Semantic Role Labelling with Similarity-Based Generalization Using EM-based Clustering

Ulrike Baldewein, Katrin Erk, Sebastian Padó
Saarland University
Saarbrücken, Germany
{ulrike,erk,pado}@coli.uni-sb.de

Detlef Prescher
University of Amsterdam
Amsterdam, The Netherlands
prescher@science.uva.nl

Abstract

We describe a system for semantic role assignment built as part of the Senseval III task, based on an off-the-shelf parser and Maxent and Memory-Based learners. We focus on generalisation using several similarity measures to increase the amount of training data available and on the use of EM-based clustering to improve role assignment. Our final score is Precision=73.6%, Recall=59.4% (F=65.7).

1 Introduction

This paper describes a study in semantic role labelling in the context of the Senseval III task, for which the training and test data were both drawn from the current FrameNet release (Johnson et al., 2002). We concentrated on two questions: first, whether role assignment can be improved by generalisation over training instances using different similarity measures; and second, the impact of EM-based clustering, both in deriving more informative selectional preference features and in the generalisations mentioned above. The basis of our experiments was formed by off-the-shelf statistical tools for data processing and modelling.

After listing our data preparation steps (Sec. 2) and features (Sec. 3), we describe our classification procedure and the learners we used (Sec. 4). Sec. 5 outlines our experiments in similarity-based generalisations, and Section 6 discusses our results.

2 Data and Instances

Parsing. To tag and parse the data, we used LoPar (Schmid, 2000), a probabilistic context-free parser, which comes with a Head-Lexicalised Grammar for English (Carroll and Rooth, 1998). We considered only the most probable parse for each sentence and simplified parse trees by eliminating unary nodes. The resulting nodes form the instances of our classification. We used the Stuttgart TreeTagger (Schmid, 1994) to lemmatise constituent heads.

Projection of role labels. FrameNet provides semantic roles as character offsets. We labelled those instances (i.e. nodes in the parse tree) with gold standard semantic roles which corresponded to roles' *maximal projections*. 13.95% of roles in the training corpus spanned more than one parse tree node. Figure 1 shows an example sentence for the AWARENESS frame. The nodes' respective semantic role labels are given in small caps, and the target predicate is marked in boldface.

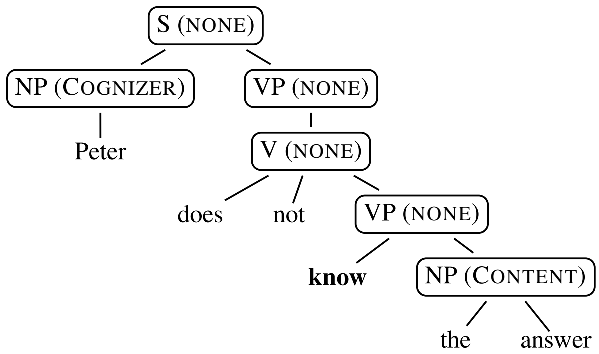

Figure 1: Example parse tree with role labels

Semantic clustering. We used clustering to generalise over possible fillers of roles. In a first model, we derived a probability distribution $p(y)$ for pairs $y = (y_1, y_2)$, where y_1 is a target:role combination and y_2 is the head lemma of a role filler. The key idea is that y_1 and y_2 are mutually independent, but conditioned on an unobserved class $c \in C$. In this manner, we define the probability of $y = (y_1, y_2) \in \mathcal{Y}_1 \times \mathcal{Y}_2$ as:

$$\begin{aligned} p(y) &= \sum_{c \in C} p(c, y) = \sum_{c \in C} p(c) p(y|c) \\ &= \sum_{c \in C} p(c) p(y_1|c) p(y_2|c) \end{aligned}$$

Estimation was performed using a variant of the expectation-maximisation algorithm (Prescher et al., 2000). We used this model both as a feature and in the generalisation described in Sec. 5. In a second model, we clustered pairs of target:role and the syntactic properties of the role fillers; the resulting model was only used for generalisation.

3 Features

Constituent features. The first group of features represents properties of instances (i.e. constituents). We used the phrase type and head lemma of each constituent, its preposition, if any (otherwise NONE), its relative position with respect to the target (left, right, overlapping), the phrase type of its mother node, and the simplified path from the target to the constituent: all phrase types encountered on the way, and whether each step was *up* or *down*. Two further features stated whether this path had been seen as a frame element in the training data, and whether the constituent was subcategorised for (determined heuristically).

Sentence level features. The second type of feature described the context of the current instance: The target word was characterised by its lemma, POS, voice, subcat frame (determined heuristically), and its governing verb; we also compiled a list of all prepositions in the sentence.

Semantic features. The third type of features made use of EM-based clustering, stating the most probable label assigned to the constituent by the clustering model as well as a confidence score for this decision.

4 Classification

We first describe our general procedure, then the two different machine learning systems we used.

Classification Procedure. As the semantic role labels of FrameNet are frame-specific, we decided to train one classifier for each frame. To cope with the large amount of constituents bearing no role label, we divided the procedure into two steps, distinguishing *argument identification* and *argument labelling*. First, argument identification decides for all constituents whether they are role-bearers or not. Then, argument labelling assigns semantic roles to those sequences classified as role-bearing. In our example (Fig. 1), the first step of classification ideally would single out the two NPs as possible role fillers, while the second step would assign the COGNIZER and CONTENT roles.

Maximum Entropy Learning. Our first classifier was a log-linear model, where the probability of a class c given an feature vector $\vec{v}$ is defined as

$$p(c|\vec{v}) = \frac{1}{Z} \prod_i e^{\alpha_i f_i(v,c)}$$

where Z is a normalisation constant, $f_i(v, c)$ the value of feature v_i for class c, and α_i the weight assigned to f_i. The model is trained by optimising the weights α_i subject to the *maximum entropy* constraint which ensures that the *least committal* optimal model is learnt. Maximum Entropy (Maxent) models have been successfully applied to semantic role labelling (Fleischman et al., 2003). We used the `estimate` software for estimation, which implements the LMVM algorithm (Malouf, 2002) and was kindly provided by Rob Malouf.

Memory-based Learning. Our second learner implements an instance of a memory-based learning (MBL) algorithm, namely the k-nearest neighbour algorithm. This algorithm classifies test instances by assigning them the label of the most similar examples from the training set. Its parameters are the number of training examples to be considered, the similarity metric, and the feature weighting scheme. We used the implementation provided by TiMBL (Daelemans et al., 2003) with the default parameters, i.e. k=1 and the weighted overlap similarity metric with gain ratio feature weighting.

5 Similarity-based Generalisation over Training Instances

FrameNet role labels are frame-specific. This makes it necessary to either train individual classifiers with little training data per frame, or train a large classifier with many sparse classes. So one important question is whether we can *generalise*, i.e. exploit similarities between frame elements, to gain more training data.

We experimented with different generalisation methods, all following the same basic idea: If frame element A1 of frame A and frame element B1 of frame B are similar, we re-use A1 training data as B1 instances. In this process, we mask out features which might harm learning for A1, such as targets or sentence level features, or semantic features in case of syntactic similarities (and vice versa). We explored three types of role similarities, two based on symbolic information from the FrameNet database, and one statistical.

Frame Hierarchy. FrameNet specifies frame-to-frame relations, among them three that order frames hierarchically: *Inheritance*, the *Uses* relation of partial inheritance, and the *Subframe* relation linking larger situation frames to their individual stages. All three indicate semantic similarity between (at least some) frame elements; in some cases corresponding frame elements are also syntactically similar, e.g. the Victim role of Cause_harm and the Evaluee role of Corporal_punishment are both typically realised as direct objects.

Peripheral frame elements. FrameNet distinguishes core, extrathematic, and peripheral frame elements. Peripheral frame elements are frame-independent adjuncts; however the same frame element may be peripheral to one frame and core to another. So we took a peripheral frame element as similar to the same peripheral frame element in other frames: Given an instance of a peripheral frame element, we used it as training instance for all frames for which it was marked as peripheral in the FrameNet database.

Group 6: puzzle:Experiencer_obj.Stimulus, increase:Change_position_on_a_scale.Item, praise:Judgment_communication.Communicator, travel:Travel. Traveler, …
Group 11: lodge:Residence.Location, scoff:Judgment_communication.Evaluee, chug:Motion_noise.Path, emerge:Departing.Source, …

Figure 2: EM-based syntactic clustering: excerpts of 2 clusters

EM-based clustering. The EM-based clustering methods introduced in Sec. 2 measure the "goodness of fit" between a target word and a potential role filler. We now say that two frame elements are similar if they are appropriate for some common cluster. For the head lemma clustering model, we define the appropriateness $w_c(tr)$ of a target:role pair tr for a cluster c as follows:

$$w_c(tr) = \sum_{\ell \text{ with } f(\ell,tr)>0} f(\ell)p(c|\ell)$$

where $w_c(tr)$ is the total frequency of all head lemmas ℓ that have been seen with tr, weighted by the class-membership probability of ℓ in c. This appropriateness measure $w_c(tr)$ is built on top of the class-based frequencies $f(\ell)p(c|\ell)$ rather than on the frequencies $f(\ell)$ or the class-membership probabilities $p(c|\ell)$ in isolation: For some tasks the combination of lexical and semantic information has been shown to outperform each of the single information sources (Prescher et al., 2000). Our similarity notion is now formalised as follows: With a threshold θ as a parameter, two frame elements tr_1, tr_2 count as similar if for some class c, $w_c(tr_1) > \theta$ and $w_c(tr_2) > \theta$.

In the syntactic clustering model, a role filler was described as a combination of the path from instance to target, the instance's preposition, and the target voice. The appropriateness of a target:role pair is defined as for the above model. For time reasons, only verbal targets were considered.

Figure 2 shows excerpts of two "syntactic" clusters in the form of target:frame.role members. Group 6 is a very homogeneous group, consisting of roles that are usually realised as subjects. Group 11 contains roles realised as prepositional phrases, but with very diverse prepositions, including *in*, *at*, *along*, and *from*.

6 Results and Discussion

We first give the final results of our systems on the test set according to the official evaluation software. Then we discuss detailed results on a development set we randomly extracted from the training data.

6.1 Final Results

We submitted the results of two models. One was produced using the maximum entropy learner, including all features of Sec. 3 and with the three most helpful generalisation techniques (EM head lemma, EM path, and Peripherals). For the second model we used the MBL learner trained on all features, with no additional training data[1]. The performance of the two models is shown in Table 1.

	Maxent	MBL
Precision	73.6%	65.4%
Recall	59.4%	47.1%
F-score	65.7	54.8
Coverage	80.7%	72.0%
Overlap	67.5%	60.2%

Table 1: Test set results (official scoring scheme)

6.2 Detailed Results

For a detailed evaluation, we randomly split off 10% of the training data to form development sets. In this section, we report results of two such splits to take chance variation into account.

For time reasons, this detailed evaluation was performed using our own evaluation software, which is based on our internal constituent-based representation. This software gives the same tendencies (improvements / deteriorations) as the official software, but absolute values differ; so we restrict ourselves to reporting relative figures.

Basis for Comparison. All following models are compared against a set of basic models trained on *all* features of Sec. 3. Table 2 gives the results for these models, using our own scoring software.

Contribution of Features. We computed the contribution of individual features by leaving out each feature in turn. Table 3 shows the results, averaging

[1]For time reasons, we were not able to test generalisation in the Memory-Based Learning paradigm.

	1st split	2nd split
Maxent	F=80.02	F=80.86
MBL	F=86.43	F=85.66

Table 2: Devel set results (own scoring scheme)

	Δ F-score	
Feature	MBL	Maxent
head lemma	0	0.6
emmc label	**3.9**	**3.9**
emmc prob	-0.3	**1.8**
mother phrase type	-0.7	-0.3
governing verb	-0.1	-0.5
is subcategorized	-0.1	-0.5
path	0.2	0.5
path length	-0.5	-0.5
path seen	**1.6**	**3.4**
preposition	0	-0.3
all preps	-0.2	-0.7
phrase type	**1.2**	**2.2**
position	0.5	0.3
sc frame	0.1	-0.2
target lemma	0	-0.6
target POS	0.1	-0.3
voice	0.1	-0.3

Table 3: Contribution of each feature

FN hierarchy (sem): ~ 10,000 instances head lemma
FN hierarchy (syn): ~ 10,000 instances phrase type, path, prep., path seen, is subcategorised, voice, target POS
Peripherals: ~ 55,000 instances head lemma, phrase type, path, prep., path seen, is subcategorised, voice, target POS
EM head: ~ 1,000,000 instances head lemma
EM path: ~ 433,000 instances phrase type, mother phrase type, path, path length, prep., path seen, is subcategorised, voice, target POS

Table 4: Similarity-based generalisation: Features retained and number of generated instances

	Δ F-score	
Strategy	Split 1	Split 2
FN hierarchy (sem)	0.3	-0.5
FN hierarchy (syn)	-0.2	-0.4
Peripherals	0.2	-0.1
EM head	0.4	0.5
EM path	1.0	0.2

Table 5: Contribution of generalization strategies

over the two splits. The features that contributed most to the performance were the same for both learners: the label assigned by the EM-based model, the phrase type, and whether the path had been seen to lead to a frame element. The relative position to the target helped in one MBL and one Maxent run. Interestingly, the Maxent learner profits from the probability with which the EM-based model assigns its label, while MBL does not.

Generalisation. To measure the effect of each of the similarity measures listed in Sec. 5, we tested them individually using the Maximum Entropy learner with all features.

As mentioned above, training instances of one frame were generalised and then added to the training instances of another, retaining only part of the features in the generalisation. Table 4 shows the features retained for each similarity measure, as well as the number of additional instances generated, summed over all frames. We empirically determined the optimal parameter values as: For *FN-h (sem)* and *FN-h (syn)*, 1 level in the hierarchy; for *EM head*, a weight threshold of $\theta = 20$, and for *EM path*, a weight threshold of $\theta = 10$.

Table 5 gives the improvements made over the baseline through adding data gained by each generalisation strategy. Results are shown in points F-score and individually for both training/development splits. EM-based clustering proved to be helpful, showing both the highest single improvement (*EM path*) and the highest consistent improvement (*EM head*), while all other generalisations show mixed results.

Combining the three most promising generalisation techniques (Peripherals, EM head, and EM path) led to an improvement of 0.7 points F-score for split 1 and 1.1 points F-score for split 2.

6.3 Discussion.

Feature quality. The features that improved the learners' performance most are EM-based label, phrase type and the "path seen as FE". The other features did not show much impact for us. The Maxent learner was negatively affected by sentence-level features such as the subcat frame and "is subcategorised".

Comparing the learners. In a comparable basic setting (all features, no generalisation), the Memory-Based learner easily outperforms the Maxent learner, according to our scoring scheme. However, the official scoring scheme determines the Memory-based learner's performance at more than

10 points F-score below the Maxent learner. We intend to run the Memory-based learner with generalisation data for a more comprehensive comparison.

Generalisation. Gildea and Jurafsky (2002) report an improvement of 1.6% through generalisation, which is roughly comparable to our figures. The two strategies share the common idea of exploiting role similarities, but the realisations are converse: Gildea and Jurafsky manually compact similar frame elements into 18 abstract, frame-independent roles, whereas we keep the roles frame-specific but augment the training data for each by automatically discovered similarities.

One reason for the disappointing performance of the FrameNet hierarchy-based generalisation strategies may be simply the amount of data, as shown by Table 4: *FN-h (sem)* and *FN-h (syn)* each only yield 10,000 additional instances as compared to around 1,000,000 for *EM head*. That the reliability of the results roughly seems to go up with the number of additional instances generated (Peripherals: ca. 50,000, EM-Path: ca. 400,000) fits this argumentation well.

The input to the *EM path* clusters is a tuple of the path, target voice and preposition information. In the resulting model, generalisation over voice worked well, yielding clusters containing both active and passive alternations of similar frame elements. However, prepositions were distributed more arbitrarily. While this may indicate problems of clustering with more structured forms of input, it may also just be a consequence of noisy input, as the preposition feature has not had much impact either on the learners' performance.

The *EM head* strategy adds large amounts of head lemma instances, which probably alleviates the sparse data problem that makes the head lemma feature virtually useless. Another way of capitalising on this type of information would be to use the FN hierarchy generalisation to derive more input for EM-based clustering and see if this indirect use of generalisation still improves semantic role assignment. Interestingly, the *EM head* strategy and the EM-based clustering feature, both geared at solving the same sparse data problem, do not cancel each other out. In future work, we will try to combine the *EM head* strategy with the FrameNet hierarchy to derive more input for the clustering model to see if this can improve the present generalisation results.

Comparison with CoNLL. We recently studied semantic role labelling in the context of the CoNLL shared task (Baldewein et al., 2004). The two key differences to this study were that the semantic roles in question were PropBank roles and that only shallow information was available. Our system there showed two main differences to the current system: the overall level of accuracy was lower, and EM-based clustering did not improve the performance. While the performance difference is evidently a consequence of only shallow information being available, it remains an interesting open question why EM-based clustering could improve one system, but not the other.

References

U. Baldewein, K. Erk, S. Pado, and D. Prescher. 2004. Semantic role labelling with chunk sequences. In *Proceedings of CoNLL-2004*.

G. Carroll and M. Rooth. 1998. Valence induction with a head-lexicalized PCFG. In *Proceedings of EMNLP-1998*.

W. Daelemans, J. Zavrel, K. van der Sloot, and A. 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`.

M. Fleischman, N. Kwon, and E. Hovy. 2003. Maximum entropy models for FrameNet classification. In *Proceedings of EMNLP-2003*.

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

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`.

R. Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of CoNLL-2002*.

D. Prescher, S. Riezler, and M. Rooth. 2000. Using a probabilistic class-based lexicon for lexical ambiguity resolution. In *Proceeedings of COLING-2000*.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of NeMLP-1994*.

H. Schmid, 2000. *LoPar – Design und Implementation*. Institute for Computational Linguistics, University of Stuttgart.