirre et al. (2007) for full system descriptions.

Four of the five teams (FUH, GYDER, up13, UTD) used supervised machine learning, including instance-based learning (FUH), maximum entropy (GYDER) and rule-based learning (up13), as well as voting between different classifiers (UTD). In contrast, XRCE-M is a hybrid system. Trigger environments for the target classes (such as that the subject of an economic action verb should be metonymic) were derived manually from a parsed version of the training corpus. These triggers were then generalised automatically via measuring distributional similarity of environments in the BNC.

The teams up13 and FUH used solely shallow features such as co-occurrences and collocations: up13 used plain word forms only, while FUH also used prefixes, lemmata, parts-of-speech and WordNet synsets as co-occurrences/collocations. All other systems used syntactic relations: XRCE-M via deep parsing and GYDER and UTD via the manually annotated head-modifier relations we provided. UTD and GYDER also used other feature types, such as collocations (UTD only), occurrence of determiners, number of the name to be classified (GYDER only), the individual name form (GYDER only) and quotation marks around the name (UTD only).

All systems except up13 used external knowledge resources for feature generalisation to capture regularities between instances such as *BMW says* and *BMW announces*. These included WordNet (UTD, GYDER, FUH), Verbnet (Schuler 2005) in UTD, Levin verb classes (Levin 1993) in GYDER, and the BNC for computing distributional similarity (XRCE-M). Only FUH used additional training material explicitly annotated for metonymies, i.e. the Mascara corpus (Markert and Nissim 2006).

5 Results and discussion

Table 3 reports accuracy for all systems.[5] The task seemed extremely difficult, with two of the five systems (up13 and FUH) not beating MFS. Although all the other systems perform significantly better than MFS, no system achieves a significantly better accuracy than the other baselines (GRAMM and SUBJ) on the location data for the coarse-grained setup. On organisations, only GYDER significantly beats SUBJ. However, when we get into more detailed interpretations, especially GYDER outperforms all baselines significantly for both locations and organisations.

In a highly skewed data distribution such as ours, MFS is advantaged when using simple accuracy for evaluation. Therefore, for the coarse classification, we also calculated the balanced error rate (BER), which averages the error rate on positive (non-literal) examples and that on negative (literal) ones. The balanced error rate for MFS is 50%. On locations, both up13 and FUH show an improvement with a lower BER of 40%. However, FUH and up13 are still outperformed by the other systems (BER is 30% for GYDER and XRCE-M, 27% for UTD). For organisations, GYDER performs best (BER = 26%), followed by UTD (BER = 29%) and XRCE-M (BER = 31%).

5.1 Target category difficulty

Only few fine-grained categories could be distinguished with reasonable success. These include literal readings and the most frequent metonymic patterns place-for-people (highest f-score: 0.589), org-for-members (highest f-score: 0.630), and org-for-product (highest f-score: 0.5). The only rare metonymic pattern that two systems (XRCE-M and UTD) could distinguish with good success (highest f-scores: 0.667 for

[5] FUH results are slightly different from the FUH system paper due to a preprocessing problem in the system, fixed only after the run submission deadline.

locations and 0.8 for organisations) is object-for-name. No system could identify unconventional metonymies correctly as their non-regularity does not lend itself easily to a paradigm that learns from similar examples.

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 correctly identified some mixed readings of organisations (f-score = 0.34) We did not grant systems credit for the recognition of one of the two readings as this would be an oversimplification of the category, which specifically asks for the joint recognition of two readings. In addition, for all mixed instances in the test set one of the two readings involved is "literal" (the other one is place-for-people in 18 out of 20 cases for locations, and org-for-members in 58 out of 61 cases for organisations). Thus, an all-literal baseline over the mixed cases would achieve top performance in a partial credit scenario.

Regarding the agreement between the three top systems on the location task (GYDER, XRCE-M, and UTD), 675 out of 908 location names (74.3%) were correctly classified by all three systems. Interestingly, but perhaps not so surprisingly, only 42 of them (5.6%) are non-literal readings. Given that non-literal instances make up 20.6% of the whole dataset, their identification was clearly more difficult. Similarly, among the 57 country names that no system could classify correctly, 53 (93%) are non-literal. Similar results can be observed for the organisation task.

5.2 Feature analysis

All three top scoring systems used head-modifier relations. Previous work has also shown such relations to play a crucial role in metonymy resolution, allowing to beat an MFS baseline relatively easily (Markert and Nissim 2002; Nissim and Markert 2003). Unfortunately, performance is not equally convincing on all relation types. Most of the systems' gains are made on subjects and objects (14.7% of locations and 31.9% of organisations, see also Table 4) with low performance on all other relations. Table 5 shows the performance of the three top systems on subjects and objects (extension .so) and on the set of all other relations (extension .notso) for the coarse-granularity tasks. The systems do not outperform MFS on the .notso subset for locations. They do slightly better on the .notso set for organisations via the use of determiner features (which help identify metonymies such as *a BMW*) and number features (which help identify metonymies such as *two BMWs*).

The systems up13 and FUH, which relied on shallow features such as co-occurrences and collocations only, did not achieve high results. Similarly, GYDER

Table 4 Subjects and objects in our datasets

Dataset	subj		obj		Other
	# inst	# rels	# inst	# rels	# inst
countries.train	94	100	29	29	802
countries.test	94	100	40	43	774
companies.train	344	374	53	56	693
companies.test	238	291	31	33	573

Table 5 Performance of baselines and systems on the subjects and objects subset (so), and on the remaining instances (notso)

Data subset	MFS	SUBJ	GYDER	UTD	XRCE-M
countries.test.so	0.395	0.664	0.773	0.657	0.694
countries.test.notso	0.867	0.867	0.873	0.873	0.878
companies.test.so	0.316	0.684	0.729	0.695	0.617
companies.test.notso	0.757	0.757	0.785	0.789	0.785

report in their system paper that the addition of shallow features to their system did not improve its performance. However, as pointed out that the beginning of this Section, the balanced error rate of up13 and FUH is better than for MFS and showed that co-occurrences and collocations seem to have some value for metonymy recognition but would need further exploration.

6 Related work: exploiting selectional restrictions

Classic work on metonymy resolution carried out only small-scale evaluations, on either artificially created examples or datasets annotated by a single annotator only (Pustejovsky 1995; Fass 1997; Hobbs et al. 1993; Stallard 1993; Copestake and Briscoe 1995; Briscoe and Copestake 1999; Markert and Hahn 2002). However, this flaw does not mean that the algorithmic approaches used in previous work cannot yield interesting or high results on our larger, reliably annotated dataset.

Most of these approaches furnish their algorithms with (manually modelled) selectional restrictions (SRs), in a lexicon (Pustejovsky 1995; Copestake and Briscoe 1995; Briscoe and Copestake 1999) or in a knowledge base (Hobbs et al. 1993; Fass 1997; Stallard 1993; Markert and Hahn 2002). These are normally not seen as preferences but as absolute constraints. If and only if such an absolute constraint is violated, a non-literal reading is proposed. In Ex. 2, an organisation can normally not *slip*, so that a non-literal reading of *BMW* might be stipulated.[6]

This differs from the approaches of Nissim and Markert (2003), Peirsman (2006) and the systems submitted to the SemEval-2007 competition, none of which used explicitly represented SRs, whether hand-modelled or automatically acquired. Instead, they use machine learning and example similarity to recognise metonymies with a wide set of features. In the experiment described below we simulate the traditional approach with hand-annotated selectional restriction violations (SRVs) in order to compare it to the current approaches.

6.1 Experiment

As SRs in the above approaches are normally defined for subjects and direct objects only, we limited this empirical study to such instances. Table 4 shows the number of

[6] This is sometimes enhanced with morphological/syntactic violations such as the plural use for proper names (Copestake and Briscoe 1995) or anaphoric information (Markert and Hahn 2002). However, the basic model relies to a large degree on SRs.

instances in each dataset that have at least one subject or object relation or none of them.

Three native speakers of English annotated subject-verb and object-verb tuples for SRVs. All annotators had a linguistic background, with Annotator1 being an expert on SRs, but they were not involved in metonymy annotation or research. They were given simple instructions, such as that a location is a spatial region that cannot perform actions that humans/animals perform. Annotator1 annotated all four datasets. To measure task feasibility, all subject–verb and object–verb tuples in the training sets countries.train and companies.train were in addition annotated by Annotator2 and Annotator3, respectively. Their agreement with Annotator1 was satisfactory, although not extremely high, with a percentage agreement of 84.5% and a kappa of 0.688 on countries.train and a percentage agreement of 83.3% and a kappa of 0.650 on companies.train. We then simulated a metonymy recognition algorithm SELRES based on the expert Annotator1, postulating a non-literal reading for an instance if and only if an SRV for one of its relations was annotated. Evaluation measures for SELRES, MFS and SUBJ for the coarse-grained task restricted to the subject/object instances of the test sets (indicated by the extension so) are summarised in Table 6.[7]

For both datasets, SELRES significantly outperforms MFS but not SUBJ. Therefore, the SRs of the verb do not necessarily add consistently useful information to the knowledge of the syntactic role alone. If we combine SELRES with a literal baseline for all instances which are not subjects/objects, we get the potential best results for the whole datasets in Table 7. These results outperform MFS but not the other baselines SUBJ/GRAMM. The best three submitted systems achieve comparable results to SELRES in the coarse evaluation framework, with GYDER significantly outperforming SELRES for organisations.

6.2 Discussion

Even for a human gold standard of hand-annotated head-modifier relations and SRVs, the results that can be achieved with an SRV approach are limited. Submitted systems were able to perform equally or better than an SRV approach without explicit modelling of verb preferences. One problem for SRVs is that their application to figurative language in prior research is limited to subjects and objects. In our datasets, only 13–15% of location and 32–36% of organisation instances (depending on training/test set) are subjects or objects (see Table 4). In addition, SRs are strong for some grammatical relationships and word combinations, but not for others (McCarthy and Carroll 2003). They are therefore unlikely to achieve high accuracy without using other knowledge sources as well. Selectional restrictions can also differ for different verb senses. An optimal approach would therefore need sense disambiguation of the verb before or joint with metonymy recognition.

However, there are also two main advantages to an SRV approach. First, SRs can sometimes indicate a fine-grained interpretation. Thus, *drive a BMW* would indicate a vehicle interpretation, due to the selectional preferences of *drive* for its direct object. However, in most cases we encountered, this interpretation is not more

[7] The SUBJ and GRAMM baselines are equal on this subset.

Table 6 Results for SRVs for subjects and objects, reported as accuracy (acc), precision (P), recall (R), and f-score (F) for non-literal (nonlit) and literal (lit) readings

Data	Classifier	acc	P_{nonlit}	R_{nonlit}	F_{nonlit}	P_{lit}	R_{lit}	F_{lit}
countries.test.so	MFS	0.395	n/a	0	n/a	0.395	1.00	0.566
countries.test.so	SUBJ	0.664	0.691	0.802	0.742	0.600	0.452	0.516
countries.test.so	SELRES	0.769	0.847	0.753	0.797	0.678	0.793	0.730
companies.test.so	MFS	0.316	n/a	0	n/a	0.316	1.00	0.480
companies.test.so	SUBJ	0.684	0.705	0.913	0.796	0.419	0.150	0.221
companies.test.so	SELRES	0.691	0.762	0.799	0.779	0.513	0.459	0.484

Table 7 Best possible results on the full corpora using SRVs

Dataset	acc	$prec_{nonlit}$	rec_{nonlit}	F_{nonlit}	$prec_{lit}$	rec_{lit}	F_{lit}
countries.test	0.849	0.847	0.326	0.471	0.849	0.984	0.912
companies.test	0.739	0.766	0.459	0.574	0.732	0.913	0.812

comprehensive than the interpretation given by our metonymic patterns as metonymies are often used in situations where the referent is deliberately left underspecified. Second, an SRV approach is unsupervised and therefore a possibly cheaper way to recognise metonymies than using training data. Obviously, the feasibility of this unsupervised approach in a non-simulation environment depends on automatic computation of selectional preferences. Algorithms exist (McCarthy and Carroll 2003; Clark and Weir 2002) but have not achieved high performance yet. In addition, they build on frequencies of word tuples in corpora. Frequent metonymies such as "<organisation> says" will therefore be included in the original countings and might be included in the selectional preference for that verb.[8] We would also need to learn a threshold to indicate when an unusual word combination might suggest a metonymic reading, which might again require training material.

7 Conclusions and future work

The first shared task on figurative language resolution organised within SemEval-2007 has made it possible to compare different systems on the same data, thus allowing us to see more clearly what features contribute chiefly to a successful approach to metonymy resolution.

Specifically, baseline performance indicates that grammatical roles play a crucial role in the identification of non-literal readings, to the point that simply using this information enables our SUBJ/GRAMM baselines to achieve a reasonably high performance on the recognition task, although not on a more detailed interpretation task. Participating systems that use grammatical roles plus the head/modifier

[8] We thank Diana McCarthy for pointing that problem out to us.

lemmata as well as additional syntactic features can beat such baselines for detailed interpretation tasks. In contrast, collocations and cooccurrences have not achieved such good performance although different use of these features might lead to improvements in future systems.

We also presented an experiment where human judges simulated a selectional restriction approach, similar to traditional approaches to figurative language recognition. Due to some intrinsic features of this approach, the results that can be achieved are limited and do not improve on the baselines that use grammatical roles alone. As violations were manually annotated, we can assume that automatic detection would bring performance figures even lower.

Instead, learning approaches to resolution, which can exploit the regularity of metonymic readings, appear to be more promising, at least for regular metonymic patterns and for fine-grained interpretation. These have been used by the participating systems. However, these systems also have up to now not achieved very high accuracies, illustrating the difficulty of the task. One reason is the data sparseness problem that we have witnessed in our dataset. Indeed, the SemEval-2007 corpus was collected in such a way that 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 develop a stratified corpus. One avenue of future work is to explore acquisition strategies for such a corpus, including active learning.

There are also several options for expanding the scope of the task, to a wider range of semantic classes, from proper names to common nouns, and from lexical sample to an all-words task. In addition, a broader task to include figurative language phenomena other than metonymy could be organised within future evaluation campaigns.

Acknowledgements We thank the BNC Consortium for allowing us to distribute the extracted samples. We are also grateful to the annotators for the selectional restriction simulations: Ben Hachey, Tim O'Donnell and especially Stephen Clark, who bore the brunt of the annotation. We also had valuable discussions with Diana McCarthy during the preparation of this work.

References

Agirre, E., Màrquez, L., & Wicentowski, R. (Eds.). (2007). *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics.

Barnden, J., Glasbey, S., Lee, M., & Wallington, A. (2003). Domain-transcending mappings in a system for metaphorical reasoning. In *Proc. of EACL-2003*, pp. 57–61.

Birke, J., & Sarkaar, A. (2006). A clustering approach for the nearly unsupervised recognition of nonliteral language. In *Proceedings of EACL-2006*.

Briscoe, T., & Copestake, A. (1999). Lexical rules in constraint-based grammar. *Computational Linguistics, 25*(4), 487–526.

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

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*(2), 249–254.

Clark, S., & Weir, D. (2002). Class-based probability estimation using a semantic hierarchy. *Computational Linguistics, 28*(2), 187–206.

Copestake, A., & Briscoe, T. (1995). Semi-productive polysemy and sense extension. *Journal of Semantics, 12*, 15–67.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. of ACL-2002*.

Fass, D. (1997). *Processing metaphor and metonymy*. Stanford, CA: Ablex.

Fellbaum, C. (Ed.). (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

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

Hobbs, J. R., Stickel, M. E., Appelt, D. E., & Martin, P. (1993). Interpretation as abduction. *Artificial Intelligence, 63*, 69–142.

Kamei, S.-I., & Wakao, T. (1992). Metonymy: Reassessment, survey of acceptability and its treatment in machine translation systems. In *Proc. of ACL-1992*, pp. 309–311.

Krishnakamuran, S., & Zhu, X. (2007). Hunting elusive metaphors using lexical resources. In *Proc. of the NAACL-2007 Workshop on Computational Approaches to Figurative Language*.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: Chicago University Press.

Lapata, M., & Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics, 29*, 263–317.

Leveling, J., & Hartrumpf, S. (2006). On metonymy recognition for gir. In *Proc. of GIR-2006*.

Levin, B. (1993). *English verb classes and alternations*. Chicago: University of Chicago Press.

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

Markert, K., & Nissim, M. (2002). Metonymy resolution as a classification task. In *Proc. of EMNLP-2002*, pp. 204–213.

Markert, K., & Nissim, M. (2006). Metonymic proper names: A corpus-based account. In A. Stefanowitsch (Ed.), *Corpora in cognitive linguistics. Vol. 1: Metaphor and metonymy*. Berlin: Mouton de Gruyter.

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

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

McCarthy, D., & Carroll, J. (2003). Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics, 29*(4), 639–654.

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

Nunberg, G. (1995). Transfers of meaning. *Journal of Semantics, 12*, 109–132.

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

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.

Schuler, K. K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon*. Dissertation, University of Pennsylvania.

Stallard, D. (1993). Two kinds of metonymy. In *Proc. of ACL-1993*, pp. 87–94.