Towards Better Understanding of Automatic Semantic Role Assignment

Sebastian Padó
SALSA
Dept. of Computational Linguistics
Saarland University
Saarbrücken
pado@coli.uni-sb.de

Gemma Boleda Torrent
GLiCom
Dept. de Traducció i Interpretació
Universitat Pompeu Fabra
Barcelona
gemma.boleda@upf.edu

Abstract

We investigate whether automatic semantic role assignment can benefit from taking argument structure into consideration.
Our case study on FrameNet data shows huge variability between error rates for different frames that can be explained by the different uniformity of the frames' argument structures, but not by other factors like the amount of training data.

1 Introduction

After the efforts of the last decade to create large syntactically annotated corpora, recent years have witnessed growing interest in extending annotation to the semantic level. Most of these efforts have concentrated on *role semantics*, like FrameNet (Baker et al., 1998), the Prague Dependency Treebank (Hajičová, 1998), or PropBank (Kingsbury et al., 2002). These approaches aim at finding a level of representation that is genuinely semantic, but avoids the pitfalls of exhaustive meaning representation.

The present study concentrates on FrameNet[1]. FrameNet is designed as an ontology of *frames*, representations of prototypical situations. Each frame provides a set of predicates (nouns, verbs or adjectives) by which it can be introduced and a set of *semantic roles* which correspond to categories of entities or concepts that occur in the situation. FrameNet distinguishes between roles central to the situation (*core* roles) and marginal roles (*non-core* roles). As an example, the AWARENESS frame offers the core roles *Cognizer* (Cogn) and *Content* (Cont), while non-core roles are *Evidence* and *Topic*. The following examples show the kind of annotation provided in FrameNet for some of the verbs in the AWARENESS frame:

(1) [$_{Cogn}$ I] **know** [$_{Evidence}$ from personal experience] [$_{Cont}$ how difficult it is to write a book].

(2) [$_{Cogn}$ Rose] will **believe** [$_{Cont}$ anything that man tells her].

(3) [$_{Cont}$ Its functions] are not always fully **understood** [$_{Cogn}$ by academics and funding bodies]

Like other projects, FrameNet has predominantly concentrated on building a large manually annotated corpus. The corpus, a subset of the British National Corpus, currently contains about 125.000 instances of 480 frames. However, for role semantics to become relevant for language technology, robust and accurate methods for *automatic semantic role assignment* are needed. In recent years, a number of studies has investigated this task on the FrameNet corpus (Gildea and Jurafsky, 2002; Fleischman et al., 2003; Chen and Rambow, 2003; Thompson et al., 2003).

Role assignment has generally been modelled as a classification task: A statistical model is trained on manually annotated data and later assigns a role label out of a fixed set to every constituent in new, unlabelled sentences[2]. The existing studies have used different statistical frameworks, but have largely converged on a common set of features to base their decisions on, namely syntactic information (path from predicate to constituent, phrasal type of constituent) and lexical

[1] http://www.icsi.berkeley.edu/~framenet/

[2] Constituents are assigned a special tag if they do not realise a semantic role.

information (head word of the constituent, predicate). With this methodology, the studies report error rates in the range of 15% to 25%.

Since the number of feature combinations is huge, it is impossible to learn the appropriate role for each combination of features from the training data. Therefore, the question of how to *generalise* across examples is crucial: if no generalisation takes place, then unseen feature combinations cannot be classified; overgeneralisation, on the other hand, leads to misclassification.

However, all studies use their statistical models as black boxes, and to our knowledge, no thorough error analysis has been carried out for existing models. Such an analysis can determine how well generalisation works in existing models, which in turn can lead to better models that perform more accurate generalisation. Our work starts with the observation that formal linguistics has faced this very problem for several decades now, attempting at a *linking theory* that explains the mapping between semantic arguments and syntactic constituents.

In this study, we perform an error analysis of a role assignment model that uses a simple concept of *argument structure* (Section 2), and show that performance figures for frames correlate with the uniformity of their argument structures (Section 3). We conclude that grouping verbs according to argument structure could be a key ingredient in improving performance of automatic semantic role assignment.

2 A Case Study in Role Assignment

2.1 Argument Structure and Patterns

In order to investigate the influence of argument structure on role assignment, we had to define argument structure in a way compatible with our FrameNet data. Our definition has two aims: (a), to be coherent with the literature on the matter (work initiated, among others, by Gruber (1965), Fillmore (1968) or Jackendoff (1972)), and (b), to be able to use the data at our disposal, namely the annotated corpus.

To meet the first aim, we only took into account FrameNet roles that could be considered arguments, and ignored "adjunct-like" roles. In practice, we disregarded most *non-core* roles (see Section 1) as well as very infrequent roles. For instance, in the frame AWARENESS we treated only the roles *Cognizer* and *Content* as arguments and ignored *Evidence* and *Topic*.

Since we had no additional lexical information about the predicates like subcategorisation restrictions, we modelled argument structure in terms of the data in the corpus. The basis of our definition is the concept of *pattern*, a corpus-attested instance of a predicate with its realised roles. We formalise a pattern as a set of pairs, each pair consisting of a semantic role and the grammatical function (SUBJ, COMP, MOD) of the constituent that realises the role in a sentence.[3] We then define the *argument structure of a predicate* as the set of all corpus-attested patterns for this predicate, together with their respective frequencies. Finally, the *argument structure of a frame* is the set of argument structures of its predicates.

As an example, consider sentences (1-3) from Section 1. If we only had these three sentences in the AWARENESS corpus, the argument structure for this frame would be the following:

(4) {(Cogn, SUBJ), (Cont, COMP)}, 2
{(Cont, SUBJ), (Cogn, COMP)}, 1

(4) shows that in this frame it is possible that *Cognizer* be realised as subject and *Content* as complement (attested twice) or, alternatively, the *Content* can be realised as the subject, and the *Cognizer* as the complement (attested once in (3), a passive sentence). Note that *Evidence* is not included in the patterns in (4): it is not directly relevant to argument structure, as has already been discussed.

2.2 The Experiment

The task. We modelled role assignment as a classification task, using *patterns* (in the sense defined in the last section) as classes. The learner had to predict the correct pattern for a sentence on the basis of the following feature set: the lemma of the predicate, the POS-Tag of the predicate and

[3] We did not take information into account regarding the phrasal type or, in case of PP complements, the prepositional head. The model is thus very simple, but sufficient for the purposes of the paper.

the grammatical functions of the constituents in the sentence.

We used the rule extraction system RIPPER (Cohen, 1995) as machine learner. We chose RIPPER because decision rules offer a compact, human-readable representation of the role assignment regularities, as will be shown below. Still, RIPPER derives rules probabilistically and is sufficiently similar to other statistical frameworks that the results are comparable.

Experimental material and baseline. As experimental material, we used all annotated sentences for all verbal predicates of the 16 frames listed in Table 1. Each sentence was converted into a feature vector as input to RIPPER, and the corresponding pattern was constructed as presented in Section 2.1, resulting in between 1 and 17 patterns (i.e. classes) per frame. This gives rise to two baselines: A random assignment of patterns to sentences results in an error rate of 89.6%, and assigning the most frequent pattern per frame reduces it to 53.0%.

Procedure. For each frame, we split the data randomly into ten different training sets (90%) and test sets (10%) to perform 10-fold cross validation. Preliminary tests with different feature combinations showed that the lemma information did not improve performance, probably due to sparse data, and we discarded the feature for the final run.

Results. Figure 1 shows some of the decision rules RIPPER derives for the frame AWARENESS.

```
DEFAULT Cogn-SUBJ, Cont-COMP
IF Comp=no AND
POS_Pred=participle
THEN Cont-SUBJ
IF Comp=no THEN Cogn-SUBJ
```

Figure 1: Decision Rules for AWARENESS

RIPPER correctly captures that the default realisation is *Cognizer* as subject and *Content* as complement, as we saw in sentences (1-2) in Section 1. It also captures the mechanism for passive: the second line states that if the verb is a participle, then it is the *Content* which is realised as the subject.[4] The last condition states that in any other case than passives, if no complement is present in the sentence, the subject will be assigned the role *Cognizer* (as in *Jill believes*).

The final error rates for role assignment with RIPPER are reported in the ER column of Table 1. Our model attains an average error rate of 22.8% over all frames, which is within the range found by other studies.

Note however that the figures cannot be directly compared, since our model solves a somewhat different (and simplified) task: some of the previous models had to first identify the constituents which had to be assigned roles, and then attempt to assign the adequate role. Our model focuses on the second part of the problem, for the relevant constituents are given in the annotation.

We nevertheless regard the performance of our model (well above baseline) as a confirmation that our model is not trivial, and that its inspection can reveal phenomena of relevance for the task at hand. Namely, our results highlight a phenomenon that has not received attention so far, a huge variance in error rate across frames. Among our 16 frames, we found error rates as low as 0% (FRUGALITY) and as high as 77% (DESIRING).

Frame	#I	ER	U
ABUNDANCE	25	45.00 ± 9.46	0.63
AWARENESS	541	15.56 ± 1.60	0.89
BUNGLING	58	40.17 ± 6.29	0.27
COMPATIBILITY	70	12.00± 3.82	0.91
EXPERIENCER-SUBJ	1264	8.23 ± 0.94	0.93
CATASTROPHE	54	18.50 ± 4.97	0.23
CONTRITION	21	0.00 ± 0.00	0.91
CURE	201	59.50 ± 3.09	0.65
DEATH	456	20.67 ± 2.47	0.41
DESIRING	19	76.67 ± 13.91	0.00
EXPECTATION	282	18.85 ± 1.37	0.96
FRUGALITY	15	0.00 ± 0.00	N/A
IMITATION	73	16.83 ± 4.95	0.49
JUDGMENT	2192	12.14 ± 0.76	0.81
JC[5]	1791	43.80 ± 1.39	0.83
PR[5]	204	36.41 ± 2.93	0.59

Table 1: Results of RIPPER study. #I: Number of instances (annotated FrameNet examples), ER: RIPPER error rate with standard deviation (10-fold CV), U: frame argument structure uniformity (see Section 3.3)

[4] It does not capture that the *Cognizer* can be realised as a complement, because there are more examples of passive sentences in the corpus where only the subject is realised.

3 Explaining the Variance

We have seen that the difficulty of automatically assigning semantic roles varies enormously across frames. Moreover, the existence of frames with a large number of instances among the difficult cases (like JUDGMENT_COMMUNICATION) shows that this variance has a direct impact on the overall performance. In the following sections, we consider three plausible sources of the variance.

3.1 Non-representative Data

The aim of the FrameNet project is primarily lexicographic, and its corpus consists of "prototypical" cases for each frame that illustrate the use of semantic roles. In other words, it is not guaranteed to be a representative sample, condition assumed by statistical models. Yet, it is rather unlikely that non-representativity affects individual frames differently; instead, it should affect the upper bound for role assignment performance.

3.2 Amount of Training Data

Another natural explanation would be that the variance is caused by the different amounts of training data for different frames, listed in the #I column of Table 1. If that were true, then there should be a correlation between #I, the number of instances, and ER, the error rate. To test this hypothesis, we performed a number of correlation analyses. Since we could not guarantee normal distribution of the data, we chose to compute Kendall's τ, a rank-based nonparametric correlation test:

$$\tau = 1 - \frac{2(\text{Number of rank inversions})}{\text{Number of pairs of objects}}$$

τ values range between 1 (perfect positive correlation) and -1 (perfect negative correlation).

To control for noise introduced by frames with a very small number of instances, we ran the correlation analysis on different datasets, namely the set of all frames and the set of frames with more than 20, 50, 100, 200 or 500 instances. All tests were for positive correlation (one-tailed t tests). The results are shown in the first column of Table 2. τ values range around zero (no correlation), and the p values confirm that, assuming the usual significance level of p=0.05, no significant correlation exists between #I and ER.

#I	τ(ER, #I)	τ(U, ER)	τ(U, #I)
all	0.01 (p=0.54)	-0.58 (p<0.01)	0.33 (p=0.05)
>20	0.19 (p=0.84)	-0.54 (p<0.01)	0.21 (p=0.17)
>50	0.03 (p=0.58)	-0.52 (p=0.01)	0.43 (p=0.31)
>100	-0.14 (p=0.36)	-0.79 (p<0.01)	0.22 (p=0.27)
>200	0.14 (p=0.72)	-0.71 (p=0.01)	-0.05 (p=0.61)
>500	0.67 (p=0.96)	-1.00 (p=0.04)	-0.67 (p=0.96)

Table 2: Results of correlation analyses for different datasets with Kendall's τ

3.3 Uniformity of Argument Structure

According to our hypothesis from Section 1, models for role assignment should be sensitive to the regularity of the argument structure. We are now in a position to test a more specific wording of this hypothesis, namely that instances of frames with *less uniform* argument structures should be more difficult to label.

Since we found no established measures for the uniformity of argument structure, we defined a measure ourselves. Recall from Section 2.1 that we defined a predicate's argument structure as the set of patterns the predicate was attested to realise, together with their frequencies, $\{(p_1, f_1), \ldots, (p_n, f_n)\}$. We can regard these data as vectors of frequencies $\vec{v} = \langle f_1, \ldots, f_n \rangle$ whose dimensions are labelled with patterns. This makes the argument structure of a frame a vector space, and we can model the *uniformity* of the argument structure as the *similarity* of the vectors in the space.

Work on vector space semantics (Lee, 1999) has established many similarity metrics for two vectors. One that has turned out to be suitable for a broad range of linguistic applications is cosine similarity, ranging between 1 (most similar) and 0 (least similar):

$$cos(\vec{v}_1, \vec{v}_2) = \frac{\sum_i v_{1,i} v_{2,i}}{\sqrt{\sum_i {v_{1,i}}^2}\sqrt{\sum_i {v_{2,i}}^2}}$$

However, there appear to be no similarity measures for more than two vectors, and we decided to use a generalisation of cosine similarity. Our new

[5] JC: JUDGMENT_COMMUNICATION, PR: PERSONAL_RELATIONSHIP

uniformity measure U is defined as the weighted average of the cosine similarity of every pair of vectors:

$$U(\vec{v}_1, \ldots, \vec{v}_n) = \frac{\sum_i \sum_{j \neq i} |\vec{v}_i||\vec{v}_j| cos(\vec{v}_i, \vec{v}_j)}{\sum_i \sum_{j \neq i} |\vec{v}_i||\vec{v}_j|}$$

The weight is defined as the product of the length of the two vectors. This gives more weight to longer vectors, that is, in our application, predicates that have occurred with higher frequencies. At the same time, this limits the amount of noise introduced by very infrequent predicates. The normalisation in the denominator guarantees that U values can be interpreted like cosine values.

We computed U scores for all frames[6], shown in the right column of Table 1. Using the same procedure as in Section 3.2, we tested for negative correlation between U and ER, because we expected higher uniformity to yield lower error rates.

Results are reported in the second column of Table 2. The τ values show that there is a very strong negative correlation between the two variables, and the p scores confirm that it is highly significant at $p = 0.01$ for all datasets, except for the #I>500. For this dataset, there is a perfect negative correlation (τ=-1), too, but it contains too few datapoints (4 frames) for the correlation to be significant.

To verify that the correlation between ER and U was not the result of an indirect relationship mediated by #I, we finally tested for a positive correlation between U and #I, using the same procedure as before. The results are shown in the third column of Table 2. The correlation borders significance at $p = 0.05$ for the "all frames" dataset, but is was not significant for all other datasets. At any rate, the correlation between U and #I is much weaker than the one between ER and U.

4 Conclusions and further work

Our study has shown that automatic semantic role assignment is of largely varying difficulty in different frames. We have established that the error rate is negatively correlated with the argument structure uniformity within frames. This shows that frames containing predicates with very dissimilar argument structures are more difficult to annotate automatically.

Frames can vary so much in argument structure uniformity because predicates are grouped into frames according to ontological considerations, not to linguistic principles. Since semantically related predicates can exhibit completely opposite linking patterns (cf. *buy* vs. *sell*), frames are not necessarily consistent with respect to argument structure. It would be thus mistaken to look for a one-to-one correspondence between a classification based on argument structure and the classification in FrameNet.

However, knowledge about argument structure could be used to guide generalisation. As an example, consider the frame CATASTROPHE, which is evoked by verbs such as *betide, befall* and *suffer*. It is easy to see that *betide* and *befall* have linking patterns which are identical and completely orthogonal to those of *suffer*. An independent classification of verbs could be used to exploit *betide* data (and not *suffer* data) to tag unseen *befall* examples.

One study has attempted to use Levin classes (Levin, 1993) as such an independent classification (Gildea and Jurafsky, 2002), but this decreased performance. A plausible reason for that is that is that Levin's classes may not adequate to generalise across predicates in a frame. They can indeed be used to predict different possible mappings for a single predicate, for example to account for the causative and noncausative constructions of *break*: *John broke the window* vs. *the window broke*. However, as explained above, it should not be expected that the other predicates in the relevant frame (CAUSE_TO_FRAGMENT) will exhibit exactly the same alternation behavior. For instance, other verbs in the frame CAUSE_TO_FRAGMENT are *rip* or *smash*, which do not have a noncausative reading (*Mary ripped the envelope* vs. **the envelope ripped*).

Levin's classes thus seem to be at a too fine-grained level, so that overgeneralisation takes place. For the purposes of automatic role assignment, an alternative scheme should developed that could be sufficiently broad so as to be generally applicable but at the same time fine-grained

[6] U(FRUGALITY) is not defined, since the frame contains only one verbal predicate.

enough to capture the relevant information (e.g. the *befall* vs. *suffer* asymmetry). The development of such a constrained classification is the topic of ongoing research.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL-98*, Montreal, Canada.

John Chen and Owen Rambow. 2003. Use of deep linguistic features for the recognition and labeling of semantic arguments. In *Proceedings of EMNLP-03*, Sapporo, Japan.

William W. Cohen. 1995. Fast effective rule induction. In *Proceedings of ICML-95*, pages 115–123, Tarragona, Spain.

Charles J. Fillmore, 1968. *Universals in Linguistic Theory*, chapter The Case for Case, pages 1–88. Holt, Rinehart, and Winston, New York.

Michael Fleischman, Namhee Kwon, and Ed Hovy. 2003. Maximum entropy models for FrameNet classification. In *Proceedings of EMNLP-03*, Sapporo, Japan.

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

Jeffrey Gruber. 1965. *Studies in lexical relations*. MIT Working Papers in Linguistics, Cambridge, MA.

Eva Hajičová. 1998. Prague Dependency Treebank: From Analytic to Tectogrammatical Annotation. In *Proceedings of TSD-98*, Brno, Czech Republic.

Ray Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press.

Paul Kingsbury, Martha Palmer, and Mitch Marcus. 2002. Adding semantic annotation to the Penn TreeBank. In *Proceedings of HLT-02*, San Diego, CA.

Lillian Lee. 1999. Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Cynthia A. Thompson, Roger Levy, and Christopher Manning. 2003. A generative model for FrameNet semantic role labeling. In *Proceedings of the 14th European Conference on Machine Learning*, Cavtat, Croatia.