Mapping WordNet Senses to a Lexical Database of Verbs

Rebecca Green, Lisa Pearl, Bonnie J. Dorr

LAMP TR-061, UMIACS TR-2001-02, CS TR-4206

Abstract:

This paper describes automatic techniques for mapping 9611 semantically classified English verbs to WordNet senses. The verbs were initially grouped into 491 semantic classes based on syntactic categories; they were then mapped into WordNet senses according to three pieces of information: (1) prior probability of WordNet senses; (2) semantic similarity of WordNet senses for verbs within the same category; and (3) probabilistic correlations between WordNet relationship and verb frame data. Our techniques make use of a training set of 1791 disambiguated entries, representing 1442 verbs occurring in 167 of the categories. The best results achieved .58 recall and .72 precision, versus a lower bound of .38 recall and .62 precision for assigning the most frequently occurring WordNet sense, and an upper bound of .75 recall and .87 precision for human judgment.

Acknowledgements:

All three authors are supported, in part, by PFF/PECASE Award IRI-9629108, DOD Contract MDA904-96-C-1250, and DARPA Contracts N66001-97-C-8540 and N66001-00-2-8910. Rebecca Green is supported, in part, by a National Science Foundation Graduate Research Fellowship. We are indebted to Philip Resnik for his assistance with experimental runs of his algorithm on the data and his useful commentary in the preparation of this document.

(Submitted to 39th Annual Meeting of the Association for Computational Linguistics, Toulouse France.)

1 Introduction

Our goal is to map entries in a lexical database of 4076 verbs automatically to WordNet senses (Miller and Fellbaum, 1991), (Fellbaum, 1998). The task of mapping each entry to a set of WordNet senses involves word sense disambiguation (WSD), but with several twists that distinguish it from the standard WSD efforts represented by, for example, SENSEVAL (Kilgarriff and Rosenzweig, 2000). WSD research typically involves disambiguating words in context, using either an all-words (all words in a set of texts are to be disambiguated) or lexical-sample (only a specific set of words from a set of corpus instances are to be disambiguated) approach; SENSEVAL takes the lexical-sample approach, in which it is possible to take advantage of detailed knowledge of specific word senses for the sample being investigated. While an underlying assumption holds that only one word sense is accurate for each token, provisions are made for evaluation purposes to give partial credit if the correct sense is included among a disjunctive set of word senses.

In contradistinction to the circumstances of SENSEVAL, the task reported here takes an all-words approach, since the goal is to map all entries in the database to their respective word senses. Another difference resides in the lack of contextual data (i.e., corpus instances) for the words to be disambiguated. In place of context, information about verb senses, encoded in terms of thematic grids and lexical-semantic representations from (Olsen et al., 1997), has been used. Moreover, our task differs from that of SENSEVAL in that it involves finding the single corresponding WordNet sense, as several may be equally appropriate. WordNet has been criticized for making distinctions that are too fine-grained (Palmer, 2000), such that, even in context, it may be unclear which of several WordNet senses is invoked or even if only one sense is invoked. When words occur out of context, it is even more likely that more than one sense will be applicable.

One might argue that the task is more one of *lexicon integration*[1] than word sense disambiguation, since word sense distinctions are recognized in both the lexical database and WordNet. Nevertheless, at each turn, it really is a WSD task: For each entry in the lexical database, the task is to find the corresponding WordNet sense. As opposed to integration efforts, which are symmetric, the present task is asymmetric: There is no attempt to map the senses of the lexical database to WordNet or to unify the senses of the two lexicons. Moreover, the approach used is one that could be adopted as a part of an online application, e.g., for lexical selection in machine translation—the same sort of disambiguation is required, and has been used, for the translation of multiply ambiguous words such as Chinese verb *la* (which has several possible English translations, including *slash*, *cut*, *chat*, *pull*, *drag*, *transport*, *move*, *raise*, *help*, *implicate*, *involve*, *defecate*, *pressgang*) (Dorr et al., 2000).

[1]The phrase is meant to be reminiscent of the vocabulary integration efforts of the National Library of Medicine's Unified Medical Language System (UMLS) project, in which 40-odd medical vocabularies have been unified into a single "metathesaurus" (Bodenreider and Bean, 2000), as well as of ontology integration efforts (Hovy, 2001). See (Bean and Green, 2000) for more details.

2 Semantic Resources

In text processing, WSD relies on contextual information—yet we have no such information for mapping the verbs into their appropriate WordNet senses. Happily, the two semantic resources we use for this work *both* record a certain amount of information about the word senses they contain. We describe each of these resources, in turn.

Our main semantic resource is an existing classification of 4076 English verbs—called here *Levin+ classes*—based initially on English Verbs Classes and Alternations (Levin, 1993) and extended through the splitting of some classes into subclasses and the addition of new classes (Dorr and Olsen, 1997). As verbs may be assigned to multiple classes, the number of entries in the database is rather larger, viz., 9611. These classes represent semantic groupings with largely shared syntactic behavior, captured in both *thematic roles* and *lexical conceptual structures* (LCS) (Jackendoff, 1990; Dorr, 1993) shared by all members of the class. We distinguish between two types of semantic classes: (1) The Levin+ Class, based on thematic roles coupled with their verb groups organized by syntactically frame; (2) The LCS Class, based on the fundamental LCS representation (e.g., ACT_{Perc}, GO_{Loc}, BE_{Ident}). There is a one-to-many relationship between Levin+ and LCS Classes: A Levin+ class belongs to one LCS Class, while an LCS class usually includes more than one Levin+ Class.

Our second semantic resource, WordNet, associates syntactic data with semantic information, recorded as patterns ("frames") (e.g., Somebody ___s something; Something ___s; Somebody ___s somebody into V-ing something). There are 35 such verb frames in WordNet and a synset may have only one or as many as a half dozen or so frames assigned to it. Further information about WordNet senses is

derived from SEMCOR, a semantic concordance incorporating tagging of the Brown corpus with WordNet senses.[2]

Our mapping between these two resources relies on an implicit relation between thematic roles in Levin+ and verb frames in WordNet. Both reflect how many and what kinds of arguments a verb may take. However, they take rather different approaches in conveying this information. Levin+ makes use of *thematic grids* (henceforth abbreviated θ-grids), i.e., a listing of arguments and their types in an integrated unit. An example is the θ-grid `_ag_th,instr(with)` (i.e., *Agent,Theme,Instrument*) as in the verb *stock* in *I stocked the fridge with coke*.[3] WordNet encodes this information indirectly, listing all the frames a verb sense may be found in. As a result, the θ-grids in the lexical database distinguish 67 individual argument components (e.g., `_ag`, `_th`, `,instr(with)`), combined into 106 distinct full θ-grids, while, by way of contrast, WordNet's smallest syntactic unit is the frame, of which 35 are used, combined into 217 different ways for the verbs being examined. This suggests that the integration of argument components into full θ-grids is tighter, more systematic, and more informative than the combination of verb frames in WordNet: A θ-grid is a unified structure, but a set of frames in WordNet is simply a disjunction of individual structures. One may also surmise that the θ-grid assignments are more accurate, as they are based on work in which syntactic behavior was a major focus of research rather than a peripheral component of the work. Despite these differences, it is reasonable to assume there should be some correlation between θ-grids and verb frames.

3 Training Data

Our sense-linking task relies on knowledge about correlations between information in each of our two semantic resources. While parallel data are available for both, it is not clear *a priori* that the correlations between the two sets of data are very direct. For example, an attempt to construct a mapping between WordNet frames and θ-grids revealed that the underlying classifications differ in significant ways. Training data are thus necessary.

[2]For further information, see the manual page for SEMCOR, available from http://www.cogsci.princeton.edu, under WordNet manuals, section 7, SEMCOR.

[3]Commas (,) between thematic roles indicate optionality; underscores (_) indicate obligatoriness. This is a distinction WordNet can make only indirectly by assigning multiple frames.

At the time this research effort was started, WordNet senses had been assigned manually to a significant number of the lexical database entries (Dorr and Jones, 1996). However, some of the assignments were in doubt, since class splitting had occurred subsequent to those assignments, with all WordNet senses having been carried over to all new subclasses. New classes had also been added since the manual tagging. It was determined that the tagging for only 1791 entries—including 1442 verbs in 167 classes—could be considered stable; for these entries, 2756 assignments of WordNet senses had been made. Data for these entries, taken from both WordNet and from Levin+, constitute the training data for this study.

The following probabilities were generated from the training data:

1. LCS probability: The probability that a WordNet verb sense, related to another WordNet verb sense through a particular relationship type, would be mapped to the same LCS class. This probability was computed separately for each relationship type and was limited to those relationship instances where the WordNet senses on both sides of the relationship were in the training data; the computed values generally ranged between .3 and .35.

2. Levin+ probability: The probability that a WordNet verb sense, related to another WordNet verb sense through a particular relationship, would be mapped to the same Levin+ class (with the same θ-grid). As above, this probability was computed separately for each relationship type and limited to WordNet senses in the training data; the computed values generally ranged between .25 and .3.

3. Combination frame probability: The probability that a verb in a class with a particular θ-grid would be mapped to a WordNet verb sense with some specific combination of frames. Values average only .11, but in some cases the probability is 1.0.

4. Individual frame probability: The probability that a verb in a class with a particular θ-grid component would be mapped to a WordNet verb sense assigned a specific frame (possibly among others). Values average .20, but in some cases the probability is 1.0.

5. Prior WordNet sense probability: Probability of a prior WordNet verb sense, based on the tagging in SEMCOR. Values average .11, but in some cases the probability is 1.0.

6. Semantic similarity probability: Probability that a verb, given all the other verbs assigned with it to a single class, would be mapped to a specific WordNet sense. This represents an implementation of a class disambiguation algorithm (Resnik, 1999), modified to run against the WordNet verb hierarchy.[4]

In addition, a rather powerful assumption (referred to hereafter as the *same synset assumption*) was made: When a WordNet sense has been assigned to a verb in a Levin+ class, if another verb from that WordNet synset has been assigned to the same class, it should have the sense corresponding to that synset assigned for it. (Since Levin+ verbs are mapped into WordNet senses through the synsets to which they belong, the resulting action is more easily captured by saying: It should also have the same synset assigned.)

The fact that the training data represent only 167 of the 491 Levin+ classes makes the task especially non-trivial, since some of the available data on the lexical database side derives from these classes. The need to extend data from 167 classes for the making of assignments across 491 classes has been addressed in several ways:

- The one-to-many relationship between LCS and Levin+ classes opens up the possibility that some assignments in Levin+ classes for which no training data is available could be made on the basis of data for Levin+ classes in the same LCS class for which training data is available.
- Since several classes might use the same θ-grids, training data correlating verb frames and θ-grids (taken as wholes or as parts thereof) could serve to map WordNet senses into new Levin+ classes.
- Class-independent data, such as the prior probability of a WordNet sense, might also warrant the assignment of WordNet senses for verbs in classes outside those in the training data.

4 Evaluation

It would be fruitless to undertake the mapping task without having some way of evaluating its effectiveness. Subsequent to the culling of the training set, several processes were undertaken that resulted in full tagging of the lexical database. First, WordNet assignments for verbs in classes that had been split after manual tagging had some of their senses automatically filtered out based on incompatibilities between θ-grids and WordNet verb frames. Note that this did not involve a full set of correspondences between θ-grids and WordNet verb frames, but simply a set of (fairly egregious) incompatibilities. Second, tagging of entries for which no WordNet senses had yet been made was accomplished through a second manual assignment process. Thus, for a not insignificant portion of the lexical database, WordNet senses have been assigned manually, each entry having been considered by at least two coders.

In both manual tagging exercises, if a WordNet sense was considered correct by any of the coders, it was assigned. In the first round of tagging and in the first phase of the second round of tagging, one coder worked through the entire set of verbs under consideration, while two other coders split responsibilities for parallel tagging of the same set of verbs. In the first of these exercises, there was a relatively high degree of agreement: Of the senses identified as being correct, .5465 were identified independently by both coders; the kappa coefficient (K) of intercoder agreement was .4668.[5] The second round of manual

[4]The assumption underlying this measure is that the appropriate word senses for a group of semantically related words should themselves be semantically related. Given WordNet's hierarchical structure, the semantic similarity between two WordNet senses corresponds to the degree of informativeness of the most specific concept that subsumes them both.

[5]The kappa statistic measures the degree to which pairwise agreement of coders on a classification task surpasses what would be expected by chance; the standard definition of this coefficient is: $K = (P(A) - P(E))/(1 - P(E))$, where $P(A)$ is the actual percentage of agreement and $P(E)$ is the expected percentage of agreement, averaged over all pairs of assignments. Several adjustments in the computation of the kappa coefficient were made necessary by the possible assignment of multiple senses for each verb in a Levin+ class. Without prior knowledge of how many senses are to be assigned, there is no basis on which to compute $P(E)$. As a first adjustment, the computation used here employed knowledge of how many senses were assigned by each coder. $P(E)$ was then computed as the sum of the probability that two coders would agree on assigning a specific sense and the probability that two coders would agree on not assigning a specific sense, i.e., $((A/S)(B/S)) + (((S - A)/S)((S - B)/S))$, where A equals the number of senses assigned by one coder, B equals the number of senses assigned by the other coder, and S equals the total number of WordNet senses. As a second adjustment, when $S = 1$, there is no need to take the number of senses assigned into account (indeed, it produces anomalous results) and $P(E)$ was automatically set at .5. As a third adjustment, since the number of senses possible per verb in a Levin+ class varies, the value computed for each such verb was weighted by the number of senses possible, so that the average was computed over number of senses considered rather than over number of verbs being tagged.

tagging included verbs that had been automatically assigned to the Levin+ classes to begin with and was considered a harder task by the coder involved in both tagging rounds. Here the rate of identification by multiple coders was lower, .3589; the kappa coefficient of intercoder agreement was .2434. (Over the two rounds of parallel tagging, .5067 of the assigned senses were identified by multiple coders.) A final independent coder tagged all entries for which no WordNet sense had been identified by the coding teams. Overall, 13452 WordNet sense assignments were made.

While the full tagging of the lexical database may make the automatic tagging task appear superfluous, the low rate of agreement between coders and the automatic nature of some of the tagging suggest that there is still room for adjustment of WordNet sense assignments in the lexical database. On the one hand, even the higher of the kappa coefficients mentioned above is significantly lower than the standard suggested for good reliability ($K > .8$) or even the level where tentative conclusions may be drawn ($.67 < K < .8$) (Carletta, 1996), (Krippendorff, 1980). On the other hand, if the automatic assignments agree with human coding at levels comparable to the degree of agreement among the humans, it may be used, on the one hand, to identify assignments that are in question for review, and, on the other hand, to suggest other assignments for further consideration. Moreover, there are consistency checks that can be made much more easily by the automatic process than can be made manually. For example, the premise that when a WordNet sense is assigned for a verb in a Levin+ class, if another verb from the synset also occurs in that class, it should also have the same synset assigned is much more easily enforced automatically than manually. When such WordNet sense assignments are made automatically on the basis of the 2756 senses in the training set, another 967 sense assignments are generated, only 131 of which were assigned manually. Similarly, when such a premise is enforced on the entirety of the lexical database of 13452 assignments, another 5059 sense assignments are generated. If the premise is valid and if the senses assigned in the database are accurate, then the human tagging has a recall of no more than .7267. Alternatively, some of the senses that have been assigned may not be valid (recall that a sense was assigned even if only one coder felt it applied, which applies to .4535 of the senses assigned in the first round of manual tagging and to .6411 of the senses assigned in the second round, or .4933 of the senses assigned in the two rounds taken together).

In a task of this sort, the typical approach would be to set both upper bound and lower bound baselines, with the upper bound set by human performance and the lower bound set by application of the simplest algorithm, such as always assigning the most probable word sense. For purposes of this comparison, it will be assumed that all senses assigned by any coder are correct, although this is almost certainly not the case. This would give as an upper bound a recall ratio of .7534 against a precision ratio of 1.0. (However, if only sense assignments on which there was agreement were considered correct, the sense assignments without agreement would be precision failures rather than recall failures, so that the upper bound would then be a precision ration of .7534 at a recall ratio of 1.0). The lower bound, based on prior probability of WordNet senses, has a recall ratio of .3768 at a precision ratio of .6181.

Since a word sense was assigned even if only one coder judged it to apply, all human judgments are credited as being correct. On this basis, human coding has a precision of 1.00. However, it is reasonable to assume that some of the solo judgments were idiosyncratic. To determine what proportion of such judgments were in reality precision errors, a random sample of 50 WordNet senses in the database that were supported by only one of the two original judges were investigated further by a team of three judges. In this round, judges rated the senses for the verbs, as assigned to a specific Levin+ class, as falling into one of three categories: (1) definitely correct, (2) definitely incorrect, and (3) arguable whether correct. If any one of the judges rated a sense 'definitely correct,' another judge independently judged it likewise; this accounts for thirty-one instances. Thirteen of the instances were judged 'definitely incorrect' by at least two of the judges. No consensus was reached–other than lack of certainty!–on the remaining six instances. Extrapolating from this sample to the full set of judgments in the database supported by only one coder leads to the assumption that approximately 1725 (26% of 6636 solo judgments) of those senses are incorrect. This puts the precision of the database at .8718.

5 Mapping Strategies

Recent work (Van Halteren et al., 1998) has demonstrated improvement in part-of-speech tagging when the outputs of multiple taggers are combined. When the errors of multiple classifiers are not significantly correlated, the result of combining votes from a set of individual classifiers often outperforms the best result from any single classifier. Using a voting strategy seems especially appropriate here: Most of the data available for picking out WordNet senses for

entries in the lexical database function as only weak indicators of correct senses; on average, they identify correct senses from the training data about 40% of the time. At the same time, there is significant variation in which senses they pick out.

The investigations undertaken here used both simple and aggregate voters, combined using various voting strategies. The simple voters were the measures introduced above in Section 3: LCS probability, Levin+ probability, combination frame probability, individual frame probability (breaking off into two measures: one, the maximum probability for any frame associated with a WordNet sense; the other, the average probability for all frames associated with a sense), prior probability of a WordNet sense, and semantic similarity measure. In addition, three aggregate voters were generated: (1) the product of the seven simple measures (smoothed so that zero values wouldn't totally offset all other measures); (2) the weighted sum of the seven simple measures, with weights representing the percentage of the training set assignments correctly identified by the highest score of the simple probabilities; and (3) the maximum score of the seven simple measures.

Using these data, two different sorts of voting schemes were investigated. The two sets of schemes differ most significantly on: (1) the circumstances under which a voter casts its vote for a WordNet sense, (2) the size of the vote cast by each voter, and (3) under what circumstances a WordNet sense was selected, i.e., was declared a "winner." Using their differences on this last variable to characterize them, we will refer to one set as *Majority Voting Schemes* and to the other set as *Threshold Voting Schemes*.

5.1 Majority Voting Schemes

Although we do not know in advance how many WordNet senses should be assigned to an entry in the lexical database, we assume that, in general, there is at least one. In line with this intuition, one strategy we investigated was to have both simple and aggregate measures cast a vote for whichever sense(s) of a verb in a semantic class received the highest (non-zero) value for that measure. This general strategy underlay several rounds of votes:

Round 1. Non-combination voting. Each simple and aggregate measure voted on its own. The most effective of these are:

1a Prior probability of WordNet senses

1b Semantic similarity measure

1c Product of simple measures

Round	Recall	Precision
1a	.3768	.6181
1b	.5575	.7089
1c	.5092	.7413
1d	.5283	.7675
2	.2301	.7133
3	.3808	.5960
4	.5769	.7228
5	.5197	.7823
6	.4422	.7406
7	.4949	.7724

Table 1: Recall and precision measures (prior to implementation of "same synset" assumption) for Majority Voting Schemes

1d Weighted sum of simple measures

Round 2. Majority vote of all (seven) simple voters.

Round 3. Majority vote of all (twenty-one) pairs of simple voters, where the pair cast a vote for a sense if, among all the senses of a verb, a specific sense had the highest value for both measures.

Round 4. Majority vote of two (product and weighted sum) of the aggregate voters.

Round 5. Majority vote of the three most effective single measures (1b + 1c + 1d)

Round 6. Majority vote of simple voters and (more heavily weighted) aggregate voters (2 + 4)

Round 7. Majority vote of pairs of simple voters and (more heavily weighted) aggregate voters (3 + 4)

Table 1 gives recall and precision measures for all of these voting schemes. Table 2 gives recall and precision measures for the same schemes, but after the "same synset" assumption has subsequently been implemented.

In rounds 2–7, a majority vote is achieved in a case where a sense receives half or more of all the votes cast, where all voters cast a single vote, except in rounds 6 and 7, where the aggregate voters' votes were weighted to have as much voting clout as the whole ensemble of (pairs of) simple voters combined. This means the number of WordNet senses chosen for a verb in a specific semantic class would be at most two, which would happen only if each of the two senses received exactly half of the votes. As this may be too restrictive, future work with these voting schemes will explore the use of a lower percentage of all votes cast for the selection cutoff. So far, the best

Round	Recall	Precision
1a	.4531	.4640
1b	.6001	.5509
1c	.5676	.5542
1d	.5791	.5613
2	.2957	.4813
3	.4495	.4349
4	.6252	.5344
5	.5726	.5732
6	.5041	.5421
7	.5517	.5664

Table 2: Recall and precision measures (after implementation of "same synset" assumption) for Majority Voting Schemes

voting scheme appears to be that used in Round 4, based on the product and weighted-sum aggregate voters.

5.2 Threshold Voting Schemes

The second voting strategy commenced by identifying, for each simple and aggregate measure, the threshold value at which the product of recall and precision scores in the training set has the highest value if that threshold is used to select WordNet senses. During the voting, if a WordNet sense has a higher score for a measure than its cut-off threshold, the measure votes for the sense; otherwise, it votes against it. The weight of the measure's vote is the precision-recall product at the cut-off threshold. This voting strategy has the advantage of taking into account each individual attribute's strength of prediction.

Five variations on this basic voting scheme were investigated. In each, senses were selected if their vote total exceeded a variation-specific threshold. Table 3 summarizes recall and precision for these variations at their optimal vote thresholds.

The first variation (Automatic Mapping and Abstaining Votes) implements the same synset assumption (automatic mapping); it also had two of the simple measures (LCS probability and Levin+ probability) abstain from ever voting against WordNet senses, inasmuch as only a small percentage of the overall set of test data had non-zero values for these measures. The second variation (Automatic Mapping and Non-votes) differed from this first variation in completely disregarding the LCS and Levin+ probabilities.

A third variation (Triples Voting) placed the simple and composite measures into three groups, the three with the highest weights, the three with the lowest weights, and the middle or remaining three. Voting first occurred within the group, and the group's vote was brought forward with a weight equaling the sum of the group members' weights. This variation also added to the vote total if the sense had been assigned in the training data.

The fourth variation (Combination Attributes) kept the assumptions from the third variation 3 and altered the method of voting. Rather than using the weights and thresholds calculated for the single measures from the training data, this variation calculated weights and thresholds for combinations of two, three, four, five, six, and, seven measures. Each combination attribute would vote either for or against a sense based on whether the sum of the values for the individual measures was above the threshold for that combination. The fifth variation (Combination Attributes with Automatic Mapping) added implementation of the same synset assumption to the previous variation.

6 Conclusions and Future Work

A practical use of these voting schemes would be to assist in verifying the quality of WordNet assignments in the current database. The following sets might be profitably brought forward for investigation, both to determine if senses in the database are correct and to determine if correct senses are missing from the database:

- Word sense assignments in the lexical database for which the same synset assumption does not hold; that is, verbs that have a sense assigned, when other verbs that are in that synset and also in that Levin+ class do not have that sense/synset assigned.
- Word sense assignments in the lexical database that receive substantially fewer votes than would have been needed for selection by the voting scheme.
- Word sense assignments not in the lexical database that receive substantially more than the votes needed for selection by the voting scheme.

On a more theoretical level, this research shows that it is possible to undertake all-words word sense disambiguation without relying on contextual data. In place of such data we have used syntactic information, knowledge of semantic relationships, and non-contextual corpus data (e.g., prior probability of WordNet senses). This demonstrates that the lexical resources developed within the computational linguistics community are beginning to bear fruit in ways not previously anticipated.

Variation	Recall	Precision
Automatic Mapping with Abstaining Votes	.6094	.5395
Automatic Mapping with Non-votes	.6099	.5424
Triples Voting	.6319	.5185
Combination Attributes	.5263	.4401
Combination Attributes with Automatic Mapping	.5871	.4544

Table 3: Recall and precision for Threshold Voting Schemes

References

C.A. Bean and R. Green. 2000. *Relationships in the Organization of Knowledge*. Kluwer, Dordrecht.

Olivier Bodenreider and Carol A. Bean. 2000. Relationships among Knowledge Structures: Vocabulary Integration within a Subject Domain. In C.A. Bean and R. Green, editors, *Relationships in the Organization of Knowledge*, pages 81–98. Kluwer, Dordrecht.

Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Lingustics*, 22(2):249–254, June.

Bonnie J. Dorr and Douglas Jones. 1996. Robust Lexical Acquisition: Word Sense Disambiguation to Increase Recall and Precision. Technical report, University of Maryland, College Park, MD. Submitted to Computational Linguistics.

Bonnie J. Dorr and Mari Broman Olsen. 1997. Deriving Verbal and Compositional Lexical Aspect for NLP Applications. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pages 151–158, Madrid, Spain, July 7-12.

Bonnie J. Dorr, Gina-Anne Levow, and Dekang Lin. 2000. Chinese-English Machine Translation: Building a Conceptual Verb Hierarchy. In *Fourth Conference of the Association for Machine Translation in the Americas, AMTA-2000*, pages 1–12, Cuernavaca, Mexico.

Bonnie J. Dorr. 1993. *Machine Translation: A View from the Lexicon*. The MIT Press, Cambridge, MA.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press. Further information: http://www.cogsci.princeton.edu/w̃n [2000, September 7].

Eduard Hovy. 2001. Comparing Sets of Semantic Relations in Ontologies. In R. Green, C.A. Bean, and S. Myaeng, editors, *The Semantics of Relationships: An Interdisciplinary Perspective*, page Book manuscript not yet submitted for final review. Kluwer, Dordrecht.

Ray Jackendoff. 1990. *Semantic Structures*. The MIT Press, Cambridge, MA.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and Results for English SENSEVAL. *Computers and the Humanities*, 34:15–48.

Klaus Krippendorff. 1980. *Content analysis: An Introduction to its Methodology*. Sage, Beverly Hills.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

George A. Miller and Christiane Fellbaum. 1991. Semantic Networks of English. In Beth Levin and Steven Pinker, editors, *Lexical and Conceptual Semantics, Cognition Special Issue*, pages 197–229. Elsevier Science Publishers, B.V., Amsterdam, The Netherlands.

Mari Broman Olsen, Bonnie J. Dorr, and David J. Clark. 1997. Using WordNet to Posit Hierarchical Structure in Levin's Verb Classes. In *Proceedings of the Workshop on Interlinguas in MT, MT Summit, New Mexico State University Technical Report MCCS-97-314*, pages 99–110, San Diego, CA, October. Also available as UMIACS-TR-97-85, LAMP-TR-011, CS-TR-3857, University of Maryland.

Martha Palmer. 2000. Consistent Criteria for Sense Distinctions. *Computers and the Humanities*, 34:217–222.

Philip Resnik. 1999. Disambiguating noun groupings with respect to wordnet senses. In S. Armstrong, K. Church, P. Isabelle, E. Tzoukermann S. Manzi, and D. Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, pages 77–98. Kluwer Academic, Dordrecht.

Hans Van Halteren, Jakub Zavrel, and Walter Daelemans. 1998. Improving data-driven wordclass tagging by system combination. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, pages 491–497. Available: http://xxx.lanl.gov/ps/cmp-lg/9807013 [2000, September 8].