Flexible, Corpus-Based Modelling of Human Plausibility Judgements

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

Katrin Erk
Dept. of Linguistics
University of Texas at Austin
Austin, Texas
katrin.erk@mail.utexas.edu

Abstract

In this paper, we consider the computational modelling of human plausibility judgements for verb-relation-argument triples, a task equivalent to the computation of selectional preferences. Such models have applications both in psycholinguistics and in computational linguistics.

By extending a recent model, we obtain a completely corpus-driven model for this task which achieves significant correlations with human judgements. It rivals or exceeds deeper, resource-driven models while exhibiting higher coverage. Moreover, we show that our model can be combined with deeper models to obtain better predictions than from either model alone.

1 Introduction

One fundamental and intuitive finding in experimental psycholinguistics is that humans judge the plausibility of a verb-argument pair vastly differently depending on the semantic relation in the pair. Table 1 lists example human judgements which McRae et al. (1998) elicited by asking about the plausibility of, e.g., a hunter shooting (relation *agent*) or being shot (relation *patient*). McRae et al. found that "hunter" is judged to be a very plausible *agent* of "shoot" and an implausible *patient*, while the reverse is true for "deer". In linguistics, this phenomenon is explained by *selectional preferences* on verbs' argument positions; we use *plausibility* and *fit with selectional preferences* interchangeably.

Verb	Relation	Noun	Plausibility
shoot	agent	hunter	6.9
shoot	patient	hunter	2.8
shoot	agent	deer	1.0
shoot	patient	deer	6.4

Table 1: Verb-relation-noun triples with plausibility judgements on a 7-point scale (McRae et al., 1998)

In this paper, we consider computational models that predict human plausibility ratings, or the fit of selectional preferences and argument, for such $(verb, relation, argument)$, in short, (v, r, a), triples. Being able to model this type of data is relevant in a number of ways. From the point of view of psycholinguistics, selectional preferences have an important effect in human sentence processing (e.g., McRae et al. (1998), Trueswell et al. (1994)), and models of selectional preferences are therefore necessary to inform models of this process (Padó et al., 2006). In computational linguistics, a multitude of tasks is sensitive to selectional preferences, such as the resolution of ambiguous attachments (Hindle and Rooth, 1993), word sense disambiguation (McCarthy and Carroll, 2003), semantic role labelling (Gildea and Jurafsky, 2002), or testing the applicability of inference rules (Pantel et al., 2007).

A number of approaches has been proposed to model selectional preference data (Padó et al., 2006; Resnik, 1996; Clark and Weir, 2002; Abe and Li, 1996). These models generally operate by generalising from seen (v, r, a) triples to unseen ones. By relying on resources like corpora with semantic role annotation or the WordNet ontology, these models

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 400–409, Prague, June 2007.

generally share two problems: (a), limited coverage; and (b), the resource (at least partially) predetermines the generalisations that they can make.

In this paper, we investigate whether it is possible to predict the plausibility of (v, r, a) triples in a completely corpus-driven way. We build on a recent selectional preference model (Erk, 2007) that bases its generalisations on word similarity in a vector space. While that model relies on corpora with semantic role annotation, we show that it is possible to predict plausibility ratings solely on the basis of a parsed corpus, by using shallow cues and a suitable vector space specification.

For evaluation, we use two balanced data sets of human plausibility judgements, i.e., datasets where each verb is paired both with a good agent and a good patient, and where both nouns are presented in either semantic relation (as in Table 1). Using balanced test data is a particularly difficult task, since it forces the models to account reliably both for the influence of the semantic relation (*agent/patient*) and of the argument head ("hunter"/"deer").

We obtain three main results: (a), our model is able to match the superior performance of the model proposed by Padó et al. (2006), while retaining the high coverage of the model proposed by Resnik (1996); (b), using parsing as a preprocessing step improves the model's performance significantly; and (c), a combination of our model with the Padó model exceeds both individual models in accuracy.

Plan of the paper. In Section 2, we give an overview of existing selectional preferences and vector space models. Section 3 introduces our model and discusses its parameters. Sections 4 and 5 present our experimental setup and results. Section 6 concludes.

2 Related Work

Modelling Selectional Preferences with Grammatical Functions. The idea of inducing selectional preferences from corpora was introduced by Resnik (1996). He approximated the semantic verb-argument relations in (v, r, a) triples by grammatical functions, which are readily available for large training corpora. His basic two-step procedure was followed by all later approaches: (1), extract argument headwords for a given predicate and relation from a corpus; (2), generalise to other, similar words using the WordNet noun hierarchy. Other models also relying on the WordNet resource include Abe and Li (1996) and Clark and Weir (2002).

We present Resnik's model in some detail, since we will use it for comparison below. Resnik first computes the overall selectional preference strength for each verb-relation pair, i.e. the degree of "constrainedness" of each relation. This quantity is estimated as the difference (in terms of the Kullback-Leibler divergence D) between the distribution over WordNet argument classes given the relation, $p(c|r)$, and the distribution of argument classes given the current verb-relation combination, $p(c|v, r)$. The intuition is that a verb-relation pair that only allows for a limited range of argument heads will have a probability distribution over argument classes that strongly diverges from the prior distribution.

Next, the selectional association of the triple, $A(v, r, c)$, is computed as the ratio of the selectional preference strength for this particular class, divided by the overall selectional preference strength of the verb-relation pair. This is shown in Equation 1.

$$A(v, r, c) = \frac{p(c|v, r) log \frac{p(c|v,r)}{p(c|r)}}{D(p(c|r)||p(c|v, r))} \tag{1}$$

Finally, the selectional preference between a verb, a relation, and an argument head is taken to be the selectional association of the verb and relation with the most strongly associated WordNet ancestor class of the argument.

WordNet-based approaches however face two problems. One is a coverage problem due to the limited size of the resource (see the task-based evaluation in Gildea and Jurafsky (2002)). The other is that the shape of the WordNet hierarchy determines the generalisations that the models make. These are not always intuitive. For example, Resnik (1996) observes that $(answer, obj, tragedy)$ receives a high preference because "tragedy" in WordNet is a type of written communication, which is a preferred argument class of "answer".

Rooth et al. (1999) present a fundamentally different approach to selectional preference induction which uses soft clustering to form classes for generalisation and does not take recourse to any hand-crafted resource. We will argue in Section 6 that our model allows more control over the generalisations made.

Modelling Selectional Preferences with Thematic Roles. Padó et al. (2006) present a deeper model for the plausibility of (v, r, a) triples that approximates the relations with thematic roles. It estimates the selectional preferences of a verb-role pair with a generative probability model that equates the plausibility of a (v, r, a) triple with the joint probability of seeing the thematic role with the verb-argument pair. In addition, the model also considers the verb's sense s and the grammatical function gf of the argument; however, since the model is generative, it can make predictions even when not all variables are instantiated. The final model is shown in Equation 2.

$$Plausibility_{v,r,a} = P(v, s, r, a, gf) \qquad (2)$$

The induction of this model from the FrameNet corpus of semantically annotated training data (Fillmore et al., 2003) encounters a serious sparse data problem, which is approached by the application of word-class-based and Good-Turing re-estimation smoothing. The resulting model's plausibility predictions are significantly correlated to human judgements, but because of the use of verb-specific thematic roles, the model's coverage is still restricted by the verb coverage of the training corpus.

Vector Space Models. Another class of models that has found wide application in lexical semantics is the family of vector space models. In a vector space model, each *target word* is represented as a vector, typically constructed from co-occurrence counts with context words in a large corpus (the so-called *basis elements*). The underlying assumption is that words with similar meanings occur in similar contexts, and will be assigned similar vectors. Thus, the distance between the vectors of two target words, as given by some distance measure (e.g., Cosine or Jaccard), is a measure of their *semantic similarity*.

Vector space models are simple to construct, and the semantic similarity they provide has found a wide range of applications. Examples in NLP include information retrieval (Salton et al., 1975), automatic thesaurus extraction (Grefenstette, 1994), and predominant sense identification (McCarthy et al., 2004). In cognitive science, they have been used to account for the influence of context on human lexical processing (McDonald and Brew, 2004), and to model lexical priming (Lowe and McDonald, 2000).

A drawback of vector space models is the difficulty of interpreting what some degree of "generic semantic similarity" between two target words means in linguistic terms. In particular, this similarity is not sensitive to selectional preferences over *specific* semantic relations, and thus cannot model the plausibility data we are interested in. The next section demonstrates how the integration of ideas from selectional preference induction makes this distinction possible.

3 The Vector Similarity Model: Corpus-Based Modelling of Plausibility

3.1 Model Architecture

Our model builds on the architecture of Erk (2007). It combines the idea underlying the selectional preference models from Section 2, namely to predict plausibility by generalising over head words, with vector space similarity. The fundamental idea of our model is to model the plausibility of the triple (v, r, a) by comparing the argument head a to other headwords a' which we have already seen in a corpus for the same verb-relation pair (v, r), and which we therefore assume to be plausible. We write $\text{Seen}_r(v)$ for the *set of seen headwords*. Our intuition is that if a is similar to the words in $\text{Seen}_r(v)$, then the triple (v, r, a) is plausible; conversely, if it is very dissimilar, then the triple is implausible.

Concretely, we judge the plausibility of the triple by averaging over the similarity of the vector for a to all vectors for the seen headwords in $\text{Seen}_r(v)$:

$$Pl(v, r, a) = \sum_{a' \in \text{Seen}_r(v)} \frac{w(a') \cdot sim(a, a')}{|\text{Seen}_r(v)|} \qquad (3)$$

where w is a weight factor specific to each a'. w can be used to implement different weighting schemes that encode prior knowledge, e.g., about the reliability of different words in $\text{Seen}_r(v)$. In this paper, we only consider a very simple weighting factor, namely the frequency of the seen headwords. This encodes the assumption that similarity to frequent head words is more important than similarity to infrequent ones:

$$Pl(v, r, a) = \sum_{a' \in \text{Seen}_r(v)} \frac{f(a') \cdot sim(a, a')}{|\text{Seen}_r(v)|} \qquad (4)$$

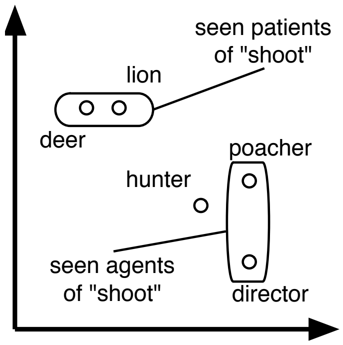

Figure 1: A vector space for estimating the plausibilities of $(shoot, agent, hunter)$ and $(shoot, patient, hunter)$.

This model can be seen as a straightforward implementation of the selectional preference induction process of generalising from seen headwords to other, similar words. By using vector space representations to judge the similarity of words, we obtain a completely corpus-driven model that does not require any additional resources and is very flexible. A complementary view on this model is as a generalisation of traditional vector space models that computes similarity not between two vectors, but between a vector and a set of other vectors. By using the vectors for seen headwords of a given relation as this set, the similarity we compute is specific to this relation.

Example. Figure 1 shows an example vector space. Consider $v =$ "shoot", $r =$ *agent*, and $a =$ "hunter". In order to judge whether a hunter is a plausible agent of "shoot", the vector space representation of "hunter" is compared to all representations of known agents of "shoot", namely "poacher" and "director". Due to the nearness of the vector for "hunter" to these two vectors, "hunter" will be judged a fairly good agent of "shoot". Compare this with the result for the role *patient*: "hunter" is further away from "lion" and "deer", and will therefore be found to be a rather bad patient of "shoot". However, "hunter" is still more plausible as a patient of "shoot" than e.g., "director".

3.2 Instantiating the Model: Unparsed vs. Parsed Corpora

The two major tasks which need to be addressed to obtain an instance of this model are (a), determining the sets of seen head words $\text{Seen}_r(v)$, and (b), the construction of a vector space. Erk (2007) extracted the set of seen head words from corpora with semantic role annotation, and used only a single vector space representation. In this paper, we eliminate the reliance on special annotation by considering shallow approximations of the semantic relations in question. In addition, we discuss in detail which properties of the vector space are crucial for the prediction of plausibility ratings, a much more fine-grained task than the pseudo-word disambiguation task presented in Erk (2007) that is more closely related to semantic role labelling. The goal of our exposition is thus to develop a model that can use more training data, and represent the corpus information optimally in order to obtain superior coverage.

In fact, tasks (a) and (b) can be solved on the basis of unparsed corpora, but we would expect the results to be rather noisy. Fortunately, the state of the art in broad-coverage (Lin, 1993) and unsupervised (Klein and Manning, 2004) dependency parsing allows us to treat dependency parsing merely as a preprocessing step. We therefore describe two instantiations of our model: one based on an unprocessed corpus, and one based on a dependency-based parsed corpus. By comparing the models, we can gauge whether syntactic preprocessing improves model performance. In the following, we describe the strategies the two models adopt for (a) and (b).

Identifying seen head words for relations. Recall that the set $\text{Seen}_r(v)$ is supposed to contain known head words a that are observed in the corpus as triples (v, r, a). In a parsed corpus, we can approximate the relation *agent* by the dependency relation of `subject` provided by the parser, and the relation *patient* by the dependency relation of `object`. In an unparsed corpus, these grammatical relations are unavailable, and the only straightforward evidence we can use is word order. In this case, we assume that words directly adjacent to the left of a predicate are subjects, and therefore *agent*s, whereas words directly to its right are objects, and thus *patient*s.

Vector space topology. The success of our method depends directly on the topology of the vector space. More specifically, two words should only be assigned similar vectors if they are in fact of similar plausibility. If this is not the case, there is no guarantee that a word a that is similar to the words in $\text{Seen}_r(v)$ forms

Basis elements \ Target	deer	hunter
shoot	10	10
escape	12	12

Basis elements \ Target	deer	hunter
shoot-SUBJ	0	8
shoot-OBJ	10	2
escape-SUBJ	10	5
escape-OBJ	2	7

Figure 2: Two vector spaces, using as basis elements either context words (above) or words paired with grammatical functions (below)

a plausible triple (v, r, a) itself (cf. Figure 1).

The topology, in turn, is related to the choice of basis elements. Traditional vector space models use context words as basis elements of the space. The top table in Figure 2 illustrates our intuition that such spaces are problematic: "deer" and "hunter" receive identical vectors, even though they show complementary plausibility ratings (cf. Table 1). The reason is that "deer" and "hunter" often co-occur quite closely to one another (e.g., in the vicinity of "shoot"), and thus show a very similar profile in terms of context words. In preliminary experiments, we found that vector spaces with context words as basis elements are in fact unable to distinguish such word pairs reliably.

In contrast, the bottom table in Figure 2 indicates that this problem can be alleviated by using context words *combined with the grammatical relation to the target word* as basis elements. Target words now receive different representations, depending on the grammatical function in which they occur with context words. In consequence, resulting spaces can distinguish, for example, between "hunter" and "deer".

We adopt word-function pairs as basis elements for the vector spaces in all our models. In a dependency-parsed corpus, the basis elements can be directly read off the syntactic structure. In an unparsed corpus, we again fall back on word order, appending to each context word its relative position to the target word.

4 Experimental Setup

Experimental Materials. In order to make our evaluation comparable to the earlier modelling study by Padó et al. (2006), we present evaluations on the two plausibility judgement datasets used there.[1]

The first dataset consists of 100 data points from McRae et al. (1998). Our example in Table 1, which is taken from this dataset, demonstrates its *balanced* structure: 25 verbs are paired with two arguments and two relations each, such that each argument is highly plausible in one relation, but implausible in the other. The resulting distribution of ratings is thus highly bimodal. Models can only reliably predict the human ratings in this data set if they can capture the difference between verb argument slots as well as as between individual fillers.

The second, larger dataset is less strictly balanced, since its triples are constructed on the basis of corpus co-occurrences (Padó et al., 2006). 18 verbs are combined with the three most frequent subjects and objects from both the Penn Treebank and the FrameNet corpus. Each verb-argument pair was rated both as an agent and as a patient, which leads to a total of 24 rated triples per verb. The dataset contains ratings for a total of 414 triples, due to overlap between corpora. The resulting judgements show a more even distribution of ratings than the McRae data.

Vector Similarity Models. Following our exposition in the last section, we construct two instantiations of our vector similarity model, one using unparsed and one parsed data. Both are trained on the complete British National Corpus (Burnard, 1995, BNC) with more than six million sentences.

The unparsed model (Unparsed) uses the BNC without any pre-processing. We first construct the set of known headwords, $\text{Seen}_r(v)$, as follows: All words up to 2 words to the left of instances of v are assumed to be `subjects`, and thus agents; vice versa for patients to the right. Then, we construct semantic space representations for the experimental arguments and known headwords, adopting optimal parameter settings from the literature (Padó and Lapata, 2007). This means a context window of 5 words to either side and 2,000 basis elements (dimensions), which are formed by the most frequent 1,000 words

[1] We are grateful to Ken McRae for his dataset.

in the BNC, combined with each of the relations agent and patient. All counts are log-likelihood transformed (Lowe, 2001).

To construct the parsed model (Parsed), we dependency-parsed the BNC with Minipar (Lin, 1993). We first obtain the seen headwords $\text{Seen}_r(v)$ by using all subjects and objects of v as agents and patients, respectively. We then construct a vector space for the experimental arguments and known headwords.[2] We use 2,000 dimensions again, but adopt the most frequent $(head, grammatical\ function)$ pairs in the BNC as basis elements. The context window is formed by `subject` and `object` dependencies. All counts are log-likelihood transformed.

We experiment with two distance measures to compute vector similarity, namely the Jaccard Coefficient and Cosine Distance, both of which have been shown to yield good performance in NLP tasks (Lee, 1999; McDonald and Lowe, 1998).

Evaluation Procedure. We evaluate our models by correlating the predicted plausibility values with the human judgements, which range between 1 and 7. Since the human judgement data is not normally distributed, we use Spearman's ρ, a non-parametric rank-order test. We determine the statistical significance of differences in correlation strength using the method described in Raghunathan (2003). This method can deal with missing values and thus allows us to compare models with different coverage.

It is difficult to specify a straightforward baseline for our correlation-based evaluation. In contrast to classification tasks, where models choose one out of a fixed number of classes, our model predicts continuous data. This task is more difficult to approximate, e.g., using frequency information.

With respect to upper bounds, we hold that automatic models of plausibility cannot be expected to surpass the typical agreement on the plausibility judgement task between human participants. Thus, we assume an upper bound of $\rho \approx 0.7$.

Comparison against Other Models. We compare our performance to two models from the literature discussed in Section 2. The first model (Pado) is the thematic role-based model by Padó et al. (2006) trained on the FrameNet (Fillmore et al., 2003) release 1.2 example sentences, a subset of the BNC annotated with semantic roles. This corpus contains about 57,000 sentences, which corresponds to roughly 1% of the BNC data.

The second model (Resnik) is the WordNet-based selectional preference model by Resnik (1996), trained on the dependency-parsed BNC (see above).

[2] This space was computed using the `DependencyVectors` software described in Padó and Lapata (2007). This software can be downloaded from `http://www.coli.uni-saarland.de/~pado/dv.html`.

Model	Coverage	Spearman's ρ
Unparsed Cosine	90%	0.023, ns
Unparsed Jaccard	90%	0.044, ns
Parsed Cosine	91%	0.218, *
Parsed Jaccard	91%	0.129, ns
Resnik	94%	0.028, ns
Pado	56%	0.415, **

Table 2: Model performance on McRae data. *: $p < 0.05$, **: $p < 0.01$

5 Experimental Evaluation

The McRae Dataset. Table 2 summarises our results on the McRae dataset. The upper part shows the results for our two vector similarity models (Parsed/Unparsed), combined with the two distance measures (Cosine/Jaccard). The lower part shows the two resource-based models we use for comparison.

We find that all vector similarity models exhibit high coverage (above 90%), and one model (Parsed Cosine) can predict human judgements with a significant correlation. The instantiation of the model has a significant impact on the performance: The Parsed models clearly outperform the Unparsed models. The effect of the distance measure is less clear-cut, since the Unparsed models perform better with Jaccard, while the Parsed models prefer Cosine.

The deep semantic plausibility model (Pado) makes predictions only for slightly more than half of the data. This low coverage is a direct result of the small overlap in verbs between the McRae dataset and the FrameNet corpus. However, on the data points it covers, it achieves a significant correlation to human judgements. The correlation coefficient is numerically much higher than that of the Parsed Cosine model, but due to the large coverage difference, the two models are not statistically distinguishable.

Model	Coverage	Spearman's ρ
Unparsed Cosine	98%	0.117, *
Unparsed Jaccard	98%	0.149, **
Parsed Cosine	98%	0.479, ***
Parsed Jaccard	98%	0.120, *
Resnik	98%	0.237, ***
Pado	97%	0.515, ***

Table 3: Model performance on Pado data. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$

Resnik's WordNet-based model shows a coverage that is comparable to the vector similarity models, but does not achieve a significant correlation to the human judgements.

The Pado Dataset. Table 3 summarises the results for the Pado dataset. Since all verbs in this dataset are covered in FrameNet, the deep Pado model shows a coverage comparable to all other models, at >95%.

The main difference to the McRae dataset lies in the models' performance. We find that all models, including the Unparsed vector models and Resnik, manage to achieve significant correlations with the human judgements. Within the vector similarity models, the same trends hold as for the McRae dataset: Parsed outperforms Unparsed, and the best combination is Parsed Cosine. The models fall into two clearly separated groups: The Pado and Parsed Cosine models achieve a highly significant correlation, and are statistically indistinguishable. They significantly outperform the second group ($p < 0.001$), formed by all other models. Within this second group, Resnik is numerically the best model and shows a significant correlation with human data; nevertheless, the difference to the first group is evident from its substantially lower correlation coefficient.

The construction of the Pado dataset allows a further analysis. As mentioned in Section 4, the dataset consists of verb-argument pairs drawn from two different corpora. Therefore, each verb is combined both with some arguments that are seen in FrameNet, and some that are not. Our hypothesis is that the FrameNet-trained Pado model performs considerably better on the 216 "FN-Seen" data points (verb-argument pairs observed in FrameNet in at least one relation) than on the 198 "FN-Unseen" data points (verb-argument pairs unseen in both relations).

Table 4 shows the results of this analysis for the best-performing models. We observe a pattern corresponding to our expectations: The performance of the Pado model is clearly worse for FN-Unseen than for FN-Seen, while the Resnik and Parsed Cosine models perform more evenly across both datasets. While the Pado model is significantly better on the FN-Seen dataset, it is numerically outperformed by the Parsed Cosine model for the FN-Unseen data points. We conclude that the deep model is more accurate within the coverage of its resources, but loses its advantage when it has to resort to smoothing.

Model combination. Our last analysis indicates that the models have complementary strengths: the thematic role-based Pado model is the best plausibility predictor on the data points it has seen, while the Parsed cosine model overall predicts human data only numerically worse, and with better coverage. We therefore suggest to combine the predictions of the two models to combine their respective strengths.

For the moment, we only consider a naive backoff scheme: For each data point, we use the prediction of the Pado model if the data point is "FN-Seen" (cf. the last paragraph), and the prediction of the Parsed Cosine model otherwise. Note that this criterion does not consider the predictions of the models themselves, only properties of the underlying training set.

The actual combination requires a normalisation of the respective predictions, since one of the models (Pado) is probabilistic, while the other one (Parsed Cosine) is similarity-based, and their predictions are not directly comparable. We perform a simple normalisation by z-transforming the complete predictions of each model.[3] The combination of the scaled predictions in fact results in an improved correlation with the human data. The correlation coefficient of ρ=0.552 numerically exceeds either base model, and the coverage of 98% corresponds to the coverage of the more robust Parsed Cosine model.

We take this result as evidence that even a simple combination technique can lead to improved predictions. Unfortunately, our naive backoff scheme does not directly carry over to the McRae dataset, where only 2 out of 100 data points are "FN-Seen", and the Pado model would thus hardly contribute.

[3] The z transformation scales a dataset to a mean of 0 and a standard deviation of 1.

Model	FN-Seen Data			FN-Unseen Data		
Parsed Cosine	94%	0.426,	***	100%	0.461,	***
Resnik	96%	0.217,	**	100%	0.263,	***
Pado	97%	0.569,	***	96%	0.383,	***

Table 4: Performance on data points seen and unseen in FrameNet (Pado dataset). **: $p < 0.01$ ***: $p < 0.001$

Discussion. We have verified experimentally that our vector similarity model is able to match the performance of a deep plausibility model, exceeding it in coverage, and to outperform a WordNet-based selectional preference model. We conclude that a completely corpus-driven approach constitutes a viable alternative to resource-based models.

One insight from our experiments is that vector similarity models constructed from dependency-parsed corpora perform significantly better than unparsed models. This indicates that dependency relations like `subject` and `object` are reliable syntactic correlates of semantic relations like *agent* and *patient*, but that their approximation in terms of word order introduces considerable noise. The Parsed models are best combined with Cosine Distance. We surmise that Cosine, which tends to consider low-frequency words more than Jaccard, is more susceptible to the additional noise in unparsed corpora.

Furthermore, the choice of basis elements for the vector space is vital: Plausibilities could only be predicted successfully with word-relation pairs as basis elements. This is in contrast to recent results on predominant sense acquisition, the task of identifying the most frequent sense for a given word in an unsupervised manner (McCarthy et al., 2004). On that task, Padó and Lapata (2007) found vector spaces with words as basis elements are in fact competitive with models using word-relation pairs. This divergence underlines an interesting difference between the two tasks. Evidently, predominant senses identification, as a WSD-related task, can succeed on the basis of topical information, which is represented well in word-based spaces. In contrast, plausibility judgments can only be predicted by a space based on word-relation pairs which can represent the finer-grained distinctions arising from different *relations* between verb and noun.

A second important finding is that the relative performance of the different models is the same on the McRae and Pado datasets. The Pado model performs best, followed by our Parsed Cosine vector similarity model, followed by the Unparsed and Resnik models.

The McRae dataset, however, is much more difficult to account for than the Pado data, independent of the model. This effect was already noted by Padó et al. (2006), who attributed it to the very limited overlap between the McRae dataset and FrameNet. While this explanation can account for the difference for the Pado model, we observe the same pattern across all models. This suggests that a more general frequency effect is at work here: The median frequency of the hand-selected McRae nouns is 1,356 in the BNC, as opposed to 8,184 for the corpus-derived Pado nouns. The resulting sparseness affects all model families, since all ultimately rely on co-occurrences.

The performance difference between the two datasets is particularly large for the WordNet-based selectional preference model (Resnik). A further analysis of the model's predictions shows that the model has difficulty in distinguishing between verb-relation-argument triples that differ only in the argument, such as $(shoot, agent, hunter)$ and $(shoot, agent, deer)$. Recall that it is crucial for the prediction of the McRae data to make this distinction, since the arguments for each relation are chosen to differ widely in plausibility. The reason for the Resnik model's difficulty is that arguments are mapped onto WordNet synsets, and whenever two arguments are mapped onto closely related synsets, their plausibility ratings are similar. This problem is graver for the McRae test set, where all arguments are animates, and thus more similar in terms of WordNet, than for the Pado set, which also contains a portion of inanimate arguments with animate counterparts. This analysis highlights again the fundamental problem of resource-based models, where design decisions of the underlying resource may limit, or even mislead, the models' generalisations.

Finally, we have shown in a first experiment that the syntax-based vector similarity model can be combined with the role-base model to obtain a combined model that performs superior to both. In this combined model, the shallow model's better coverage supplements the accurate predictions of the deep model.

6 Conclusions

In this paper, we have considered the computational modelling of human plausibility judgements for verb-relation-argument triples, a task equivalent to the computation of selectional preferences. We have extended a recent proposal (Erk, 2007) which combines ideas from selectional preference induction and vector space models. Our model can be constructed from a large corpus with partial syntactic information (specifically, subject and object relations) from which it builds an optimally informative vector space.

We have demonstrated that the successful evaluation of the model in Erk (2007) on the coarse-grained pseudo-word disambiguation task carries over to the prediction of human plausibility judgments which requires relatively fine-grained, relation-based distinctions. Our model is competitive with existing "deep" models while exhibiting a higher coverage. We have also shown that our vector similarity model can be combined with a "deep" model so that the combined model outperforms both base models. A thorough investigation of strategies for prediction combination and scaling remains future work.

The strategy of our model to derive generalisations directly from corpus data, without recourse to resources, is similar to another family of corpus-driven selectional preference models, namely EM-based clustering models (Rooth et al., 1999). However, we believe that our model has a number of advantages. (1), It is conceptually simple and implements the intuition behind selectional preference models, "generalise from known headwords to unknown ones", particularly directly through the comparison of new headwords to known ones according to a given definition of similarity. (2), The separation of the similarity computation and the acquisition of seen headwords gives the experimenter fine-grained control over the types and sources of information which inform the construction of the model. (3), The instantiation of the similarity computation with a vector space makes it possible to integrate additional linguistic information beyond verb-argument co-occurrences into the model, building on a large body of work in vector space construction. In sum, our modular model provides a higher degree of control than one-step models like the EM-based proposal.

An important avenue of further research is the ability of the vector plausibility model to model finer-grained distinctions between semantic relations beyond the agent/patient dichotomy, as thematic role-based models are able to. Excluding the direct use of role-annotated corpora like FrameNet for coverage reasons, the most promising strategy is to extend our present scheme of approximating semantic relations by grammatical realisations. How much noise this approximation introduces when finer role sets are used is an open research question.

Acknowledgments. The work presented in this paper was supported by the financial support of DFG (grants Pi-154/9-2 and IRTG "Language Technology and Cognitive Systems").

References

Naoki Abe and Hang Li. 1996. Learning word association norms using tree cut pair models. In *Proceedings of ICML 1996*, pages 3–11.

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

Stephen Clark and David Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206.

Katrin Erk. 2007. A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th ACL*, Prague, Czech Republic.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

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

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42th ACL*, pages 478–485, Barcelona, Spain.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th ACL*, pages 25–32, College Park, MA.

Dekang Lin. 1993. Principle-based parsing without overgeneration. In *Proceedings of the 31st ACL*, pages 112–120, Columbus, OH.

Will Lowe and Scott McDonald. 2000. The direct route: Mediated priming in semantic space. In *Proceedings of the 22nd CogSci*, pages 675–680, Philadelphia, PA.

Will Lowe. 2001. Towards a theory of semantic space. In *Proceedings of the 23rd CogSci*, pages 576–581, Edinburgh, UK.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computatinal Linguistics*, 29(4):639–654.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42th ACL*, pages 279–286, Barcelona, Spain.

Scott McDonald and Chris Brew. 2004. A distributional model of semantic context effects in lexical processing. In *Proceedings of the 42th ACL*, pages 17–24, Barcelona, Spain.

Scott McDonald and Will Lowe. 1998. Modelling functional priming and the associative boost. In *Proceedings of the 20th CogSci*, pages 675–680, Madison, WI.

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

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2).

Ulrike Padó, Frank Keller, and Matthew W. Crocker. 2006. Combining syntax and thematic fit in a probabilistic model of sentence processing. In *Proceedings of the 28th CogSci*, pages 657–662, Vancouver, BC.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning inferential selectional preferences. In *Proceedings of NAACL 2007*, Rochester, NY.

Trivellore Raghunathan. 2003. An approximate test for homogeneity of correlated correlations. *Quality and Quantity*, 37:99–110.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–159.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing an semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th ACL*, pages 104–111, College Park, MA.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector-space model for information retrieval. *Journal of the American Society for Information Science*, 18:613–620.

John Trueswell, Michael Tanenhaus, and Susan Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33:285–318.