The Influence of Argument Structure on Semantic Role Assignment

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

Gemma Boleda
GLiCom
Dept. of Translation and Interpreting
Pompeu Fabra University
Barcelona
gemma.boleda@upf.edu

Abstract

We present a data and error analysis for semantic role labelling. In a first experiment, we build a generic statistical model for semantic role assignment in the FrameNet paradigm and show that there is a high variance in performance across frames. The main hypothesis of our paper is that this variance is to a large extent a result of differences in the underlying argument structure of the predicates in different frames. In a second experiment, we show that *frame uniformity*, which measures argument structure variation, correlates well with the performance figures, effectively explaining the variance.

1 Introduction

Recent years have witnessed growing interest in corpora with semantic annotation, especially on the semantic role (or argument structure) level. A number of projects are working on producing such corpora through manual annotation, among which are FrameNet (Baker et al., 1998), the Prague Dependency Treebank (Hajičová, 1998), PropBank (Kingsbury et al., 2002), and SALSA (Erk et al., 2003).

For semantic role annotation to be widely useful for NLP, however, robust and accurate methods for *automatic semantic role assignment* are necessary. Starting with Gildea and Jurafsky (2000), a number of studies have developed (almost exclusively statistical) models of this task, e.g. Thompson et al. (2003) and Fleischman et al. (2003). This year (2004), semantic role labelling served as the shared task at two conferences, CoNLL[1] and SENSEVAL[2].

However, almost all studies have concentrated on the technical aspects of the models – identifying informative feature sets and suitable statistical frameworks – with the goal of optimising the performance of the models on the complete dataset. The only study we are aware of with a more detailed evaluation is Fleischman et al. (2003), who nevertheless come to the conclusion that either *"new features"*, *"more data"*, or *"more sophisticated models"* are needed.

The present study is a first step in pursuing the third alternative, presenting a data and error analysis for semantic role assignment in the FrameNet paradigm. We first build two different, generic statistical models for semantic role assignment, which are fairly representative for the span of models investigated in the literature. A frame-wise evaluation shows that the models exhibit a large variance in performance across frames.

Our hypothesis is that this variance is to a large extent caused by differences in the underlying *argument structure* of the predicates: Frames which are less *uniform*, i.e. whose predicates have a more heterogeneous mapping between semantic roles and syntactic functions, are more difficult to label automatically. In order to put this hypothesis, which is intuitively very plausible, on a a firm empirical footing, we investigate the relationship between frame uniformity and the variance in the data and show that the two variables correlate. Since argument structure has been investigated mostly for verbs, we restrict our study to verbal predicates.

Structure of the paper. In Section 2 we give a brief introduction to FrameNet. Section 3 outlines the first experiment and discusses the variance in performance across frames. In Section 4, we define two measures of *frame uniformity* based on argument structure, and show in our second experiment (Section 5) that they correlate with the performance figures. Finally, Section 6 discusses the implications of our results for semantic role assignment.

2 FrameNet

FrameNet is a lexical resource based on Fillmore's Frame Semantics (Fillmore, 1985). It is designed as an ontology of *frames*, representations of prototypical situations. Each frame provides a set of *predicates* (nouns, verbs or adjectives) which can introduce the frame. The semantic roles are frame-specific, since they are defined as categories of enti-

[1] http://www.lsi.upc.es/~conll04st/

[2] http://www.clres.com/SensSemRoles.html

ties or concepts pertaining to the particular situation a predicate evokes.

The following sentences are examples for the semantic annotation provided in the FrameNet corpus for verbs in the IMPACT frame, which describes a situation in which typically "an *Impactor* makes sudden, forcible contact with the *Impactee*, or two *Impactors* both ... [make] forcible contact"[3].

(1) a. [$_{\text{Impactee}}$ His car] was **struck** [$_{\text{Impactor}}$ by a third vehicle].
b. [$_{\text{Impactor}}$ The door] **slammed** [$_{\text{Result}}$ shut].
c. [$_{\text{Impactors}}$ Their vehicles] **collided** [$_{\text{Place}}$ at Pond Hill].

Note that the frame-specificity of semantic roles in FrameNet has important consequences for semantic role assignment, since there is no direct way to generalise across frames. Therefore, the learning for automatic assignment of semantic roles has to proceed frame-wise. Thus, the data sparseness problem is especially acute, and automatic assignment for frames with no training data is very difficult (see Gildea and Jurafsky (2002)).

3 Experiment 1: Frame-Wise Evaluation of Semantic Role Assignment

In our first experiment, we perform a detailed (frame-wise) evaluation of semantic role assignment to discover general patterns in the data. Our aim is not to outperform existing models, but to replicate the workings of existing models so that our findings are representative for the task as it is currently addressed. To this end, we (a) use a standard dataset, the FrameNet data, (b) model the task with two different statistical frameworks, and (c) keep our models as generic as possible.

3.1 Data and experimental setup

For this experiment, we use 57758 manually annotated sentences from FrameNet (release 2), corresponding to all the sentences with verbal predicates (2228 lemmata from 196 frames). Gildea and Jurafsky (2000) and Fleischman et al. (2003) used a previous release of the dataset with less annotated instances, but covered all predicates (verbs, nouns and adjectives).

Data preparation. After tagging the data with TnT (Brants, 2000), we parse them using the Collins parsing model 3 (Collins, 1997). We consider only the most probable parse for each sentence and simplify the resulting parse tree by removing all unary nodes. We lemmatise the head of each constituent with TreeTagger (Schmid, 1994).

Gold standard. We transform the FrameNet character-offset annotations for semantic roles into our constituent format by determining the *maximal projection* for each semantic role, i.e. the set of constituents that exactly covers the extent of the role. A constituent is assigned a role iff it is in the maximum projection of a role.

Classification procedure. The instances to be classified are all parse tree constituents. Since direct assignment of role labels to instances fails due to the preponderance of unlabelled instances, which make up 86.7% of all instances, we follow Gildea and Jurafsky (2000) in splitting the task into two sequential subtasks: first, *argument recognition* decides for each instance whether it bears a semantic role or not; then, *argument labelling* assigns a label to instances recognised as role-bearers. For the second step, we train frame-specific classifiers, since the frame-specificity of roles does not allow to easily combine training data from different frames.

Statistical modelling. We perform the classification twice, with two learners from different statistical frameworks, in order to make our results more representative for the different statistical models employed so far for the task. The first learner uses the maximum entropy (Maxent) framework, which has been applied e.g. by Fleischman et al. (2003). The model is trained with the `estimate` software, which implements the LMVM algorithm (Malouf, 2002)[4]. The second learner is an instance of a memory-based learning (MBL) algorithm, the k-nearest neighbour algorithm. We use the implementation provided by TiMBL (Daelemans et al., 2003) with the recommended parameters, namely $k = 5$, adopting modified value difference with gain ratio feature weighting as similarity metric.

3.2 Features

In accordance with our goal of keeping our models generic, we use a set of vary (syntactic and lexical) features which more than one study in the literature has found helpful, without optimising the features for the individual learners.

Constituent features: The first type of feature represents properties of the constituent in question. We use the phrase type and head lemma of each constituent; its preposition (if available); its position

[3] From the definition of the frame at `http://www.icsi.berkeley.edu/~framenet/`. Examples adapted from the FrameNet data, release 2.

[4] Software available for download at `http://www-rohan.sdsu.edu/ malouf/pubs.html`

relative to the predicate (left, right or overlapping); the phrase type of its mother constituent; whether it is an argument of the target, according to the parser; and the path between target and constituent as well as its length.

Sentence level features: The second type of feature describes the context of the current instance. The predicate is represented by its lemma, its part of speech, its (heuristic) subcategorisation frame, and its governing verb. We also compile a list of all the prepositions in the sentence.

3.3 Results

All results in this section are averages over F scores obtained using 10-fold cross validation. For each frame, we perform two evaluations, one in *exact match* and one in *overlap* mode. In exact match mode, an assignment only counts as a true positive if it coincides exactly with the gold standard, while in overlap mode it suffices that they are not disjoint. F scores are then computed in the usual manner.

Table 1 shows the performance of the different configurations over the complete dataset, and the standard deviation of these results over all frames. To illustrate the results for individual frames, Table 2 lists frame-specific performances for five randomly selected frames and how they varied over cross validation runs.

	Maxent	MBL
Exact Match	53.3 ± 10.8	56.9 ± 10.1
Overlap	70.0 ± 11.0	74.2 ± 10.0

Table 1: Overall F scores and standard deviation across frames for Experiment 1.

3.4 Analysis and Discussion

In terms of overall results, the MBL model outperforms the Maxent model by 3 to 4 points F-score. However, all our results lie broadly in the range of existing systems with a similar architecture (i.e. sequential argument identification and labelling): Gildea and Jurafsky (2002) report $F = 55.1$, and Fleischman et al. (2003) $F = 57.4$ for exact match evaluation. We assume that our feature formulation is more suitable for the MBL model. Also, we do not smooth the Maxent model, while we use the recommended optimised parameters for TiMBL.

Our most remarkable finding is the high amount of variance presented by the numbers in Table 1. Computed across frames, the standard deviation amounts to 10% to 11%, consistently across evaluation measures and statistical frameworks. Since these figures are results of a 10-fold cross validation run, it is improbable that the effect is solely due to chance splits into training and test data. This assessment is supported by Table 2, which shows that, while the performance on individual frames can vary largely (especially for small frames like ROBBERY), the average performance on all frames varies less than 0.5% over the cross validation runs.

Exact match	Maxent	MBL
APPEARANCE	50.5 ± 4.5	60.1 ± 7.3
AVOIDING	47.9 ± 5.0	51.3 ± 6.9
JUDGM._COMM.	57.0 ± 1.5	57.5 ± 3.4
ROBBERY	38.4 ± 19.1	37.9 ± 16.2
WAKING_UP	60.5 ± 11.4	64.4 ± 11.8
Overlap	**Maxent**	**MBL**
APPEARANCE	68.3 ± 4.0	75.0 ± 5.6
AVOIDING	68.6 ± 4.3	72.7 ± 5.9
JUDGM._COMM.	76.9 ± 1.6	77.6 ± 1.8
ROBBERY	61.2 ± 20.6	55.2 ± 17.6
WAKING_UP	75.1 ± 9.1	77.6 ± 7.8
Total Exact Match	53.3 ± 0.5	56.9 ± 0.4
Total Overlap	70.0 ± 0.4	74.2 ± 0.5

Table 2: F scores and standard deviations over cross validation runs for five random frames (Exp. 1).

The reasons which lead to the across-frames variance warrant investigation, since they may lead to new insights about the nature of the task in question, answering Fleischman et al.'s (2003) call for better models. Some of the plausible variables which might explain the variance are the number of semantic roles per frame, the amount of training data, and the number of verbs per frame.

However, we suggest that a fourth variable might have a more decisive influence. Seen from a linguistic perspective, semantic role assignment is just an application of *linking*, i.e. learning the regularities of the relationship between semantic roles and their possible syntactic realisation and applying this knowledge. Therefore, our main hypothesis is: The more varied the realisation possibilities of the verbs in a frame, the more difficult it is for the learner to learn the correct linking patterns, and therefore the more error-prone semantic role assignment. Even though this claim appears intuitively true, it has never been explicitly made nor empirically tested, and its consequences might be relevant for the design of future models of semantic role assignment.

As an example, compare the frame IMPACT, as exemplified by the instances in (1), with the frame INGESTION, which contains predicates such as *drink*, *consume* or *nibble*. While every sentence in (1) shows a different linking pattern, linking for

INGESTION is rather straightforward: the subject is usually the Ingestor, and the direct object is an Ingestible. This is reflected in the scores: $F = 57.0$ for IMPACT and $F = 70.4$ for INGESTION (exact match scores for the MBL model).

The most straightforward strategy to test for the different variables would be to perform multiple correlation analyses. However, this approach has a serious drawback: The results are hard to interpret when more than one variable is significantly correlated with the data, and this is increasingly probable with higher amounts of data points. Instead, we adopt a second strategy, namely to design a new data set in which all variables but one are controlled for and correlation can be tested unequivocally. The new experiment is explained in Section 5. Section 4 describes the quantitative model of argument structure required for the experiment.

4 Argument Structure and Frame Uniformity

In this section, we define the concepts we require to test our hypothesis quantitatively. First, we define argument structure for our data in a corpus-driven way. Then, we define the uniformity of a frame according to its variance in argument structure.

4.1 An Empirical Model of Argument Structure

Work in theoretical linguistics since at least Gruber (1965) and Jackendoff (1972) has attempted to account for the regularities in the syntactic realisation of semantic arguments. Models for role assignment also rely on these regularities, as can be seen from the kind of features used for this task (see Section 3.2), which are either syntactic or lexical. Thus, current models for automatic role labelling rely on the regularities at the syntax-semantics interface. Unlike theoretical work, however, they do not explicitly represent these regularities, but extract statistical properties about them from data.

The model of argument structure we develop in this section retains the central idea of linking theory, namely to model argument structure symbolically, but deviates in two ways from traditional work in order to bridge the gap to statistical approaches: (1), in order to emulate the situation of the learners, we use only the data available from the FrameNet corpus; this excludes e.g. the use of more detailed lexical information about the predicates. (2), to be able to characterise not only the possibility, but also the probability of linking patterns, we take frequency information into account.

Our definition proceeds in three steps. First, we define the concept of a *pattern*, then we define the *argument structure* of a predicate, and finally the argument structure of a frame.

Patterns. A pattern encodes the argument structure information present in one annotated corpus sentence. It is an unordered set of pairs of semantic role and syntactic function, corresponding to all roles occurring in the sentences and their realisations. The syntactic functions used in the FrameNet corpus are as follows[5]: COMP (complement), EXT (subject in a broad sense, which includes controlling subjects), OBJ (object), MOD (modifier), GEN (genitive modifier, as 'John' in *John's hat*). For example, Sentence (1-a) gives rise to the pattern

$$\{(\text{Impactee}, \textsc{Ext}), (\text{Impactor}, \textsc{Comp})\}$$

which states that the Impactee is realised as subject and the Impactor as complement.

Argument Structure for Predicates and Frames. For each verb, we collect the set of all patterns in the annotated sentences. The argument structure of a verb is then a vector $\vec{v}$, whose dimensionality is the number of patterns found for the frame. Each cell v_i is filled with the frequency with which pattern i occurs for the predicate, so that the vector mirrors the distribution of the occurrences of the *verb* over the possible patterns. Finally, the set of all vectors for the predicates in a frame is a model for the argument structure of the *frame*.

The intuition behind this formalisation is that two verbs which realise their arguments alike will show a similar distribution of patterns, and conversely, if they differ in their linking, these differences will be mirrored in different pattern distributions.

Example. If we only had the three sentences in (1) for the IMPACT corpus, the three occurring patterns would be {(Impactee, EXT), (Impactor, COMP)}, {(Impactor, EXT), (Result, COMP)}, and {(Impactors, EXT), (Place, MOD)}. The argument structure of the frame would be

$$\left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

containing the information for the predicates *strike*, *slam* and *collide*, respectively. The variation arises from differences in syntactic construction (e.g. passive vs. active), but also, more significantly, from lexical differences: *collide* accepts a reciprocal plural subject, i.e. an *Impactors* role, while *strike* does not. This model is very simple, but achieves the

[5] See Johnson et al. (2002) for details.

goal of highlighting the differences and similarities in the mapping between semantics and syntax for different verbs in a frame.

4.2 Uniformity of Argument Structure

At this point, we can define a measure to compute the uniformity of a frame from the frame's argument structure, which is defined as a set of integer-valued vectors.

Similarity metrics developed for vector space models are obvious candidates, but work in this area has concentrated on metrics for comparing two vectors, whereas we may have an arbitrary number of predicates per frame. Therefore, we borrow the concept of *cost function* from clustering, as exemplified by the well known sum-of-squares function used in the k-means algorithm (see e.g. Kaufman and Rousseeuw (1990)), which estimates the "cost" of a cluster as the sum of squared distances d between each vector $\vec{v}_i$ and the cluster centroid $\vec{c}$: [6]

$$C(\vec{v}_1, \ldots, \vec{v}_n) = \sum_{i=1}^{n} d(\vec{v}_i, \vec{c})^2$$

Under this view, a good cluster is one with a low cost, and the goal of the clustering algorithm is to minimise the average distance to the centroid. However, for our purposes it is more convenient for a good cluster to have a high rating. Therefore, we turn the cost function into a "quality" function. By replacing the distance function with a similarity function s, we say that a good cluster is one with a high average similarity to the centroid:

$$Q(\vec{v}_1, \ldots, \vec{v}_n) = \sum_{i=1}^{n} s(\vec{v}_i, \vec{c})^2$$

If we consider each frame to be a cluster and each predicate to be an object in the cluster, represented by the argument structure vector, the values of Q can be interpreted as a measure for frame uniformity: Verbs with a similar argument structure will have similar vectors, resulting in high values of Q for the frame, and vice versa.

What intuitively validates this formalisation is that frames are clusters of predicates grouped together on semantic grounds, i.e. predicates in a frame share a common set of arguments. What Q checks is whether the mapping from semantics to syntax is also similar.

In order to obtain an actual measure for frame uniformity, we take two further steps. First, we instantiate s with the cosine similarity cos, which has been found to be appropriate for a wide range of linguistic tasks (see e.g. Lee (1999)) and ranges between 0 (least similar) and 1 (identity):

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

Second, we normalise the values of Q, which grow in $O(n)$, the number of vectors, to $[0; 1]$, to make them interpretable analogously to values of the cosine similarity. Since this is possible in two different ways, we obtain two different measures for frame uniformity. The first one, which we call *normalised quality-based uniformity* (qU), simply divides the values by n:

$$qU(\vec{v}_1, \ldots, \vec{v}_n) = \frac{1}{n}\sum_{i=1}^{n}[cos(\vec{v}_i, \vec{c})]^2$$

The second measure, *weighted quality-based uniformity* (wqU), is a weighted average of the similarities. The weights are given by the vector sizes – in our case, the frequency of the predicates:

$$wqU(\vec{v}_1, \ldots, \vec{v}_n) = \frac{1}{\sum_{j=1}^{n} |\vec{v}_j|}\sum_{i=1}^{n} |\vec{v}_i|[cos(\vec{v}_i, \vec{c})]^2$$

The weighting lends more importance to well-attested predicates, limiting the amount of noise introduced by infrequent predicates. Therefore, our intuition is that wqU should be a better measure than qU for argument structure uniformity.

5 Experiment 2: Explaining the Variance With Argument Structure

With two measures for the uniformity of argument structure at hand, we now proceed to test our main hypothesis.

5.1 Data and Experimental Setup

As argued in Section 3.4, our aim in this experiment is to control for the most plausible sources of performance variance and isolate the influence of argument structure.

To meet this condition, we perform both the experiments and the uniformity measure calculation on a controlled subset of the data, with the condition that both the number of verbs and the number of sentences are the same for each frame.

Following the methodology in Keller and Lapata (2003), we divide the verbs into four frequency bands, frequency being absolute number of

[6] The centroid of a cluster is "a point in p-dimensional space found by averaging the measurement values along each dimension" (Kaufman and Rousseeuw, 1990, p. 112), so that it is the point situated at the "center" of the cluster.

annotated sentences: low (5), medium-low (12), medium-high (22), and high (38). We set the boundaries between the bands as the quartiles of all the verbs containing at least 5 annotated examples[7]. For each frame, 2 verbs in each frequency band are randomly chosen. This reduces our frame sample from 196 to 40. We furthermore randomly select a number of sentences for each verb which matches the boundaries between frequency bands, that is, all verbs in each frequency bands are artificially set to have the same number of annotated sentences. This method assures that all frames in the experiment have 8 verbs and 154 sentences, so that both the performance figures and the uniformity measures were acquired under equal conditions.

The models for semantic role assignment were trained in the same way as for Experiment 1 (see Section 3.1), using the same features. We also performed 10-fold cross validation as before. The uniformity measures qU and wqU were computed according to the definitions in Section 4.2.

5.2 Results and Discussion

Table 3 shows the overall results and variance across frames for the new dataset. Table 4 contains detailed performance results (Columns 1 and 2) and uniformity figures (Columns 3 and 4) for five randomly drawn frames.

	Maxent	MBL
Exact Match	47.5 ± 11.0	53.4 ± 11.1
Overlap	66.4 ± 11.0	72.4 ± 9.9

Table 3: Overall F scores and standard deviation across frames for Experiment 2.

The overall results for the new, controlled dataset are 3 to 5 points F-score worse than in Experiment 1, which is a result of the artificial limitation of larger frames to fewer training examples. Otherwise, the same tendencies hold: The memory-based learner again performs better than the maximum entropy learner, and overlap evaluation returns higher scores than exact match. More relevantly, the data show the same amount of variance across frames as before (between 10 and 11%), even though the most plausible sources of variance are controlled for. The variation over cross validation runs is somewhat larger, but still small (2.0%/1.9% for Maxent and 0.9%/0.8% for MBL, respectively).

We can now test our main hypothesis through an analysis of the correlation between performance and uniformity figures. We log-transformed both variables to guarantee normal distribution and used the standard Pearson product-moment correlation coefficient, testing for positive correlation (higher uniformity – higher performance). The results in Table 5 show that all correlation tests are significant, and most are highly significant. This constitutes very good empirical support for our hypothesis.

Exact match	Maxent	MBL	qU	wqU
BODY_MOVMT.	51.2	57.5	33.0	39.0
COMMERCE	25.7	41.9	27.4	31.1
MOTION	54.6	58.1	57.2	60.8
PERC._ACTIVE	52.1	51.5	30.0	35.4
REMOVING	59.3	60.1	58.7	64.2

Overlap	Maxent	MBL	qU	wqU
BODY_MOVMT.	56.4	64.8	33.0	39.0
COMMERCE	48.9	66.4	27.4	31.1
MOTION	68.1	71.9	57.2	60.8
PERC._ACTIVE	69.3	69.0	30.0	35.4
REMOVING	76.1	77.2	58.7	64.2

Table 4: F scores and frame uniformities for data from Exp. 2. qU = normalised uniformity, wqU = weighted uniformity (in percentages).

Exact match	Maxent	MBL
qU	0.39 (p=0.007)	0.33 (p=0.04)
wqU	0.45 (p=0.002)	0.35 (p=0.01)

Overlap	Maxent	MBL
qU	0.54 (p<0.001)	0.50 (p<0.001)
wqU	0.58 (p<0.001)	0.55 (p<0.001)

Table 5: Pearson coefficients ρ^2 and significance levels for correlating frame performance and frame uniformity for the dataset from Experiment 2.

We find that wqU yields consistently higher correlation measures (and therefore more significant correlations) than qU, which supports our hypothesis from Section 4 that wqU is a better measure for argument structure uniformity. Recall that the intuition behind the weighting is to let well-attested predicates (those with higher frequency) have a larger influence upon the measure. However, an independent experiment for the adequacy of the measures should be devised to verify this hypothesis.

A comparison of the evaluation modes shows that frame uniformity correlates more strongly with the overlap evaluation measures than with exact match. We presume that this is due to the evaluation figures in exact match mode being somewhat noisier. All other things being equal, random errors introduced

[7] We consider 5 to be the (very) minimum number of instances necessary to construct a representative argument structure for a predicate.

during the different processing stages (e.g. parsing errors) are more likely to influence the exact match outcome: A processing error which leads to a partially right argument assignment will influence the outcome of the exact match evaluation, but not of the overlap evaluation.

As for the two statistical frameworks, uniformity is better correlated with the Maxent model than with the MBL model, even though MBL performs better on the evaluation. However, this does not mean that the correlation will become weaker for semantic role labelling systems performing at higher levels of accuracy. We compared our current models with an earlier version, which had an overall lower performance of about 5 points F-score. Using the same data, the correlation coefficients ρ^2 were on average 0.09 points lower, and the p-values were not significant for the Maxent model in exact match mode. This indicates that correlations tend to increase for better models.

Therefore, we attribute the difference between the Maxent and the MBL model to their individual properties, or more specifically to differences in the distribution of the performance figures for the individual frames around the mean. While they are more evenly distributed in the MBL model, they present a higher peak with more outliers in the Maxent model, which is also reflected in the slightly higher standard deviation of the Maxent model (cf. Tables 1 and 3). In short, the Maxent model appears to be more sensitive to differences in the data.

Nevertheless, both models correlate strongly with each other in both evaluation modes ($\rho^2 = 0.79$, p<0.001 for exact match, $\rho^2 = 0.84$, p<0.001 for overlap). Thus, they agree to a large extent on which frames are easy or difficult to label.

Our present results, thus, seem to indicate that the influence of argument structure cannot be solved by simply improving existing systems or choosing other statistical frameworks. Instead, there is a systematic relationship between the uniformity of the argument structures of the predicates in the frames and the performance of automatic role assignment.

6 Conclusion and Outlook

In this paper, we have performed an error analysis for semantic role assignment, concentrating on the relationship between argument structure and semantic role assignment. To obtain general results, we kept our models as general as possible and verified our results in two different statistical frameworks.

In our first experiment, we showed that there is considerable variance across frames in the performance of semantic role assignment, and hypothesised that the effect was due to the varying "difficulty" of the underlying argument structure. To test the hypothesis, we defined a measure of frame uniformity which modelled the variability of argument structure. In a second experiment, in which we controlled for other plausible sources of variance, we showed a reliable correlation between performance and uniformity figures.

The underlying reason for the difficulty of semantic role assignment is that FrameNet is essentially an ontological classification. While the predicates of one frame share the same semantic arguments, they can vary widely in their linking patterns. Without unlimited training data, automatic role assignment has to find and exploit regularities in linking to achieve good results. A priori, this can only be done within frames, since roles are frame-specific, and there is no unique right mapping between roles.

Consequently, as observed by Fleischman et al. (2003), relatively rare constructions, such as passives, are frequent error sources. Because such constructions have to be learnt individually for each frame, data sparseness is a serious issue. A similar problem arises for lexical differences in the linking properties of predicates in a frame, as with the *collide* vs. *strike* case discussed above. Here, the learning has to take into account that the relevant linking properties differ between individual predicates.

Our results suggest that the variance caused by argument structure will not disappear with better classifiers, but that the problem of inadequate generalisations should be addressed in a principled way. There are several possible approaches to do so.

First, the classic statistical approach: Combining evidence from different frame-specific roles to alleviate data sparseness. To this end, Gildea and Jurafsky (2002) developed a mapping from frame-specific to syntactic roles, but results did not improve much. Baldewein et al. (2004) experiment with EM-driven generalisation, and obtain also only modest improvements.

A second approach is to identify other levels, different from frames, at which regularities can be learnt better. One possibility is to identify smaller units within frames which have a more uniform structure and which can be learnt more easily. Since uniformity is defined in terms of a quality function, clustering would be the natural method to employ for this task. However, this method is only viable for frames with a large amount of annotation.

A more general idea in this spirit is to construct an independent classification of verbs motivated at the argument structure level (transitive, intransitive, unaccusative, etc.), e.g. using data sources like Levin's verb classes (Levin, 1993). This would allow models to learn class-specific regularities and diathesis alternations more easily. However, it is unclear if there is a unique level at which all relevant regularities can be stated. A more realistic variant might be to map FrameNet roles to an existing, more syntactically oriented role set, such as PropBank. These roles can serve as an intermediate level to capture mapping regularities, and can be translated back to semantically defined FrameNet roles when the mapping has been accomplished.

A third, different approach to semantic role assignment is presented by Frank (2004), who presents a syntax-semantics interface to extract symbolic frame element projection rules from an LFG-annotated corpus and discusses strategies to generalise over these rules. Such an approach is, due to the finer control over the generalisation, not as susceptible to the problem described in this study as purely statistical models. However, it has yet to be tested on large-scale semantic role assignment.

Acknowledgements

We are grateful to Katrin Erk, Alexander Koller and three anonymous reviewers for helpful comments on previous versions of the paper. We also thank the audiences at the *Prospects and Advances in the Syntax/Semantics Interface* Workshop in Nancy and the Computational Linguistics Seminar at UPF for their suggestions. This work is supported by the Departament d'Universitats, Recerca i Societat de la Informació (grant 2001FI-00582, Gemma Boleda).

References

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

U. Baldewein, K. Erk, S. Pado, and D. Prescher. 2004. Semantic role labelling with similarity-based generalisation using EM-based clustering. In *Proceedings of SENSEVAL-3*, Barcelona, Spain.

Thorsten Brants. 2000. TnT - a statistical part-of-speech tagger. In *Proceedings of ANLP-2000*, Seattle, WA.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of ACL-97*, Madrid, Spain.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2003. Timbl: Tilburg memory based learner, version 5.0, reference guide. Technical Report ILK 03-10, Tilburg University. Available from `http://ilk.uvt.nl/downloads/pub/papers/ilk0310.ps.gz`.

Katrin Erk, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. 2003. Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. In *Proceedings of ACL-03*, Sapporo, Japan.

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, IV(2).

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

Anette Frank. 2004. Generalisations over corpus-induced frame assignment rules. In *Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora*, pages 31–38, Lissabon, Portugal.

Dan Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *Proceedings of ACL-00*, pages pages 512–520, Hong Kong.

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, Cambridge, MA.

C. R. Johnson, C. J. Fillmore, M. R. L. Petruck, C. F. Baker, M. J. Ellsworth, J. Ruppenhofer, and E. J. Wood. 2002. FrameNet: Theory and Practice. `http://www.icsi.berkeley.edu/~framenet/book/book.html`.

L. Kaufman and P. J. Rousseeuw. 1990. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, New York City, NY.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29:459–484.

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 *Proceedings of ACL-99*, pages 25–32, College Park, MD.

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

Rob Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proc. of CoNLL-02*, Taipei, Taiwan.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of NeMLaP*, Manchester, UK.

Cynthia A. Thompson, Roger Levy, and Christopher Manning. 2003. A generative model for FrameNet semantic role labeling. In *Proceedings of ECML-03*, Cavtat, Croatia.