Use of Machine Readable Dictionaries for Word-Sense Disambiguation in SENSEVAL-2

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

Abstract

CL Research's word-sense disambiguation (WSD) system is part of the DIMAP dictionary software, designed to use any full dictionary as the basis for unsupervised disambiguation. Official SENSEVAL-2 results were generated using WordNet, and separately using the New Oxford Dictionary of English (NODE). The disambiguation functionality exploits whatever information is made available by the lexical database. Special routines examined multiword units and contextual clues (both collocations, definition and example content words, and subject matter analyses); syntactic constraints have not yet been employed. The official coarse-grained precision was 0.367 for the lexical sample task and 0.460 for the all-words task (these are actually recall, with actual precision of 0.390 and 0.506 for the two tasks). NODE definitions were automatically mapped into WordNet, with precision of 0.405 and 0.418 on 75% and 70% mapping for the lexical sample and all-words tasks, respectively, comparable to WordNet. Bug fixes and implementation of incomplete routines have increased the precision for the lexical sample to 0.429 (with many improvements still likely).

Introduction

CL Research's participation in SENSEVAL-2 was designed to (1) extend WSD techniques from SENSEVAL-1 (Litkowski, 2000), (2) generalize WSD mechanisms to rely on a full dictionary rather than a small set of entries where individual crafting might intrude, and (3) investigate WSD using one dictionary mapped into another (WordNet). Results indicate positive achievements for each of these goals. Time constraints precluded a complete assessment of the upper limits that can be achieved. In particular, although the general architecture from SENSEVAL-1 was retained, several specific WSD routines were not reimplemented. Incomplete testing, debugging, and implementation of new routines significantly affected the official results. Several of these problems are investigated more fully below.

CL Research's WSD functionality is implemented in DIMAP[1], designed primarily for creation and maintenance of lexicons for natural language processing. In particular, DIMAP is designed to make machine-readable dictionaries (MRDs) tractable and to create semantic networks (similar to WordNet (Fellbaum, 1998) and MindNet (Richardson, 1997)) automatically by analyzing and parsing definitions. Section 1 describes the dictionary preparation techniques for WordNet and NODE (The New Oxford Dictionary of English, 1998), as well as the mapping from NODE to WordNet. Section 2 describes the WSD techniques used in SENSEVAL-2. Section 3 describes the SENSEVAL-2 results and section 4 discusses these results..

1 Dictionary Preparation

DIMAP can disambiguate any text against WordNet or any other dictionary converted to DIMAP, with a special emphasis on corpus instances for specific lemmas. 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,

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

arbitrary other semantic relations, and feature structures containing arbitrary information.

WordNet is already integrated in DIMAP in several ways, but for SENSEVAL-2, WordNet 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.

A separate "phrase" dictionary was constructed from all noun and verb multiword units (MWUs), using WordNet's sense index file. For nouns, 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 verbs, 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 and more formal structure (e.g., lexical preferences, grammar fields, and subsensing). 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. E.g., "**(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.

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 the superordinate or hypernym, holonyms, meronyms, satellites, telic roles, and frame elements. After parsing was completed, a phrase dictionary was also created for NODE.[2]

The SENSEVAL tasks were run separately 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 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 *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.) 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.

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.[3] 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

[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.

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

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.[4]

2 Disambiguation Techniques

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. 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 a single line. 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 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). Subject labels for all senses of all content words in the context were tallied.

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.[5]

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 Results

As shown in Table 1, using WordNet as the disambiguation dictionary resulted in an overall precision (and recall) of 0.293 at the fine-grained level and 0.367 at the coarse-grained level. 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. 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. Overall, about 75% of the senses were mapped into WordNet.

Table 2. All-Words

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

For the all-words task, the disambiguation results

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

[5] 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 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).

Table 1. Lexical Sample Precision

Run	Adjectives			Nouns			Verbs			Total		
	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 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

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.[6]

4 Discussion

Because of the usual bugs and incomplete implementation, the official results do not adequately indicate the potential of our approach. The official results are actually recall rather than precision, since an answer was submitted when it shouldn't have been, as distinguished from cases where the parser picked the wrong part of speech or was unable to select a sense. The actual precision for the lexical sample task is 0.311 for the fine grain and 0.390 for the coarse grain, and for the all-words task, 0.496 and 0.506 for fine and coarse grains, respectively.

Minimal debugging and inability to implement several routines significantly affected the scores. Examining the reasons for failures in the test runs and making program fixes has thus far resulted in increasing precision (and recall) to 0.340 and 0.429 for the lexical sample. Further improvements are likely, although it is not clear whether the SENSEVAL-1 precision of 0.67 is achievable using only the information available in WordNet.

It is more likely that using NODE will achieve better results. Improvements in automatic mapping have now reached 90% mapping; it is also relatively easy to make manual adjustments to the maps to achieve even higher performance from the lexicographically-based lexical resource. Since the automatic mapping is inaccurate to an unknown degree (perhaps 25-30%), improving the maps will achieve better results using NODE via WordNet, rather than WordNet alone. Using NODE also provides a much richer set of data upon which to make improvements in WSD. Finally, since NODE is lexicographically-based and with an arguably better sense inventory, we are confident that our WSD would have scored much higher if the taggers had used this inventory.

Conclusion

Given the very preliminary implementation of the disambiguation routines and lack of adequate debugging, the results indicate that using MRDs (and even mapping from one into another) shows considerable potential for unsupervised and general word-sense disambiguation.

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

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

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.

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

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.

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