Metonymy Resolution as a Classification Task

Katja Markert and Malvina Nissim
Division of Informatics
University of Edinburgh
2 Buccleuch Place EH8 9LW
Edinburgh, Scotland, UK
{markert|malvi}@cogsci.ed.ac.uk

Abstract

We reformulate metonymy resolution as a classification task. This is motivated by the regularity of metonymic readings and makes general classification and word sense disambiguation methods available for metonymy resolution. We then present a case study for location names, presenting both a corpus of location names annotated for metonymy as well as experiments with a supervised classification algorithm on this corpus. We especially explore the contribution of features used in word sense disambiguation to metonymy resolution.

1 Introduction

Metonymy is a figure of speech, in which one expression A is used to refer to the standard referent of a related one B (Lakoff and Johnson, 1980). In Example (1),

(1) *"The **ham sandwich** is waiting for his check."*

expression A *"ham sandwich"* refers to the customer who ordered the sandwich (B) (Nunberg, 1978).

The importance of resolving metonymies has been shown for a variety of natural language processing tasks, e.g., machine translation (Kamei and Wakao, 1992), question answering (Stallard, 1993) and anaphora resolution (Harabagiu, 1998; Markert and Hahn, 2002).

Most approaches to automatic metonymy resolution are structured in two phases: metonymy *recognition* (distinguishing *literal* usage from *metonymic* usage) and then metonymy *intepretation* (identifying the intended referent (B)).

Thus, metonymy recognition can be seen as a *classification task*, making it comparable to classic *word sense disambiguation* (WSD), which is also concerned with distinguishing between possible word senses/interpretations. However, whereas standard WSD deals with a *fixed* set of possible senses among which to disambiguate, metonymy resolution must face a potentially open-ended set of possible metonymic readings (Nunberg, 1978).

Nevertheless, it has long been recognised that many metonymic readings are actually quite regular (Lakoff and Johnson, 1980). In Example (2), *"Vietnam"*, the name of a location, refers to an event (a war) that happened there.

(2) *"The broadcast covered **Vietnam**."*

Similar examples can be regularly found for many other location names. Therefore, names belonging to the *semantic class* 'location' can undergo the *metonymic pattern* `place-for-event`. Other semantic classes also have metonymic patterns applicable to them, which are in general much more frequent than unconventional metonymies (Verspoor, 1997).

This allows metonymy recognition to be treated as a disambiguation task between literal readings and a fixed set of metonymic patterns for a particular semantic class. Thus, whereas a classic (supervised) WSD classification algorithm takes a set of labelled training instances

Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 204-213. Association for Computational Linguistics.

of *one particular word* as input and assigns word senses to new test instances of the *same word* as output, (supervised) metonymy recognition can take a set of labelled training instances of *different words belonging to one semantic class* as input and assign literal readings and possible metonymic patterns to new test instances of *possibly different words of the same semantic class*. Thus, it needs to infer from training instances like Example (3) (when labelled as a `place-for-people` metonymy) that Examples (4) and (5) are also metonymic, a task which poses no problems for most humans.

(3) *"**Bosnia**'s view of"*

(4) *"**Hungary**'s view of"*

(5) *"**Hungary**'s position on"*

In this paper, we explore this view of metonymy recognition as a class-based WSD task for the semantic class of locations.[1]

The corpus data we use is described in Section 2. As resources reliably annotated for metonymy do not exist (see also Section 6), we constructed a corpus of location names annotated for literal and metonymic readings.

The supervised classification algorithm we use are decision lists as they have been successfully used in several classification tasks (Yarowsky, 1995; Collins and Singer, 1999) (see Section 3). In Section 4, we explore whether features traditionally used in WSD carry over to metonymy resolution, concentrating on (i) *cooccurrences*; (ii) *collocations*; and (iii) *grammatical features*. Results are discussed in Section 5. We show that cooccurrences are in general not appropriate for metonymy resolution; collocations are useful but suffer from data sparseness when used as simple word forms; the grammatical relations subject and object perform well but are only applicable to a small part of the data. We then compare our algorithm to metonymy recognition approaches based on selectional restriction violations in Section 6.

[1] At the moment we restrict ourselves to location names.

2 Experimental Data

We present a short overview of the collection of a corpus of location names and its annotation for literal and metonymic readings. A more detailed description can be found in (Markert and Nissim, 2002) and the annotation scheme is downloadable from `http://www.ltg.ed.ac.uk/~malvi/mascara/publications.html`.

2.1 Corpus Collection

We extracted all country names from WordNet (Fellbaum, 1998) and the CIA factbook (`http://www.cia.gov/cia/publications/factbook/`). This collection of names forms our sampling frame CountryList.

We built a corpus of text samples that contains 1000 occurrences of country names, randomly extracted from the British National Corpus (`http://info.ox.ac.uk/bnc`), henceforth abbreviated as BNC. Any country name in CountryList was a *Possibly Metonymic Word* (PMW, henceforth) and was allowed to occur in the samples extracted. We searched the BNC using Gsearch (Corley et al., 2001). All samples include a PMW surrounded by three sentences of context. All examples introduced from now on are from the BNC.[2]

2.2 Annotation Scheme for Location Names

After excluding some undesired examples (i.e., *noise*) which our extraction method collected (e.g. homonyms, such as "*Greenland*" in *"Professor **Greenland**"*), the annotation can proceed to identify *literal*, *metonymic*, and *mixed* readings. Our annotation scheme built on lists of metonymic patterns in the literature (Lakoff and Johnson, 1980; Fass, 1997; Stern, 1931), but diverted from these patterns when they did not provide full coverage or could not be distinguished reliably (Markert and Nissim, 2002).

The `literal` reading for location names comprises a locative (see Example (6)) and a political entity interpretation (see Example (7)).

(6) *"coral coast of **Papua New Guinea**"*

[2] An exception is Example (12).

(7) *"**Britain**'s current account deficit"*

For metonymic readings, we distinguish between *general* patterns (valid for all physical objects) and *location-specific* ones. As general patterns were never encountered in our corpus, we describe here only the latter.

- `place-for-people`: a place stands for any persons/organisations associated with it. In Example (8), "***San Marino***" stands for one of its sports teams.

  (8) *"a 29th-minute own goal from **San Marino** defender"*

  Often, the explicit referent is underspecified, as in Example (9), where the reference could be to the government, an organisation or the whole population.

  (9) *"The ... group expressed readiness to provide **Albania** with food aid"*

  We therefore adopt a *hierarchical* approach, and assign a pattern (`place-for-people`) at a higher level (*supertype*), as well as a more specific pattern (*subtype*), if identifiable, at a lower level. This deviates from common practice in the linguistic literature, but has the great advantage of 'punishing' disagreement only at a later stage and allowing fall-back options for automatic systems. We also experienced a drop in human annotation agreement from supertype to subtype classifications (see (Markert and Nissim, 2002)). In this paper, we evaluate our system on supertype classification.

- `place-for-event`: a location name stands for something that happened there (see Example (2)).

- `place-for-product`: a place stands for a product manufactured there (e.g., "*Bordeaux*" can refer to the wine produced there).

The category `othermet` covers unconventional metonymies. Since they are open-ended and context-dependent, no specific category indicating the intended semantic class can be introduced. In Example (10), *"New Jersey"* metonymically refers to the local typical tunes.

(10) *"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"*

The category `othermet` is only used if none of the other categories fits.

In addition to literal and metonymic readings, we found examples where two predicates are involved, triggering a different reading each, thus yielding a *mixed* reading. This often occurs with *coordinations* and *appositions*.

(11) *"they arrived in **Nigeria**, hitherto a leading critic of ... "*

In Example (11), both a literal (triggered by "arriving in") and a `place-for-people` reading (triggered by "leading critic") are invoked. We therefore introduced the category `mixed` to deal with these cases (not treated as a category in the literature).

2.3 Annotation Reliability, Distribution and Data Preparation

The 1000 examples of our corpus have been independently annotated by two computational linguists, who are the authors of this paper. Reproducibility of results (Krippendorff, 1980) yielded a percentage agreement of .95 and a *kappa* (Carletta, 1996) of .88. The annotation can therefore be considered reliable. In the corpus data used for our classification experiments, we only included the samples which both annotators could agree on and which were not marked as noise. Therefore our corpus for testing and training the algorithm includes 925 samples. The resulting distribution of readings is described in Table 1.

The data was further stripped of all punctuation and capitalisation was removed. No stemming or lemmatisation was performed.

Table 1: Distribution of readings in our corpus

reading	N	%
`literal`	737	79.7
`place-for-people`	161	17.4
`place-for-event`	3	.3
`place-for-product`	0	.0
`mixed`	15	1.6
`othermet`	9	1.0
total	925	100.0

3 Decision lists for metonymy resolution

The distribution in the corpus shows that metonymic readings that do not follow established metonymic patterns (`othermet`) are very rare. This seems to be the case for other kinds of metonymies, too (Verspoor, 1997). This strengthens our case for viewing metonymy recognition as a classification task between literal readings and metonymic patterns that can be identified in advance for particular semantic classes. We therefore explore the usage of a classification algorithm and features used in WSD for metonymy recognition. The target readings for the algorithm to distinguish are `literal`, `place-for-people`, `place-for-event`, `place-for-product`, `othermet` and `mixed`.

As an algorithm we use decision lists.[3] The advantage of decision lists for a first exploration of a feature space is that their choices are easy to follow as they make use of the most informative feature only instead of a combination of features. All features encountered in the training data are ranked in the decision list (best evidence first) according to a log-likelihood ratio calculated as follows (Yarowsky, 1995; Martinez and Agirre, 2000):

$$Log\left(\frac{Pr(reading_i|feature_k)}{\sum_{j\neq i} Pr(reading_j|feature_k)}\right)$$

When applying the decision list to a test example, the winning reading is selected by the feature in the test example with the highest rank in the decision list.

We estimated probabilities via maximum likelihood, adopting a simple smoothing method: 0.1 is added to both the denominator and numerator.

[3] All experiments reported here have also been repeated using a Naive Bayes classifier. The results have not improved on decision lists.

4 Exploration of feature space

We investigated the following feature types. Examples are given in Table 2, together with examples of their distribution and the reading they trigger.

Cooccurrences. They have proved useful for WSD (Gale et al., 1993; Pedersen, 2000). We used left and right windows of context of 8 different sizes: 0, 1, 2, 3, 4, 5, 10, and 25 words, thus yielding 64 possible combinations of left and right sizes (e.g., l3r1 for 3 words to the left and 1 to the right). Any content word in the window considered was included as a feature.

Collocations. We selected 4 different collocations frequently used in WSD (Ng and Lee, 1996; Martinez and Agirre, 2000). The word to the right of the PMW, the word to the left, two words to the left and the word to the right and the left. The first two features consist of a single word form, the latter two of a sequence of two word forms. Function words were allowed as collocations, as e.g., the presence of a preposition directly to the left of the PMW can be indicative (see also Table 2).

Grammatical features. Following some WSD approaches (Ng and Lee, 1996; Yarowsky, 1995) we use grammatical features, namely,

Table 2: Examples of features and their decision list score

feature	example	assigned reading$_i$	Log	reading$_i$	readings$_{j \neq i}$
cooccurrences					
content-word-in-window-l4.r4	states +/- 4 <country>	`literal`	4.709	11	0
content-word-in-window-l4.r4	win +/- 4 <country>	`place-for-people`	3.434	3	0
collocations					
word-to-left	in <country>	`literal`	7.314	150	0
word-to-right	<country> seemed	`place-for-people`	3.434	3	0
two-words-to-left	one of <country>	`literal`	4.263	7	0
word-to-left-and-right	provide <country> with	`place-for-people`	3.044	2	0
grammar					
role	<country> = subj	`place-for-people`	0.863	57	24
role-of-verb	<country> = subj-of-have	`literal`	3.714	4	0

(i) the grammatical role (role) of the PMW, distinguishing between subjects, direct objects and any other grammatical role (including e.g. prepositional phrases, NP modifiers); (ii) both the grammatical role and the stemmed form of the corresponding verb for subjects and direct objects (role-of-verb).

5 Results and Discussion

We have tested the decision list algorithm on our annotated corpus, employing 10-fold cross-validation. Results as reported in Table 3 are averaged over all 10 folds. The first column describes the feature used in the experiment. Then we report accuracy and coverage.[4]

$$\text{coverage} = \frac{\text{number of decisions made}}{\text{number of test data}}$$

$$\text{accuracy} = \frac{\text{number of correct decisions made}}{\text{number of decisions made}}$$

We also used a backing-off strategy to the most frequent sense `literal` for the cases where no decision could be made (increasing coverage to 1) and report these results as *accuracy/coverage-backoff*. As it is of particular interest to us to see how many non-literal readings (metonymies and mixed readings) can be correctly identified we compute precision and recall (based on the algorithm including backing off strategy). Let A be the number of correctly identified non-literal readings and B the number of incorrectly identified non-literal readings.

$$\text{precision} = \frac{A}{A+B}$$

$$\text{recall} = \frac{A}{\#\text{non-literal examples in the test data}}$$

When significance claims are made they are based on a 10-fold cross-validated t-test, using significance level 0.05.

The baseline used for comparison is the assignment of the most frequent reading `literal` (see Table 1). It has a coverage of 1 as it is applicable to all examples. Recall is 0 as no metonymies can be recognised.

5.1 Cooccurrences

For all 64 window size combinations (for example results see Table 3), the accuracy never significantly beats the baseline. Both precision and recall are unsatisfactory and get steadily worse with increasing window sizes. We identified the following reasons for such a behaviour.

Topical v. fine-grained sense distinctions. Cooccurrences and large window sizes traditionally work well for topical distinctions (Gale et al., 1993). Metonymy, though, does often not cross topical boundaries — thus, whether a location name is used as a literal (political) reading or as a reading for the government often does not change coocurrence features. This is especially true for large window sizes.

[4] Please note that a test example might not be covered because of either the absence of a feature value in the training set or because the highest ranked feature gives equal evidence for two different readings.

Table 3: Results for feature types

feature	accuracy/coverage		accuracy/coverage-backoff		precision/recall	
	acc	cov	acc-backoff	cov-backoff	prec	rec
baseline	.797	1.00	.797	1.00	n/a	0.00
cooccurrences						
content-word-in-window-l2.r1	.770	.443	.780	1.00	.510	.204
content-word-in-window-l3.r1	.783	.588	.806	1.00	.554	.249
content-word-in-window-l4.r1	.790	.686	.803	1.00	.538	.226
content-word-in-window-l4.r4	.794	.959	.790	1.00	.458	.180
content-word-in-window-l10.r10	.779	1.00	.779	1.00	.250	.043
collocations						
word-to-left	.843	.780	.809	1.00	.677	.112
word-to-right	.831	.740	.810	1.00	.650	.139
two-words-to-left	.858	.297	.801	1.00	.625	.053
word-to-right-and-left	.870	.426	.795	1.00	.471	.087
all-collocations	.819	.944	.810	1.00	.607	.185
grammar						
role	.837	.995	.837	1.00	.703	.344
role-of-verb+role	.843	.999	.843	1.00	.750	.339

Pruning and decision list ordering. Every content word encountered in the training set is included in the decision list, even if it occurred infrequently. The simple smoothing method we used did not fully take this problem into account. Therefore, for example, a content word w_1 occurring only once and with a metonymic reading can be ranked higher than a content word w_2 occurring 10 times, 8 times with a literal reading and twice with a metonymic reading. A test example containing both content words will therefore use w_1 to decide in favour of a metonymic reading, despite the weak evidence. This might explain the low precision. We therefore tested the effect of deleting all *non-informative* features from the decision list, using the G^2 test (Dunning, 1993) to measure *independence* between cooccurrence features and readings. Using pruned decision lists yielded some improvement in precision, but a significant drop in coverage, given the lower number of features used (for window l4.r1: precision=.609; recall=.210; coverage=.098). The general tendency to prefer smaller windows over larger ones still holds.

5.2 Collocations

The one-word collocations had in general a high coverage as function words were included. Accuracy for collocations is quite good (ranking from 81.9% to 87.0%). But increasing coverage to 1.00 (*coverage-backoff*) causes *accuracy-backoff* to drop. Recall is very low. We discuss here two reasons why collocations do worse in metonymy recognition than in WSD (Yarowsky, 1995; Ng and Lee, 1996; Pedersen, 2001).

Target readings. Readings like `othermet` and `mixed` are unsuited for a collocation-based approach.

Sparse data. When we inspected the decision lists, we found that strong collocations are mostly found for literal readings (e.g. spatial prepositions to the left of the PMW), so that a high percentage of literal examples can be identified correctly. Some good collocations for metonymic readings were found only once or twice in the training data and then not again in the test data, thus causing low recall and accuracy-backoff. One reason for this is that the training data for literal readings is about 5

times as big as for metonymic readings. This is aggravated by the use of the BNC that includes a wide variety of genres using different style, register and vocabulary.[5] Often, though, a "similar" collocation was seen (compare e.g., *"view"* and *"position"* in Example (4) and (5)). Using word forms as collocations can only make the generalisation from Example (3) to Example (4), not the one to Example (5). Thus, we will in the future explore semantic generalisation of collocations by e.g., using synonym information from Wordnet.

5.3 Grammatical roles

Grammatical roles yield significant improvements in *accuracy-backoff* over the baseline and good precision. One reason is that they do not suffer as much from sparse data and generalise well over the whole semantic class.

Regarding the classifier based on the feature role only, e.g., being a subject can be learned as a good indicator for `place-for-people` metonymies regardless of country name or verb.[6] Recall (.344) is also promising considering that the roles of subject and object, which give good hints for metonymic readings, are relatively rare (only 120 of 925 examples in our corpus were subjects or direct objects). The classifier learns to assign literal readings to all other instances, whose grammatical roles are not further distinguished as feature values. Inclusion of more grammatical roles might further improve recall.

Precision can be improved without sacrificing recall by also considering the verb, if present in the training data (classifier role-of-verb+role). So, whereas considering the role only will lead to assigning a `place-for-people` metonymy to all subjects, this is avoided in some cases when considering the verb in addition (e.g., for being the subject of the full verb "have"; see also Table 2). If the grammatical role with this particular verb has not been seen in the training data, the classifier will default to role, thus keeping coverage high.

Please note that the grammatical roles have been annotated by hand as we wanted to measure the contribution of different features to metonymy classification without encountering error chains from e.g., parsing or tagging processes. Therefore the results we present are an upper bound to what can be achieved with subject/object roles.

6 Related Work

We compared our approach and results to WSD in Section 1 and 5, stressing word-to-word vs. class-to-class inference.

Most *traditional approaches to metonymy recognition* use *violations* of selectional restrictions (plus sometimes syntactic violations) for recognition (Pustejovsky, 1995; Hobbs et al., 1993; Fass, 1997; Copestake and Briscoe, 1995; Stallard, 1993).[7] Thus they furnish their algorithms with (mostly hand-modelled) selectional or grammatical restrictions. Note that selectional restrictions in these approaches 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 those experiments in which we also use grammatical knowledge, our system does not have **any** a priori knowledge of semantic argument-verb restrictions. Rather it refers to previously seen training data of country names as verb arguments and their labelled senses and computes the likelihood of each sense using this distribution. This is advantageous for the following reasons:

- There are many verbs with weak selectional restrictions (e.g., the verb "seem"). Both literal (see Example (12)) and metonymic (see Example (13)) readings of a location ocurring as subject of "seem" are therefore possible, although one of the readings might be more frequent given these features.

  (12) *"**Hungary** seemed far away."*

[5] (Martinez and Agirre, 2000) also achieved better results with the use of collocations on the Wall Street Journal than on the balanced Brown Corpus.

[6] Obviously the usefulness of grammatical roles will also depend on the kind of metonymy prevalent in the semantic class.

[7] (Markert and Hahn, 2002) enhance this with anaphoric information.

(13) *"**Britain** seemed close to intervention."*

Selectional restrictions as used in most metonymy recognition approaches therefore do not detect any violation. In contrast, the training data we use supplies the information that the metonymic `place-for-people` reading is more frequent given these grammatical features, leading the classifier to assign the correct reading in the majority of cases.[8]

- Our algorithm does not need to make any assumptions about the sense of the verb. Selectional restrictions, instead, must assume that the verb is disambiguated beforehand as they can vary between different verb senses (compare, e.g., the "confront" reading and the "to be opposite" reading of the verb "face").

To compare our decision list algorithm role-for-verb+role to a selectional restriction violations approach we limited our next empirical study to the 120 examples in our data that had the grammatical role of subjects or direct objects (SETGRAMM).

Two native speakers of English (both linguists) were asked to annotate the 120 subj-verb/obj-verb tuples in SETGRAMM for selectional restriction violations. Agreement between the two subjects was satisfactory (*kappa*=.70). We then simulated two metonymy recognition algorithms based on the annotations of subject1 and subject2, postulating a non-literal reading when a selectional restriction violation was annotated and literal otherwise and computed corresponding evaluation measures.

We also computed the evaluation measures for our role-of-verb+role classifier, limited to SETGRAMM.

Results are summarised in Table 4. Our classifier has higher recall, but lower precision than subject2 and subject1 . To compare the trade-off between precision and recall we computed the F-measure for all algorithms, where our algorithm performed best.

We also evaluate our approach more rigorously than other metonymy resolution algorithms. Some researchers use constructed examples only (Fass, 1997; Hobbs et al., 1993; Copestake and Briscoe, 1995; Pustejovsky, 1995; Verspoor, 1996), and do not report any numerical results. Others (Markert and Hahn, 2002; Harabagiu, 1998; Stallard, 1993) use naturally-occurring data that, however, seem to be analysed according to subjective intuitions of one individual only, not assessing the reliability of their annotation. We, instead, use a reliably annotated corpus that we can make available to other researchers. In addition, most previous evaluations report only recall figures for metonymy recognition, neglecting the question of precision and false positives as well as baseline comparisons and accuracy. Evaluations of metonymy *interpretation* (Lapata, 2001) include more disciplined evaluations, but do not handle metonymy recognition yet.

7 Conclusions

We argued for viewing metonymy recognition as a WSD task based on semantic classes instead of individual words. This is motivated by the regularity of most metonymic readings. We presented a corpus reliably annotated for metonymic and literal usage which supports this claim. We also conducted several experiments with a decision list algorithm to explore the usefulness of common WSD features for metonymy recognition. We showed that coocurrence features are not useful for metonymy resolution, whereas collocation features need to be generalised from word forms to semantic classes to have wide application. Grammatical features perform well. We also compared our grammatical features to a selectional restriction based approach to recognition with promising results.

In the future, we will explore two avenues for improvement: Firstly, we will experiment with more sophisticated machine learning algorithms, starting with improving on our smoothing pro-

[8] (Briscoe and Copestake, 1999) propose using frequency information in addition to syntactic and semantic restrictions, but use only a priori sense frequencies without feature integration.

Table 4: Comparison of human subjects and decision list for grammatical roles

classifier	accuracy	coverage	precision	recall	F-measure
subject1	.625	1.00	.857	.525	.651
subject2	.708	1.00	.846	.687	.758
role-of-verb+role	.706	.992	.750	.830	.788

cedure, which we experienced as too simplistic (see also (Yarowsky, 1997)). Secondly, we will generalise the collocation features we use, incorporate more grammatical relations and explore other feature types and feature combination.

Acknowledgements. Katja Markert is funded by an Emmy Noether Fellowship of the Deutsche Forschungsgemeinschaft (DFG) and Malvina Nissim by ESRC Project R000239444. We thank our colleagues Stephen Clark and Tim O'Donnell for their help with annotation as well as two anonymous reviewers for their comments and suggestions.

References

Ted Briscoe and Ann Copestake. 1999. Lexical rules in constraint-based grammar. *Computational Lingusitics*, 25(4):487–526.

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

M. Collins and Y. Singer. 1999. Unsupervised models for named entity classification. In *Proc. of the 1999 Joint SIGDAT Conference, College Park, MD*, pages 100–110.

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

Steffan Corley, Martin Corley, Frank Keller, Matthew Crocker, and Shari Trewin. 2001. Finding syntactic structure in unparsed corpora: The Gsearch corpus query system. *Computers and the Humanities*, 35(2):81–94.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74.

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

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, Mass.

William Gale, Kenneth Church, and David Yarowsky. 1993. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439.

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

Jerry R. Hobbs, Mark E. Stickel, Douglas E. Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63:69–142.

Shin-ichiro Kamei and Takahiro Wakao. 1992. Metonymy: Reassessment, survey of acceptability and its treatment in machine translation systems. In *Proc. of the 30^{th} Annual Meeting of the Association for Computational Linguistics; Newark, Del., 28 June – 2 July 1992*, pages 309–311.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology.* Sage Publications.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By.* Chicago University Press, Chicago, Ill.

Maria Lapata. 2001. A corpus-based account of regular polysemy: The case of context-sensitive adjectives. In *Proc. of the 2nd Meeting of the North American Chapter of the ACL*, Pittsburgh, PA.

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

Katja Markert and Malvina Nissim. 2002. Towards a corpus annotated for metonymies: the case of location names. In *Proc. of the 3^{rd} International Conference on Language Resources and Evaluation; Las Palmas, Canary Islands, 2002.*

David Martinez and Eneko Agirre. 2000. One sense per collocation and genre/topic variations. In *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and very large corpora.*

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics; Santa Cruz, Cal., 23–28 June 1996*, pages 40–47, Santa Cruz, Ca.

Geoffrey Nunberg. 1978. *The Pragmatics of Reference.* Ph.D. thesis, City University of New York, New York.

Ted Pedersen. 2000. A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation. In *Proc. of the 1st Conference of the North American Chapter of the ACL; 2000*, pages 63–69.

Ted Pedersen. 2001. A decision tree of bigrams is an accurate predictor of word sense. In *Proc. of the 2nd Conference of the North American Chapter of the ACL; 2001*, pages 79–86.

James Pustejovsky. 1995. *The Generative Lexicon.* MIT Press, Cambridge, Mass.

David Stallard. 1993. Two kinds of metonymy. In *Proc. of the 31st Annual Meeting of the Association for Computational Linguistics; Columbus, Ohio, 22-26 June 1993*, pages 87–94, Columbus, Ohio.

Gustav Stern. 1931. *Meaning and Change of Meaning.* Göteborg: Wettergren & Kerbers Förlag.

Cornelia Verspoor. 1996. Lexical limits on the influence of context. In *Proc. of the 18th Annual Conference of the Cognitive Science Society; La Jolla, Cal., 12–15 July 1996*, pages 116–120.

Cornelia Verspoor. 1997. Conventionality-governed logical metonymy. In H. Bunt, L. Kievit, R. Muskens, and N. Verlinden, editors, *Proc. of the 2nd International Workshop on Computational Semantics*, pages 300–312, Tilburg, The Netherlands.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics; Cambridge, Mass., 26–30 June 1995*, pages 189–196.

David Yarowsky. 1997. Homograph disambiguation in speech synthesis. In R. Sproat, J. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, pages 159–175. Springer-Verlag.