

Extracting Large-Scale Lexical Resources for LFG from the Penn-II Treebank



Aoife Cahill, Mairéad McCarthy, Ruth O' Donovan, Josef van Genabith, Andy Way
National Centre for Language Technology, School of Computing, Dublin City University, Dublin 9, Ireland
{acahill, mccarthy, rdonovan, josef, away}@computing.dcu.ie

Introduction and Motivation

In this work, we present an approach to automating the large-scale acquisition of subcategorisation details for LFG-based NLP systems. The induced semantic forms are required to convert the extracted probabilistic grammar of Cahill et al. (2002a) to a stand-alone LFG grammar and permit the resolution of long distance dependencies. Our choice to extract the semantic forms automatically is motivated by the fact that the manual construction of such an extensive resource may be time consuming and error prone.

Subcategorisation in LFG

In LFG subcategorisation requirements are expressed at f-structure level, in *functional* rather than *phrasal* terms. These requirements are denoted using a *semantic form*, the value of the PRED attribute, which is represented in the following way:

$$\pi < gf, gf, \dots, gf >$$

where π is a lemma and gf is a *governable grammatical function*. In conjunction with the conditions of *completeness* and *coherence*, the value of the argument list of the semantic form ensures the well-formedness of the f-structure within which it occurs.

Our Methodology

Our methodology requires a treebank annotated with LFG f-structure information. We utilise the automatic annotation algorithm of (Cahill et al., 2002b) to derive a version of the Penn-II Treebank where each node in each tree is annotated with an LFG functional annotation. These functional equations are then collected and used by the constraint solver to produce an f-structure for the tree in question. The following extraction algorithm is used to compile semantic forms from the f-structure annotated treebank:

Extraction Algorithm

For each f-structure, for each level of embedding, the local pred is determined and all subcategorisable grammatical functions present at that level of embedding are collected for that predicate. (van Genabith, Way and Sadler, 1999)

For example the f-structure in **Figure 1** produces the following semantic forms:

- `accept ([obj, subj, obl:as])`
- `as ([obj])`
- `john ([])`
- `mary ([])`
- `replacement ([])`
- `a ([])`

NOTE: While the argument lists of LFG semantic forms are arranged based on the obliqueness hierarchy, ours are ordered alphabetically by Prolog. obl: appears last as Prolog converts it to : (obl,as) and letters precede punctuation in alphabetical ordering.

The semantic forms are associated with conditional probabilities $P(s|l)$ (derived from the corpus) where l is a lemma and s a semantic form. **Table 1** shows the extracted semantic forms for the verb *accept* with their associated frequency counts and probabilities.

Semantic Form	Frequency	Probability
<code>accept ([obj, subj])</code>	122	0.813
<code>accept ([subj])</code>	11	0.073
<code>accept ([comp, subj])</code>	5	0.033
<code>accept ([obj, subj, obl:as])</code>	3	0.020
<code>accept ([obj, subj, obl:from])</code>	3	0.020
<code>accept ([subj, obl:as])</code>	3	0.020
Others	3	0.021

Table 1 Semantic Forms for accept with associated probabilities.

```

subj : num : sing
      pers : 3
      cat : nnp
      pred : 'John'
obj : num : sing
     pers : 3
     cat : nnp
     pred : 'Mary'
obl : obj : spec : det : pred : a
      pred : replacement
      num : sing
      pers : 3
      cat : nn
      pred : as
tense : past
cat : vbd
pred : accept
    
```

Figure 1 F-Structure for "John accepted Mary as a replacement"

Further Extensions

- The `cat` feature included in the f-structures contains the syntactic category of the lexical item whose lemma is the value of the `pred` feature at that particular level of embedding. This allows us to classify words and their semantic forms based on their syntactic category.
- This feature also allows us to read off the syntactic category of the head of each of the subcategorised grammatical functions.
- Due to information contained in the f-structures we can now distinguish verbs occurring in the *passive*. This information can be used to adjust the conditional probabilities. Of the 11 occurrences of the verb *accept* with the semantic form `[subj]` in Figure 1, 9 of these are in a passive context. The probability of *accept* occurring with this semantic form in an active context is therefore reduced to 0.014. Its probability of occurring with `[subj, obl:as]` is reduced to 0 for the same reason.

	Precision	Recall	F-Score
Without PP	75.2%	69.1%	72.0%
With PP	65.5%	63.1%	64.3%
With Preposition	71.8%	16.8%	27.3%

Table 2 Threshold of 1%

	Precision	Recall	F-Score
Without PP	80.2%	63.6%	70.9%
With PP	69.6%	56.9%	62.7%
With Preposition	76.7%	13.9%	23.5%

Table 3 Threshold of 5%

Evaluation

We extracted over 16,000 unique non-empty semantic forms. These include semantic forms for 3287 different verbs. The automatically extracted semantic forms were evaluated against the COMLEX Dictionary. Frequency thresholds were set to filter the semantic forms associated with each verb. Based on the approach of Schulte im Walde (2002) and for the sake of evaluation we had three levels of prepositional detail in the semantic forms. The results with varying thresholds are shown in **Tables 2** and **3**.

Conclusions and Further Work

- Our approach allows the large-scale extraction of a lexical resource for LFG.
- The high precision figures achieved in evaluation demonstrate the accuracy of this resource.
- Cahill et al. (2003) show the manner in which the extracted semantic forms may be used effectively for the resolution of long distance dependencies.
- A possible extension to our work is to merge our entries with those of the COMLEX Resource.
- We wish to further examine the usefulness of our semantic forms in parsing and continue our evaluation work.
- We also hope to test our extraction procedure on large quantities of raw rather than annotated text. We believe this to be possible using the LFG parser of Cahill et al. (2002).

References:

Cahill, A., M. McCarthy, J. van Genabith and A. Way. 2002a. Parsing Text with a PCFG derived from Penn-II with an Automatic F-Structure Annotation Procedure. In M. Butt and T. Holloway-King (eds.) *Proceedings of the Seventh International Conference on LFG*, CSLI Publications, Stanford, CA., pp. 55-71.

Cahill, A., M. McCarthy, J. van Genabith and A. Way. 2002b. Automatic Annotation of the Penn-Treebank with LFG F-Structure Information. In *Proceedings of the LREC Workshop on Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data*, Las Palmas, Spain, pp. 8-15.

Cahill, A., M. McCarthy, R. O'Donovan, J. van Genabith and A. Way. 2003. Lexicalisation of Long Distance Dependencies in a Treebank-Based, Wide-Coverage, Statistical LFG Grammar. In *Proceedings of the Eighth International Conference on LFG*. (to appear)

Schulte im Walde, S. 2002. Evaluating Verb Subcategorisation Frames learned by a German Statistical Grammar against Manual Definitions in the Duden Dictionary. In *Proceedings of the 10th EURALEX International Congress*, Copenhagen, Denmark.

Van Genabith, J., A. Way and L. Sadler. 1999. Data-driven Compilation of LFG Semantic Forms. In *EACL-99 Workshop on Linguistically Interpreted Corpora*, Bergen, Norway, pp. 69-76.