SemEval'07 Task 19: Frame Semantic Structure Extraction

Collin Baker, Michael Ellsworth
International Computer Science Institute
Berkeley, California
{collinb,infinity}
@icsi.berkeley.edu

Katrin Erk
Computer Science Dept.
University of Texas
Austin
katrin.erk@mail.utexas.edu

Abstract

This task consists of recognizing words and phrases that evoke semantic **frames** as defined in the FrameNet project (http://framenet.icsi.berkeley.edu), and their semantic dependents, which are usually, but not always, their syntactic dependents (including subjects). The training data was FN annotated sentences. In testing, participants automatically annotated three previously unseen texts to match gold standard (human) annotation, including predicting previously unseen frames and roles. Precision and recall were measured both for matching of labels of frames and FEs and for matching of semantic dependency trees based on the annotation.

1 Introduction

The task of labeling frame-evoking words with appropriate frames is similar to WSD, while the task of assigning frame elements is called **Semantic Role Labeling** (**SRL**), and has been the subject of several shared tasks at ACL and CoNLL. For example, in the sentence "Matilde said, 'I rarely eat rutabaga,'" *said* evokes the Statement frame, and *eat* evokes the Ingestion frame. The role of SPEAKER in the Statement frame is filled by *Matilda*, and the role of MESSAGE, by the whole quotation. In the Ingestion frame, *I* is the INGESTOR and *rutabaga* fills the INGESTIBLES role. Since the ingestion event is contained within the MESSAGE of the Statement event, we can represent the fact that the message conveyed was about ingestion, just by annotating the sentence with respect to these two frames.

After training on FN annotations, the participants' systems labeled three new texts automatically. The evaluation measured precision and recall for frames and frame elements, with partial credit for incorrect but closely related frames. Two types of evaluation were carried out: **Label matching evaluation**, in which the participant's labeled data was compared directly with the gold standard labeled data, and **Semantic dependency evaluation**, in which both the gold standard and the submitted data were first converted to semantic dependency graphs in XML format, and then these graphs were compared.

There are three points that make this task harder and more interesting than earlier SRL tasks: (1) while previous tasks focused on role assignment, the current task also comprises the identification of the appropriate FrameNet frame, similar to WSD, (2) the task comprises not only the labeling of individual predicates and their arguments, but also the integration of all labels into an overall **semantic dependency graph**, a partial semantic representation of the overall sentence meaning based on frames and roles, and (3) the test data includes occurrences of frames that are not seen in the training data. For these cases, participant systems have to identify the closest known frame. This is a very realistic scenario, encouraging the development of robust systems showing graceful degradation in the face of unknown events.

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 99–104,
Prague, June 2007.

2 Frame semantics and FrameNet

The basic concept of Frame Semantics is that many words are best understood as part of a group of terms that are related to a particular type of event and the participants and "props" involved in it (Fillmore, 1976; Fillmore, 1982). The classes of events are the semantic **frames** of the **lexical units** (**LU**s) that evoke them, and the roles associated with the event are referred to as **frame elements** (**FE**s). The same type of analysis applies not only to events but also to relations and states; the frame-evoking expressions may be single words or multi-word expressions, which may be of any syntactic category. Note that these FE names are quite frame-specific; generalizations over them are expressed via explicit FE-FE relations.

The Berkeley FrameNet project (hereafter FN) (Fillmore et al., 2003) is creating a computer- and human-readable lexical resource for English, based on the theory of frame semantics and supported by corpus evidence. The current release (1.3) of the FrameNet data, which has been freely available for instructional and research purposes since the fall of 2006, includes roughly 780 frames with roughly 10,000 word senses (lexical units). It also contains roughly 150,000 annotation sets, of which 139,000 are lexicographic examples, with each sentence annotated for a single predicator. The remainder are from full-text annotation in which each sentence is annotated for all predicators; 1,700 sentences are annotated in the full-text portion of the database, accounting for roughly 11,700 annotation sets, or 6.8 predicators (=annotation sets) per sentence. Nearly all of the frames are connected into a single graph by frame-to-frame relations, almost all of which have associated FE-to-FE relations (Fillmore et al., 2004a)

2.1 Frame Semantics of texts

The ultimate goal is to represent the lexical semantics of all the sentences in a text, based on the relations between predicators and their dependents, including both phrases and clauses, which may, in turn, include other predicators; although this has been a long-standing goal of FN (Fillmore and Baker, 2001), automatic means of doing this are only now becoming available.

Consider a sentence from one of the testing texts:

(1) This geography is important in understanding Dublin.

In the frame semantic analysis of this sentence, there are two predicators which FN has analyzed: *important* and *understanding*, as well as one which we have not yet analyzed, *geography*. In addition, *Dublin* is recognized by the NER system as a location. In the gold standard annotation, we have the annotation shown in (2) for the Importance frame, evoked by the target *important*, and the annotation shown in (3) for the Grasp frame, evoked by *understanding*.

(2) [$_{\text{FACTOR}}$ This geography] [$_{\text{COP}}$ is] IMPORTANT [$_{\text{UNDERTAKING}}$ in understanding Dublin]. [$_{\text{INTERESTED\_PARTY}}$ INI]

(3) This geography is important in UNDERSTANDING [$_{\text{PHENOMENON}}$ Dublin]. [$_{\text{COGNIZER}}$ CNI]

The definitions of the two frames begin like this:

Importance: A FACTOR affects the outcome of an UNDERTAKING, which can be a goal-oriented activity or the maintenance of a desirable state, the work in a FIELD, or something portrayed as affecting an INTERESTED_PARTY...

Grasp: A COGNIZER possesses knowledge about the workings, significance, or meaning of an idea or object, which we call PHENOMENON, and is able to make predictions about the behavior or occurrence of the PHENOMENON...

Using these definitions and the labels, and the fact that the target and FEs of one frame are subsumed by an FE of the other, we can compose the meanings of the two frames to produce a detailed paraphrase of the meaning of the sentence: Something denoted by *this geography* is a factor which affects the outcome of the undertaking of understanding the location called "Dublin" by any interested party. We have not dealt with *geography* as a frame-evoking expression, although we would eventually like to. (The preposition *in* serves only as a marker of the frame element UNDERTAKING.)

In (2), the INTERESTED_PARTY is not a label on any part of the text; rather, it is marked INI, for "indefinite null instantiation", meaning that it is conceptually required as part of the frame definition, absent from the sentence, and not recoverable from the context as being a particular individual–meaning

that *this geography* is important for anyone in general's understanding of Dublin. In (3), the COGNIZER is "constructionally null instantiated", as the gerund *understanding* licenses omission of its subject. The marking of null instantiations is important in handling text coherence and was part of the gold standard, but as far as we know, none of the participants attempted it, and it was ignored in the evaluation.

Note that we have collapsed the two null instantiated FEs, the INTERESTED_PARTY of the importance frame and the COGNIZER in the Grasp frame, since they are not constrained to be distinct.

2.2 Semantic dependency graphs

Since the role fillers are dependents (broadly speaking) of the predicators, the full FrameNet annotation of a sentence is roughly equivalent to a dependency parse, in which some of the arcs are labeled with role names; and a dependency graph can be derived algorithmically from FrameNet annotation; an early version of this was proposed by (Fillmore et al., 2004b)

Fig. 1 shows the semantic dependency graph derived from sentence (1); this graphical representation was derived from a semantic dependency XML file (see Sec. 5). It shows that the top frame in this sentence is evoked by the word *important*, although the syntactic head is the copula *is* (here given the more general label "Support"). The labels on the arcs are either the names of frame elements or indications of which of the daughter nodes are semantic heads, which is important in some versions of the evaluation. The labels on nodes are either frame names (also colored gray), syntactic phrases types (e.g. NP), or the names of certain other syntactic "connectors", in this case, Marker and Support.

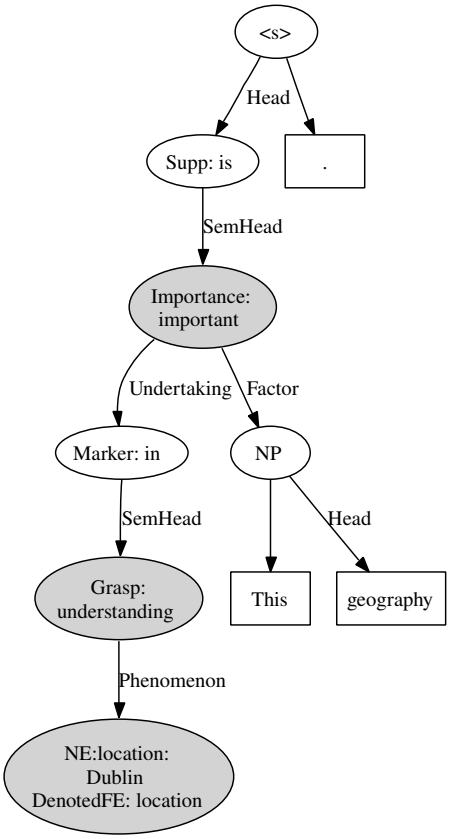

Figure 1: Sample Semantic Dependency Graph

3 Definition of the task

3.1 Training data

The major part of the training data for the task consisted of the current data release from FrameNet (Release 1.3), described in Sec.2 This was supplemented by additional training data made available through SemEval to participants in this task. In addition to updated versions of some of the full-text annotation from Release 1.3, three files from the ANC were included: from Slate.com, "Stephanopoulos Crimes" and "Entrepreneur as Madonna", and from the Berlitz travel guides, "History of Jerusalem".

3.2 Testing data

The testing data was made up of three texts, none of which had been seen before; the gold standard consisted of manual annotations (by the FrameNet team) of these texts for all frame evoking expressions and the fillers of the associated frame elements. All annotation of the testing data was carefully reviewed by the FN staff to insure its correctness. Since most of the texts annotated in the FN database are from the NTI website (`www.nti.org`), we decided to take two of the three testing texts from there also. One, "China Overview", was very similar to other annotated texts such as "Taiwan Introduction", "Russia Overview", etc. available in Release 1.3. The other NTI text, "Work Advances", while in the same domain, was shorter and closer to newspaper style than the rest of the NTI texts. Finally, the "Introduction to

	Sents	NEs	Frames	
			Tokens	Types
Work	14	31	174	77
China	39	90	405	125
Dublin	67	86	480	165
Totals	120	207	1059	272

Table 1: Summary of Testing Data

Dublin", taken from the American National Corpus (ANC, `www.americannationalcorpus.org`) Berlitz travel guides, is of quite a different genre, although the "History of Jerusalem" text in the training data was somewhat similar. Table 1 gives some statistics on the three testing files. To give a flavor of the texts, here are two sentences; frame evoking words are in boldface:

From "Work Advances": "The **Iranians** are **now willing** to **accept** the **installation** of cameras only **outside** the **cascade halls**, which will not enable the IAEA to **monitor** the **entire uranium enrichment process**," the **diplomat said**.

From "Introduction to Dublin": And **in** this **city**, **where literature** and **theater** have **historically dominated** the scene, visual **arts** are **finally** coming into their own with the **new Museum** of Modern **Art** and the **many galleries** that display the work of **modern Irish artists**.

4 Participants

A number of groups downloaded the training or testing data, but in the end, only three groups submitted results: the UTD-SRL group and the LTH group, who submitted full results, and the CLR group who submitted results for frames only. It should also be noted that the LTH group had the testing data for longer than the 10 days allowed by the rules of the exercise, which means that the results of the two teams are not exactly comparable. Also, the results from the CLR group were initially formatted slightly differently from the gold standard with regard to character spacing; a later reformatting allowed their results to be scored with the other groups'.

The LTH system used only SVM classifiers, while the UTD-SRL system used a combination of SVM and ME classifiers, determined experimentally. The CLR system did not use classifiers, but hand-written symbolic rules. Please consult the separate system papers for details about the features used.

5 Evaluation

The labels-only matching was similar to previous shared tasks, but the dependency structure evaluation deserves further explanation: The XML semantic dependency structure was produced by a program called fttosem, implemented in Perl, which goes sentence by sentence through a FrameNet full-text XML file, taking LU, FE, and other labels and using them to structure a syntactically unparsed piece of a sentence into a syntactic-semantic tree. Two basic principles allow us to produce this tree: (1) LUs are the sole syntactic head of a phrase whose semantics is expressed by their frame and (2) each label span is interpreted as the boundaries of a syntactic phrase, so that when a larger label span subsumes a smaller one, the larger span can be interpreted as a the higher node in a hierarchical tree. There are a fair number of complications, largely involving identifying mismatches between syntactic and semantic headedness. Some of these (support verbs, copulas, modifiers, transparent nouns, relative clauses) are annotated in the data with their own labels, while others (syntactic markers, e.g. prepositions, and auxiliary verbs) must be identified using simple syntactic heuristics and part-of-speech tags.

For this evaluation, a non-frame node counts as matching provided that it includes the head of the gold standard, whether or not non-head children of that node are included. For frame nodes, the participants got full credit if the frame of the node matched the gold standard.

5.1 Partial credit for related frames

One of the problems inherent in testing against unseen data is that it will inevitably contain lexical units that have not previously been annotated in FrameNet, so that systems which do not generalize well cannot get them right. In principle, the decision as to what frame to add a new LU to should be helped by the same criteria that are used to assign polysemous lemmas to existing frames. However, in practice this assignment is difficult, precisely because, unlike WSD, there is no assumption that all the senses of each lemma are defined in advance; if

the system can't be sure that a new use of a lemma is in one of the frames listed for that lemma, then it must consider all the 800+ frames as possibilities. This amounts to the automatic induction of fine-grained semantic similarity from corpus data, a notoriously difficult problem (Stevenson and Joanis, 2003; Schulte im Walde, 2003).

For LUs which clearly do not fit into any existing frames, the problem is still more difficult. In the course of creating the gold standard annotation of the three testing texts, the FN team created almost 40 new frames. We cannot ask that participants hit upon the new frame name, but the new frames are not created in a vacuum; as mentioned above, they are almost always added to the existing structure of frame-to-frame relations; this allows us to give credit for assignment to frames which are not the precise one in the gold standard, but are close in terms of frame-to-frame relations. Whenever participants' proposed frames were wrong but connected to the right frame by frame relations, partial credit was given, decreasing by 20% for each link in the frame-frame relation graph between the proposed frame and the gold standard. For FEs, each frame element had to match the gold standard frame element and contain at least the same head word in order to gain full credit; again, partial credit was given for frame elements related via FE-to-FE relations.

6 Results

Text	Group	Recall	Prec.	F1
Dublin	UTD-SRL	0.4188	0.7716	0.5430
China	UTD-SRL	0.5498	0.8009	0.6520
Work	UTD-SRL	0.5251	0.8382	0.6457
Dublin	LTH	0.5184	0.7156	0.6012
China	LTH	0.6261	0.7731	0.6918
Work	LTH	0.6606	0.8642	0.7488
Dublin	CLR	0.3984	0.6469	0.4931
China	CLR	0.4621	0.6302	0.5332
Work	CLR	0.5054	0.7452	0.6023

Table 2: Frame Recognition only

The strictness of the requirement of exact boundary matching (which depends on an accurate syntactic parse) is compounded by the cascading effect of semantic classification errors, as seen by comparing the F-scores in Table 3 with those in Table 2. The difficulty of the task is reflected in the F-scores of around 35% for the most difficult text in the most difficult condition, but participants still managed to reach F-scores as high as 75% for the more limited task of Frame Identification (Table 2), which more closely matches traditional Senseval tasks, despite the lack of a full sense inventory. The difficulty posed by having such an unconstrained task led to understandably low recall scores in all participants (between 25 and 50%). The systems submitted by the teams differed in their sensitivity to differences in the texts: UTD-SRL's system varied by around 10% across texts, while LTH's varied by 15%.

Text	Group	Recall	Prec.	F1
Label matching only				
Dublin	UTD-SRL	0.27699	0.55663	0.36991
China	UTD-SRL	0.31639	0.51715	0.39260
Work	UTD-SRL	0.31098	0.62408	0.41511
Dublin	LTH	0.36536	0.55065	0.43926
China	LTH	0.39370	0.54958	0.45876
Work	LTH	0.41521	0.61069	0.49433
Semantic dependency matching				
Dublin	UTD-SRL	0.26238	0.53432	0.35194
China	UTD-SRL	0.31489	0.53145	0.39546
Work	UTD-SRL	0.30641	0.61842	0.40978
Dublin	LTH	0.36345	0.54857	0.43722
China	LTH	0.40995	0.57410	0.47833
Work	LTH	0.45970	0.67352	0.54644

Table 3: Results for combined Frame and FE recognition

There are some rather encouraging results also. The participants rather consistently performed better with our more complex, but also more useful and realistic scoring, including partial credit and grading on semantic dependency rather than exact span match (compare the top and bottom halves of Table 3). The participants all performed relatively well on the frame-recognition task, with precision scores averaging 63% and topping 85%.

7 Discussion

The testing data for this task turned out to be especially challenging with regard to new frames, since, in an effort to annotate especially thoroughly, almost

40 new frames were created in the process of annotating these three specific passages. One result of this was that the test passages had more unseen frames than a random unseen passage, which probably lowered the recall on frames. It appears that this was not entirely compensated by giving partial credit for related frames.

This task is a more advanced and realistic version of the Automatic Semantic Role Labeling task of Senseval-3 (Litkowski, 2004). Unlike that task, the testing data was previously unseen, participants had to determine the correct frames as a first step, and participants also had to determine FE boundaries, which were given in the Senseval-3.

A crucial difference from similar approaches, such as SRL with PropBank roles (Pradhan et al., 2004) is that by identifying relations as part of a frame, you have identified a gestalt of relations that enables far more inference, and sentences from the same passage that use other words from the same frame will be easier to link together. Thus, the FN SRL results are translatable fairly directly into formal representations which can be used for reasoning, question answering, etc. (Scheffczyk et al., 2006; Frank and Semecky, 2004; Sinha and Narayanan, 2005).

Despite the problems with recall, the participants have expressed a determination to work to improve these results, and the FN staff are eager to collaborate in this effort. A project is now underway at ICSI to speed up frame and LU definition, and another to speed up the training of SRL systems is just beginning, so the prospects for improvement seem good.

This material is based in part upon work supported by the National Science Foundation under Grant No. IIS-0535297.

References

Charles J. Fillmore and Collin F. Baker. 2001. Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop*, Pittsburgh, June. NAACL.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16.3:235–250.

Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. 2004a. FrameNet as a "Net". In *Proceedings of LREC*, volume 4, pages 1091–1094, Lisbon. ELRA.

Charles J. Fillmore, Josef Ruppenhofer, and Collin F. Baker. 2004b. FrameNet and representing the link between semantic and syntactic relations. In Churen Huang and Winfried Lenders, editors, *Frontiers in Linguistics*, volume I of *Language and Linguisitcs Monograph Series B*, pages 19–59. Inst. of Linguistics, Acadmia Sinica, Taipei.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280:20–32.

Charles J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Anette Frank and Jiri Semecky. 2004. Corpus-based induction of an LFG syntax-semantics interface for frame semantic processing. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora (LINC 2004)*, Geneva, Switzerland.

Ken Litkowski. 2004. Senseval-3 task: Automatic labeling of semantic roles. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 9–12, Barcelona, Spain, July. Association for Computational Linguistics.

Sameer S. Pradhan, Wayne H. Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. 2004. Shallow semantic parsing using support vector machines. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 233–240, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Jan Scheffczyk, Collin F. Baker, and Srini Narayanan. 2006. Ontology-based reasoning about lexical resources. In Alessandro Oltramari, editor, *Proceedings of ONTOLEX 2006*, pages 1–8, Genoa. LREC.

Sabine Schulte im Walde. 2003. Experiments on the choice of features for learning verb classes. In *Proceedings of the 10th Conference of the EACL (EACL-03)*.

Steve Sinha and Srini Narayanan. 2005. Model based answer selection. In *Proceedings of the Workshop on Textual Inference, 18th National Conference on Artificial Intelligence*, PA, Pittsburgh. AAAI.

Suzanne Stevenson and Eric Joanis. 2003. Semi-supervised verb class discovery using noisy features. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL-03)*, pages 71–78.