Analyzing models for semantic role assignment using confusability

Katrin Erk and **Sebastian Padó**
Computational Linguistics
Saarland University
Saarbrücken, Germany
{erk,pado}@coli.uni-sb.de

Abstract

We analyze models for semantic role assignment by defining a *meta-model* that abstracts over features and learning paradigms. This meta-model is based on the concept of *role confusability*, is defined in information-theoretic terms, and predicts that roles realized by less specific grammatical functions are more difficult to assign. We find that confusability is strongly correlated with the performance of classifiers based on syntactic features, but not for classifiers including semantic features. This indicates that syntactic features approximate a description of grammatical functions, and that semantic features provide an independent second view on the data.

1 Introduction

Semantic roles have become a focus of research in computational linguistics during the recent years. The driving force behind this interest is the prospect that semantic roles, as a shallow meaning representation, can improve many NLP applications, while still being amenable to automatic analysis. The benefit of semantic roles has already been demonstrated for a number of tasks, among others for machine translation (Boas, 2002), information extraction (Surdeanu et al., 2003), and question answering (Narayanan and Harabagiu, 2004).

Robust and accurate *automatic semantic role assignment*, a prerequisite for the wide-range use of semantic roles in NLP, has been investigated in a number of studies and shared tasks. Typically, role assignment has been modeled as a classification task, with models being estimated from large corpora (Gildea and Jurafsky, 2002; Moschitti, 2004; Xue and Palmer, 2004; Surdeanu et al., 2003; Pradhan et al., 2004; Litkowski, 2004; Carreras and Màrquez, 2005).

Within this framework, there is a number of architectural parameters which lend themselves to optimization: the machine learning framework, the feature set, pre- and postprocessing, each of which has been investigated in the context of semantic role assignment. The current paper concentrates on feature engineering, since the feature set is a pivotal component of any kind of machine learning system, and allows us to incorporate and test linguistic intuitions on the role assignment task.

We approach feature engineering not by directly optimizing system performance. Instead, we proceed by *error analysis*, like Pado and Boleda (2004). Our aim is to form a *global hypothesis* that explains the distribution of errors across classes. Insofar as the model does not contain model-specific information, following this methodology can provide a *meta-model* of a model family which abstracts over concrete features and over the learning paradigm.

The concrete global hypothesis we test is: (1) All features of current models approximate a description of grammatical functions, and the complete systems approximate an assignment based on grammatical functions. (2) System performance for a given role depends on how easily it is confused with other roles. We will give this concept of *role confusability* a formal, information-theoretic definition.

The present study specifically analyzes models for semantic role assignment in the FrameNet

paradigm (Fillmore et al., 2003). We are going to show that our hypothesis indeed holds for a variety of models – but only models that comprise exclusively syntactic features. We conclude that syntactic features approximate a description of grammatical functions, but that semantic features model a different aspect of the role assignment mapping. Together with the reasonable performance of a solely semantics-based system, this leads us to suggest a closer investigation of semantic features – and in particular, a co-training approach with syntactic and semantic features as different views on the role assignment data.

Plan of the paper. In Section 2, we give a brief introduction to FrameNet, the semantic role paradigm and corpus we are using in this study. Our first experiment, described in Section 3, establishes that there is a high variance in performance across roles, and that this variance is itself stable across models and learners. In Section 4, we state our hypothesis, namely that this variance can be explained through role confusability, and formalize the concept . In Section 5, we perform detailed correlation tests to verify our hypothesis and discuss our findings. Section 6 concludes the paper.

2 FrameNet

This section presents the semantic role paradigm and the role-annotated corpus on which the present study is based. FrameNet[1] is a lexical resource based on Fillmore's Frame Semantics (Fillmore, 1985). It describes *frames*, representations of prototypical situations. Each frame provides its set of semantic roles, the entities or concepts pertaining to the prototypical situation. Each frame is further associated with a set of *target predicates* (nouns, verbs or adjectives), occurrences of which can introduce the frame.

FrameNet provides manually annotated examples for each predicate, sampled from the British National Corpus (Burnard, 1995). The size of this corpus exceeds 135,000 sentences. The following sentences are examples 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":

(1) [$_{\text{Impactee}}$ His car] was **struck** [$_{\text{Impactor}}$ by a third vehicle].

(2) [$_{\text{Impactor}}$ The door] **slammed** [$_{\text{Result}}$ shut].

(3) [$_{\text{Impactors}}$ Their vehicles] **collided** [$_{\text{Place}}$ at Pond Hill].

FrameNet manual annotation also comprises a layer of grammatical functions: For example, the subject of finite verbs is labeled `Ext`, and `Mod` is a label used for modifiers of heads, e.g. an adjective modifying a noun. The grammatical functions used in FrameNet are listed in Fillmore and Petruck (2003).

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

3 Experiment 1: Variance in role assignment

Several studies have established that there is considerable variance in semantic role assignment performance across different semantic roles within systems (Carreras and Màrquez, 2004; Carreras and Màrquez, 2005; Pado and Boleda Torrent, 2004). However, these studies used either the PropBank semantic role paradigm (Carreras and Màrquez) or a limited of experimental conditions (Pado and Boleda). For this reason, we perform a first experiment to replicate this phenomenon in our setting.

Note that the vast majority of participant systems in recent shared tasks divides semantic role assignment into multiple sequential steps. The maximal decomposition is as follows: *preprocessing*, e.g. removal of unlikely argument candidates; *argument recognition*, the distinction between role-bearing and non-role-bearing instances; *argument labeling*, the actual classification of role-bearing instances; and *postprocessing*, e.g. by inference over probable role sequences.

Following this distinction, we concentrate in this study on the argument labeling step, i.e. distinguishing between roles, rather than distinguishing roles

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

from non-roles. This is justified by earlier empirical results, namely that the argument labeling step requires more training data than argument recognition (Fleischmann and Hovy, 2003), and that it calls for more sophisticated feature construction (Xue and Palmer, 2004). We take this as evidence that the quality of the argument labeling step is central to a good semantic role assignment system.

In order to isolate the effects of argument labeling, we assume perfect argument recognition by using gold standard role boundaries; however, we do not use gold standard parse trees, but rather automatically computed ones, which realistically introduces some noise (see the following paragraph).

Data and preprocessing. As experimental material, we used the same data that was used in the Senseval-3 semantic role assignment task: 40 frames from FrameNet version 1.1, comprising 66,777 instances. The number of roles per frame ranged from 2 to 22, and the number of role instances ranged from 593 to 8,378. The data was randomly split into training (90%) and test instances (10%).

The data was parsed with the Collins model 3 (1996) parser; in addition, all tokens were lemmatized with TreeTagger (Schmid, 1994).

Modeling. We model role assignment as a classification task, with parse tree constituents as instances to be classified. We repeated the classification with two different learners: The first learner, TiMBL (Daelemans et al., 2003) is an implementation of nearest-neighbor classification algorithms in the memory-based learning paradigm[2]. The second learner, Malouf's probabilistic maximum entropy (Maxent) system (Malouf, 2002), uses the LMVM algorithm to estimate log-linear models. We did not perform smoothing.

Table 5 shows the features we use. Here as in the system setup, we keep close to current existing models for semantic role assignment in order to make our results as representative as possible. We investigate different feature sets in order to verify our results. In Exp. 1, we limit ourselves to two feature sets, *Syn* (syntactic features) and *Sem* (lexical features) from the bottom of Table 5. The feature sets were exactly the same for both learners.

[2]TiMBL was set to k-NN classification, using the MVDM distance metric and 5 neighbors.

	Syn/Sem	Syn
MBL	87.1 ± 12.7	82.2 ± 17.8
Maxent	87.5 ± 13.4	$82.4 \pm 18.\ 2$

Table 1: Exp. 1: Overall results (F-scores and standard deviation across roles).

	Syn/Sem		Syn	
Role	F_{MBL}	F_{Maxent}	F_{MBL}	F_{Maxent}
Frame: CHANGE_POSITION_ON_A_SCALE				
ATTR	79.0	80.7	57.6	66.1
CO_VAR	55.6	64.0	22.2	31.6
DIFF	87.1	84.9	75.0	66.7
ITEM	68.6	70.3	48.0	61.3
VALUE_1	88.0	91.7	78.3	72.7
VALUE_2	93.3	90.9	89.3	85.2
Frame: KINSHIP				
ALTER	87.0	89.2	87.8	87.4
EGO	96.7	98.8	96.7	95.5
Frame: PART_ORIENTATIONAL				
PART	98.2	96.4	97.6	97.0
WHOLE	100	100	98.2	100
Frame: TRAVEL				
AREA	31.6	52.6	25.0	45.5
GOAL	74.4	71.4	68.3	62.2
MODE	46.2	72.7	12.5	15.4
PATH	66.7	53.3	50.0	40.0
SOURCE	66.7	72.7	66.7	66.7
TIME	77.8	66.7	15.4	40.0
TRAVELER	90.9	90.6	90.9	90.6

Table 2: Exp. 1: Role-specific figures of system performance for four example frames.

Results. Table 1 shows the systems' overall F-scores and standard deviation across roles. Table 2 illustrates the differences in performance across roles on four frames: It lists all roles with ≥ 5 occurrences for each frame. PART_ORIENTATIONAL shows very little variance, while the roles of CHANGE_POSITION_ON_A_SCALE and especially TRAVEL differ widely. For KINSHIP, the system shows good performance for both roles, but the F-scores still differ by around 9 points.

Discussion. Table 1 shows that there is considerable variance across roles, with a standard deviation in the range of 18% for the syntax-only model. We note that the deviation decreases to 13% for the combined syntax-semantics model. Table 2 confirms that this is not purely between-frames, but also within-frames variance. This confirms the phenomenon described at the beginning of this section.

fr	frame
fe	role (frame element)
$fes(fr)$	roles of a frame
$gfs(fr)$	gramm. functions of a frame
$gfs_{fr}(fe)$	gramm. functions realizing a role in a frame

Table 3: Notation summary

4 A meta-model for role assignment: Confusability

The experiment of the previous section has shown a considerable variance in system performance across roles. The aim of this section is to develop a meta-model which can explain this variance.

The models we have explored in Exp. 1 rely mainly on syntactic features: Even in the combined syntax-semantics model, 24 of the 31 features describe syntactic structure. This predominance of syntactic features can be observed in many current models for semantic role assignment. Accordingly, our meta-model focuses on the uniformity of the mapping from syntactic structure to semantic roles. We formalize the variance in this mapping by the *confusability* of a semantic role. It implements the following hypothesis:

(1) The semantic role assignment systems we study approximate role assignment through grammatical functions.

(2) System performance for a given role depends on the role's *confusability*: A role is highly confusable if the grammatical functions that instantiate it often also instantiate other roles.

By using the ideal, manually assigned grammatical functions that are available from the FrameNet data – and which are not passed on to the learner – our meta-model abstracts over concrete feature sets.

Our definition of confusability proceeds in two steps. First we model the informativity of a grammatical function by the entropy of semantic roles that it maps to. Then we compute the confusability of a role as a weighted average of the entropies of the grammatical functions that realize it.

Grammatical function entropy. Viewing a grammatical function as a random variable with semantic roles as values, we define the entropy of a grammatical function *gf* within the frame *fr* as

$$H_{fr}(gf) = \sum_{fe \in fes(fr)} -p(fe|gf) \log p(fe|gf)$$

where $p(fe|gf) = \frac{f(gf,fe)}{f(gf)}$ is the conditional probability of roles *fe* given *gf* (cf. the notation in Table 3).

Grammatical function entropy					
GF	DEG	THM	DEP	LOC	H
Mod	69	43	24	0	1.46
Comp	18	491	12	41	0.72
Ext	0	17	0	561	0.16
Head	0	0	0	273	0.0
Obj	0	0	0	3	0.0

Role Confusability						
Role	Mod	Comp	Ext	Head	Obj	Conf
DEG	69	18	0	0	0	1.31
THM	43	491	17	0	0	0.76
DEP	24	12	0	0	0	1.22
LOC	0	41	561	273	3	0.16

Table 4: Grammatical function entropy and role confusability for the frame ABUNDANCE

Role confusability. The confusability of a role is the sum of its grammatical function entropies, weighted by the conditional probabilities $p(gf|fe) = \frac{f(gf,fe)}{f(fe)}$ of grammatical functions *gf* given *fe*.

$$c_{fr}(fe) = \sum_{gf \in gfs(fr)} p(gf|fe) H_{fr}(gf)$$

An example. Table 4 shows the grammatical function entropies and role confusabilities for the frame ABUNDANCE, both computed on the training data. The upper part of Table 4 lists the entropies of the grammatical functions `Mod`, `Comp`, `Ext`, `Head` and `Obj`[3] and the counts $f(gf, fe)$ of occurrences of the grammatical functions together with the roles DEGREE (DEG), THEME (THM), DEPICTIVE (DEP) and LOCATION (LOC). The entropy of `Mod`, with similar numbers of occurrences for three different roles, is relatively high, while `Ext` occurs almost exclusively for one role and has a much lower entropy. The lower part of Table 4 shows the confusability for the same set of roles. The confusability of

[3] See Fillmore and Petruck (2003) for a glossary of FrameNet's grammatical functions.

DEGREE is relatively high even though it is mostly realized by `Mod` because `Mod` has a high entropy, i.e. it indicates multiple roles; LOCATION on the other hand is not very confusable even though it occurs frequently as both `Ext` and `Head`, since both grammatical functions indicate this role.

Related work. Our approach is similar to Pado and Boleda (2004) in that they also use the uniformity of linking as an explanation for performance variations in semantic role assignment. However, their analysis is located at the frame level. We examine individual roles, which allows us to derive a simpler and more intuitive formalization of linking uniformity. Also, our model will ultimately lead us to a different conclusion: the uniformity of linking is a good predictor of the performance of role assignment systems, but only for exclusively syntactic models (see Section 5).

5 Experiment 2: Relating confusability and system performance

In this section, we test the validity of our metamodel. We assess whether confusability, defined in Section 4, can explain the variance in role assignment that we have found in Section 3, by testing the correlation between the two variables.

Experimental setup. We use the same data set (Senseval-3) and the same two classifiers (memory-based and maximum entropy classification) as in Exp. 1. To cover a wider range of models and thus increase the validity of our analysis, we split up the *Syn* feature set from Exp. 1 into the four smaller sets described in the upper part of Table 5. We use these sets individually, combined, and together with the lexical features in the *Sem* set. This results in a total of 20 different models (10 for each classifier), for which we computed role-specific F-scores.

In parallel, we estimated the confusability as described in Section 4, with FrameNet's manually assigned grammatical functions as a basis, using only the training portion of our data. We did not smooth, but omitted roles occurring less than 5 times to avoid sparse and thus unreliable data points. Recall that confusability does not vary with the feature set, since its central asset is to abstract over concrete model parameters and feature sets.

Feature set	F_{MBL}	F_{Maxent}
Path0	70.9	71.3
Path	73.3	72.6
Pt	78.8	79.0
Path/Pt	80.8	79.8
Path/Sibling	76.7	76.6
Pt/Sibling	78.8	79.1
Syn	82.2	82.4
Sem	80.3	80.7
Syn/Sem	87.1	87.5

Table 6: Exp. 2: Results for different feature sets

Results. The F-scores for the subdivided *Syn* feature set are shown in the upper part of Table 6, with the complete *Syn* and *Sem* sets and their combination below. There is a clear relationship between features and F-score: additional features are consistently rewarded with higher performance. Interestingly, phrase type information appears to be a better role predictor than path (compare models *Path* and *Pt*). Also, the semantic feature set alone (*Sem*) performs at over 80% F-Score, slightly better any of the individual syntactic feature groups.

The high F-score variance between individual roles which we have shown for the feature sets *Syn* and *Syn/Sem* in Exp. 1 generalizes to the other feature sets; all individual syntactic feature sets exhibit a higher variance than *Syn*, and *Sem* shows a higher variance than the *Syn/Sem* combination. This does not come as a surprise, since the two models of Exp. 1 use the two richest feature sets, and we would expect less robust behavior for weaker models. Another point to note is that the performance of the two learners is remarkably similar.

The high variance in the F-scores is mirrored in the confusability figures; we obtain an average confusability for our semantic roles of 1.79 with a high standard deviation of 0.84. A scatter plot of F-scores against confusability figures (Fig.1) suggests a linear correlation analysis.

Analysis 1: Correlating confusability and F-score. Since the data does not appear to be normally distributed, we apply Kendall's nonparametric rank test. The results, which are listed in Table 7, show an extremely significant negative correlation between confusability and F-score: higher confus-

Path0	These are features centered around the path from the target lemma to the constituent: the path itself, its length, partial path up to the lowest common ancestor, the grammatical rule that expands the target predicate's parent, relative position of constituent to target
Path	Feature set Path, plus target lemma
Pt	These are features related to phrase type and part of speech: the phrase type of the constituent and its parent, the POS of the constituent first word, last word and head as well as the POS of an informative content word of the constituent (for PP and SBar constituents only: the head of the head's complement), as well as the target lemma
Sibling	Phrase type and POS of the head of the left and right sibling constituent, and the Collins parser's judgment on the argumenthood of the constituent
Syn	This set combines Path, Sibling and Pt. Additional features are: target voice; the constituent's preposition; a feature combining path with target voice and target POS; and two rule-based features judging argumenthood and grammatical function of the constituent
Sem	These are lexical features: Head words of the constituent and of its left and right siblings; leftmost and rightmost word of the constituent; informative content word lemma (see set Pt for details); and the governing verb of the target predicate

Table 5: Feature groups used in the experiments

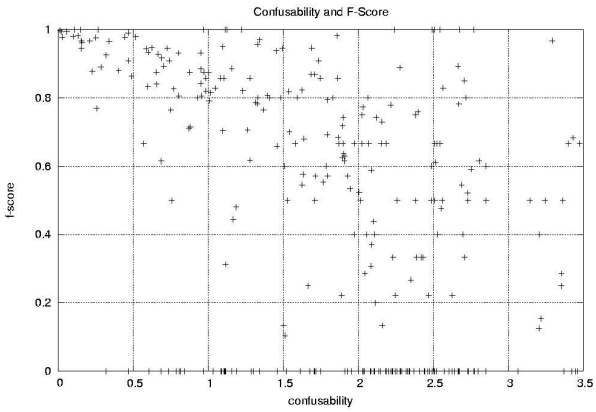

Figure 1: Scatter plot: F-score against confusability (Feature set *Syn*).

	MBL		MaxEnt	
Feature set	z	p	z	p
Path0	-11.72	10^{-15}	-11.76	10^{-15}
Path	-12.29	10^{-15}	-11.23	10^{-15}
Pt	-10.64	10^{-15}	-11.12	10^{-15}
Path/Pt	-11.19	10^{-15}	-10.45	10^{-15}
Path/Sibling	-12.65	10^{-15}	-11.76	10^{-15}
Pt/Sibling	-10.58	10^{-15}	-9.90	10^{-15}
Syn	-9.47	10^{-15}	-9.38	10^{-15}
Sem	-6.90	10^{-11}	-8.23	10^{-15}
Syn/Sem	-8.30	10^{-15}	-8.29	10^{-15}

Table 7: Exp. 2, Analysis 1: Correlation between F-Score and confusability. z: Kendall's tau coefficient, p: significance level

ability appears to be related to lower F-score.

However, note that the correlation is extremely significant even for the model which only uses semantic features. This is unexpected at best and makes a strong interpretation of this correlation doubtful: it is rather likely that there is a third variable with which both F-score and confusability are correlated. The most obvious candidate for such a *confounding variable* is the size of the training set – clearly, we expect our models to perform better with larger training sets. In order to get a more realistic assessment of the relationship between confusability and F-score, we perform an additional analysis to disconfound confusability and frequency.

Analysis 2: Disconfounding confusability and frequency. One way of factoring out the influence of a confounding variable is to perform a partial correlation analysis, which explicitly removes the effects of a third variable when determining the strength of a correlation between two variables. Like a normal correlation analysis, it yields a *partial correlation coefficient*.

Features	MBL		MaxEnt	
	r_c	r_f	r_c	r_f
Path0	-.29***	-.03	-.29***	-.03
Path	-.30***	-.02	-.27***	-.07**
Pt	-.19***	-.11**	-.21***	-.12**
Path/Pt	-.22***	-.07*	-.19***	-.16***
Path/Sibl	-.31***	+.01	-.28***	-.06*
Pt/Sibl	-.20***	-.10**	-.18***	-.16***
Syn	-.10*	-.17***	-.12*	-.19***
Sem	+.01	-.27***	-.02	-.24***
Syn/Sem	+.02	-.25***	-.01	-.25***

Table 8: Exp. 2, Analysis 2: Partial correlation coefficients. r_c: correlation between F-score and confusability, controlling for training set size. r_f: correlation between F-score and training set size, controlling for confusability. Significance levels: ***: p<0.001; **: p<0.01; *: p<0.05.

We first compute partial correlation coefficients between F-score and confusability, controlling for training set size. The results, which indicate the "true" relationship between performance and confusability, are shown in the r_c columns of Table 8. For both learners, confusability is significantly correlated with F-score for all syntactic feature sets, but not for the semantic feature set and for the combined set *Syn/Sem*.

We also compute the partial correlation coefficients between F-score and training set size, controlling for confusability. These figures are reported in the r_f columns of Table 8 and show the "true" relationship between performance and training set size. There is no significant correlation between training set size and performance for simple syntax based-models, but the correlation is highly significant for complex syntactic models and all semantic models.

Discussion. The partial correlation analysis confirms that confusability is a meta-model that can explain the performance of a range of different models for semantic role assignment, namely those models which rely exclusively on syntactic features. Since we used the gold standard features provided by FrameNet and did not introduce implementation- or feature-specific knowledge, this points to a general limitation of syntax-based models. In contrast, semantic features behave completely differently; their contribution is not limited by a role's confusability. At the very least, it cannot be captured by our current meta-model, but the absolute increase in performance indicates that integrating semantics is the way forward, which is surprising given that the purely lexical features we use the present study are usually extremely sparse.

The analysis of the partial correlation between F-score and training set size also allows interesting conclusions. The correlation is not significant for small syntactic feature sets like *Path*, indicating that models for such features can be learned satisfactorily from relatively small training sets (but which are also limited in expressivity). This is markedly different for richer feature sets. Arguably, these feature sets are sparser and can therefore profit more from an increased amount of training data. Again, the effect is most pronounced for the semantic feature set.

6 Conclusion

In this paper, we have formulated a *meta-model* for semantic role assignment. We have used the *confusability* of roles to predict classification performance independently of the classification framework and feature sets used. We have defined role confusability in two steps: First, we have formalized the certainty with which we can predict a semantic role from a given grammatical function with *grammatical function entropy*. Then, we have defined the *confusability of a role* as a weighted sum of grammatical function entropies.

We have found that role confusability is highly significantly correlated with system performance for models based solely on syntactic features. We conclude that syntactic features approximate a description of grammatical functions, but that semantic features model a different aspect of the world.

Much of current research in semantic role assignment is centered on the refinement of syntactic features. Our study suggests that it may be worthwhile to explore the refinement of semantic features as well. The most obvious choice is to investigate features related to selectional preferences. Possible features include goodness of fit relative to pre-computed preferences (Baldewein et al., 2004), named entities (Pradhan et al., 2004), or broad ontological classes like "animate" or "artifact". Fol-

lowing up on this idea, a natural continuation of the present study would be to create a meta-model that subsumes *semantic* features. Such a model could use optimal selectional restrictions as a predictor. The next step would then be to construct a combined meta-model that describes the behavior of systems with both syntactic and semantic features.

Another interesting research direction that our study suggests is the combination of syntactic and semantic models in co-training. Co-training can be sensibly applied only when conditional independence holds for the two target functions and the distribution (Blum and Mitchell, 1998), i.e. when it uses two independent views on the instance set. By pointing out a highly significant distinction between syntactic and semantic features with respect to role confusability, our study provides empirical evidence that syntactic and semantic features model different aspects of the role assignment mapping, and that co-training may be feasible by using syntactic and semantic features as views.

Acknowledgments. We are grateful to the Deutsche Forschungsgemeinschaft (DFG) for funding the SALSA-II project (grant PI-154/9-2).

References

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

A. Blum, T. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*.

H. C. Boas. 2002. Bilingual framenet dictionaries for machine translation. In *Proceedings of LREC 2002*, 1364–1371, Las Palmas, Canary Islands.

L. Burnard, 1995. *User's guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Services, 1995.

X. Carreras, L. Màrquez. 2004. Introduction to the CoNLL-2004 shared task: semantic role labeling. In *Proceedings of CoNLL 2004*, Boston, MA.

X. Carreras, L. Màrquez. 2005. Introduction to the CoNLL-2005 shared task: semantic role labeling. In *Proceedings of CoNLL 2005*, Ann Arbor, MI.

M. J. Collins. 1996. A new statistical parser based on bigram lexical dependencies. In A. Joshi, M. Palmer, eds., *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, 184–191, San Francisco. Morgan Kaufmann Publishers.

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

C. J. Fillmore, M. R. Petruck. 2003. FrameNet glossary. *International Journal of Lexicography*, 16:359–361.

C. J. Fillmore, C. R. Johnson, M. R. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

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

M. Fleischmann, E. Hovy. 2003. A maximum entropy approach to framenet tagging. In *Proceedings of HLT/NAACL 2003*, Edmonton, Canada.

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

K. Litkowski. 2004. Senseval-3 task: Automatic labeling of semantic roles. In R. Mihalcea, P. Edmonds, eds., *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain.

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

A. Moschitti. 2004. A study on convolution kernel for shallow semantic parsing. In *Proceedings of the ACL 2004*, Barcelona, Spain.

S. Narayanan, S. Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of COLING 2004*, Geneva, Switzerland.

S. Pado, G. Boleda Torrent. 2004. The influence of argument structure on semantic role assignment. In *Proceedings of EMNLP 2004*, Barcelona, Spain.

S. Pradhan, W. Ward, K. Hacioglu, J. H. Martin, D. Jurafsky. 2004. Shallow semantic parsing using support vector machines. In *Proceedings of HLT/NAACL 2004*, Boston, MA.

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

M. Surdeanu, S. Harabagiu, J. Williams, P. Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL 2003*, Sapporo, Japan.

N. Xue, M. Palmer. 2004. Calibrating features for semantic role labeling. In *Proceedings of EMNLP 2004*, Barcelona, Spain.