Sense Information for Disambiguation: Confluence of Supervised and Unsupervised Methods

Kenneth C. Litkowski
CL Research
9208 Gue Road
Damascus, MD 20872
ken@clres.com

Abstract

For SENSEVAL-2, we disambiguated the lexical sample using two different sense inventories. Official SENSEVAL-2 results were generated using WordNet, and separately using the New Oxford Dictionary of English (NODE). Since our initial submission, we have implemented additional routines and have now examined the differences in the features used for making sense selections. We report here the contribution of default sense selection, idiomatic usage, syntactic and semantic clues, subcategorization patterns, word forms, syntactic usage, context, selectional preferences, and topics or subject fields. We also compare the differences between WordNet and NODE. Finally, we compare these features to those identified as significant in supervised learning approaches.

1 Introduction

CL Research's official submission for SENSEVAL-2 used WordNet as the lexical inventory. Separately, we also used a machine-readable dictionary (The New Oxford Dictionary of English, 1998) (NODE), mapping NODE senses automatically into WordNet senses. We did not submit these results, since we were not sure of the feasibility of using one dictionary mapped into another. Our initial results (Litkowski, 2001) proved to be much better than anticipated, achieving comparable levels of precision although at lower levels of recall, since not all senses in NODE mapped into WordNet senses. Subsequently, we examined our results in more detail (Litkowski, 2002), primarily focusing on the quality of the mapping and its effect on our performance using NODE. This led us to the conclusion that we had likely performed at a considerably higher level using the NODE inventory, with an opportunity for even better performance as we were able to exploit much more information available in NODE.

We have now identified what features (i.e., sense information) were used in our disambiguation. In particular, we have examined the role of (1) default sense selection, (2) idiomatic usage, (3) typing (e.g., transitivity), (4) syntactic and semantic clues, (5) subcategorization patterns, (6) word form (e.g., capitalization, tense, or number), (7) selectional preferences (for verbs and adjectives), (8) syntactic usage (e.g., nouns as modifiers), (9) context (in definitions and in examples), and (10) topic area (e.g., subject fields associated with definitions).

Our methodology enables us to compare the features relevant to disambiguation in WordNet and in NODE, allowing us to pinpoint differences between the two sense inventories.[1] In addition, comparing our findings with those identified in supervised machine learning algorithms, we can see patterns of similarity with our features.

In the following sections, we describe our methods of dictionary preparation, our disambiguation techniques, our methodology for analyzing features and our results. We discuss these findings in terms of what they say about the differences between the information available in each of the two sense inventories, the possible generalizability of our analysis technique, and how our features relate to those used by other SENSEVAL participants who used supervised learning techniques. Finally, we describe our future plans of analysis, based on attempting to merge

[1] We have not yet determined how decisive these features are in making correct sense selections. The present study should be viewed as an examination of sense distinctions in lexical resources.

supervised and unsupervised word-sense disambiguation.

2 Dictionary Preparation

CL Research's DIMAP (Dictionary Maintenance Programs) disambiguates open text against WordNet or any other dictionary converted to DIMAP. The dictionaries used for disambiguation operate in the background (as distinguished from the foreground development and maintenance of a dictionary), with rapid lookup to access and examine the multiple senses of a word after a sentence has been parsed. DIMAP allows multiple senses for each entry, with fields for definitions, usage notes, hypernyms, hyponyms, other semantic relations, and feature structures containing arbitrary information.

For SENSEVAL-2, WordNet was entirely converted to alphabetic format for use as the disambiguation dictionary. Details of this conversion (which captures all WordNet information) and the creation of a separate "phrase" dictionary for all noun and verb multiword units (MWUs) are described in Litkowski (2001). In disambiguation, the phrase dictionary is examined first for a match, with the full phrase then used to identify the sense inventory rather than a single word.

NODE was prepared in a similar manner, with several additions. A conversion program transformed the MRD files into various fields in DIMAP, the notable difference being the much richer and more formal structure (e.g., lexical preferences, grammar fields, and subsensing). Conversion also automatically created "kind" and "clue" regular expression phrases under individual headwords, e.g., "**(as) happy as a sandboy (**or **Larry** or **a clam)**" under *happy* was converted into a collocation pattern for a sense under *happy*, written "(as|?) ~ as (a sandboy | Larry | a clam)", with the tilde marking the target word. Further details on this conversion and definition parsing to enrich the sense information are also provided in Litkowski (2001). After parsing was completed, a phrase dictionary was also created for NODE.[2]

[2]WordNet definitions were not parsed. An experiment showed the semantic relations identifiable through parsing were frequently inconsistent with those in WordNet.

The SENSEVAL lexical sample tasks (disambiguating one of 73 target words within a text of several sentences) were run independently against the WordNet and NODE sense inventories, with the WordNet results submitted. To investigate the viability of mapping for WSD, subdictionaries were created for each of the lexical sample words. For each word, the subdictionaries consisted of the main word and all entries identifiable from the phrase dictionary for that word. (For *bar*, in NODE, there were 13 entries where **bar** was the first word in an MWU and 50 entries where it was the head noun; for *begin*, there was only one entry.)

The NODE dictionaries were then mapped into the WordNet dictionaries (see Litkowski, 1999), using overlap among words and semantic relations. The 73 dictionaries for the lexical sample words gave rise to 1372 WordNet entries and 1722 NODE entries. Only 491 entries (of which, 418 were MWUs) were common (i.e., no mappings were available for the remaining 1231 NODE entries, all of which were MWUs); 881 entries in WordNet were therefore inaccessible through NODE. For the entries in common, there was an average of 5.6 senses, of which only 64% were mappable into WordNet, thus creating our initial impression that use of NODE would not be feasible.[3]

3 Disambiguation Techniques

Details of the disambiguation process are provided in Litkowski (2001). In general, for the lexical sample, the sentence containing the target word was first parsed and the part of speech of the target word was used to select the sense inventory. If the tagged word was part of an MWU, the MWU's sense inventory was used. The dictionary entry for the word was then accessed. Before evaluating the senses, the topic area of the context provided by the sentence was "established". Subject labels for all senses of all content words in the context were tallied.

Each sense of the target was then evaluated, based on the available information for the sense, including type restrictions such as transitivity, presence of accompanying grammatical constituents such as infinitives or complements, selectional

[3]Note that a mapping from WordNet to NODE generates similar mismatch statistics.

Table 1. Lexical Sample Precision												
	Adjectives			Nouns			Verbs			Total		
Run	Items	Fine	Coarse	Items	Fine	Coarse	Items	Fine	Coarse	Items	Fine	Coarse
WordNet Test	768	0.354	0.354	1726	0.338	0.439	1834	0.225	0.305	4328	0.293	0.367
NODE Test	420	0.288	0.288	1403	0.402	0.539	1394	0.219	0.305	3217	0.308	0.405
WordNet Test (R)	768	0.435	0.435	1726	0.430	0.535	1834	0.267	0.387	4328	0.368	0.462
NODE Test (R)	684	0.472	0.472	1567	0.429	0.537	1605	0.189	0.300	3856	0.337	0.427

preferences for verbs and adjectives, form restrictions such as number and tense, grammatical roles, collocation patterns, contextual clue words, contextual overlap with definitions and examples, and topical area matches. Points were given to each sense and the sense with the highest score was selected; in case of a tie, the first sense was selected.

The top line of Table 1 shows our official results using WordNet as the disambiguation dictionary, with an overall precision (and recall) of 0.293 at the fine-grained level and 0.367 at the coarse-grained level. Disambiguating with NODE immediately after the official submission and mapping its senses into WordNet senses achieved comparable levels of precision, with a coverage of 75% based on the senses that could be mapped into WordNet, even though the NODE coverage was 100%.

Since our original submission, we have implemented many additional routines and improved our NODE mapping to WordNet; our revised precision shown in Table 1 are now 0.368 at the fine-grained level and 0.462 at the coarse-grained level using WordNet and 0.337 and 0.427 using NODE. Of particular note are the facts that the mapping from NODE to WordNet is now 89% and that precision is comparable except for the verbs.

In Litkowski (2002), we examined the mapping from NODE to WordNet in considerable detail. Several of our findings are pertinent to our analysis of the features affecting disambiguation. Table 1 reflects changes to the automatic mapping along with hand changes. The automatic mapping changes account for the change in coverage. The hand mapping shows that the automatic mapping was about 70% accurate. Interestingly, the hand changes did not affect precision. In general, the fact that we were able to achieve a level of precision comparable to WordNet suggests the most frequent senses of the lexical sample words were able to be disambiguated and mapped correctly into WordNet.

The significant discrepancy between the entries (all MWUs, 1231 entries in NODE not in WordNet and 871 entries in WordNet not in NODE) in part reflects the usual editorial decisions that would be found in examining any two dictionaries. However, since WordNet is not lexicographically based, many of the differences are indicative of the idiosyncratic development of WordNet. WordNet may identify several types of an entity (e.g., *apricot bar* and *nougat bar*), where NODE may use one sense ("an amount of food or another substance formed into a regular narrow block") without creating separate entries that follow this regular lexical rule.

For the most part, verb phrases containing particles are equally present in both dictionaries (e.g., *draw out* and *draw up*), but NODE contains several more nuanced phrases (e.g., *draw in one's horns*, *draw someone aside*, *keep one's figure*, and *pull oneself together*). NODE also contains many idioms where a noun is used in a verb phrase (e.g., *call it a day*, *keep one's mouth shut*, and *go back to nature*). About 100 of our disambiguations using NODE were to MWUs not present in WordNet (20% of our coverage gap).

Of most significance to the sense mapping is the classical problem of **splitting** (attaching more importance to differences than to similarities, resulting in more senses) and **lumping** (attaching more significance to similarities than to differences, resulting in fewer senses). Splitting accounts for the remaining 80% gap in our coverage (where NODE identified senses not present in WordNet). The effect of lumping is more difficult to assess. When a NODE definition corresponds to more than one sense in WordNet, we may disambiguate correctly in NODE, but receive no score since we have mapped into the wrong definition; the WordNet sense groupings may allow us to receive credit at the coarse grain, but not at the fine grain. We have examined this issue in more detail in Litkowski (2002), with the conclusion that lumping reduces our NODE score since we are unable to pick out the single WordNet sense answer.

More problematic for our mapping was the absence of crucial information in WordNet. Delfs

Table 2. Comparative Analysis of Features Used in WordNet and NODE Disambiguation											
Instance	Default	Idiom	Kind	Clue	Context	Topics	Form	With	As	Prefs	POS
WordNet											
768	556	79	0	0	190	0	0	15	0	1	Adjectives
1754	1140	293	0	0	536	0	0	29	0	0	Nouns
1804	436	161	0	2	576	0	0	984	0	0	Verbs
4326	2132	533	0	2	1302	0	0	1028	0	1	Total
NODE											
768	324	81	0	2	249	168	14	11	11	33	Adjectives
1754	456	269	14	94	546	364	317	28	136	3	Nouns
1804	175	105	61	124	564	285	353	573	187	108	Verbs
4326	955	455	75	220	1359	817	684	612	334	144	Total

(2001) described a sense for *begin* that has an infinitive complement, but present only in an example sentence and not explicitly encoded with the usual WordNet verb frame. Similarly, for *train*, two sentences were "tagged to transitive senses despite being intransitive because again we were dealing with an implied direct object, and the semantics of the sense that was chosen fit; we just pretended that the object was there." In improving our disambiguation routines, it will be much more difficult to glean the appropriate criteria for sense selection in WordNet without this explicit information than to obtain it in NODE and map it into WordNet. Much of this information is either not available in WordNet, available only in an unstructured way, only implicitly present, or inconsistently present.

4 Feature Analysis Methodology

4.1 Identifying Disambiguation Features

As indicated above, our disambiguation routines assign point values based on a judgment of how important each feature seems to be. The weighting scheme is ad-hoc. For the feature analysis, we simply recorded a binary variable for each feature that had made a contribution to the final sense selection. In particular, we identified the following features: (1) whether the sense selected was the default (first) sense (i.e., no other features were identified in examining any of the senses), (2) whether the identified sense was based on the occurrence of the target word in an idiom, (3) whether a type (specifically, transitivity) factored into the sense selection, (4) whether the selected sense had any syntactic or semantic clues, (5) whether a subcategorization pattern figured into the sense selection, (6) whether the sense had a specified word form (e.g., capitalization, tense, or number), (7) whether a syntactic usage was relevant (e.g., nouns as modifiers or an adjective being used as a noun, such as "the blind"), (8) whether a selectional preference was satisfied (for verb subjects and objects and adjective modificands), (9) whether we were able to use a Lesk-style context clue from the definitions or an example, and (10) topic area (e.g., subject fields, usage labels, or register labels associated with definitions).

As the disambiguation algorithm proceeded, we recorded each of the features associated with each sense. After a sense was selected, the features associated with that sense were written to a file (as a hexadecimal number) for subsequent analysis. We sorted the senses for each target word in the lexical sample and summarized the features that were used for all instances that had the same sense. We then summarized the features over all senses and further summarized them by part of speech. These results are shown in Table 2.

The first column shows the number of instances for each part of speech and overall. The second column shows the number of instances where the disambiguation algorithm selected the default sense. These cases indicate the absence of positive information for selecting a sense and may be construed as indicating that the sense inventory may not make sufficient sense distinctions. The default numbers are somewhat misleading for verbs, where the mere presence of an object (recorded in the "with" column) sufficed to make a selection "non-default". As well, the default selections may indicate that our disambiguation does not yet make full use of the distinctions that are available. As we make improvements in our algorithm, we would expect the number of default selections to decrease.

The significant difference in the number of default selections between WordNet and NODE is a broad indicator that there is more information available in NODE than in WordNet. In examining the results for individual words, even in cases where the "default" (or first) sense was being selected, the decision was being made in NODE based on positive information rather than the absence of information.

Generally (but not absolutely), the intent of the compilers of both WordNet and NODE is that the first sense correspond to the most frequent sense. The relative importance of the default sense indicated by our results suggests the importance of ensuring that this is the case. In a few instances, the first NODE sense did not correspond to the first WordNet sense, and we were able to obtain a much better result disambiguating in NODE than in WordNet by using an appropriate mapping from NODE to a second or third WordNet sense. The significance of the default sense is important in the selection of instances in an evaluation such as SENSEVAL; if the instances do not reflect common usage, WSD results may be biased simply because of the instance selection.

The "idiom" column indicates those cases where a phrasal entry was used to provide the sense inventory. As pointed out above, these correspond to the MWUs that were created and account for over 10% of the lexical instances.

The "kind" and "clue" columns correspond to either strong or slightly weaker collocational patterns that have been associated with individual senses. These correspond to similarly named sense attributes used in the Hector database for SENSEVAL-1, which was the experimental basis for NODE. As can be seen in the table, these were relevant to the sense selection for about 6.5 percent of the instances for NODE. We converted several of WordNet's verb frames into clue format; however, they did not show up as features in our analysis, probably because our implementation needs to be improved. We expect that further improvements will obtain some cases where these are relevant in the WordNet disambiguation (as well as increasing the number of cases where these are relevant to NODE senses).

The context column reflects the significance of Lesk-style information available in the definitions and examples. In general, it appears that about a third of the lexical instances were able to use this information. This reflects the extent to which the dictionary compilers are able to provide good examples for the individual senses. Since space is limited for such examples, our results indicate that there will an inevitable upper limit of the extent to which disambiguation can rely on such information (a conclusion also reached by (Haynes 2001)).

The potential significance of subject or topic fields associated with individual senses is indicated by the number of cases where NODE was able to use this information (nearly 20 percent of the instances). NODE makes extensive use of subject labels, particularly in the MRD. We included many subject labels, usage labels, and register labels in our WordNet conversion, but these did not surface in our disambiguation with WordNet. They were very rare for the lexical items used in SENSEVAL. The value shown here is similar to the results obtained by Magnini, et al. (2001), but their low recall suggests that for more common words, there will be a lower opportunity for their use.

The word form of a lexical item also emerged as being of some significance when disambiguating with NODE, slightly over 16 percent. In NODE, this is captured by such labels as "often capitalized" or "often in plural form". No comparable information is available in WordNet.

Subcategorization patterns (indicated under the "with" column) were very important in both WordNet (based on the verb frames) and NODE, relevant in 55% and 32% of the sense selections, respectively. As indicated, the "with" category is also important for nouns. For the most part, this indicates that a given noun sense is usually accompanied by a noun modifier (e.g., "metal fatigue").

The "as" column corresponds to nouns used as modifiers, verbs used as adjectives, and adjectives used as nouns. These were fairly important for nouns (7.7%) and verbs (10.3%).

The final column, "prefs", corresponds to selectional preferences for verb subjects and objects and adjective modificands. In these cases, a match occurred when the head noun in these positions either matched literally or was a synonym or within two synsets in the WordNet hierarchy. Although the results were relatively small, this demonstrates the viability of using such preferences.

Finally, anomalous entries in the table (e.g., nouns having subcategorization patterns used in the sense selection) generally correspond to our parser

incorrectly assigning a part of speech (i.e., treating the noun as a verb sense).

4.2 Variation in Disambiguation Features

Space precludes showing the variation in features by lexical item. The attributes in NODE for individual items varies considerably and the differences were reflected in which features emerged as important.

For adjectives, idiomatic usages were significant for *free*, *green*, and *natural*. Topics were important for *fine*, *free*, *green*, *local*, *natural*, *oblique*, and *simple*, indicating that many senses of these words have specialized meanings. Form was important for *blind*, arising from the collocation "the blind". The default sense was most prominent for *colorless*, *graceful* (with only one sense in NODE), and *solemn*. Context was important for *blind*, *cool*, *fine*, *free*, *green*, *local*, *natural*, *oblique*, and *simple*, suggesting that these words participate in common expressions that can be captured well in a few choice examples. Selectional preferences on the modificands were useful in several instances.

For nouns, idioms were important for *art*, *bar*, *channel*, *church*, *circuit*, and *post*. Clues (i.e., strong collocations) were important for *art*, *bar*, *chair*, *grip*, *post*, and *sense*. Topics were important for *bar*, *channel*, *church*, *circuit*, *day*, *detention*, *mouth*, *nation*, *post*, *spade*, *stress*, and *yew* (even though *yew* had only one sense in NODE). Context was important for *art*, *authority*, *bar*, *chair*, *channel*, *child*, *church*, *circuit*, *day*, *detention*, *facility*, *fatigue*, *feeling*, *grip*, *hearth*, *lady*, *material*, *mouth*, *nature*, *post*, and *restraint*. The presence of individual lexical items in several of these groupings shows the richness of variations in characteristics, particularly into specialized usages and collocations.

For verbs, idioms were important for *call*, *carry*, *draw*, *dress*, *live*, *play*, *pull*, *turn*, *wash*, and *work*, a reflection of the many entries where these words were paired with a particle. Form was an important feature for *begin* (over 50% of the instances), *develop*, *face*, *find*, *leave*, *match*, *replace*, *treat*, and *work*. Subcategorization patterns were important for all the verbs. However, many verb senses in both WordNet and NODE do not show wide variation in their subcategorization patterns and are insufficient in themselves to distinguish senses. Strong ("kind") and weak ("clue") collocations are relatively less important, except for a few verbs (*collaborate*, *serve*, and *work*). Topics are surprisingly significant for several verbs (*call*, *carry*, *develop*, *dress*, *drive*, *find*, *play*, *pull*, *serve*, *strike*, and *train*), indicating the presence of specialized senses. Context does not vary significantly among the set of verbs, but it is a feature in one-third of the sense selections. Finally, selectional preferences on verb subjects and objects emerged as having some value.

5 Generalizability of Feature Analysis, Relation to Supervised Learning, and Implications for Future Studies

The use of feature analysis has advanced our perception of the disambiguation process. To begin with, by summarizing the features used in the sense selection, the technique identifies overall differences between sense inventories. While our comments have focused on information available in NODE, they reflect only what we have implemented. Many opportunities still exist and the results will help us identify them.

In developing our feature analysis techniques, we made lists of features available for the senses of a given word. This gradually gave rise to the notion of a "feature signature" associated with each sense. In examining the set of definitions for each lexical item, an immediate question is how the feature signatures differ from one another. This allows us to focus on the issue of adequate sense distinctions: what is it that distinguishes each sense.

The notion of feature signatures also raises the question of their correspondence to supervised learning techniques such as the feature selection of (Mihalcea & Moldovan, 2001) and the decision lists used in WASPS (Tugwell & Kilgarriff 2001). This raises the possibility of precompiling a sense inventory and revising our disambiguation strategy to identify the characteristics of an instance's use and then simply to perform a boolean conjunction to narrow the set of viable senses.

The use of feature signatures also allows us to examine our mapping functionality. As indicated above, we are unable to map 10 percent of the senses from NODE to WordNet, and of our mappings, approximately 33 percent have appeared to be inaccurate when examined by hand. When we

examine the instances where we selected a sense in NODE, but were unable to map to a WordNet sense, we can use these instances either to identify clear cases where there is no WordNet sense.

In connection with the use of supervised learning techniques, participants of other teams have provided us with the raw data with which their systems made their sense selections. The feature arrays from (Mihalcea & Moldovan, forthcoming) identify many features in common with our set. For example, they used the form and part of speech of the target word; this corresponds to our "form". Their collocations, prepositions after the target word, nouns before and after, and prepositions before and after correspond to our idioms, "clues", and "with" features.

The array of grammatical relations used with WASPS (Tugwell & Kilgarriff, 2001) (such as bare-noun, plural, passive, ing-complement, noun-modifier, PP-comp) correspond to our "form", "clue", "with", and "as" features.

The data from these teams also identifies bigrams and other context information. Pedersen (2001) also provided us with the output of several classification methods, identifying unigrams and bigrams found to be significant in sense selection. These data correspond to our "context" feature.

We have begun to array all these data by sense, corresponding to our detailed feature analysis. Our initial qualitative assessment is that there are strong correspondences among the different data set. We will examine these quantitatively to assess the significance of the various features. In addition, while several features are already present in WordNet and NODE, we fully expect that these other results will help us to identify features that can be added to the NODE sense inventory.

6 Conclusions

Our analysis has identified many characteristics of sense distinctions, but indicates many difficulties in making such distinctions in WordNet (but also NODE). It is questionable whether WSD has been fully tested without a carefully drawn sense inventory. A lexicographically-based sense inventory shows considerable promise and invites the WSD community to pool its resources to come up with such an inventory.

Acknowledgments

I wish to thank Oxford University Press for allowing me to use their data, and particularly to Rob Scriven, Judy Pearsall, Glynnis Chantrell, Patrick Hanks, Catherine Soanes, Angus Stevenson, Adam Kilgarriff, and James McCracken for their invaluable discussions, to Rada Mihalcea, Ted Pedersen, and David Tugwell for making their data available, and to the anonymous reviewers.

References

Delfs, L. (2001, 6 Sep). Verb keys. (Personal communication)

Haynes, S. (2001, July). Semantic Tagging Using WordNet Examples. In *Association for Computational Linguistics SIGLEX Workshop* (pp. 79-82). Toulouse, France.

Litkowski, K. C. (2002). *SENSEVAL Word-Sense Disambiguation Using a Different Sense Inventory and Mapping to WordNet* (CL Research No. 02-01). . Damascus, MD.

Litkowski, K. C. (2001, July). Use of Machine Readable Dictionaries for Word-Sense in SENSEVAL-2. In *Association for Computational Linguistics SIGLEX Workshop* (pp. 107-110). Toulouse, France.

Litkowski, K. C. (1999, June). Towards a Meaning-Full Comparison of Lexical Resources. In *Association for Computational Linguistics SIGLEX Workshop* (pp. 30-7). College Park, MD.

Magnini, B., Strapparava, C., Pezzulo, G., & Gliozzo, A. (2001, July). Using Domain Information for Word Sense Disambiguation. In *Association for Computational Linguistics SIGLEX Workshop* (pp. 111-4). Toulouse, France.

Mihalcea, R., & Moldovan, D. (2001, July). Pattern Learning and Active Feature Selection for Word Sense Disambiguation. In *Association for Computational Linguistics SIGLEX Workshop* (pp. 127-30). Toulouse, France.

Mihalcea, R., & Moldovan, D. (Forthcoming). Word Sense Disambiguation with Pattern Learning and Active Feature Selection. *Journal of Natural Language Engineering.*

The New Oxford Dictionary of English. (1998) (J. Pearsall, Ed.). Oxford: Clarendon Press.

Pedersen, T. (2001, July). Machine Learning with Lexical Features: The Duluth Approach to SENSEVAL-2. In *Association for Computational Linguistics SIGLEX Workshop* (pp. 139-142). Toulouse, France.

Tugwell, D., & Kilgarriff, A. (2001, July). WASP-Bench: A Lexicographic Tool Supporting Word Sense Disambiguation. In *Association for Computational Linguistics SIGLEX Workshop* (pp. 151-4). Toulouse, France.