Computers and the Humanities **34:** 153–158, 2000.

Senseval: The CL Research Experience

KENNETH C. LITKOWSKI
CL Research, 9208 Gue Road, Damascus, MD 20872, USA (E-mail: ken@clres.com)

Abstract. The CL Research Senseval system was the highest performing system among the "All-words" systems, with an overall fine-grained score of 61.6 percent for precision and 60.5 percent for recall on 98 percent of the 8,448 texts on the revised submission (up by almost 6 and 9 percent from the first). The results were achieved with an almost complete reliance on syntactic behavior, using (1) a robust and fast ATN-style parser producing parse trees with annotations on nodes, (2) DIMAP dictionary creation and maintenance software (after conversion of the Hector dictionary files) to hold dictionary entries, and (3) a strategy for analyzing the parse trees in concert with the dictionary data. Further considerable improvements are possible in the parser, exploitation of the Hector data (and representation of dictionary entries), and the analysis strategy, still with syntactic and collocational data. The Senseval data (the dictionary entries and the corpora) provide an excellent testbed for understanding the sources of failures and for evaluating changes in the CL Research system.

Key words: word-sense disambiguation, Senseval, dictionary software, analysis of parsing output

1. Introduction and Overview

The CL Research Senseval system was developed specifically to respond to the Senseval call, but made use of several existing components and design considerations. The resultant system, however, provides the nucleus for general natural language processing, with considerable opportunities for investigating and integrating additional components to assist word-sense disambiguation (WSD). We describe (1) the general architecture of the CL Research system (the parser, the dictionary components, and the analysis strategy); (2) the Senseval results and observations on the CL Research performance; and (3) opportunities and future directions.

2. The CL Research System

The CL Research system consists of a parser, dictionary creation and maintenance software, and routines to analyze the parser output in light of dictionary entries. In the Senseval categorization, the CL Research system is an "All-words" system (nominally capable of "disambiguating all content words"). We did not actually attempt to disambiguate all content words, only assigning parts of speech to these other words during parsing. A small separate program was used to convert the Hector dictionary data into a form which could be uploaded and used by the

dictionary software. As the analysis strategy evolved during development, some manual adjustments were made to the dictionary entries, but these could have been handled automatically by simple revisions to the original conversion program. Our system could in theory proceed to disambiguate any word for which Hector-style dictionary information is available.

2.1. THE PARSER

The parser used in Senseval (provided by Proximity Technology) is a prototype for a grammar checker. The parser uses an augmented transition network grammar of 350 rules, each consisting of a start state, a condition to be satisfied (either a non-terminal or a lexical category), and an end state. Satisfying a condition may result in an annotation (such as number and case) being added to the growing parse tree. Nodes (and possibly further annotations, such as potential attachment points for prepositional phrases) are added to the parse tree when reaching some end states. The parser is accompanied by an extensible dictionary containing the parts of speech (and frequently other information) associated with each lexical entry. The dictionary information allows for the recognition of phrases (as single entities) and uses 36 different verb government patterns to create dynamic parsing goals and to recognize particles and idioms associated with the verbs. These government patterns follow those used in (Oxford Advanced Learner's Dictionary, 1989).[1]

The parser output consists of bracketed parse trees, with leaf nodes describing the part of speech and lexical entry for each sentence word. Annotations, such as number and tense information, may be included at any node. The parser does not always produce a correct parse, but is very robust since the parse tree is constructed bottom-up from the leaf nodes, making it possible to examine the local context of a word even when the parse is incorrect. The parser produced viable output for almost all the texts in the evaluation corpora, 8443 out of 8448 items (99.94 percent).

2.2. DICTIONARY COMPONENT

The CL Research Senseval system relies on the DIMAP dictionary creation and maintenance software as an adjunct to the parser dictionary. This involved using the existing DIMAP functionality to create dictionary entries from the Hector data (with multiple senses, ability to use phrasal and collocational information, and attribute-value features for capturing information from Hector) and using these entries for examining the parser output. Some features were added by hand using DIMAP, rather than revising the Hector conversion program, in the interests of time; the conversion program can be easily modified to automate the process. These features formed the primary information used in making the sense assignments in Senseval.[2]

2.3. ANALYSIS STRATEGY

The CL Research system is intended to be part of a larger discourse analysis processing system (Litkowski and Harris, 1997). The most significant part of this system for WSD is a lexical cohesion module intended to explore the observation that, even within short texts of 2 or 3 sentences, the words induce a reduced ontology (i.e., a circumscribed portion of a semantic network such as WordNet (Miller et al., 1990) or MindNet (Richardson, 1997)). The implementation in Senseval does not attain this objective, but does provide insights for further development of a lexical cohesion module.

The CL Research system involves: (1) preprocessing the Senseval texts; (2) submitting the sentences to the parser; (3) examining the parse results to identify the appropriate DIMAP entry (relevant only where Hector data gave rise to distinct entries for derived forms and idioms); (4) examining each sense in the DIMAP entry to filter out non-viable senses and adding points to senses that seem preferred based on the surrounding context of a tagged item; and (5) sorting the still viable senses by score to select the answer to be returned.

The DIMAP dictionary contained all Hector senses, phrases, and collocations; step 3 particularly focused on recognizing phrases and collocations and selecting the appropriate DIMAP entry (important, for example, in recognizing Hector senses for *milk shake* and *onion dome*). Step 4 is the largest component of the CL Research system and where the essence of the sense selection is made. In this step, we iterate over the senses of the DIMAP entry, keeping an array of viable senses (each with an accompanying score), examining the features for the sense. The features were first used to filter out inappropriate senses. The parse characteristics of the tagged word were examined and flags set based on the part of speech (such as number for nouns and verbs, whether a noun modified another noun, whether a verb had an object, and whether a verb or adjective was a past tense, past participle, or present participle); these characteristics were sometimes used to retrieve a different DIMAP entry (to get an idiom, for example). The flags were then used in conjunction with the Hector grammar codes to eliminate senses for such reasons as countability of nouns, number mismatch (e.g., when a verb required a plural subject), transitivity incompatibility (an intransitive sense when a verb object was present), tense incompatibility (e.g., if a verb sense could never be passive and the past tense flag was set or when a gerundial was required and not present), when there was no modified noun for a noun-modifier sense, and when an adjective sense was required to be in the superlative form.

The system examined grammar codes indicating that a sense was to be used "with" or "after" a specific word or part of speech; if the condition was satisfied, 3 points were added to the sense's score. Hector clues specifying collocates (e.g., **experience** for *bitter*) were used to add 5 points for a sense; clues specifying semantic classes have not yet been implemented.

The **kind** feature of Hector definitions (e.g., *indie band, jazz band*) was generalized into a quasi- regular-expression recognizer for context preceeding and

Table I. Precision for major tasks

Task	Number of texts	Grain			Attempted
		Fine	Mixed	Coarse	
Overall	8448	61.6	66.0	68.3	98.13
Noun	2756	71.1	75.2	78.6	97.86
Verb	2501	53.5	57.8	59.6	98.44
Adjective	1406	61.7	65.2	69.1	98.15
Indeterminate	1785	58.4	64.0	64.2	98.10

following the tagged word (e.g., "on [prpos] =" to recognize any possessive pronoun for *on one's knees*). Many of the phrasal or idiom entries were transformed manually[3] into **kind** features in DIMAP senses, facilitating idiom recognition or serving as a backup when the parser did not pick up a phrase as an entity. This mechanism was also used for Hector clues that specified particular words or parts of speech. The **kind** features were used as strong indicators in matching a sense. When a **kind** equation was satisfied, any viable senses up to that point were dropped and only senses that satisfied a **kind** equation were then allowed as viable. Overall, this mechanism only added a couple of percentage points; however, for some words with several **kind** equations, the effect was much more significant.

After elimination of senses, the viable senses were sorted by score and the top score was the sense selected. In case of ties (such as when no points were added for any senses), the most frequent sense (as reflected in the Hector order) was chosen.

3. CL Research System Results

Table I shows the CL Research system results for the major Senseval tasks. Since most tasks have a high percent attempted, the recall for each task is only slightly lower (around one percent). The CL Research system was the top performing "All-words" system in both the initial and revised submissions for these major tasks. For the initial submission, precision was 6 percent lower and recall 9 percent lower; this was due to the fact that the percent attempted in the initial submission was 92.74 percent. Thus, most of the improvement between the initial and revised submissions resulted from simply being able to provide a guess for about 400 additional tasks.

For the initial submission, the CL Research system was the best system on 19 of the 41 individual tasks, above average for 12 more, and worst for 2 tasks. Table II shows the CL Research system results for three tasks. For *onion* and *generous*, the results changed little from the initial to the revised submission. For *onion*, the results were at the top for the initial submission and second for the revised submission; for *generous*, the results were only one above the worst performing

Table II. Precision for selected tasks

Task	Number of texts	Grain			Attempted
		Fine	Mixed	Coarse	
Onion-n	214	84.6	84.6	84.6	97.20
Generous-a	227	37.7	37.7	37.7	98.24
Shake-p	356	66.0	68.9	69.8	96.63

system. For *shake*, there was a seven percent increase at the fine-grained level, with the system as the second-best for the initial submission and the top system for the revised submission; a considerable portion of the improvement was the ability to make a guess for an additional 12 percent of the texts between the initial and revised submissions (primarily due to correcting a faulty mechanism for recognizing the phrase *shake up*).

These examples illustrate characteristics of the CL Research system. For *onion*, which has a low entropy (0.86), the high precision is due to the fact that the highest frequency sense is ordered first in the DIMAP dictionary; there was no semantic discrimination in use and the system guessed the first sense. The same is true of *generous*, where, however, the entropy was much higher (2.30). Since, again, the CL Research system had little semantic information, the most frequent sense was guessed in the largest percentage of cases. Because of the higher entropy, the guesses were more often incorrect and the performance of the CL Research system very poor.

For *shake*, there was a much higher entropy (3.70). This might have led to a lower performance, except that there was a considerable amount of additional information in the Hector definitions that permitted sense discrimination. Generally, the system was able to recognize the difference between noun and verb senses. Among the nouns, there were several "kinds" (*milk shake*, *handshake*) that were readily recognized. Among the verbs, the CL Research system was able to recognize a large number of phrases, not only specific idioms (*shake a leg*, *shake off*), but also, through the extension of the "kind" mechanism, phrases that could include optional elements, both specific words and words of a specific part of speech (*shake one's head*, *shake in one's boots*).

4. Discussion and Future Directions

The CL Research system contains many opportunities for improvement. Many of the wrong guesses were due to incorrect parses; we can expect significant improvement in overall results from parser changes. Further, we did not fully exploit the information available in the Hector data; we can expect some improvements in

this area. Finally, we can expect some improvements from semantic processing, working off a semantic network like WordNet or MindNet.

Since the level of WSD was achieved with very little semantics and with likely improvements from further exploitation of the data, the CL Research system results are consistent with the suggestion in (Wilks and Stevenson, 1997) of achieving 86 percent correct tagging from sense frequency ordering, grammar codes, and collocational data. In addition, our data suggest the WSD can be accomplished within small windows (i.e., short surrounding context) of the tagged word. Finally, the Senseval system (the dictionary entries and the corpora) provides an excellent testbed for understanding the sources of failures and for evaluating changes in the CL Research system.

Notes

[1] Source C code (8,000 lines) for the parser, which compiles in several Unix and PC environments, is available upon request from the author, along with 120 pages of documentation.

[2] An experimental version of DIMAP, containing all the functionality used in Senseval, is available for immediate download at http://www.clres.com.

[3] Most of these **kind** equations are amenable to automatic generation, but this was not developed for the current Senseval submission.

References

Litkowski, K.C. and M.D. Harris. *Category Development Using Complete Semantic Networks*, Technical Report 97-01. Gaithersburg, MD: CL Research, 1997.

Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross and K.J. Miller. “Introduction to WordNet: An On-Line Lexical Database”. *International Journal of Lexicography*, 3(4) (1990), 235–244.

Oxford Advanced Learner’s Dictionary, 4th edn. Oxford, England: Oxford University Press, 1989.

Richardson, S.D. *Determining Similarity and Inferring Relations in a Lexical Knowledge Base* [Diss]. New York, NY: The City University of New York, 1997.

Wilks, Y. and M. Stevenson. “Sense Tagging: Semantic Tagging with a Lexicon”. In: *Tagging Text with Lexical Semantics: Why, What, and How?* SIGLEX Workshop. Washington, D.C.: Association for Computational Linguistics, April 4–5 1997.