The FrameNet Data and Software

Collin F. Baker
International Computer Science Institute
Berkeley, California, USA
collinb@icsi.berkeley.edu

Hiroaki Sato
Senshu University
Kawasaki, Japan
hiroaki@ics.senshu-u.ac.jp

Abstract

The FrameNet project has developed a lexical knowledge base providing a unique level of detail as to the the possible syntactic realizations of the specific semantic roles evoked by each predicator, for roughly 7,000 lexical units, on the basis of annotating more than 100,000 example sentences extracted from corpora. An interim version of the FrameNet data was released in October, 2002 and is being widely used. A new, more portable version of the FrameNet software is also being made available to researchers elsewhere, including the Spanish FrameNet project.

This demo and poster will briefly explain the principles of Frame Semantics and demonstrate the new unified tools for lexicon building and annotation and also FrameSQL, a search tool for finding patterns in annotated sentences. We will discuss the content and format of the data releases and how the software and data can be used by other NLP researchers.

1 Introduction

FrameNet[1] (Fontenelle, 2003; Fillmore, 2002; Baker et al., 1998) is a lexicographic research project which aims to produce a lexicon containing very detailed information about the relation between the semantics and the syntax of predicators, including verbs, nouns and adjectives, for a substantial subset of English.

The basic unit of analysis is the semantic **frame**, defined as a type of event or state and the participants and "props" associated with it, which we call **frame elements** (**FE**s).[2] Frames range from highly abstract to quite specific. An example of an abstract frame would be the Replacement frame, with FEs such as OLD and NEW as in the sentence *Pat* ***replaced*** *[$_{Old}$ the curtains] [$_{New}$ with wooden blinds].* One sense of the verb *replace* is associated with the Replacement frame, thus constituting one **lexical unit** (LU), the basic unit of the FrameNet lexicon.

An example of a more specific frame is Apply_heat, with FEs such as COOK, FOOD, MEDIUM, and DURATION. as in ***Boil*** *[$_{Food}$ the rice] [$_{Duration}$ for 3 minutes] [$_{Medium}$ in water], then drain.*[3] LUs in Apply_heat include *char*, *fry*, *grill*, and *microwave*, etc.

In our daily work, we define a frame and its FEs, make lists of words that evoke the frame (its LUs), extract example sentences containing these LUs from corpora, and semi-automatically annotate the parts of the sentences which are the realizations of these FEs, including marking the phrase type (PT) and grammatical function (GF). We can then automatically create a report which constitutes a lexical entry for this LU, detailing all the possible ways in which these FEs can be syntactically realized. The

[1] http://framenet.ICSI.berkeley.edu/ framenet

[2] In similar approaches, these have been referred to as schemas or scenarios, with their associated roles or slots.

[3] In this sentence, as in most examples of *boil* in recipes, the COOK is **constructionally null-instantiated**, because of the imperative.

annotated sentences and lexical entries for approximately 7,000 LUs will be available on the FN website and the data will be released by the end of August in several formats.

2 Frame Semantics and FrameNet II

2.1 Frame Semantics in Theory and Practice

The development of the theory of Frame Semantics began more than 25 years ago (Fillmore, 1976; Fillmore, 1977), but since 1997, thanks to two NSF grants[4], we have been able to apply it in a serious way to building a lexicon which we intend to be both usable by human beings and machine-tractable, so that it can serve as a lexical database for NLP, computational lexical semantics, etc. In FrameNet II, all the data, including the definitions of frames, FEs, and LUs and all of the sentences and the annotation associated with them is stored in one relational database implemented in MySQL (Baker et al., 2003; Fillmore et al., 2001).

The FrameNet public website contains an index by frame and an index by LU which links to both the lexical entry and the full annotation for each LU. The frame-to-frame relations which are now being entered in the database will be visible on the website soon.

2.2 FrameNet II Data Release 1.0

The HTML version of the data consists of all the files on the web site, so that users can set up a local copy and browse it with any web browser. It is fairly compact, less than 100 Mb in all.

The plain XML version of the data consists of the following files:

frames.xml This file contains the descriptions of all the 450 frames and their FEs, totaling more than 3,000. Each frame also includes information as to frame-to-frame relations.

luNNN.xml There is one such file per LU (roughly 7500) which contain the example sentences and annotation (if any) for each LU.

[4]We are grateful to the National Science Foundation for funding the project through two grants, IRI #9618838 and ITR/HCI #0086132. We refer to these two three-year stages in the life of the project as FrameNet I and FrameNet II.

relations.xml A file containing information about frame-to-frame and FE-to-FE relations and meta-relations between them.

We intend to have a version of the XML that includes RDF of the DAML+OIL flavor, so that the FN frames and FEs can be related to existing ontologies and Semantic Web-aware applications can access FN data using a standard methodology. Narayanan has created such a version for the FN I data, and a new version reflecting the more complex FN II data is under construction (Narayanan et al., 2002).

3 The FrameNet Software Suite

3.1 The FrameNet Desktop tools

The FN software used for frame definition and annotation has been fundamentally rewritten since the demo at the LREC conference last summer (Fillmore et al., 2002a). The two major changes are (1) combining the frame editing tools and the annotation tools into a single GUI, making the interface more intuitive and (2) moving to a client-server model.

In the previous version, each client accessed the database directly, which made it very difficult to avoid collisions between users, and meant that each client was large, containing a lot of the logic of the application, MySQL-specific queries, etc. In the new version, the basic modules are now the MySQL database, an application server, and one or more client processes. This has a number of advantages: (1) All the database calls are made by the server, making it much easier to avoid conflicts between users. (2) The application server contains nearly all the logic, meaning that the clients are "thin" processes, concerned mainly with the GUI. (3) The separation into client and server makes it easier to set up remote access to the FN database. (4) The increased overhead caused by the more complex architecture is at least offset by the ability to cache frequently-requested data on the server, making access much faster.

The public FrameNet web pages contain static versions of several reports drawn from the database, notably, the lexical entry report, displaying all the valences of each LU. The working environment for the staff includes dynamic versions of these reports and several others, all written as java applets. Partially shared code makes these reports accessible within the desktop package as well.

3.2 API, Library, and Utilities

We are currently working on defining a FN API and writing libraries for accessing the database from other programs. We plan to distribute a command-line utility as a demonstration of this API.

4 FrameSQL and Kernel Dependency Graphs

4.1 Searching with FrameSQL

Prof. Hiroaki Sato of Senshu University has written a web-based tool which allows users to search existing FN annotations in a variety of ways. The tool also makes conveniently available several other electronic resources such as WordNet, and other on-line dictionaries. It is especially useful for doing conventional lexicography.

4.2 Kernel Dependency Graphs

The major product of the project is the lexical database of frame descriptions and annotated sentences; although these clearly are potentially very useful in many sorts of NLP task, FrameNet (at least in its present phase) remains primarily lexicographic. Nevertheless, as a an intermediate step toward applications such as automatic text summarization, we have recently begun studying **kernel dependency graphs** (**KDG**s), which provide a sort of automatic summarization of annotated sentences. KDGs consist of

- the predicator (verb, noun, or adjective),
- the lexical heads of its dependents
- the "marking" on the dependents (prepositions, complementizers, etc. if any), and
- the FEs of the dependents.

To take a simple example, (1-a), which is annotated for the target *chained* in the Attaching frame, could be represented as the KDG in (1-b).

(1) a. [$_{Agent}$ Four activists] **chained** [$_{Item}$ themselves] [$_{Goal}$ to an oil drilling rig being towed to the Barents Sea] [$_{Time}$ in early August].

b.

```
<KDG  frame="Attaching" LU="chain.v">
  <Agent>activists</Agent>
  <Item>themselves</Item>
  <Goal>to:oil\_drilling\_rig</Goal>
  <Time>in:August</Time>
</KDG>
```

The situation can be complicated by the presence of higher control verbs and "transparent" nouns which bring about a mismatch between the semantic head and the syntactic head of an FE (Fillmore et al., 2002b), as in (2), which should have the same KDG as (1-a).

(2) [$_{Agent}$ Four activists] planned to **chain** [$_{Item}$ themselves] [$_{Goal}$ to the bottom of an oil drilling rig being towed to the Barents Sea] [$_{Time}$ in early August].

5 Layered Annotation and Frame Semantic Parsing

A large majority of FEs are annotated with a triplet of labels, one for the FE name, one for the phrase type and one for the grammatical function of the constituent with regard to the target. But the FN software allows more than three layers of annotation for a single target, for situations such as when one FE contains another (e.g. in *[$_{Agent}$ You] 're* ***hurting*** *[$_{Body\_part}$ [$_{Victim}$ my] arms]*).

In addition, the FN software allows us to annotate more than one target in a sentence. A full representation of the meaning of a sentence can be built up by composing the semantics of the frames evoked by the major predicators.

6 Applications and Related Projects

In addition to the original lexicographic goal, a preliminary version of our frame descriptions and the set of more than 100,000 annotated sentences have been released to more than 80 research groups in more than 15 countries. The FN data is being used for a variety of purposes, some of which we had foreseen and others which we had not; these include uses as teaching materials for lexical semantics classes, as a basis for developing multi-lingual lexica, as an interlingua for machine translation, and as training data for NLP systems that perform question answering, information retrieval (Mohit and Narayanan, 2003), and automatic semantic parsing (Gildea and Jurafsky, 2002).

A number of scholars have expressed interest in building FrameNets for other languages. Of these, three have already begun work: In Spain, a team from several universities, led by Prof. Carlos Subirats of U A Barcelona, is building using their own extraction software and the FrameNet desktop tools to build a Spanish FrameNet (Subirats and Petruck, forthcoming 2003) http://www.gemini.es/SFN. In Saarbrücken, Germany, work is proceeding on hand-annotating a parsed corpus with FrameNet FE labels (Erk et al.,). And in Japan, researchers from Keio University and University of Tokyo are building a Japanese FrameNet in the domains of motion and communication, using a large newspaper corpus.

7 Contents of the Demo

We will demonstrate how the software can be used to create a frame, create a frame element, create a lexical unit , define a set of rules for extracting example sentences (and, optionally, marking FEs on them), open an existing LU and annotate sentences, mark an LU as finished, create a frame-to-frame relation, and attach a semantic type to an FE or an LU.

We will demonstrate the reports available on the internal web pages. We will show the complex searches against the FrameNet data that can be run using FrameSQL, including displaying the resulting sentences as KDGs. We will demonstrate how frames can be composed to represent the meaning of sentences using a (manual) frame semantic parsing of a newspaper crime report as an example.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In ACL, editor, *COLING-ACL '98: Proceedings of the Conference, held at the University of Montréal*, pages 86–90. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and Beau Cronin. 2003. The structure of the FrameNet database. *International Journal of Lexicography*.

K. Erk, A. Kowalski, and M. Pinkal. A corpus resource for lexical semantics. Submitted. Available at http://www.coli.uni-sb.de/ erk/ OnlinePapers/ LexProj.ps.

Charles J. Fillmore, Charles Wooters, and Collin F. Baker. 2001. Building a large lexical databank which provides deep semantics. In Benjamin Tsou and Olivia Kwong, editors, *Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation*, Hong Kong.

Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. 2002a. The FrameNet database and software tools. In *Proceedings of the Third International Conference on Languag Resources and Evaluation*, volume IV, Las Palmas. LREC.

Charles J. Fillmore, Collin F. Baker, and Hiroaki Sato. 2002b. Seeing arguments through transparent structures. In *Proceedings of the Third International Conference on Languag Resources and Evaluation*, volume III, Las Palmas. LREC.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32.

Charles J. Fillmore. 1977. Scenes-and-frames semantics. In Antonio Zampolli, editor, *Linguistic Structures Processing*, number 59 in Fundamental Studies in Computer Science. North Holland Publishing.

Charles J. Fillmore. 2002. Linking sense to syntax in FrameNet. In *Proceedings of 19th International Conference on Computational Linguistics*, Taipei. COLING.

Thierry Fontenelle, editor. 2003. *International Journal of Lexicography*. Oxford University Press. (Special issue devoted to FrameNet.).

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Behrang Mohit and Srinivas Narayanan. 2003. Semantic extraction with wide-coverage lexical resources. In *Proceedings of the Human Language Technology Conference (HLT-NAACL)*, Edmonton, Canada.

Srinivas Narayanan, Charles J. Fillmore, Collin F. Baker, and Miriam R.L. Petruck. 2002. FrameNet meets the semantic web: A DAML+OIL frame representation. In *Proceedings of the 18th National Conference on Artificial Intelligence*, Edmonotn, Alberta. AAAI.

Carlos Subirats and Miriam R. L. Petruck. forthcoming 2003. The Spanish FrameNet project. In *Proceedings of the Seventeenth International Congress of Linguists*, Prague.