Explorations in Disambiguation Using XML Text Representation

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

Abstract

In SENSEVAL-3, CL Research participated in four tasks: English all-words, English lexical sample, disambiguation of WordNet glosses, and automatic labeling of semantic roles. This participation was performed within the development of CL Research's Knowledge Management System, which massively tags texts with syntactic, semantic, and discourse characterizations and attributes. This System is fully integrated with CL Research's DIMAP dictionary maintenance software, which provides access to one or more dictionaries for disambiguation and representation. Our core disambiguation functionality, unchanged since SENSEVAL-2, performed at a level comparable to our previous performance. Our participation in the SENSEVAL-3 tasks was concerned primarily with text processing and representation issues and did not advance our disambiguation capabilities.

Introduction

CL Research participated in four SENSEVAL-3 tasks: English all-words, English lexical sample, disambiguation of WordNet glosses, and automatic labeling of semantic roles. We also ran the latter two tasks, but since their test sets were generated blindly, our results did not involve use of any prior information.

Our participation in these tasks is a continuation and extension of our efforts to perform NLP tasks within an integrated text processing system known as the Knowledge Management System (KMS). KMS parses and processes text into an XML representation tagged with syntactic, semantic, and discourse properties. This representation is then used for such tasks as question answering and text summarization (Litkowski, 2004a; Litkowski, 2004b).

The SENSEVAL-3 tasks were performed as part of CL Research's efforts to extend and improve the semantic characterizations in the KMS XML representations. For each SENSEVAL-3 task, the corresponding texts in the test sets were processed using the general KMS functionality. However, since the texts involved in the SENSEVAL tasks were quite small, the amount of processing was quite minimal. The descriptions below focus on the integration of disambiguation technology in a larger system and do not present any advancements in this technology.

1 The SENSEVAL-3 All-Words Task

Our procedures for performing this task and our results were largely unchanged from SENSEVAL-2 (Litkowski, 2001; Litkowski, 2002). Our system is unsupervised, instead relying on information in whatever dictionary is being used to disambiguate the words. In this case, as in SENSEVAL-2, WordNet 1.7.1 was used.

The main types of information used are default sense selection, idiomatic usage, syntactic and semantic clues, subcategorization patterns, word forms, syntactic usage, context, and topics or subject fields. As pointed out in Litkowski (2002), the amount of information available in WordNet is problematic. Additional information suitable for disambiguation is available in WordNet 2.0, but we were unable to test the effect of the changes, even though we could have easily switched our system to use this later version.

In performing this task, we spent some time cleaning the text files, removing extraneous material and creating a more natural text file (e.g., joining contractions). Use of a preprocessed file is somewhat difficult. Since some tokens to be disambiguated were unnatural (e.g., "that's" broken into two tokens, with

only the "'s" to be disambiguated), this affected the quality of our parse output.

After removing extraneous material, KMS parsed and processed the XML source file, treating the text in its ordinary manner. The first step of KMS involves splitting a text into sentences and then parsing each sentence. To customize KMS for this task, we had to create a list of tokens, advancing through this list in concert with the parse output. This process was different from the normal processing of KMS where every word is disambiguated in an integrated fashion. Our results are shown in Table 1, broken down by part of speech as indicated in the answer key.

Table 1. All-Words Results

Run	Items	Precision
Nouns	895	0.523
Verbs	731	0.361
Adjectives	346	0.413
Adverbs	13	0.077
Hyphenated/U	56	0.179
Total	2041	0.434

These results are similar to our performance in Senseval-2, where our precision was 0.451. Our recall is the same, since we attempted each item.

As indicated, several factors degraded our performance, primarily the quality of the information available in the dictionary used for disambiguation. We have not attempted to optimize our system for WordNet, but rather emphasize use of lexicographically-based dictionaries. KMS can use several dictionaries at the same time, and the additional effort to disambiguate against several sense inventories at the same time is not demanding.

Our system's performance was also degraded by a difficulty in advancing through the token list, so that we did not return a sense for 305 items (some of which were due to our parser's performance). We also did not deal properly with the adverbs (most of which were adverbial phrases) and hyphenated words (which we learned about only after downloading the test set).

As indicated in Table 1, our system's performance was lowest for verbs. We believe, based on our earlier studies, that this lower score is affected by the WordNet verb sense inventory.

2 The SENSEVAL-3 Lexical Sample Task

Disambiguation for the lexical sample task is quite similar to that used for the all-words task. The effort is somewhat easier in preparation, since the text for each instance is generally in a form that has not been preprocessed to an extensive degree. Each instance in the test set generally consisted of a paragraph which could be processed immediately within KMS. It was only necessary to modify KMS in a minor way to recognize and keep track of the target word to be disambiguated.

The major difference in the SENSEVAL-3 task from SENSEVAL-2 is the sense inventory. WordNet 1.7.1 was used for nouns and adjective, while Wordsymth provided the verb senses. As indicated above, we were able to use WordNet immediately.

For the Wordsmyth sense inventory, we had to create a new dictionary with CL Research's DIMAP dictionary maintenance software. The Wordsmyth definitions were very uncomplicated, and we were able to create this dictionary quickly after downloading the task training data. On the other hand, the Wordsmyth data is not as rich as would be found in ordinary dictionaries, particularly the machine-readable versions of these dictionaries. Nonetheless, we analyzed the dictionary data to extract nuggets of information about each sense. This included creation of synsets (as in WordNet), identification of the definition proper, creation of examples where provided, identification of "clues" (e.g., "followed by 'to'"), identification of typical subjects and objects, and identification of a sense's topical area. We also used the online version of Wordsmyth to identify the transitivity of each sense.

We ran our system first on the trial data and obtained the results shown in Table 2, essentially using the identical disambiguation routines developed for SENSEVAL-2. We intended to use the training data, not for use as in supervised systems, but to analyze our results using methods we had established for identifying factors significant in disambiguation (Litkowski, 2002). We also briefly investigated the value of using (1) the topical area characterization of preceding sentences, (2) WordNet relations among words in the sentences (including the target), and (3) prepositions following the target in examples. Our investigations indicated that only negligible changes would occur from these possibilities.

Table 2. Lexical Sample Recall (Training)			
Run	**Items**	**Fine**	**Coarse**
Adjectives	314	0.382	0.516
Nouns	3593	0.490	0.561
Verbs	3961	0.409	0.525
Total	7868	0.445	0.541

We compared the results from the training data with our performance in SENSEVAL-2 (Litkowski, 2001). In all categories, the recall was considerably improved, on average about 0.15. This suggests that the lexical sample task for SENSEVAL-3 is much easier. The improvement was relatively greater for verbs, suggesting that the sense inventory for Wordsmyth is much closer to what might be found in ordinary dictionaries.

As a result of these preliminary investigations, we did not further modify our system for the test run. Our results for the test data are shown in Table 3. As is clear, the results are nearly identical with the test data. These patterns also hold for the individual lexical items (not shown), where there is much more variation in performance. The major reason for the variations appears to lie primarily in the ordering of the senses in the dictionaries. In other words, the sense inventories provide little discriminating information, with the result that sense selection is primarily to the default first sense. This indicates that the sense inventories do not reflect the frequencies in the training and test data.

Table 3. Lexical Sample Recall (Test)			
Run	**Items**	**Fine**	**Coarse**
Adjectives	159	0.409	0.503
Nouns	1806	0.488	0.576
Verbs	1977	0.419	0.540
Total	3942	0.450	0.555

3 Disambiguation of WordNet Glosses

The SENSEVAL-3 task to disambiguate content words in WordNet glosses was a slight modification of the all-words task. One main difference was that tokens to be disambiguated were not identified, requiring the systems to identify content words and phrases. Content words were considered to be any of the four major parts of speech, i.e., words or phrases that could be found in WordNet. Another major difference was that minimal context was provided, i.e., only the gloss itself (although examples were also available). The WordNet synset was also given, providing some "context" within the WordNet network of synsets.

This task had no training data, but only test data based on the tagging of content words by the eXtended WordNet (XWN) project (Mihalcea and Moldovan, 2001). The test data consisted of only and all those glosses from WordNet for which one or more word forms (a single word or a multiword unit) had received a "gold" quality WordNet sense assignment. Scoring for this task is based only on a system's performance in assigning a sense to these word forms. The test set consisted of 9257 glosses containing 15179 "gold" assignments (out of 42491 word forms in these glosses).

To perform this task[1], we used KMS to process each gloss (treated by KMS as a "text"). Each gloss was parsed and processed and converted into an XML representation. (No gloss was a sentence, so each parse was "degenerate" in that only sentence fragments were identified.)

KMS has only recently been modified to incorporate "all-words" disambiguation in the XML representation. At present, the disambiguation has only been partially implemented. One aspect still in development is a determination of exactly which items in the representation should be given a disambiguation and represented (e.g., exactly how to treat multiword units or verbs with particles). Also, we have not yet integrated the full disambiguation machinery (as used in the all-words and lexical sample tasks) into KMS. As a result, only the first (or default) sense of a word is selected.

CL Research's DIMAP dictionary software includes considerable functionality to parse and analyze dictionary definitions. Part of the analysis functionality makes use of WordNet relations in order to propagate information to features associated with a sense. CL Research has previously parsed WordNet glosses as part of an investigation into

[1]Note that, although CL Research ran this task, and we had access to the test data beforehand, we did not actually work with the data until the date indicated for other participants to download and work with the data prior to submission. In any event, our participation in this task was primarily to investigate the parsing and processing of sentence fragments in KMS.

WordNet's internal consistency. However, we did not incorporate any of this experience in performing this task. We also did not incorporate any routines that make use of WordNet relations for disambiguation (as enabled by identification of the WordNet synset identifier). Determining the extent to which these functionalities are relevant for KMS is a matter for future investigation.

Our performance for this task reflects our somewhat limited implementation, as shown in Table 4. Among 10 participating runs, our precision was the second lowest and our recall was the third lowest. We were only able to identify 76.8 percent of the test items with our current implementation. However, in comparing our results with our performance in the all-words and lexical sample tasks, the results here are not significantly different. Moreover, these results suggest a minimum that might be obtained with a disambiguation system that relies only on picking the first sense.

Table 4. Disambiguation of WordNet Glosses			
	Items	**Precision**	**Recall**
"Gold" words	15179	0.449	0.345

4 Automatic Labeling of Semantic Roles

The SENSEVAL-3 task to label sentence constituents with semantic roles was designed to replicate the tagging and identification of frame elements performed in the FrameNet project (Johnson et al., 2003). This task was modeled on the study of automatic labeling by Gildea & Jurafsky (2002), to allow other participants to investigate methods for assigning semantic roles. That study was based on FrameNet 1.0, whereas this task used data from FrameNet 1.1, which considerably expanded the number of frames and the corpus sentences that were tagged by FrameNet lexicographers.

The test data for this task consisted of 200 sentences that had been labeled with frame elements for 40 different frames. Participants were provided with the sentences, the target word (along with its beginning and ending positions in the sentence), and the frame name (i.e., no attempt was made to determine the applicable frame). Specific training data for the task consisted of all sentences not in the test set for the individual frame (ranging from slightly fewer than 200 sentences to as many as 1500 sentences). In addition, participants could use the remainder of the FrameNet corpus for training purposes (another 447 frames and nearly 133,000 sentences). Participants could submit two types of runs: unrestricted (in which frame element boundaries, but not frame element names, could be used, i.e., essentially a classification task) and restricted (in which these boundaries could not be used, i.e., the more difficult task of segmenting constituents and identifying their semantic role). CL Research submitted only one run, for the restricted task.

To perform this task[2], we used KMS to parse and process the sentences (where each sentence was treated as a "text"). We made a slight modification to our system to enable to identify the applicable frame and to keep track of the target word. We also created a special dictionary for FrameNet frames. This dictionary was put into an XML file and consisted only of the frame name, the frame elements, the type of frame element (a classification used by FrameNet as "core", "peripheral", or "extra-thematic"), and a characterization or "definition" of the frame element.

"Definitions" of frame elements were written as specifications for the type of syntactic constituent that was expected to instantiate a frame element in a sentence. Thus, for frames usually associated with verbs, a specification for a frame element might be "subject" or "object". More generally, many frame elements specified "prepositional phrases" headed by one of a set of prepositions (such as "about" or "with"). The basic structure of the FrameNet dictionary was created automatically. The specifications for each frame element was created manually after inspecting the training set for only the 40 frames in the task (which we had processed to show what frame elements had been identified for

[2]Note that, again, although CL Research ran this task, and we had access to the test data beforehand, we did not actually work with the data until the date indicated for other participants to download and work with the data prior to submission. We used only the training data for development of our system. Our participation in this task was exploratory in nature, designed to examine the feasibility and issues involved in integrating frame semantics into KMS. This involves development of processing routines and examination of methods for including frame elements in our XML representation.

each sentence).

To process the test data and create answers, we first parsed and processed each sentence with KMS to create an XML representation using the full set of tags and attributes normally generated. Then, we used the applicable FrameNet "definition" for the frame, the XML representation of the sentence, and the identification of the target word. We iterated through the frame elements and if we had a specification for that element, we used this specification to create an XPath expression used to query the XML representation of the sentence to determine if the sentence contained a constituent of the desired type. If a frame element was labeled as a "core" element for the frame, but no constituent was identified, KMS treated this a "null" instantiation (i.e., a situation where linguistic principles allow frame elements to be omitted within a sentence). Each frame element identified in the sentence was appended to a growing list and the full list was returned as the set of labeled semantic roles for the sentence.

Our results for this task are shown in Table 5. Precision and recall reflect standard measures of how well we were able to identify frame elements. The low recall is a reflection of the small percentage of items attempted. The overlap indicates how well we were able to identify the beginning and ending positions of the constituents we identified.

Table 5. Automatic Labeling of Semantic Roles

Items	Precision	Overlap	Recall	Attempted
16279	0.583	0.480	0.111	19.0

Our poor results stem in large part from only a cursory development of our FrameNet dictionary. We only created substantial entries for 16 of the 40 frames, minimal entries for another 11, and no detailed specifications at all for the remaining 13. The minimal entries were created on the basis of frame elements with the same name (such as **time**, **manner**, and **duration**), which appear in more than one frame. In addition, our method of specification is still somewhat limiting. For example, in frames associated with both nouns and verbs, our method only permitted us to specify the subject or object for a verb and not also a prepositional phrase following a noun. Another deficiency of our system was seen in cases where a long constituent (such as a noun phrase with multiple attached prepositional phrases) was required. Notwithstanding, with only a limited time for development, we able to obtain substantial results, suggesting that simple methods may plausibly be used for a large percentage of cases.

It appears that most participants in this task used statistical methods in training their systems and achieved results better than those obtained by Gildea & Jurafsky. It is possible that these improved results stem from the much larger corpus available in FrameNet 1.1. These results suggest the possibility that it may be feasible and more appropriate to include statistical bases for identifying frame elements in KMS.

Conclusions

In participating in four tasks of SENSEVAL-3, we examined several aspects of disambiguation within the framework of massive tagging of text with syntactic, semantic, and discourse characterizations and attributes. We established basic mechanisms for integrating disambiguation and representational procedures into a larger text processing and analysis system. Our results further demonstrated difficulties in using the WordNet sense inventory, but have further illuminated a number of important issues in disambiguation and representation. At the same time, we have identified a significant number of shortcomings in our system, but with considerable opportunities for further refinement and development.

References

Gildea, Daniel, and Daniel Jurafsky. Automatic Labeling of Semantic Roles. *Computational Linguistics, 28* (3), 245-288.

Johnson, Christopher; Miriam Petruck, Collin Baker, Michael Ellsworth, Josef Ruppenhofer, and Charles Fillmore, (2003). FrameNet: Theory and Practice. Berkeley, California.

Litkowski, K. C. (2001, 5-6 July). Use of Machine-Readable Dictionaries for Word-Sense Disambiguation in SENSEVAL-2. *Proceedings of SENSEVAL-2: 2nd International Workshop on Evaluating Word Sense Disambiguation Systems*. Toulouse, France, pp. 107-110.

Litkowski, K. C. (2002, 11 July). Sense Information for Disambiguation: Confluence of Supervised and Unsupervised Methods. *Word Sense Disambiguation: Recent Successes and Future Directions*. Philadelphia, PA, pp. 47-53.

Litkowski, Kenneth. C. (2004a). Use of Metadata for Question Answering and Novelty Tasks. In E. M. Voorhees & L. P. Buckland (eds.), *The Twelfth Text Retrieval Conference (TREC 2003)*. (In press.)

Litkowski, Kenneth. C. (2004b). Summarization Experiments in DUC 2004. (In press.)

Mihalcea, Rada and Dan Moldovan. (2001). EXtended WordNet: Progress Report. In: *WordNet and Other Lexical Resources: Applications, Extensions, and Customizations*. NAACL 2001 SIGLEX Workshop. Pittsburgh, PA.: Association for Computational Linguistics.