Learning to buy a Renault and talk to BMW: A supervised approach to conventional metonymy

Malvina Nissim*
School of Informatics
University of Edinburgh

Katja Markert
School of Computing
University of Leeds

Abstract

We test the validity of a machine learning approach to metonymy resolution, presenting a case study for organisation names and comparing the results with a previous experiment on location names. We describe a reliable annotation scheme for organisation names and present a corpus annotated for metonymic usage. We then discuss experiments with a supervised classification algorithm on this corpus, focusing on feature contributions and highlighting advantages and disadvantages of a classification approach to metonymy resolution.

1 Introduction

Metonymy is a figure of speech, in which one expression is used to refer to the standard referent of a related one (Lakoff and Johnson, 1980). In (1),[1] "seat 19" refers to the person occupying seat 19.

(1) *Ask **seat 19** whether he wants to swap*

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

In order to recognise and interpret the metonymy in (1), a large amount of knowledge and contextual inference is needed (e.g. seats cannot be questioned, people occupy seats, people can be questioned). This problem is

*Malvina Nissim is supported by Scottish Enterprise Stanford-Link Grants R36766 (Paraphrase Generation) and R36759 (SEER).

[1] (1) was actually uttered by a flight attendant on a plane.

exacerbated by the potential open-endedness of metonymic readings (Nunberg, 1978). Thus, most approaches to metonymy resolution rely on manually created knowledge bases or lexica. Moreover, there are few annotation schemes and corpora annotated for metonymies to allow large-scale testing.[2]

However, in contrast to (1), many metonymies are actually quite regular (Lakoff and Johnson, 1980; Nunberg, 1995). In (2), for instance, an organisation name refers to a product the organisation manufactures.[3]

(2) *And he hasn't bought me a* ***Renault*** *[...]*

Such reference shifts occur systematically with a wide variety of organisation names. Thus, linguistic studies (Lakoff and Johnson, 1980; Fass, 1997) postulated conventionalised *metonymic patterns* (e.g., `org-for-product`) that operate on semantic *base classes* (here, ORGANISATION).[4] In another such regular shift, an organisation name refers to its representatives, as in (3).[5]

(3) *Last February* ***NASA*** *announced [...]*

Exploiting such regularities, in previous work we proposed a corpus-based paradigm for treating metonymies. We developed a framework for reliably annotating metonymies in domain-independent text (Markert and Nissim, 2002b). We also suggested viewing metonymy resolution as a *classification task* (Markert and Nissim, 2002a), distinguishing between literal readings and a prespecified set of metonymic patterns for a particular base class.

We tested the annotation framework and experimented with a supervised machine learning approach to the classification task in a case study for location names.[6] For this base class, we showed that our annotation categories have high coverage, lead to reliable annotation and that treating metonymy resolution as a classification task yields promising results. However, the generalisability of this approach to other semantic classes was left open. In the current paper, we test and refine the paradigm for the semantic class ORGANISATION. In Section 2, we describe a reliable annotation scheme

[2]For a discussion of related work see Section 4.

[3]All examples except (1) are from the British National Corpus (BNC, (Burnard, 1995)).

[4]Not all instances of a semantic class can necessarily undergo given metonymic patterns. For example, "pig" cannot undergo the `animal-for-meat` pattern due to lexical blocking (Briscoe et al., 1995). We do not handle such exceptions in this paper.

[5]Conventional metonymy is also known as *regular polysemy* (Copestake and Briscoe, 1995). We use "metonymy" to encompass both conventional and unconventional readings.

[6]An example pattern for locations is `place-for-people`, where a place stands for any persons/organisations associated with it, as "America" in "***America*** *did once try to ban alcohol*". For our experiments for the class LOCATION see (Markert and Nissim, 2002b).

for organisation names and a corpus annotated for metonymic usage. In Section 3, discuss a supervised classification algorithm for metonymy resolution for organisation names, focusing on the features useful for treating this base class. Overall, the results confirm the reliability of our annotation framework and the feasibility of supervised learning for handling conventional metonymy. In Section 4, we discuss the advantages and limitations of our work in comparison to other approaches to metonymy resolution.

2 Annotating Metonymies

Our general framework (Markert and Nissim, 2002b) distinguishes between `literal`, `metonymic` and `mixed` readings for each base class.[7]

2.1 Annotation scheme for organisation names

The `literal` reading refers to the organisation as a legal entity which has a charter or defined aims. This includes descriptions of the organisation's structure (see (4)), and relations between organisations and products/services they offer (see (5)).

(4) ***NATO*** *members [. . .]*

(5) ***Intel****'s Indeo video compression hardware [. . .]*

Metonymic readings cover metonymies that follow regular patterns as well as metonymies that do not. Most patterns are *base class-specific*. We distinguish the *organisation-specific* patterns in Table 1 (see (2-3) and (6-9)).

Table 1: Organisation-specific metonymic patterns

`org-for`	<org> stands for	Ex.
`members`	an official who acts for <org>, or all members of <org>	(3),(6)
`product`	the product(s) an <org> produces	(2)
`facility`	the facility that houses <org> or one of its branches	(7)
`index`	an index, like a stock index, indicating the value of <org>	(8)
`event`	an event associated with <org>, e.g., a scandal	(9)

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

[7] We follow the MUC Named Entity guidelines (Chinchor, 1997) for assigning base classes. For example, hospitals are facilities, whereas companies are organisations.

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

(8) ***Eurotunnel** was the most active stock*

(9) *[...] the resignation of Leon Brittan from Trade and Industry in the aftermath of **Westland***[8]

Besides class-specific patterns, two *class-independent* metonymic patterns can be applied to most nouns. In `object-for-name` shifts, a word is used as a mere signifier rather than referentially, as *Chevrolet* and *Ford* in (10).

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

In `object-for-representation` metonymies, a name refers to a representation (such as a photo or painting) of the standard referent. We regard the logo of an organisation as its representation, as in (11).

(11) *BT's pipes-of-Pan motif was, for him, somehow too British. Graphically, it lacked what King calls the "world class" of **IBM**, **Apple Computer**, **Ford**, **Sony**, and **Shell**.*

The category `other` covers unconventional metonymies (see (1)). Since they are open-ended and context-dependent, no specific category indicating the intended class can be introduced.

In some examples, two predicates trigger a different reading each, thus yielding a *mixed* reading (cf. Nunberg's (1995) account of *co-predication*). This often occurs with *coordinations*, *appositions* and *gerunds*. (12) evokes an `org-for-index` as well as an `org-for-members` reading (triggered by "slipped" and "confirming", respectively).

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

2.2 Annotation Experiment

We now describe an annotation exercise for the class ORGANISATION, restricting ourselves to names of companies.

Data. Our sampling frame, CompList, consists of the Fortune500 company names (`http://www.fortune.com/fortune/fortune500`), plus alternative spellings, acronyms, and abbreviations, for a total of 528 different company names. We randomly extracted 3100 instances of these names from

[8] The British helicopter company Westland was involved in a 1980s economic scandal.

the BNC, allowing any name in CompList to occur. All samples include three sentences of context. The extracted dataset contained many instances of common noun homographs, such as the celestial body "sun", which were ignored during annotation.

Method and Results. The annotation followed written guidelines, whose main features are *replacement tests* (e.g., if an occurrence of "BP" can be replaced by "shares of BP", we annotate it as `org-for-index`), examples for each category and instructions for ambiguous cases.

We measured reproducibility of the distinction between the categories `literal`, `org-for-members`, `org-for-product`, `org-for-facility`, `org-for-index`, `object-for-name`, `object-for-representation`, `other`, and `mixed`, using the kappa statistic (K) (Carletta, 1996). Good quality annotation of discourse phenomena normally yields a K of about .80.

The annotators are the authors of this paper and were trained by independently annotating 400 samples, which included 125 instances of company names. The annotation of the training set yielded $K = .804$ ($N = 125, k = 2$) (N is the number of examples annotated, k is the number of annotators). We used the remaining 2700 samples as an annotation test set, containing 984 instances that both annotators marked as companies. K was measured at .894 ($N = 984, k = 2$), showing that the annotation is highly reliable.

Gold Standard Corpus. After the annotation, we discussed all samples in the test set and created a gold standard. In addition to the 984 samples which both annotators had marked as companies, we included three cases where one of the annotators had originally not understood the context (and had therefore not recognised them as companies) but could understand the samples after the joint discussion. This yielded a total of 987 instances of annotated company names. In 20 cases we could not agree on the reading even after discussion. In the remainder of this paper these are not considered, yielding a gold standard corpus of 967 annotated instances. The distribution of readings is given in Table 2.

Comparison to the LOCATION *class.* We compare these results to our previous annotation exercise on 931 location names from the BNC (Markert and Nissim, 2002b). First, we achieved a similar degree of reliability ($K = .870$ for locations, $K = .894$ for organisations), showing the validity of our approach and suggesting extensibility to yet other base classes. Second, literal readings are the majority both in the location and organisation corpora (79.7% and 64.3%, respectively). Third, organisation names are used metonymically more frequently than location names. Fourth, whereas the `place-for-people` pattern accounts for 93.1% of location metonymies, there is a wider variety of patterns for organisation names. This, together

Table 2: Distribution of readings in the gold standard corpus (967 instances of company names). All percentages are rounded to the first decimal.

READING	FREQ	% of all	% of met
`literal`	622	64.3	n/a
`mixed`	50	5.2	n/a
`metonymies`	295	30.5	100
`org-for-members`	188	19.4	63.7
`org-for-product`	66	6.8	22.4
`org-for-facility`	14	1.4	4.7
`org-for-index`	6	0.6	2.0
`org-for-event`	1	0.1	0.3
`object-for-name`	6	0.6	2.0
`object-for-representation`	1	0.1	0.3
`other`	13	1.3	4.4

with the extreme rarity of some metonymic patterns, should make supervised learning for organisations names more difficult than for location names. Fifth, unconventional metonymies (category `other`) are very rare for both locations and organisations (1.0% and 1.3% of all readings, respectively).

3 Supervised Learning for Metonymy Resolution

The rarity of unconventional metonymies strengthens the case for viewing metonymy resolution as a *classification task* between the literal reading and a fixed set of metonymic patterns that can be identified in advance for particular semantic classes (Markert and Nissim, 2002a). This approach makes the task comparable to classic *word sense disambiguation* (WSD). However, whereas a classic (supervised) WSD algorithm is trained on a set of labelled instances of *one particular word* and assigns word senses to new test instances of the *same word*, (supervised) metonymy resolution can be trained on a set of labelled instances of *different words of one semantic class* and assign literal and metonymic readings to new test instances of *possibly different words of the same semantic class*.

Thus, a supervised metonymy classification algorithm needs to infer from (13a) (when labelled as `org-for-members`) that (13b) and (13c) are also `org-for-members` metonymies. Similarly, it needs to infer from (14a), labelled as `org-for-product`, that (14b) is an instance of the same pattern.

(13) a. *Last February **NASA** announced [...]*

b. ***IBM Corp.*** *was able to announce that [...]*

c. *While* ***Olivetti*** *did say it was a joint decision [...]*

(14) a. *press-men hoisted their notebooks and their* ***Kodaks***

b. *little old ladies in small* ***Renaults*** *[...]*

In order to perform such inferences, the algorithm needs to (i) generalise over organisation names and (ii) capture the morphosyntactic and lexical similarities between training and test instances. The first requirement is fulfilled by using the class-based training/testing mentioned above, while the second depends on appropriate feature selection.

Features and Model The *target readings* for the algorithm are the annotation categories. The *feature selection* is based on linguistic properties of organisation metonymies as well as previous experiments for location names that, for example, showed the limited usefulness of co-occurrences for metonymy recognition (Markert and Nissim, 2002a). We used the 6 features in Table 3 to describe each example in the corpus. Table 4 shows example feature vectors.

Table 3: Feature Set

Feat.	Description	Values
f1	grammatical role of <org>	subj, obj, ...
f2	lemmatised head/modifier of <org>	announce, shiny, ...
f3	determiner of <org>	def, indef, bare, demonst, other
f4	grammatical number of <org>	sing, plural
f5	# grammatical roles of <org>	1, more than 1
f6	# words in <org>	1,2,3, ...

Features f1-f4 were annotated manually.[9] f1 is annotated within a dependency grammar framework and includes both head and modifier relations to content words, whose lemmatised form is annotated via f2. Feature f5 indicates whether more than one role is annotated for a training instance. When an example has n grammatical roles, it is represented by n feature vectors, yielding n training/testing instances. This is frequent for mixed readings but can also occur with other target categories (see (14b)). Each

[9] Manual annotation allows us to measure the contribution of different features without encountering error chains from parsing or tagging processes. Obviously, automatic feature extraction will lower the classifier's performance (see also Nissim and Markert (2003)).

of these feature vectors is assigned the same feature values for f3-6 and the same target reading (see (14b) in Table 4). Features f3, f4, and f6 capture morphological and surface attributes.

Table 4: Example feature vectors

Example	f1	f2	f3	f4	f5	f6
Ex (13a)	subj	announce	bare	sing	1	1
Ex (13b)	subj	announce	bare	sing	1	2
Ex (13c)	subj	say	bare	sing	1	1
Ex (14a)	obj	hoist	other	plural	1	1
Ex (14b)	prep	in	bare	plural	2	1
	hasadj	small	bare	plural	2	1

Given a training and a test set, our algorithm proceeds as follows. First, we remove all examples annotated as `mixed` from the training set. Mixed readings are composed of two different readings and we have found that they do not give clear evidence during training in our representation framework. Examples with mixed readings remain in the test set. Second, we represent each example as one or more feature vectors (see above), and use a machine learning classifier to predict the reading of each instance in the test set. Third, we postprocess the classifier output to assign a single target reading to examples represented by more than one instance. If two instances representing the same example are assigned conflicting readings by the classifier, we assign a `mixed` reading. This mirrors the involvement of different predicates triggering different readings. If the classifier assigns the same reading to all instances representing an example, no further decision is needed.

Results We used the Naive Bayes learner of the Weka machine learning library (Witten and Frank, 2000). All results are obtained using 10-fold cross-validation on our corpus. All significance claims are based on a t-test, using significance level 0.05. We evaluated the overall accuracy (*acc*) of an algorithm as the percentage of all examples that were assigned the correct reading. In order to see which target readings our algorithms can handle best, we also compute precision (P), recall (R) and F-measure (F) for each target category. The baseline base uses the assignment of the most frequent reading `literal` to all corpus instances and has an accuracy of 64.3%.

The results of several algorithm variations are summarised in Table 5. Variation all uses all features described in Table 3. The variations −f1 to −f6 are leave-one-out classifiers using all but the specified feature. The variations f1 to f6 are single-feature classifiers.

Table 5: Results for all algorithm variations..

f	*acc*	literal *P*	literal *R*	literal *F*	members *P*	members *R*	members *F*	product *P*	product *R*	product *F*	mixed *P*	mixed *R*	mixed *F*
base	**.643**	.643	1.00	.783	na	0	na	na	0	na	na	0	na
all	**.760**	.794	.903	.845	.670	.691	.681	.853	.439	.580	.467	.280	.350
−f1	.731	.717	.984	.830	.833	.319	.462	.879	.440	.586	.750	.120	.207
−f2	.726	.794	.854	.823	.555	.702	.620	.811	.456	.583	.391	.180	.247
−f3	.734	.783	.883	.830	.667	.691	.679	.467	.212	.292	.415	.340	.374
−f4	.756	.792	.902	.844	.667	.691	.679	.812	.394	.531	.452	.280	.346
−f5	.759	.792	.907	.846	.663	.691	.677	.906	.439	.592	.423	.220	.289
−f6	.766	.799	.908	.850	.687	.734	.709	.862	.379	.526	.433	.260	.325
f1	.703	.787	.833	.809	.557	.782	.650	na	0	na	.333	.300	.316
f2	.689	.678	.997	.807	.927	.202	.332	na	0	na	.727	.160	.262
f3	.663	.658	.998	.793	na	0	na	.869	.303	.449	na	0	na
f4	.650	.648	1.00	.787	na	0	na	1.00	.106	.192	na	0	na
f5	.643	.643	1.00	.783	na	0	na	na	0	na	na	0	na
f6	.643	.643	1.00	.783	na	0	na	na	0	na	na	0	na

The classifier all, all leave-one-out variations, and the single-feature classifiers that use grammatical roles (f1) or lemmatised heads/modifiers (f2) significantly beat the baseline in accuracy. Overall, grammatical and lexical information about head-modifier relations is crucial for distinguishing between `literal` and `org-for-members` (and `mixed`) readings: for example, a subject role or a head such as "announce" often indicate `org-for-members` metonymies. These features were also important for the distinction between `literal` and `place-for-people` readings for location names (Markert and Nissim, 2002a). As for location names, the lemmata of heads/modifiers (f2) yield a high precision feature but many of them are encountered only once in the corpus, leading to data sparseness. Therefore, f2 can make the inference from (13a) to (13b) but not to (13c), lowering recall. This can be partially remedied by generalising over lemmata to semantic classes (see for instance the similarity of "say" and "announce"). We have explored the integration of a similarity-based thesaurus for this purpose for location names, yielding good results (Nissim and Markert, 2003).

Morphological features are mostly important for `org-for-product` metonymies. For example, leaving out the determiner feature (−f3) yields a significant drop in their classification. A plural organisation name, although rare, also clearly points to an `org-for-product` pattern, thus explaining that the single classifier f4 can recognise some metonymies of this pattern.

We report P, R, and F for `literal`, `org-for-members`, `org-for-product` and `mixed` only. Because all other categories are rare and/or our features might not fully capture the regularity of the pattern, they were never assigned by the algorithms.

4 Related work

Most previous work on metonymy resolution relies on hand-encoded domain knowledge bases (Hobbs et al., 1993; Fass, 1997; Horacek, 2002; Markert and Hahn, 2002, among others) or lexica (Pustejovsky, 1995; Copestake and Briscoe, 1995; Verspoor, 1996; Lascarides and Copestake, 1998). These resources are expensive to build and the resulting methods are not easily scalable or portable to new domains. Building on the results of our corpus studies, we instead show that metonymy resolution can be reformulated as a classification task. This allows us to develop machine learning algorithms that do not require a previously constructed knowledge base. Extending our methods from location to organisation names was relatively fast within the previously developed framework. Our probabilistic algorithm can integrate several feature types and incorporate inconclusive feature evidence. For organisation names lexical as well as morphological features needed to be incorporated to achieve optimal performance.

Some previous approaches (Hobbs et al., 1993; Pustejovsky, 1995; Fass, 1997) evaluate their resolution algorithms in comparison to constructed examples only, disregarding the range of phenomena in realistic settings. Others (Verspoor, 1997; Markert and Hahn, 2002; Harabagiu, 1998) use naturally-occurring data that, however, seem to be analysed according to subjective intuitions of one individual only. The latter is also true for corpus linguistic studies such as (Marinelli, 2004). In contrast, we use a reliably annotated corpus for evaluation, moving the field towards more objective evaluation procedures. As far as we know, the ACE project (`http://www.itl.nist.gov/iad/894.01/tests/ace/`) developed the only other annotation scheme for metonymies in real occurring texts. However, to our knowledge, no agreement data has been published. In comparison to our scheme, theirs includes a larger number of base classes including both proper and common nouns. On the other hand, for each of these classes, they only consider a limited number of metonymic patterns. For example, for the class ORGANISATION, they only annotate `org-for-facility` metonymies.

The main drawbacks of our method are the lack of coverage of unconventional metonymies that do not lend themselves easily to supervised learning, and data sparseness for rare metonymic patterns and feature values. The latter problem can be partially overcome by generalisation over feature values,as we have shown for location names (Nissim and Markert, 2003). In addition, many knowledge-based techniques (Markert and Hahn, 2002; Hobbs et al., 1993; Fass, 1997) deliver a full interpretation of the metonymic noun. A knowledge base that includes which companies produce which goods allows to identify the referent of "Renault" in (2) as a car. In contrast, our approach classifies "Renault' as a product by Renault only. However, our metonymy recognition and partial interpretation algorithm provides a hyperonym class of the referent useful in applications like anaphora resolution and can be expanded to a full interpretation via subsequent processing. Thus, for example, lexical association measures could provide a list of products that most frequently co-occur with Renault in a corpus.

5 Conclusions and Future Work

We presented a reliable annotation scheme for organisation names that complements a previous scheme for location names and showed the generalisability of the underlying annotation principles (Markert and Nissim, 2002b). A Naive Bayes classifier for metonymy resolution on a corpus annotated with this scheme performed significantly better than a most frequent sense baseline. We discussed the contribution of morphological, grammatical, and lexical features, as well as the importance of feature integration for handling a variety of metonymic patterns. In future work, we will address the data sparseness problem for rare feature values and extend our approach to full interpretation of metonymic referents.

References

T. Briscoe, A. Copestake, and A. Lascarides. Blocking. In P. Saint-Dizier and E. Viegas, eds, *Computational Lexical Semantics*. CUP, 1995.

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

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

N. Chinchor. MUC-7 NE Task definition. In *Proc. of MUC-7*, 1997.

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

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

S. Harabagiu. Deriving metonymic coercions from WordNet. In *Workshop on the Usage of WordNet in NLP Systems, COLING-ACL-98*, 1998.

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

H. Horacek. Varying cardinality in metonymic extensions to nouns. In *COLING 2002*, pp. 356–362, 2002.

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

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

A. Lascarides and A. Copestake. Pragmatics and word meaning. *Journal of Linguistics*, 34(2):387–414, 1998.

R. Marinelli. Proper names and polysemy: from a lexicographic experience. In *Proc. of LREC 2004*, pp. 157–160, 2004.

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

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

K. Markert and M. Nissim. Towards a corpus annotated for metonymies: the case of location names. In *Proc. of LREC 2002*, pp. 1385–1392, 2002b.

M. Nissim and K. Markert. Syntactic features and word similarity for supervised metonymy resolution. In *Proc. of ACL-03*, pp. 56–63, 2003.

G. Nunberg. *The Pragmatics of Reference.* PhD thesis, City University of New York, New York, 1978.

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

J. Pustejovsky. *The Generative Lexicon.* MIT Press, Cambridge, MA, 1995.

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

C. Verspoor. Lexical limits on the influence of context. In *Proc. of CogSci 1996*, pp. 116–120, 1996.

C. Verspoor. Conventionality-governed logical metonymy. In H. Bunt et al., editors, *Proc. of IWCS-2*, pp. 300–312, 1997.

I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, 2000.