Modelling Semantic Role Plausibility in Human Sentence Processing

Ulrike Padó and **Matthew Crocker**
Computational Linguistics
Saarland University
66041 Saarbrücken
Germany
{ulrike,crocker}@coli.uni-sb.de

Frank Keller
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
keller@inf.ed.ac.uk

Abstract

We present the psycholinguistically motivated task of predicting human plausibility judgements for verb-role-argument triples and introduce a probabilistic model that solves it. We also evaluate our model on the related role-labelling task, and compare it with a standard role labeller. For both tasks, our model benefits from class-based smoothing, which allows it to make correct argument-specific predictions despite a severe sparse data problem. The standard labeller suffers from sparse data and a strong reliance on syntactic cues, especially in the prediction task.

1 Introduction

Computational psycholinguistics is concerned with modelling human language processing. Much work has gone into the exploration of sentence comprehension. Syntactic preferences that unfold during the course of the sentence have been successfully modelled using incremental probabilistic context-free parsing models (e.g., Jurafsky, 1996; Crocker and Brants, 2000). These models assume that humans prefer the *most likely* structural alternative at each point in the sentence. If the preferred structure changes during processing, such models correctly predict processing difficulty for a range of experimentally investigated constructions. They do not, however, incorporate an explicit notion of semantic processing, while there are many phenomena in human sentence processing that demonstrate a non-trivial interaction of syntactic preferences and semantic plausibility.

Consider, for example, the well-studied case of reduced relative clause constructions. When incrementally processing the sentence *The deer shot by the hunter was used as a trophy*, there is a local ambiguity at *shot* between continuation as a main clause (as in *The deer shot the hunter*) or as a reduced relative clause modifying *deer* (equivalent to *The deer which was shot* ...). The main clause continuation is syntactically more likely.

However, there is a second, semantic clue provided by the high plausibility of deer being shot and the low plausibility of them shooting. This influences readers to choose the syntactically dispreferred reduced relative reading which interprets the *deer* as an object of *shot* (McRae *et al.*, 1998). Plausibility has overridden the syntactic default. On the other hand, for a sentence like *The hunter shot by the teenager was only 30 years old*, semantic plausibility initially reinforces the syntactic main clause preference and readers show difficulty accommodating the subsequent disambiguation towards the reduced relative.

In order to model effects like these, we need to extend existing models of sentence processing by introducing a semantic dimension. Possible ways of integrating different sources of information have been presented e.g. by McRae *et al.* (1998) and Narayanan and Jurafsky (2002). Our aim is to formulate a model that reliably predicts human plausibility judgements from corpus resources, in parallel to the standard practice of basing the syntax component of psycholinguistic models on corpus probabilities or even probabilistic treebank grammars. We can then use both the syntactic likelihood and the semantic plausibility score to predict the preferred syntactic alternative, thus accounting for the effects shown e.g. by McRae *et al.* (1998).

Independent of a syntactic model, we want any semantic model we define to satisfy two criteria: First, it needs to be able to make predictions in-

Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 345–352. Trento, 2006.

crementally, in parallel with the syntactic model. This entails dealing with incomplete or unspecified (syntactic) information. Second, we want to extend to semantics the assumption made in syntactic models that the most probable alternative is the one preferred by humans. The model therefore must be probabilistic.

We present such a probabilistic model that can assign roles incrementally as soon as a predicate-argument pair is seen. It uses the likelihood of thematic role assignments to model human interpretation of verb-argument relations. Thematic roles are a description of the link between verb and argument at the interface between syntax and semantics. Thus, they provide a shallow level of sentence semantics which can be learnt from annotated corpora.

We evaluate our model by verifying that it indeed correctly predicts human judgements, and by comparing its performance with that of a standard role labeller in terms of both judgement prediction and role assignment. Our model has two advantages over the standard labeller: It does not rely on syntactic features (which can be hard to come by in an incremental task) and our smoothing approach allows it to make argument-specific role predictions in spite of extremely sparse training data. We conclude that (a) our model solves the task we set, and (b) our model is better equipped for our task than a standard role labeller.

The outline of the paper is as follows: After defining the prediction task more concretely (Section 2), we present our simple probabilistic model that is tailoured to the task (Section 3). We introduce our test and training data in Section 4. It becomes evident immediately that we face a severe sparse data problem, which we tackle on two levels: By smoothing the distribution and by acquiring additional counts for sparse cases. The smoothed model succeeds on the prediction task (Section 5). Finally, in Section 6, we compare our model to a standard role labeller.

2 The Judgement Prediction Task

We can measure our intuitions about the plausibility of *hunters shooting* and *deer being shot* in terms of plausibility judgements for verb-role-argument triples. Two example items from McRae *et al.* (1998) are presented in Table 1. The judgements were gathered by asking raters to assign a value on a scale from 1 (not plausible) to 7 (very plausible) to questions like *How common is it for a hunter to shoot something?* (subject reading: *hunter* must be agent) or *How common is it for a hunter to be shot?* (object reading: *hunter* must be patient). The number of ratings available in each of our three sets of ratings is given in Table 2 (see also Section 4).

Verb	Noun	Role	Rating
shoot	hunter	agent	6.9
shoot	hunter	patient	2.8
shoot	deer	agent	1.0
shoot	deer	patient	6.4

Table 1: Test items: Verb-noun pairs with ratings on a 7 point scale from McRae *et al.* (1998).

The task for our model is to correctly predict the plausibility of each verb-role-argument triple. We evaluate this by correlating the model's predicted values and the judgements. The judgement data is not normally distributed, so we correlate using Spearman's ρ (a non-parametric rank-order test). The ρ value ranges between 0 and 1 and indicates the strength of association between the two variables. A significant positive value indicates that the model's predictions are accurate.

3 A Model of Human Plausibility Judgements

We can formulate a model to solve the prediction task if we equate the plausibility of a role assignment to a verb-argument pair with its probability, as suggested above. This value is influenced as well by the verb's semantic class and the grammatical function of the argument. The plausibility for a verb-role-argument triple can thus be estimated as the joint probability of the argument head a, the role r, the verb v, the verb's semantic class c and the grammatical function gf of a:

$$Plausibility_{v,r,a} = P(r,a,v,c,gf)$$

This joint probability cannot be easily estimated from co-occurrence counts due to lack of data. But we can decompose this term into a number of subterms that approximate intuitively important information such as syntactic subcategorisation ($P(gf|v,c)$), the syntactic realisation of a semantic role ($P(r|v,c,gf)$) and selectional preferences ($P(a|v,c,gf,r)$):

$$\begin{aligned} Plausibility_{v,r,a} = P(r,a,v,c,gf) = \\ P(v) \cdot P(c|v) \cdot P(gf|v,c) \cdot \\ P(r|v,c,gf) \cdot P(a|v,c,gf,r) \end{aligned}$$

shoot.02: [The hunter $_{Arg0}$] shot [the deer $_{Arg1}$].
Killing: [The hunter $_{Killer}$] shot [the deer $_{Victim}$].

Figure 1: Example annotation: PropBank (above) and FrameNet (below).

Each of these subterms can be estimated more easily from the semantically annotated training data simply using the maximum likelihood estimate. However, we still need to smooth our estimates, especially as the $P(a|v,c,gf,r)$ term remains very sparse. We describe our use of two complementary smoothing methods in Section 5.

Our model fulfils the requirements we have specified: It is probabilistic, able to work incrementally as soon as a single verb-argument pair is available, and can make predictions even if the input information is incomplete. The model generates the missing values if, e.g., the grammatical function or the verb's semantic class are not specified. This means that we can immediately evaluate on the judgement data without needing further verb sense or syntactic information.

4 Test and Training data

Training Data To date, there are two main annotation efforts that have produced semantically annotated corpora: PropBank (PB) and FrameNet (FN). Their approaches to annotation differ enough to warrant a comparison of the corpora as training resources. Figure 1 gives an example sentence annotated in PropBank and FrameNet style. The PropBank corpus (c. 120,000 propositions, c. 3,000 verbs) adds semantic annotation to the Wall Street Journal part of the Penn Treebank. Arguments and adjuncts are annotated for every verbal proposition in the corpus. A common set of argument labels *Arg0* to *Arg5* and *ArgM* (adjuncts) is interpreted in a verb-specific way. Some consistency in mapping has been achieved, so that *Arg0* generally denotes agents and *Arg1* patients/themes.

The FrameNet corpus (c. 58,000 verbal propositions, c. 1,500 verbs in release 1.1) groups verbs with similar meanings together into frames (i.e. descriptions of situations) with a set of frame-specific roles for participants and items involved (e.g. a killer, instrument and victim in the Killing frame). Both the definition of frames as semantic verb classes and the semantic characterisation of frame-specific roles introduces a level of information that is not present in PropBank. Since corpus annotation is frame-driven, only some senses of a verb may be present and word frequencies may not be representative of English.

Test Data Our main data set consists of 160 data points from McRae *et al.* (1998) that were split randomly into a 60 data point development set and a 100 data point test set. The data is made up of two arguments per verb and two ratings for each verb-argument pair, one for the subject and one for the object reading of the argument (see Section 2). Each argument is highly plausible in one of the readings, but implausible in the other (recall Table 1). Human ratings are on a 7-point scale.

In order to further test the coverage of our model, we also include 76 items from Trueswell *et al.* (1994) with one highly plausible object per verb and a rating each for the subject and object reading of the argument. The data were gathered in the same rating study as the McRae et al. data, so we can assume consistency of the ratings. However, in comparison to the McRae data set, the data is impoverished as it lacks ratings for plausible agents (in terms of the example in Table 1, this means there are no ratings for *hunter*). Lastly, we use 180 items from Keller and Lapata (2003). In contrast with the previous two studies, the verbs and nouns for these data were not hand-selected for the plausibility of their combination. Rather, they were extracted from the BNC corpus by frequency criteria: Half the verb-noun combinations are seen in the BNC with high, medium and low frequency, half are unseen combinations of the verb set with nouns from the BNC. The data consists of ratings for 30 verbs and 6 arguments each, interpreted as objects. The human ratings were gathered using the Magnitude Estimation technique (Bard *et al.*, 1996). This data set allows us to test on items that were not hand-selected for a psycholinguistic study, even though the data lacks agenthood ratings and the items are poorly covered by the FrameNet corpus.

All test pairs were hand-annotated with FrameNet and PropBank roles following the specifications in the FrameNet on-line database and the PropBank frames files.[1]

The judgement prediction task is very hard to solve if the verb is unseen during training. Backing off to syntactic information or a frequency

[1] Although a single annotator assigned the roles, the annotation should be reliable as roles were mostly unambiguous and the annotated corpora were used for reference.

	Total	Revised	
Source		FN	PB
McRae *et al.* (1998)	100	64 (64%)	92 (92%)
Trueswell *et al.* (1994)	76	52 (68.4%)	72 (94.7%)
Keller and Lapata (2003)	180	–	162 (90%)

Table 2: Test sets: Total number of ratings and size of revised test sets containing only ratings for seen verbs (% of total ratings). –: Coverage too low (26.7%).

baseline only works if the role set is small and syntactically motivated, which is the case for PropBank, but not FrameNet. We present results both for the complete test sets and and for revised sets containing only items with seen verbs. Excluding unseen verbs seems justified for FrameNet and has little effect for the PropBank corpus, since its coverage is generally much better. Table 2 shows the total number of ratings for each test set and the sizes of the revised test sets containing only items with seen verbs. FrameNet always has substantially lower coverage. Since only 27% of the verbs in the Keller & Lapata items are covered in FrameNet, we do not test this combination.

5 Experiment 1: Smoothing Methods

We now turn to evaluating our model. It is immediately clear that we have a severe sparse data problem. Even if all the verbs are seen, the combinations of verbs and arguments are still mostly unseen in training for all data sets.

We describe two complementary approaches to smoothing sparse training data. One, Good-Turing smoothing, approaches the problem of unseen data points by assigning them a small probability. The other, class-based smoothing, attempts to arrive at semantic generalisations for words. These serve to identify equivalent verb-argument pairs that furnish additional counts for the estimation of $P(a|v,c,gf,r)$.

5.1 Good-Turing Smoothing and Linear Interpolation

We first tackle the sparse data problem by smoothing the distribution of co-occurrence counts. We use the Good-Turing re-estimate on zero and one counts to assign a small probability to unseen events. This method relies on re-estimating the probability of seen and unseen events based on knowledge about more frequent events.

Adding Linear Interpolation We also experimented with the linear interpolation method, which is typically used for smoothing n-gram models. It re-estimates the probability of the n-gram in question as a weighted combination of the n-gram, the n-1-gram and the n-2-gram. For example, $P(a|v,c,gf,r)$ is interpolated as

$$P(a|v,c,gf,r) = \lambda_1 P(a|v,c,gf,r) + \lambda_2 P(a|v,c,r) + \lambda_3 P(a|v,c)$$

The λ values were estimated on the training data, separately for each of the model's four conditional probability terms, by maximising five-fold cross-validation likelihood to avoid overfitting.

We smoothed all model terms using the Good-Turing method and then interpolated the smoothed terms. Table 3 lists the test results for both training corpora and all test sets when Good-Turing smoothing (GT) is used alone and with linear interpolation (GT/LI). We also give the unsmoothed coverage and correlation. The need for smoothing is obvious: Coverage is so low that we can only compute correlations in two cases, and even for those, less than 20% of the data are covered.

GT smoothing alone always outperforms the combination of GT and LI smoothing, especially for the FrameNet training set. Maximising the data likelihood during λ estimation does not approximate our final task well enough: The log likelihood of the test data is duly improved from -797.1 to -772.2 for the PropBank data and from -501.9 to -446.3 for the FrameNet data. However, especially for the FrameNet training data, performance on the correlation task diminishes as data probability rises. A better solution might be to use the correlation task directly as a λ estimation criterion, but this is much more complex, requiring us to estimate all λ terms simultaneously. Also, the main problem seems to be that the λ interpolation smoothes by de-emphasising the most specific (and sparsest) term, so that, on our final task, the all-important argument-specific information is not used efficiently when it is available. We therefore restrict ourselves to GT smoothing.

Train	Smoothing	Test	Smoothed Coverage		Smoothed ρ	Unsmoothed Coverage		Unsmoothed ρ
PB	GT	McRae	93.5%	(*86%*)	0.112, ns	2%	(*2%*)	–
		Trueswell	100%	(*94.7%*)	0.454, **	17%	(*16%*)	ns
		Keller&Lapata	100%	(*90%*)	0.285, **	5%	(*4%*)	0.727, *
	GT/LI	McRae	93.5%	(*86%*)	0.110, ns	2%	(*2%*)	–
		Trueswell	100%	(*94.7%*)	0.404, **	17%	(*16%*)	ns
		Keller&Lapata	100%	(*90%*)	0.284, **	5%	(*4%*)	0.727, *
FN	GT	McRae	87.5%	(*56%*)	0.164, ns	6%	(*4%*)	–
		Trueswell	76.9%	(*52.6%*)	0.046, ns	6%	(*4%*)	–
	GT/LI	McRae	87.5%	(*56%*)	0.042, ns	6%	(*4%*)	–
		Trueswell	76.9%	(*52.6%*)	0.009, ns	6%	(*4%*)	–

Table 3: Experiment 1, GT and Interpolation smoothing. Coverage on seen verbs (*and on all items*) and correlation strength (Spearman's ρ for PB and FN data on all test sets. –: too few data points, ns: not significant, *: $p < 0.05$, **: $p < 0.01$.

Model Performance Both versions of the smoothed model make predictions for all seen verbs; the remaining uncovered data points are those where the correct role is not accounted for in the training data (the verb may be very sparse or only seen in a different FrameNet frame). For the FrameNet training data, there are no significant correlations, but for the PropBank data, we see correlations for the Trueswell and Keller&Lapata sets. One reason for the good performance of the PB-Trueswell and PB-Keller&Lapata combinations is that in the PropBank training data, the object role generally seems to be the most likely one. If the most specific probability term is sparse and expresses no role preference (which is the case for most items: see Unsmoothed Coverage), our model is biased towards the most likely role given the verb, semantic class and grammatical function. Recall that the Trueswell and Keller&Lapata data contain ratings for (plausible) objects only, so that preferring the patient role is a good strategy. This also explains why the model performs worse for the McRae et al. data, which also has ratings for good agents (and bad patients). On FrameNet, this preference for "patient" roles is not as marked, so the FN-Trueswell case does not behave like the PB-Trueswell case.

5.2 Class-Based Smoothing

In addition to smoothing the training distribution, we also attempt to acquire more counts to estimate each $P(a|v,c,gf,r)$ by generalising from tokens to word classes. The term we estimate becomes $P(class_a|class_v,gf,r)$. This allows us to make argument-specific predictions as we do not rely on a uniform smoothed term for unseen $P(a|v,c,gf,r)$ terms. We use lexicographic noun classes from WordNet and verb classes induced by soft unsupervised clustering, which outperform lexicographic verb classes.

Noun Classes We tested both the coarsest and the finest noun classification available in WordNet, namely the top-level ontology and the noun synsets which contain only synonyms of the target word.[2] The top-level ontology proved to overgenerate alternative nouns, which raises coverage but does not produce meaningful role predictions. We therefore use the noun synsets below.

Verb Classes Verbs are clustered according to linguistic context information, namely argument head lemmas, the syntactic configuration of verb and argument, the verb's semantic class, the gold role information and a combined feature of gold role and syntactic configuration. The evaluation of the clustering task itself is task-based: We choose the clustering configuration that produces optimal results in the the prediction task on the McRae development set. The base corpus for clustering was always used for frequency estimation.

We used an implementation of two soft clustering algorithms derived from information theory (Marx, 2004): the Information Distortion (ID) (Gedeon *et al.*, 2003) and Information Bottleneck (IB) (Tishby *et al.*, 1999) methods. Soft clustering allows us to take verb polysemy into account that is often characterised by different patterns of syntactic behaviour for each verb meaning.

[2]For ambiguous nouns, we chose the sense that led to the highest probability for the current role assignment.

A number of parameters were set on the development set, namely the clustering algorithm, the smoothing method within the algorithms and the number of clusters within each run. For our task, the IB algorithm generally yielded better results.

We decided which clustering parametrisations should be tried on the test sets based on the notion of *stability*: Both algorithms increase the number of clusters by one at each iteration. Thus, each parametrisation yields a series of cluster configurations as the number of iterations increases. We chose those parametrisations where a series of at least three consecutive cluster configurations returned significant correlations on the development set. This should be an indication of a generalisable success, rather than a fluke caused by peculiarities of the data. On the test sets, results are reported for the configuration (characterised by the iteration number) that returned the first significant result in such a series on the development set, as this is the most general grouping.

5.3 Combining the Smoothing Methods

We now present results for combining the GT and class-based smoothing methods. We use induced verb classes and WordNet noun synsets for class-based smoothing of $P(a|v,c,gf,r)$, and rely on GT smoothing if the counts for this term are still sparse. All other model terms are always smoothed using the GT method. Table 4 contains results for three clustering configurations each for the PropBank and FrameNet data that have proven stable on the development set. We characterise them by the clustering algorithm (IB or ID) and number of clusters. Note that the upper bound for our ρ values, human agreement or inter-rater correlation, is below 1 (as indicated by a correlation of Pearson's $r = .640$ for the seen pairs from the Keller and Lapata (2003) data).

For the FrameNet data, there is a marked increase in performance for both test sets. The human judgements are now reliably predicted with good coverage in five out of six cases. Clearly, equivalent verb-argument counts have furnished accurate item-specific estimates. On the PropBank data set, class-based smoothing is less helpful: ρ values generally drop slightly. Apparently, the FrameNet style of annotation allows us to induce informative verb classes, whereas the PropBank classes introduce noise at most.

6 Experiment 2: Role Labelling

We have shown that our model performs well on its intended task of predicting plausibility judgements, once we have proper smoothing methods in place. But since this task has some similarity to role labelling, we can also compare the model to a standard role labeller on both the prediction and role labelling tasks. The questions are: How well do we do labelling, and does a standard role labeller also predict human judgements?

Beginning with work by Gildea and Jurafsky (2002), there has been a large interest in semantic role labelling, as evidenced by its adoption as a shared task in the Senseval-III competition (FrameNet data, Litkowski, 2004) and at the CoNLL-2004 and 2005 conference (PropBank data, Carreras and Márquez, 2005). As our model currently focuses on noun phrase arguments only, we do not adopt these test sets but continue to use ours, defining the correct role label to be the one with the higher probability judgement. We evaluate the model on the McRae test set (recall that the other sets only contain good patients/themes and are therefore susceptible to labeller biases).

We formulate frequency baselines for our training data. For PropBank, always assigning *Arg1* results in $F = 45.7$ (43.8 on the full test set). For FrameNet, we assign the most frequent role given the verb, so the baseline is $F = 34.4$ (26.8).

We base our standard role labelling system on the SVM labeller described in Giuglea and Moschitti (2004), although without integrating information from PropBank and VerbNet for FrameNet classification as presented in their paper. Thus, we are left with a set of fairly standard features, such as *phrase type, voice, governing category* or *path through parse tree from predicate*. These are used to train two classifiers, one which decides which phrases should be considered arguments and one which assigns role labels to these arguments. The SVM labeller's F score on an unseen test set is $F = 80.5$ for FrameNet data when using gold argument boundaries. We also trained the labeller on the PropBank data, resulting in an F score of $F = 98.6$ on Section 23, again on gold boundaries.

We also evaluate the SVM labeller on the correlation task by normalising the scores that the labeller assigns to each role and then correlating the normalised scores to the human ratings.

In order to extract features for the SVM labeller, we had to present the verb-noun pairs in full sen-

Train	Test	Verb Clusters	Coverage	ρ
PB	McRae	ID 4	93.5% (*86%*)	0.097, ns
		IB 10	93.5% (*86%*)	0.104, ns
		IB 5	93.5% (*86%*)	0.107, ns
	Trueswell	ID 4	100% (*94.7%*)	0.419, **
		IB 10	100% (*94.7%*)	0.366, **
		IB 5	100% (*94.7%*)	0.439, **
	Keller&Lapata	ID 4	100% (*90%*)	0.300, **
		IB 10	100% (*90%*)	0.255, **
		IB 5	100% (*90%*)	0.297, **
FN	McRae	ID 4	87.5% (*56%*)	0.304, *
		IB 9	87.5% (*56%*)	0.275, *
		IB 10	87.5% (*56%*)	0.267, *
	Trueswell	ID 4	76.9% (*52.6%*)	0.256, ns
		IB 9	76.9% (*52.6%*)	0.342, *
		IB 10	76.9% (*52.6%*)	0.365, *

Table 4: Experiment 1: Combining the smoothing methods. Coverage on seen verbs (*and on all items*) and correlation strength (Spearman's ρ) for PB and FN data. WN synsets as noun classes. Verb classes: IB/ID: smoothing algorithm, followed by number of clusters. ns: not significant, *: p<0.05, **: p<0.01

tences, as the labeller relies on a number of features from parse trees. We used the experimental items from the McRae et al. study, which are all disambiguated towards a reduced relative reading (object interpretation: *The hunter shot by the ...*) of the argument. In doing this, we are potentially biasing the SVM labeller towards one label, depending on the influence of syntactic features on role assignment. We therefore also created a main clause reading of the verb-argument pairs (subject interpretation: *The hunter shot the ...*) and present the results for comparison. For our model, we have previously not specified the grammatical function of the argument, but in order to put both models on a level playing field, we now supply the grammatical function of *Ext* (external argument), which applies for both formulations of the items.

Table 5 shows that for the labelling task, our model outperforms the labelling baseline and the SVM labeller on the FrameNet data by at least 16 points F score while the correlation with human data remains significant. For the PropBank data, labelling performance is on baseline level, below the better of the two SVM labeller conditions. This result underscores the usefulness of argument-specific plausibility estimates furnished by class-based smoothing for the FrameNet data. For the PropBank data, our model essentially assigns the most frequent role for the verb.

The performance of the SVM labeller suggests a strong influence of syntactic features: On the PropBank data set, it always assigns the *Arg0* label if the argument was presented as a subject (this is correct in 50% of cases) and mostly the appropriate *ArgN* label if the argument was presented as an object. On FrameNet, performance again is above baseline only for the subject condition, where there is also a clear trend for assigning agent-style roles. (The object condition is less clear-cut.) This strong reliance on syntactic cues, which may be misleading for our data, makes the labeller perform much worse than on the standard test sets. For both training corpora, it does not take word-specific plausibility into account due to data sparseness and usually assigns the same role to both arguments of a verb. This precludes a significant correlation with the human ratings.

Comparing the training corpora, we find that both models perform better on the FrameNet data even though there are many more role labels in FrameNet, and the SVM labeller does not profit from the greater smoothing power of FrameNet verb clusters. Overall, FrameNet has proven more useful to us, despite its smaller size.

In sum, our model does about as well (PB data) or better (FN data) on the labelling task as the SVM labeller, while the labeller does not solve the prediction task. The success of our model, especially on the prediction task, stems partly from the absence of global syntactic features that bias the standard labeller strongly. This also makes our model suited for an incremental task. Instead of

Train	Model	Coverage	ρ	Labelling F	Labelling Cov.
PB	Baseline	–	–	45.7 (*43.8%*)	100%
	SVM Labeller (subj)	100% (*92%*)	ns	**50** (*47.9%*)	100%
	SVM Labeller (obj)	100% (*92%*)	ns	45.7 (*43.8%*)	100%
	IB 5 (subj/obj)	93.5% (*86%*)	ns	45.7 (*43.8%*)	100%
FN	Baseline	–	–	34.4 (*26.8%*)	100%
	SVM Labeller (subj)	87.5% (*56%*)	ns	40.6 (*31.7%*)	100%
	SVM Labeller (obj)	87.5% (*56%*)	ns	34.4 (*26.8%*)	100%
	ID 4 (subj/obj)	87.5% (*56%*)	0.271, *	**56.3** (*43.9%*)	100%

Table 5: Experiment 2: Standard SVM labeller vs our model. Coverage on seen verbs (*and on all items*), correlation strength (Spearman's ρ), labelling F score and labelling coverage on seen verbs (*and on all items, if different*) for PB and FN data on the McRae test set. ns: not significant, *: p<0.05.

syntactic cues, we successfully rely on argument-specific plausibility estimates furnished by class-based smoothing. Our joint probability model has the further advantage of being conceptually much simpler than the SVM labeller, which relies on a sophisticated machine learning paradigm. Also, we need to compute only about one-fifth of the number of SVM features.

7 Conclusions

We have defined the psycholinguistically motivated task of predicting human plausibility ratings for verb-role-argument triples. To solve it, we have presented an incremental probabilistic model of human plausibility judgements. When we employ two complementary smoothing methods, the model achieves both good coverage and reliable correlations with human data. Our model performs as well as or better than a standard role labeller on the task of assigning the preferred role to each item in our test set. Further, the standard labeller does not succeed on the prediction task, as it cannot overcome the extreme sparse data problem.

Acknowledgements Ulrike Padó acknowledges a DFG studentship in the International Post-Graduate College "Language Technology and Cognitive Systems". We thank Ana-Maria Giuglea, Alessandro Moschitti and Zvika Marx for making their software available and are grateful to Amit Dubey, Katrin Erk, Mirella Lapata and Sebastian Padó for comments and discussions.

References

Bard, E. G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, **72**(1), 32–68.

Carreras, X. and Márquez, L. (2005). Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of CoNLL-2005*.

Crocker, M. and Brants, T. (2000). Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, **29**(6), 647–669.

Gedeon, T., Parker, A., and Dimitrov, A. (2003). Information distortion and neural coding. *Canadian Applied Mathematics Quarterly*, **10**(1), 33–70.

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

Giuglea, A.-M. and Moschitti, A. (2004). Knowledge discovery using FrameNet, VerbNet and PropBank. In *Proceedings of the Workshop on Ontology and Knowledge Discovering at ECML 2004*.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, **20**, 137–194.

Keller, F. and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, **29**(3), 459–484.

Litkowski, K. (2004). Senseval-3 task: Automatic labeling of semantic roles. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.

Marx, Z. (2004). *Structure-Based computational aspects of similarity and analogy in natural language*. Ph.D. thesis, Hebrew University, Jerusalem.

McRae, K., Spivey-Knowlton, M., and Tanenhaus, M. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, **38**, 283–312.

Narayanan, S. and Jurafsky, D. (2002). A Bayesian model predicts human parse preference and reading time in sentence processing. In S. B. T. G. Dietterich and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 59–65. MIT Press.

Tishby, N., Pereira, F. C., and Bialek, W. (1999). The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Trueswell, J., Tanenhaus, M., and Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, **33**, 285–318.