SENSEVAL Word-Sense Disambiguation Using a Different Sense Inventory and Mapping to WordNet

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

Abstract

In SENSEVAL-2, CL Research's unsupervised word-sense disambiguation system officially attained a coarse-grained precision of 0.367 for the lexical sample task and 0.460 for the all-words task using the WordNet sense inventory. Subsequently, an experiment investigated the viability of mapping another dictionary (the New Oxford Dictionary of English) into WordNet, disambiguating with this dictionary, and then using the maps to produce WordNet senses. The precision obtained with this intermediating dictionary was 0.402 for the lexical sample task and 0.418 for the all-words task, despite considerable mismatch in the entries and only about 70 percent mapping (which is inaccurate) for the senses in the matching entries. Results suggest that disambiguation using the dictionary was considerably better against its sense inventory than WordNet, with a better opportunity for improvement with its lexicographically-based information. Details of the mapping process provide significant insights into the issue of reuse of lexical inventories.

Introduction

The significance of the sense inventory for word-sense disambiguation (WSD) cannot be overstated. For any natural language processing (NLP) application that relies on a representation of meaning, the ability to disambiguate against the sense inventroy will affect how effective the end result will be. Considered as an end in itself, such as in SENSEVAL-2, the effectiveness of WSD may depend on the quality of the sense inventory (Kilgarriff, 2001). However, a major difficulty for sense inventories (Atkins, 1991) is that no two will be similar, for a variety of reasons, not the least of which is what counts as a sense (Kilgarriff, 1997). Finally, since any sense inventory used in WSD is ultimately a machine-readable dictionary (MRD), we must consider whether unsupervised systems relying on the MRDs can achieve the same results as other methods (Ide & Veronis, 1993).

While CL Research's participation in SENSEVAL-2 was designed primarily to (1) extend WSD (word-sense disambiguation) techniques from SENSEVAL-1 (Litkowski, 2000) and (2) generalize WSD mechanisms to rely on a full dictionary rather than a small set of entries where individual crafting might intrude, we also wanted to investigate how well WSD could be performed using an MRD of a traditional dictionary. The availability of a reference set of texts disambiguated by lexicographically-trained judges against a well-developed sense inventory (WordNet) provided a suitable opportunity. While our initial goals were met (Litkowski, 2001), we found that we were able to achieve results comparable to our official submission even with the step of using a fuzzy intermediary (i.e., the mapping between the MRD and WordNet).

CL Research's WSD functionality is implemented in DIMAP[1], designed primarily for creation and

[1] DIctionary MAintenance Programs, available from CL Research at http://www.clres.com.

maintenance of lexicons for NLP. In particular, DIMAP is designed to make MRDs tractable and to create semantic networks (similar to WordNet (Fellbaum, 1998) and MindNet (Richardson, 1997)) automatically by analyzing and parsing definitions. To place our mapping findings in perspective, we first describe the dictionary preparation techniques for WordNet and NODE (The New Oxford Dictionary of English, 1998) for use in SENSEVAL-2 (section 1). We then describe the WSD techniques used in SENSEVAL-2 (section 2) and present our results, including those achieved through mapping (section 3). In section 4, we describe the mapping from NODE to WordNet and several investigations we were able to perform in seeking to understand our performance. In section 5, we present our conclusions and future steps for improvement and in section 6, consider how these efforts correspond to other research.

1. Dictionary Preparation

DIMAP is intended to disambiguate any text against WordNet or any other dictionary converted to DIMAP, with a special emphasis on corpus instances for specific lemmas (that is, lexical samples). The dictionaries used for disambiguation operate in the background (as distinguished from the foreground development and maintenance of a dictionary), with rapid btree lookup to access and examine the characteristics of multiple senses of a word after a sentence has been parsed. DIMAP allows multiple senses for each entry, with fields for the definitions, usage notes, hypernyms, hyponyms, arbitrary other semantic relations, and feature structures containing arbitrary information, any of which can be used in disambiguation.

WordNet is already integrated in DIMAP in several ways, but for SENSEVAL-2, it was entirely converted to alphabetic format for use as the disambiguation dictionary. In this conversion, all WordNet information (e.g., verb frames and glosses) and relations are retained. Glosses are analyzed into definition, examples, usage or subject labels, and usage notes (e.g., "used with 'of'"). Verb frames are used to build collocation patterns, typical subjects and objects, and grammatical characterizations (e.g., transitivity). WordNet file and sense numbers are converted into a unique identifier for each sense. Since the glosses were intended only to serve as reminders for those constructing WordNet (Miller, 2001) and were not prepared according well-specified guidelines, the analysis into the different components is frequently inexact.

A separate "phrase" dictionary was constructed from all noun and verb multiword units (MWUs), using WordNet's sense index file. For noun MWUs, an entry was created for the last word (i.e., the head), with the first word(s) acting as a "hynonymic" indicator; an entry was also created for the first word, with the following word(s) acting as a collocation pattern (e.g., "work of art" is a hyponym of *art* and a collocation pattern under *work*, written "**~ of art**"). For verb MWUs, an entry was created for the first word, with a collocation pattern (e.g., "keep an eye on" is entered as a collocation pattern "**~ an eye on**" under *keep*). In disambiguation, this dictionary was 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 data and more formal structure (e.g., subject labels, lexical preferences, grammar fields, and subsensing) contained in well-defined tagged fields. Conversion also considerably expanded the number of entries by making headwords of all variant forms (fully duplicating the other lexical information of the root form) and phrases run on to single lemma entries. For example, "**(as) happy as a sandboy** (or **Larry** or **a clam**)" under *happy* was converted into six headwords (based on the alternatives indicated by the parentheses), as well as a collocation pattern for a sense under *happy*, written "(as|?) ~ as (a sandboy | Larry | a clam)", with the tilde marking the target word and the question mark indicating a null.

NODE was then subjected to definition processing and parsing. Definition processing consists of further expansion of the print dictionary: (1) grabbing the definitions of cross-references and (2) assigning parts of speech to phrases based on analysis of their definitions. Definition parsing puts the definition into a sentence frame appropriate to the part of speech, making use of typical subjects, objects, and modificands. The sentence parse tree was then analyzed to extract various semantic relations, including synonyms, superordinates or hypernyms, holonyms, meronyms, satellites, telic roles, and frame elements, with these elements added to the dictionary. After parsing was completed, a phrase dictionary was also created for NODE.[2]

2. Disambiguation Techniques

The SENSEVAL tasks were run separately against the WordNet and NODE sense inventories as described above, with the WordNet results submitted. The lexical sample and all-words texts were modified slightly. Satellite tags were removed and entity references were converted to an ASCII character. In the all-words texts, contraction and quotation mark discontinuities were undone.

[2]WordNet definitions were not parsed. In an experiment, the semantic relations identifiable through parsing were frequently inconsistent with those already given in WordNet, so it was decided not to confound the disambiguation.

These changes made the texts more like normal text processing conditions.

The texts were next reduced to sentences. For the lexical sample, a sentence was assumed to consist of the last single line in the text (not always the case). For the all-words texts, a sentence splitter identified the sentences, which were next submitted to the parser. The DIMAP parser produced a parse tree for each sentence, with bottom-up constituent phrases when the sentence was not parsable with the grammar, allowing the WSD phase to continue.

The first step in the WSD used the part of speech of the tagged word to select the appropriate sense inventory. Nouns, verbs, and adjectives were looked up in the phrase dictionary; if the tagged word was part of an MWU, the word was changed to the MWU and the MWU's sense inventory was used instead.

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" (only for NODE), by tallying subject labels for all senses of all content words in the context.

Each sense of the target was then evaluated. Senses in a different part of speech were dropped from consideration. The different pieces of information in the sense were assessed: 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 in the dictionary was selected.[3]

[3]Several other functions were implemented only in stub form at the time of the test runs, to evaluate: type restrictions (e.g., transitivity), presence of accompanying grammatical constituents (e.g., infinitive phrase or

Collocation pattern testing (requiring an exact match with surrounding text) was given the largest number of points (10), sufficient in general to dominate sense selection. Contextual clue words (a particle or preposition) was given a small score (2 points). Each content word of the context added two points if present in the sense's definition or examples, so that considerable overlap could become quite significant. For topic testing, a sense having a subject label matching one of the context topic areas was awarded one point for each word in the context that had a similar subject label (e.g., if four words in the context had a medical subject label, four points would be awarded if the instant sense also had a medical label).

3. SENSEVAL-2 Results

As shown in Table 1, using WordNet as the disambiguation dictionary resulted in an official precision of 0.293 at the fine-grained level and 0.367 at the coarse-grained level. The official results were actually recall, since our system erroneously generated a result in cases where it should not have (such as the cases where the assumption about the last sentence containing the tagged word was not true); the actual precision was 0.311 and 0.390, respectively. Since CL Research did not use the training data in any way, running the training data also provided another test of the system. The results are remarkably consistent, both overall and for each part of speech. Since submission, various other bug fixes and changes to the disambiguation routines has increased the precision to 0.340 and 0.429 for the two grains. It is expected that further implementation of stub routines will increase our results, although it is not clear whether we can attain the 0.67 coarse-grained precision attained in SENSEVAL-1 (Litkowski, 2000). (Kilgarriff,

complements), form restrictions (such as number and participial), grammatical role (e.g., as a modifier), and selectional restrictions (such as subject, object, modificand, and internal arguments).

2001) suggests that use of WordNet as the sense inventory could have affected performance by about 0.14.

Using NODE as the disambiguation dictionary and mapping its senses into WordNet senses achieved comparable levels of precision, although recall was somewhat lower, as indicated by the difference in the number of items on which the precision was calculated. Senses were identified for all instances using NODE, but only about 75% of the senses were mapped into WordNet in the first run immediately after the official WordNet submission. Subsequent improvements in the mapping (described in the next section) have improved recall to about 87.5% without changing precision.

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
WordNet Test (R)	768	0.422	0.422	1726	0.397	0.503	1834	0.252	0.362	4328	0.340	0.429
NODE Test	420	0.288	0.288	1403	0.402	0.539	1394	0.219	0.305	3217	0.308	0.405
NODE Test (R)	636	0.434	0.434	1568	0.393	0.520	1589	0.198	0.273	3793	0.318	0.402
WordNet Training	1533	0.365	0.365	3455	0.334	0.444	3623	0.219	0.299	8611	0.291	0.369
NODE Training	864	0.116	0.116	2848	0.366	0.483	2567	0.227	0.315	6249	0.276	0.365

For the all-words task, the disambiguation results were significantly higher than for the lexical sample, with a precision (and recall) of 0.460 for the WordNet coarse-grained level. For NODE, about 70% were mapped into WordNet (indicated by the reduced number of items), with precision on the mapped items only slightly less.[4]

Table 2. All-Words

Run	Items	Fine	Coarse
WordNet	2473	0.451	0.460
NODE	1727	0.416	0.418

[4]For both tasks, NODE senses were identified for all words, but could be mapped only for the percentages given.

4. Mapping Procedures and Explorations

To investigate the viability of mapping for WSD, subdictionaries were created for each of the lexical sample words and for each of the all-words texts. For the lexical sample words, the subdictionaries consisted of the main word and all entries identifiable from the phrase dictionary for that word. For example, *bar*, in NODE, had 13 entries where **bar** was the first word in an MWU and 50 entries where it was the head noun, compared to 16 and 40, respectively, in WordNet; for *begin*, there was only one entry in each dictionary. For the all-words texts, a list was made of all the task words to be disambiguated (including some phrases) and a subdictionary constructed from this list. For both tasks, the creation of these subdictionaries was fully automatic; no hand manipulation was involved, except for cases in the all-words entries where verb roots were added when only an inflected form appeared in the text.

Table 3. Lexical Sample Entries		
Part of Speech	**WordNet**	**NODE**
Adjectives	245	244
Nouns	648	491
Verbs	469	987
Total	1362	1722

Table 4. All-Words Entries		
Text	**WordNet**	**NODE**
d00	493	444
d01	496	439
d02	498	454
Total	1487	1356

As shown in Table 3, there were considerable differences in the number of entry words between

WordNet and NODE. Furthermore, there were some discrepancies between the number of entries for individual lexical sample words even in WordNet compared to the official number of entries (Kilgarriff, 2001). For example, the entry list for *day* constructed using our methods found 136 MWUs compared to 82 in the official inventory.

The 73 dictionaries for the lexical sample words gave rise to 1372 WordNet entries and 1722 NODE entries.[5] Only 491 entries were common (i.e., no mappings were available for the remaining 1231 NODE entries); 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. The *a priori* probability of successful mapping into the appropriate WordNet sense is 0.064, the baseline for assessing WSD via another dictionary mapped into the WordNet sense-tagged keys.[6] Even though this assumes a clearly incorrect equal likelihood of each sense appearing in the lexical sample, the low probability discouraged submitting a run using NODE mapped into WordNet.

4.1 Basic Mapping Procedures

The NODE dictionaries were mapped into the WordNet dictionaries using a selection in DIMAP for comparing definitions. This functionality allows several options. The mapping can be performed for entire dictionaries or for selected entries; the mapping can be saved to a file, where it was used with perl scripts that took the NODE WSD results and mapped them into their

[5]Entries included all parts of speech; disambiguation was required to identify the part of speech as well.

[6]Note that a mapping from WordNet to NODE is likely to generate similar mismatch statistics.

corresponding WordNet senses in the SENSEVAL-2 format, suitable for scoring against the keys. The mapping functionality uses three measures of fit: (1) word overlap between definitions (with or without a stop list to exclude frequent words and using exact matches rather than reducing inflected forms to their root forms), (2) a componential analysis that examines where a given sense fits within a semantic network (i.e., the hypernyms and other semantic relations in WordNet and the comparable relations generated during processing and parsing NODE definitions), and (3) the edit distance between two definitions (i.e., the number of insertions, deletions, and replacements necessary to convert one definition into another). (See (Litkowski, 1999) for more details on the first two methods and (Sierra & McNaught, 2000) for details on aligning definitions using edit distance, which has been implemented in DIMAP.) Edit distance was not used in the NODE to WordNet mapping.

The methods from (Litkowski, 1999) were refined considerably in working with a dictionary publisher (Macquarie Pty Ltd of Australia) who has a set of 15 dictionaries derived from its main dictionary (The Macquarie Dictionary, 1997). These dictionaries, ranging from thumbnail to children's to junior and concise versions, were developed at various times over the past 20 years, initially based on previous versions of the main dictionary. Because changes to the main dictionary were not always filtered down to the smaller dictionaries, there was some drift in the phraseology over time. We used our mapping functionality to create links between these smaller dictionaries and the main dictionary so that a tighter linkage could be maintained. We were able to develop the functionality to achieve mappings considered sufficiently accurate (around 90 percent) to map all the dictionaries in preparation for editorial work (Tardif, 2000), using word overlap with a stop list.

Subsequently, we combined the componential analysis method with the word overlap method. The extension, however, is limited to cases where no mapping is generated by the word overlap method, so that we have cases where a mapping is achieved by componential analysis alone. This was the core functionality that was used in our initial mapping between NODE and WordNet, performed immediately after we had submitted our official SENSEVAL-2 runs.

Table 5. Lexical Sample Sense Mappings

Part of Speech	Total	Mapped	Percent
Adjectives	429	215	50.1%
Nouns	652	421	64.6%
Verbs	1679	1134	67.5%
Total	2760	1770	64.1%

Table 6. All-Words Sense Mappings

Text	Total	Mapped	Percent
d00	7195	3811	53.0%
d01	6611	3368	50.9%
d02	6175	3027	49.0%
Total	19981	10206	51.1%

Tables 5 and 6 show how many senses of the common entries were mapped from NODE to WordNet. The much larger number of senses in the all-words case, where the number of total entries was about the same as the lexical sample case, reflects the fact that many of the words were very common words with many senses (e.g., *make* and *give*) and in fact, were also present in each of the three texts. Of the 1770 mappings for the lexical sample senses, 39% were based on the word overlap alone, 20.5% on the componential analysis alone, 16.7% used word overlap and were confirmed by the componential analysis, and 23.7% used word overlap but were disconfirmed by the componential analysis. No attempt has yet been made to determine the accuracy of the mappings. Cursory inspection suggests that the accuracy is much less than 100 percent. The large percentage where word overlap and componential analysis disagreed may be indicative, but there are cases where there is agreement from both methods and hand mapping would provide a different mapping.

For WordNet, there were 2516 senses in the entries that had been created (15.6% adjectives, 33.3% nouns, and 51.1% verbs). Thus, at least 746 senses were inaccessible; in fact, the number is somewhat more, since not infrequently, several NODE senses mapped into a single WordNet sense, particularly among the verbs.

Many of these mappings are not directly relevant to the lexical sample task. As indicated above, the subdictionary creation made no distinction as to part of speech. For example, the entries for the *work* subdictionary included noun senses for the primary word and included entries where *work* was a noun constituent of an MWU (such as *social work* or *work permit*). While the absolute number of these entries and senses that fall outside the part of speech of the individual tasks has not been developed, these situations apply to both WordNet and NODE and it is unlikely that the percentages quoted above would change dramatically.

Considering these mapping statistics, with many WordNet entries and senses inaccessible and many mappings likely to be incorrect, the initial results that were achieved seem quite surprising. The recall of 75 percent for the lexical sample task and 70 percent for the all-words task is a direct reflection of an inability to map either an entry or a sense that resulted from disambiguation using NODE. The precision of 40 percent also reflects inaccurate mappings and so would likely be improved quite a bit through hand manipulation of the mappings. The remaining precision would then be attributable to failure of our disambiguation. On the other hand, the fact that we were able to achieve a level of precision comparable to what was attained using WordNet suggests the most frequent senses of the lexical sample words were able to be disambiguated and mapped correctly into WordNet.

4.2 Mapping at the Entry Level

The significant discrepancy between the entries (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. Many entries arise from the needs of placing concepts into a semantic net that are not actually realized in common language (e.g., the synset {*animality, animal nature*} as a hyponym of *nature* or *natural event*), where it is unlikely that *animal nature* would occur in normal language use. Since WordNet used freely available sources, many of these may be outdated or may have been compiled from highly technical sources and not reflect current or common usage (e.g., *free pardon* or the 14 varieties of *yew*). WordNet may identify several types of an entity (e.g., *apricot bar*, *nougat bar*, and *chocolate 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.

NODE, on the other hand, is based more on lexicographic principles and also has a strong corpus base. Thus, there is an entry for *heel bar* ("a small shop or stall where shoes are repaired, especially while the customer waits”), which would not be productively formed by any lexical rule. More significantly, this difference is reflected in idiomatic verb phrases, which account for many of the entries in NODE not in WordNet. 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*).

These discrepancies between WordNet and NODE are not unique to SENSEVAL-2, but would exist when using any two sense inventories, so it would be useful to understand their implications and methods for dealing with them. For MWU entries that do not reflect common usage or have been created to provide nodes in a semantic network, no problem is likely to arise in WSD, since real-world language use is unlikely to arise. The fact that such entries are inaccessible in the NODE mapping has no effect on WSD in NODE or WordNet.

For MWUs in WordNet that are inaccessible because they are not present in the source dictionary (such as for *yew* or *bar*), a correct disambiguation in NODE is not given credit. This situation is actually a matter of grain: *Western yew* is an instance of *yew* and this could be captured in the answer key either directly by making the appropriate sense of *yew* one of the answers in the key (followed in a couple of instances, e.g., for *common sense*), or preferably, by identifying *Western yew* as a subsense of *yew* in the sense hierarchies so that at the coarse grain, *yew* would be a correct answer. In general, this suggests that MWUs in a dictionary should pay particular attention to seeing if the MWU is an instance of its head (usually by seeing if the head is the genus term of the MWU's definition). The potential number of these cases in SENSEVAL-2 was less than 100, affecting answers for *bar*, *chair*, *channel*, *church*, *circuit*, *day*, *facility*, *holiday*, *material*, and a few other isolated cases. The effect on our scores would have quite small, since some of these MWUs were also in NODE.

The reverse situation, an MWU in NODE that is not present in WordNet, occurred much more often and has a definite effect on our scores. In these cases, since there is no mapping into WordNet, an answer is not generated; these cases are a subset of our NODE disambiguated answers that lessened our recall (1111 cases in the original run with NODE). For example, 9 of 15 unknown answers for *nature* recognized phrases in the NODE disambiguation (*human nature*, *nature trail*, *by nature*); these answers are correctly disambiguated in NODE (usually with such MWUs having only one sense). Some of these cases (about 50) are due to the corresponding phrases not having been identified by the SENSEVAL-2 lexicographers (e.g., *by nature* and *United Nations*), but there are at least several hundred such MWUs recognized when using NODE as the sense inventory. Thus, in the first instance, we can say that our WSD was probably about 5 to 10 points higher than our NODE to WordNet score. Further, since we had several bugs in connection with our phrase recognition routines, we can expect a further few points improvement in this area.

In the second instance, we would like to be able to do more than just make a claim that our disambiguation would have been higher, since the problem of differing MWU sets is always going to be prevalent and problematic. For a phrase like *nature trail*, it is possible to decompose the phrase for analysis. In this solution, the definition ("a signposted path through the countryside designed to draw attention to natural features") would be examined for its correspondence to both constituent words. One NODE definition of *nature* ("the countryside, especially when picturesque") would have a strong match with the definition of *nature trail* and a definition of *trail* ("a beaten path through rough country such as a wood or moor") would match. In this case, since we are interested in mapping *nature*, we would use the mapping for the identified sense as the basis for identifying a WordNet sense. This would yield the correct answer. (One of the SENSEVAL-2 lexicographers raised a query about the feasibility of this approach, "Even in the case of genuine compounds, the dictionary should further identify where one or other element can also be assigned to a main sense (eg. 'natural history' to sense 1 of 'natural'), as that should aid the algorithm's global 'understanding' of the context." (Williams, 2001)) We have not explored the extent to which our example would generalize; however, the mapping approach outlined here does seem worthy of further exploration.

The case with verb phrases may be somewhat more problematic in devising a successful mapping strategy, particularly when they have become so idiomatic as to have lost any tie to the words that comprise them. For the phrase *keep one's mouth shut* ("not say anything, especially not reveal a secret"), there are some ties to various senses of *mouth* as an organ of speech that may give a clue to an appropriate sense in the target dictionary. However, in general, this is not the case.

4.3 Mapping at the Sense Level

Similar issues arise in mapping individual senses: WordNet senses may be inaccessible since there is no NODE sense that maps into them and NODE senses may not map into any WordNet sense.

Additional complexities arise when the mapping is incorrect; in these cases, the identified WordNet answers are wrong, even when the disambiguation in NODE may have been correct.

We examined our mappings in detail to see where improvements might be possible. We found a bug in our mapping routine for adjectives that did not pick up properly those mappings where an adjective sense was a satellite of another adjective; this accounted for a considerable portion of our increase in recall (from 55% to 83%) and precision (from 0.288 to 0.434) in the revised run shown in Table 1. We were able to make several other changes that improved the overall recall from 74.3% to 87.6%.

We observed many cases where there was only one sense for an entry in WordNet, but the mapping did not yield any hits based on the word overlap or the componential analysis. Almost all these cases were for MWUs; these changes improved recall. As noted earlier in describing the entry for *happy*, many definitions in NODE had an associated collocation pattern. We had not taken this into account in our initial mapping. We modified our mappings for these senses since they required that the exact pattern be present in WordNet. In a large number of cases, this actually removed mappings that had been based on the headword. For example, under *call*, one sense required the collocation *call collect*; the original mapping picked a sense of *call* in WordNet that pertained to calling by telephone; this mapping was undone. The review of these cases led to the removal, the addition, or the revision in the mapping. It is difficult to assess the overall effect of these changes. In some cases, it appears that a removal or revision may have changed the WordNet sense from a correct one to an incorrect one. In general, though, there was an improvement in recall, but it appears that these instances may account for the slightly lower precision.

Another change involved cases where no mapping had been made. NODE has a shallow hierarchy, so that there is a main sense and perhaps several subsenses. Although the wording of the subsenses may bear no clear relation to the supersense, we changed the “no mapping” to the mapping of the supersense when the supersense had a mapping. This is analogous to using a coarse-grained answer. In some instances, where a supersense had no mapping but had only one subsense or all its subsenses had the same mapping, we used this mapping for the supersense as well. Overall, this had the effect of improving recall and precision.

We experimented only a little with hand-mapping. Mapping for adjectives was particularly difficult, since their definitions were very short and hence having a lower likelihood of finding exact wording in WordNet. For *simple*, only 1 of 12 NODE definitions was mapped into one of the 7 WordNet definitions. Our initial recall was only 16 out of 66 and this one mapping had the bug for adjectives mentioned above, so our precision was 0.00. After correcting the bug and making hand-mappings, our recall went to 65 out of 66 and our precision increased to 0.258. For *material*, the primary sense was not mapped; our initial recall was only 25 of 69, with coarse-grained precision of 0.232; making this one change in the map increased recall to 100 percent and precision to 0.594. These cases are strongly indicative that our disambiguation with NODE was much higher than after the mapping.

Of most significance to the sense mapping is the classical problem of **lumping** (attaching more significance to similarities than to differences) and **splitting** (attaching more importance to to differences than to similarities). A single sense in NODE may correspond to several senses in WordNet (e.g., NODE has one sense of *yew* for both the tree and the wood, while WordNet has two); several senses in NODE may correspond to a single sense in WordNet. The latter problem generally does not affect our results, since it will be scored as correct regardless of which sense in NODE is identified. When an 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; if the definitions in WordNet have been related hierarchically, we may receive credit at the coarse grain, but not at the fine grain. Otherwise, it would be necessary for the lexicographer to have tagged more than one sense as being correct (e.g., when it appeared that both senses of *yew* were activated by a given context). In the case of *graceful*, there was only one sense in NODE ("having or showing grace or elegance") that did not map into one of the two senses in WordNet ("characterized by beauty of movement, style, form etc.; not awkward" and "suggesting taste, ease, and wealth"). As a result, despite having disambiguated correctly in NODE, our precision and recall in WordNet was 0.00. Choosing the first sense in WordNet would give a precision of 0.793; choosing the second sense 0.310; and choosing both senses 0.552 (in three instances, the lexicographers chose both senses as correct).

To examine this issue in more detail, we performed a mapping from WordNet to NODE for the word *develop*. NODE has 8 definitions and WordNet has 21. The first definition in NODE is a supersense with 5 subsenses; the other two are major senses. We were able to map 7 of the 8 senses into WordNet. In WordNet, there are 8 major senses, with one having 6 subsenses and another 4 subsenses. We were able to map 13 of the 21 senses into NODE. By performing this reverse mapping, we are able to identify all senses of WordNet that had positive scores in the mapping from NODE to WordNet, rather than just the one that was selected. Thus, instead of using only a single sense for the mapping, we can use all WordNet senses that mapped into the same NODE sense. In this way, the first sense of NODE was mapped into 4 WordNet senses, another sense was mapped into 3 WordNet senses, two other senses mapped into 2 WordNet senses, three others into 1 WordNet sense, and the final sense was hand-mapped into another WordNet sense that had not been mapped into any NODE sense. Since our mapping process is generally intended to operate on the basis of identifying components of meaning, the reverse mapping allows us to group senses together that have similar meaning potentials (see (Hanks, 2000) for further details).

Using this one-to-many mapping, the process begins by disambiguating the lexical sample for *develop* with NODE to identify a single sense located within the NODE hiearchy. If the selected sense maps into more than one WordNet sense, the multiple WordNet senses are returned as the answers to be judged against the answer key. When, for example, 4 WordNet senses are given as the answer, each is assumed to have a weight of 0.25 (unless specifically given different weights). At the fine-grained level, we would receive a score of only 0.25 if one of the multiple answers is correct, whereas if we had only a single answer that is correct, we would receive 1.00 (see (Kilgarriff & Rosenzweig, 2000) for SENSEVAL scoring). By diluting our answer, we receive a lower score when we have the correct sense; however, when we have a wrong answer, we may receive partial credit for having multiple answers. At the coarse grain, with multiple answers, we are more likely to hit upon one of the correct answers, even though our score for each may be diluted somewhat. Using WordNet, we had 13 and 24 out of 69 instances correct at the two grains; using NODE mapped into WordNet with only a single sense chosen, we had 3 and 12 correct; using NODE with multiple WordNet senses, our scores were 3.5 and 11.1. Tentatively, then, it appears as if this method of activating multiple senses does not make a significant difference, but this conclusion warrants further investigation.

5. Discussion and Conclusions

In general, mapping from NODE to WordNet has been shown to be viable. The availability of a large reference set (the SENSEVAL-2 corpora) has enabled the investigation of disambiguation with another sense inventory. With scores based on mapping that are comparable to those achieved using WordNet as the sense inventory, we can be confident that our disambiguation with this other sense inventory is better, perhaps significantly so.

We can also be confident that improving our disambiguation with this other sense inventory is likely to achieve even better results. As mentioned earlier, we were unable to implement many routines because of time constraints. Many of these routines are intended to take advantage of detailed lexical information contained in NODE. As we develop these routines, we can use our existing mapping to convert our disambiguations in NODE into WordNet senses and know that any improvements in our scores will be legitimate. As indicated above, these routines will examine type restrictions (e.g., transitivity), presence of accompanying grammatical constituents (e.g., infinitive phrase or complements), form restrictions (such as number and participial), grammatical role (e.g., as a modifier), and selectional restrictions (such as subject, object, modificand, and internal arguments). Much of this information is either not available in WordNet, available only in an unstructured way, only implicitly present, or even inconsistently. (Delfs, 2001) describes a sense for *begin* that has an infinitive complement, but it is present only in an example sentence and not explicitly encoded with a verb frame commonly used in WordNet. 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 NODE, it is not uncommon for a verb sense to be labelled explicitly as having both a transitive and an intransitive realization. However, in implementing further 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.

Having demonstrated the feasibility of mapping for WSD, it is possible to examine many issues in comparing entries and definitions across lexical resources. Most importantly, using SENSEVAL-2 data, it is possible to gauge changes in mapping procedures by the changes in the WSD results. We have examined several aspects of the mapping, and many more possible avenues of investigation are available. We can examine reasons for failure at both the entry and sense levels and cases where the componential analysis method gave results different from word overlap mehtod. We can explore the effect of using or not using stop lists, reducing inflected forms to their root forms, and the relative weighting of different methods (including edit distance, which was not used in the mapping). For example, by aligning definitions based on edit distance, it is possible to examine substitution of synonyms or alternative phraseologies.

6. Other Research

(Atkins, 1991) characterizes the difficulty in comparing lexical resources. She identifies some problems with current dictionary definitions, particularly lumping versus splitting, similar sense overlaps, ambiguity of components, and fuzzy sense boundaries. She suggests approaches to sense differentiation and identifies elements that need to be accounted for, including modulation (the way in which a sense is modified by context), regular polysemy (presence of lexical rules), and use of a hierarchy within the senses of a single word. She developed an experimental framework for implementing these notions within a single dictionary in the Hector project and data (Atkins, 1993), used in SENSEVAL-1 and serving as a precursor to the development of NODE (Hanks, 2001). The FrameNet project (Fillmore & Baker, 2001) also traces its roots in part to this work, specifically in the characterization of meaning components, implemented in frame semantics. Our efforts at mapping between lexical resources is strongly motivated by this earlier work, specifically attempting to use characterizations of components of meaning and the placement of an entry and sense within a sense hierarchy and semantic network as the basis for mapping. While these characterizations are still at an early stage, they have now proved sufficient for use in large-scale NLP applications.

At a recent WordNet workshop on resource integration, (Daude, et al., 2001), (Green, et al., 2001), (Burgun & Bodenreider, 2001), and (Asanoma, 2001) present various investigations for mapping between various lexical resources. Our algorithms are most similar to those of (Daude, et al., 2001) and (Asanoma, 2001). We investigated our mapping between WordNet 1.6 and WordNet 1.7 for one of subdictionaries created here (*call*, with 46 entries). As indicated earlier, our mapping first examines definitions (glosses) and then the structural relationships. Our mappings were quite robust, much stronger than for NODE to WordNet, as might be expected since the changes between the two versions of WordNet would be expected to be small relative to the overall size. Specifically, our mapping identified five additional entries (*distress call*, *call attention*, *call into question*, *call the shots*, and *call the tune*). There were 103 senses, of which 101 were mapped and 2 senses identified that were not in WordNet 1.6. Almost 90 percent of the senses mapped identically both in their glosses and in their positions in the hierarchy. The remainder were disconfirmed by our structural mapping, indicating that these senses had only been moved in their hierarchical position between WordNet 1.6 and WordNet 1.7. We expect that these results would hold in general, indicating that our mappings between the two versions (or any applications where a conversion from one version to the other would be important) would be highly reliable.

Summary

Using the gold standard sense-disambiguated data set provided by SENSEVAL-2, we have shown how it is possible to achieve wider generality in the use of lexical resources rather than relying solely on WordNet. It is first necessary to carefully prepare the lexical resources that will be used, so that they may be mapped successfully into WordNet. Analysis of the disambiguation results after the mapping can provide of the source of failures, whether they reside in the disambiguation itself or in the mapping. Such an analysis provides a deeper understanding of the disambiguation process.

In addition, the mappings themselves provide a rich source of understanding about lexical semantics when examined within the context of a specific task, namely, disambiguation. We have been able to examine more closely semantic issues that have long been the source of much difficulty. Since our methods have shown many avenues for further exploration, we can expect to characterize these issues even better.

Acknowledgements

I wish to thank Oxford University Press for making NODE available (Patrick Hanks and Rob Scriven) and for many useful discussions (Glynnis Chantrell and Judy Pearsall).

References

Asanoma, N. (2001, June 3-4). Alignment of Ontologies: WordNet and Goi-Taikei. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. NAACL 2001 SIGLEX Workshop. Pittsburgh, PA: Association for Computational Linguistics.

Atkins, B. T. S. (1991). Building a lexicon: The contribution of lexicography. *International Journal of Lexicography, 4*(3), 167-204.

Atkins, S. (1993). Tools for computer-aided lexicography: The Hector project. In *Papers in Computational Lexicography*. COMPLEX '93. Budapest.

Burgun, A., & Bodenreider, O. (2001, June 3-4). Comparing terms, concepts and semantic classes in WordNet and the Unified Medical Language System. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. NAACL 2001 SIGLEX Workshop. Pittsburgh, PA: Association for Computational Linguistics.

Daude, J., Padro, L., & Rigau, G. (2001, June 3-4). A Complete Wn 1.5 to Wn 1.6 Mapping. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. NAACL 2001 SIGLEX Workshop. Pittsburgh, PA: Association for Computational Linguistics.

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

Fellbaum, C. (1998). *WordNet: An electronic lexical database.* Cambridge, Massachusetts: MIT Press.

Fillmore, C. J., & Baker, C. F. (2001, June 3-4). Frame Semantics for Text Understanding. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. NAACL 2001 SIGLEX Workshop. Pittsburgh, PA: Association for Computational Linguistics.

Green, R., Pearl, L., Dorr, B. J., & Resnik, P. (2001, June 3-4). Lexical Resource Integration across the Syntax-Semantics Interface. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. NAACL 2001 SIGLEX Workshop. Pittsburgh, PA: Association for Computational Linguistics.

Hanks, P. (2000). Do Word Meanings Exist? *Computers and the Humanities, 34*(1-2), 205-15.

Hanks, P. (2001, May 29). Word Sense Disambiguation.

Ide, N., & Veronis, J. (1993). Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time? KB&KS93. Tokyo.

Kilgarriff, A. (1997). "I Don't Believe in Word Senses." *Computers and the Humanities, 31*(2), 91-113.

Kilgarriff, A. (2001, July). English Lexical Sample Task Description. In *SENSEVAL-2*. Association for Computational Linguistics SIGLEX Workshop. Toulouse, France.

Kilgarriff, A., & Rosenzweig, J. (2000). Frameword and Results for English SENSEVAL. *Computers and the Humanities, 34*(1-2), 15-48.

Litkowski, K. C. (2000). SENSEVAL: The CL Research Experience. *Computers and the Humanities, 34*(1-2), 153-158.

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

Litkowski, K. C. (1999, 21-22 June). Towards a Meaning-Full Comparison of Lexical Resources. Association for Computational Linguistics Special Interest Group on the Lexicon Workshop. College Park, MD.

The Macquarie Dictionary (A. Delbridge, J. R. L. Bernard, D. Blair, S. Butler, P. Peters, & C. Yallop, Eds.) (3rd). (1997). Australia: The Macquarie Library Pty Ltd.

Miller, G. A. (2001, June). *Toward WordNet 2*. WordNet and Other Lexical Resources: Applications, Extensions and Customizations. NAACL 2001 Workshop. Pittsburgh, PA.

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

Richardson, S. D. (1997). Determining similarity and inferring relations in a lexical knowledge base [Diss], New York, NY: The City University of New York.

Sierra, G., & McNaught, J. (2000). Extracting Semantic Clusters from MRDs for an Onomasiological Search Dictionary. *International Journal of Lexicography, 13*(4), 264-86.

Tardif, R. (2000, Oct. 20). Mapped databases.

Williams, J. (2001, June 29). The Hector data.