SENSEVAL-3 TASK
Automatic Labeling of Semantic Roles

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

Abstract

The SENSEVAL-3 task to perform automatic labeling of semantic roles was designed to encourage research into and use of the FrameNet dataset. The task was based on the considerable expansion of the FrameNet data since the baseline study of automatic labeling of semantic roles by Gildea and Jurafsky. The FrameNet data provide an extensive body of "gold standard" data that can be used in lexical semantics research, as the basis for its further exploitation in NLP applications. Eight teams participated in the task, with a total of 20 runs. Discussions among participants during development of the task and the scoring of their runs contributed to a successful task. Participants used a wide variety of techniques, investigating many aspects of the FrameNet data. They achieved results showing considerable improvements from Gildea and Jurafsky's baseline study. Importantly, their efforts have contributed considerably to making the complex FrameNet dataset more accessible. They have amply demonstrated that FrameNet is a substantial lexical resource that will permit extensive further research and exploitation in NLP applications in the future.

Introduction

Word-sense disambiguation has frequently been criticized as a task in search of a reason. Since a considerable portion of a sense inventory has only a single sense, the question has been raised whether the amount of effort required by disambiguation is worthwhile. Heretofore, the focus of disambiguation has been on the sense inventory and has not examined the major reason why we would have lexical knowledge bases: how the meanings would be represented and thus, available for use in natural language processing applications. At the present time, a major paradigm for representing meaning has emerged in frame semantics, specifically in the FrameNet project.

A worthy objective for the Senseval community is the development of a wide range of methods for automating frame semantics, specifically identifying and labeling semantic roles in sentences. An important baseline study of this process has recently appeared in the literature (Gildea and Jurafsky, 2002). The FrameNet project (Johnson et al., 2003) has put together a body of hand-labeled data and the Gildea and Jurafsky study has put together a set of suitable metrics for evaluating the performance of an automatic system.

1 The Senseval-3 Task

This Senseval-3 task calls for the development of systems to meet the same objectives as the Gildea and Jurafsky study. The data for this task is a sample of the FrameNet hand-annotated data. Evaluation of systems is measured using precision and recall of frame elements and overlap of a system's frame element sentence positions with those identified in the FrameNet data.

The basic task for Senseval-3 is: Given a sentence, a target word and its frame, identify the frame elements within that sentence and tag them with the appropriate frame element name.

The FrameNet project has just released a major revision (FrameNet 1.1) to its database, with 487 frames using 696 distinctly-named frame elements (although it is not guaranteed that frame elements with the same name have the same meaning). This release includes 132,968 annotated sentences (mostly taken from the British National Corpus). The Senseval-3 task used 8,002 of these sentences

selected randomly from 40 frames (also selected randomly) having at least 370 annotations (out of the 100 frames having the most annotations).[1]

Participants were provided with a training set that identified, for each of the 40 frames, the lexical unit identification number (which equates to a file name) and a sentence identification name. They were also provided with the answers, i.e., the frame element names and their beginning and ending positions. Since the training set was much larger than the test set, participants were required to use the FrameNet 1.1 dataset to obtain the full sentence, its target word, and the tagged frame elements.

For the test data, participants were provided, for each frame, with sentence instances that identified the lexical unit, the lexical unit identification number, the sentence identification number, the full sentence, and a specification of the target along with its start and end positions.

Participants were required to submit their answers in a text file, with one answer per line. Each line was to identify the frame name and the sentence identifier and then all the frame elements with their start and end positions that their systems were able to identify. For example, for the sentence

> *However, its task is made much more difficult by the fact that derogations granted to the Welsh water authority allow <Agent>it</> to <Target>pump</> <Fluid>raw sewage</> <Goal>into both those rivers</>.*

the correct answer would appear as follows:

> Cause_fluidic_motion.256263 Agent (119,120) Fluid (130,139) Goal (141,162)

The sentences provided to participants were not presegmented (as defined in the Gildea and Jurafsky study); this was left to the participants' systems. The FrameNet dataset contains considerable information that was tagged by the FrameNet lexicographers. Participants could use (and were strongly encouraged to use) any and all of the FrameNet data in developing and training their systems. In the test, participants could use any of this data, but were strongly encouraged to use only data available in the sentence itself and in the frame that is identified. (This corresponds to the "more difficult task" identified by Gildea and Jurafsky.) Participants could submit two runs, one with (non-restrictive case) and one without (restrictive case) using the additional data; these were scored separately.

FrameNet recognizes the permissibility of "conceptually salient" frame elements that have not been instantiated in a sentence; these are called null instantiations (see Johnson et al. for a fuller description). An example occurs in the following sentence (sentID="1087911") from the Motion frame: "I went and stood in the sitting room doorway, but I couldn't get any further -- my legs wouldn't move." In this case, the FrameNet taggers considered the Path frame element to be an indefinite null instantiation (INI). Frame elements that have been so designated for a particular sentence appear to be Core frame elements, but not all core frame elements missing from a sentence have designated as null instantiations. The correct answer for this case, based on the tagging, is as follows:

> Motion.1087911 Theme (82,88) Path (0,0)

Participants were instructed to identify null instantiations in submissions by giving a (0,0) value for the frame element's position.[2] Participants were told in the task description that null instantiations would be analyzed separately.[3]

For this Senseval task, participants were allowed to download the training data at any time; the 21-day

[1]The test set was generated with the Windows-based program FrameNet Explorer, available at http://www.clres.com/SensSemRoles.html. FrameNet Explorer provides several facilities for examining the FrameNet data: by frame, frame element, and lexical units. For each unit, a user can explore a frame's elements, associated lexical units, frame-to-frame relations, frame and frame element definitions, lexical units and their definitions, and all sentences.

[2]This turned out to be an incorrect method, since some frame elements (notably "I" at the beginning of a sentence) would have a position of (0,0), i.e, the beginning and ending positions are both 0. Such instances in the test set were identified and handled separately to distinguish them from null instantiations.

[3]No analysis of null instantiations has yet been performed.

restriction on submission of results after downloading the training data was waived since this is a new Senseval task and the dataset is very complex. Participants could work with the training data as long as they wished. The 7-day restriction of submitting results after downloading the test data still applied.

In general, FrameNet frames contain many frame elements (perhaps an average of 10), most of which are not instantiated in a given sentence. Systems were not penalized if they returned more frame elements than those identified by the FrameNet taggers. For the 8002 sentences in the test set, only 16212 frame elements constituted the answer set.

In scoring the runs, each frame element (not a null instantiation) returned by a system was counted as an item attempted. If the frame element was one that had been identified by the FrameNet taggers, the answer was scored as correct. In addition, however, the scoring program required that the frame boundaries identified by the system's answer had to overlap with the boundaries identified by FrameNet. An additional measure of system performance was the degree of overlap. If a system's answer coincided exactly to FrameNet's start and end position, the system received an overlap score of 1.0. If not, the overlap score was the number of characters overlapping divided by the length of the FrameNet start and end positions (i.e., **end-start+1**)[4]

The number attempted was the number of non-null frame elements generated by a system. **Precision** was computed as the number of correct answers divided by the number attempted. **Recall** was computed as the number of correct answers divided by the number of frame elements in the test set. **Overlap** was the average overlap of all correct answers. The percent **Attempted** was the number of frame elements generated divided by the number of frame elements in the test set, multiplied by 100. If a system returned frame elements not identified in the test set, its precision would be lower.

2 Results

Eight teams submitted 20 runs. Three teams submitted runs only for the restricted case (no prior knowledge about frame boundaries). The other five teams submitted at least two runs, with one team submitting 8 runs and another submitting 4 runs. Four of these five teams submitted a restricted run and an unrestricted run (frame boundaries were identified, i.e., the task was a classification task of identifying the applicable frame element).

The results for the classification task are shown in Table 1. The average precision over all these runs is 0.803 and the average recall is 0.757. The overlap in each run is almost identical to the precision, and differs slightly because there may have been some slight positional errors in either the FrameNet data or the sentence string provided in the test data.

Table 1. System Performance (Unrestricted)

Run	Prec	Over	Rec	Att
01b (HKPolyU)	0.874	0.873	0.867	99.2
01c (HKPolyU)	0.905	0.904	0.846	93.5
01d (HKPolyU)	0.859	0.858	0.852	99.2
01e (HKPolyU)	0.902	0.901	0.849	94.1
01f (HKPolyU)	0.908	0.907	0.846	93.2
01g1 (HKPolyU)	0.819	0.817	0.812	99.2
01g2 (HKPolyU)	0.819	0.817	0.812	99.2
01h (HKPolyU)	0.926	0.925	0.705	76.1
02b (InfoSciInst)	0.867	0.866	0.858	99.0
04b (UTDMorarescu)	0.946	0.946	0.907	95.8
07a (UTDMoldovan)	0.898	0.897	0.839	93.4
08b (UUtah)	0.728	0.725	0.721	99.1
08c (UUtah)	0.858	0.857	0.849	98.9

The results for the restricted case are shown in Table 2. The average precision over all these runs is 0.595 and the average recall is 0.481. The average overlap is noticeably lower than the precision, indicating the additional difficulty for these runs of identifying the frame element boundaries.

Table 2. System Performance (Restricted)

Run	Prec	Over	Rec	Att
02a (InfoSciInst)	0.802	0.784	0.654	81.5
03 (CLResearch)	0.583	0.480	0.111	19.0
04a (UTDMorarescu)	0.899	0.882	0.772	85.9
05a (USaarland)	0.654	0.602	0.471	72.0
05b (USaarland)	0.736	0.675	0.594	80.7
06 (UAmsterdam)	0.869	0.847	0.752	86.4
07b (UTDMoldovan)	0.807	0.777	0.780	96.7
08a (UUtah)	0.355	0.255	0.453	127.9
08e (UUtah)	0.387	0.295	0.335	86.7

[4]Hence the problem with an element having (0,0) as the start and end positions.

In both cases, the percent attempted is quite high, except for one system in the restricted runs. This indicates that systems were able to identify potential frame elements in quite a large percentage of the cases. Systems were allowed to return any number of frame elements for a sentence and it is possible for a system to identify more frame elements than were identified by the FrameNet taggers. For example, run 08a asserted many more frame elements than were identified in the answer key. As a result, its percent attempted was much higher than 100 percent. The number of frame elements in other runs not identified in the answer key is unknown. The effect of a higher number attempted lowers the precision for a run and increases the percent attempted.

3 Discussion

Overall, the results achieved in this SENSEVAL-3 task were quite high. Several teams achieved results much better than those obtained by Gildea and Jurafsky. The average precision of 0.80 for all runs in the unrestricted case is only slightly lower than the 82% accuracy achieved in that study when using presegmented constituents. Many teams achieved precision at or above 0.90, indicating that their routines for classifying constituents is quite good. In view of the fact that the number of frames and frame elements in FrameNet has expanded considerably since the Gildea and Jurafsky study, it appears that the methods employed have become quite accurate in classifying constituents.[5]

Results for the restricted were also quite good in comparison with the Gildea and Jurafsky study, which achieved 65% precision and 61% recall at the "more difficult task of simultaneously segmenting constituents and identifying their semantic role." In this task, four teams achieved results between 80 and 90 percent for precision and between 65 and 78 percent for recall.

The participants in this task used a wide variety of methods and data in their systems. In addition, they used the FrameNet dataset from a wide diversity of perspectives. In some cases, they developed mechanisms for grouping the FrameNet data by part of speech or making use of the nascent inheritance hierarchy in FrameNet. In some cases, they used all frames as a basis for training and in others, they employed novel grouping methods based on the similarities among different frames.

The successes of many teams seems to indicate that the FrameNet dataset is an excellent lexical resource and that the resources devoted to its development have been quite valuable. The collective efforts of the participants have contributed greatly to making this complex database more accessible and more amenable to even further development, not only for research purposes, but also for use in many NLP applications.

[5]The diversity of frame elements in the test data has not yet been investigated, so the assertion that this task is more difficult is based solely on the general expansion of FrameNet.

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.