A robust and extensible exemplar-based model of thematic fit

Bram Vandekerckhove[a], Dominiek Sandra[a], Walter Daelemans[b]
[a]Center for Psycholinguistics, [b]Center for Dutch Language and Speech (CNTS)
University of Antwerp
Antwerp, Belgium
{bram.vandekerckhove,dominiek.sandra,walter.daelemans}@ua.ac.be

Abstract

This paper presents a new, exemplar-based model of thematic fit. In contrast to previous models, it does not approximate thematic fit as argument plausibility or 'fit with verb selectional preferences', but directly as semantic role plausibility for a verb-argument pair, through similarity-based generalization from previously seen verb-argument pairs. This makes the model very robust for data sparsity. We argue that the model is easily extensible to a model of semantic role ambiguity resolution during online sentence comprehension.

The model is evaluated on human semantic role plausibility judgments. Its predictions correlate significantly with the human judgments. It rivals two state-of-the-art models of thematic fit and exceeds their performance on previously unseen or low-frequency items.

1 Introduction

Thematic fit (or semantic role plausibility) is the plausibility of a noun phrase referent playing a specific semantic role (like *agent* or *patient*) in the event denoted by a verbal predicate, e.g. the plausibility that a judge sentences someone (which makes the judge the agent of the sentencing event) or that a judge is sentenced him- or herself (which makes the judge the patient). Thematic fit has been an important concept in psycholinguistics as a predictor variable in models of human sentence comprehension, either to discriminate between possible structural analyses during initial processing in constraint-based models (see MacDonald and Seidenberg (2006) for a recent overview), or after initial syntactic processing in modular models (e.g. Frazier (1987)). In fact, thematic fit is at the core of the most-studied of all structural ambiguity phenomena, the ambiguity between a main clause or a reduced relative clause interpretation of an *NP verb-ed* sequence (the MV/RR ambiguity), which is essentially a semantic role ambiguity. If the temporarily ambiguous sentence *The judge sentenced ...* is continued as a main clause (e.g. *The judge sentenced him to 10 years in prison*), the noun phrase *the judge* would be the agent of the verb *sentenced*, while it would be the patient of *sentenced* in a reduced relative clause continuation (e.g. *The judge sentenced to 4 years in prison for indecent exposure could also lose his state pension*). Apart from its importance in psycholinguistics, the concept of thematic fit is also relevant for computational linguistics in general (see Padó et al. (2007) for some examples).

A number of models that try to capture human thematic fit preferences have been developed in recent years (Resnik, 1996; Padó et al., 2006; Padó et al., 2007). These previous approaches rely on the linguistic notion of verb selectional preferences. The plausibility that an argument plays a specific semantic role in the event denoted by a verb—in other words, that a verb, role and argument occur together—is predicted by how well the argument head fits the restrictions that the verb imposes on the argument candidates for the semantic role slot under consideration (e.g. *eat* prefers edible arguments to fill its patient slot). Therefore, what these models capture is actually not semantic role plausibility, but argument plausibility.

The model presented here takes a different approach. Instead of predicting the plausibility of an argument given a verb-role pair (e.g. the plausibility of *judge* given *sentence-patient*), it predicts the plausibility of a semantic role given a verb-argument pair (e.g. the plausibility of *patient* given *sentence-judge*), through similarity-based generalization from previously seen verb-argument pairs. In the context of modeling thematic fit as a con-

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 826–834,
Athens, Greece, 30 March – 3 April 2009.

straint in the resolution of sentence-level ambiguity problems like the MV/RR ambiguity, predicting role fit instead of argument fit seems to be the most straightforward approach. After all, when thematic fit is approached in this way, the model directly captures the semantic role ambiguity that is at stake during the analysis of sentences that are temporarily ambiguous between a main clause and a reduced relative interpretation. This means that our model of thematic fit should be very easy to extend into a full-blown model of the resolution of any sentence-level ambiguity that crucially revolves around a semantic role ambiguity. In addition, the fact that it generalizes from previously seen verb-argument pairs, based on their similarity to the target pair, should make it more robust than previous approaches.

The remainder of the paper is organized as follows: in the next section, we briefly discuss two state-of-the-art thematic fit models, the performance of which will be compared to that of our model. Section 3 introduces three different instantiations of our model. The evaluation of the model and the comparison of its performance with that of the models discussed in Section 2 is presented in Section 4. Section 5 ties everything together with some general conclusions.

2 Previous models

In this section of the paper, we look at two state-of-the-art models of thematic fit, developed by Padó et al. (2006) and Padó et al. (2007). We will not discuss the selectional preferences model of Resnik (1996), but for a comparison between the Resnik model and the Padó models, see Padó et al. (2007).

2.1 Padó et al. (2006)

In their model of thematic fit, Padó et al. (2006) use FrameNet thematic roles (Fillmore et al., 2003) to approximate semantic roles. The thematic fit of a verb-role-argument triple (v, r, a) is given by the joint probability of the role r, the argument headword a, the verb sense v_s, and the grammatical function gf of a:

$$Plausibility_{v,r,a} = P(v_s, r, a, gf) \qquad (1)$$

Since computing this joint probability from corpus co-occurrence frequencies is problematic due to an obvious sparse data issue, the term is decomposed into several subterms, including a term $P(a|v_s, gf, r)$ that captures selectional preferences. Good-Turing and class-based smoothing are used to further alleviate the remaining sparse data problem, but because of the fact that the model can only make predictions for verbs that occur in the small FrameNet corpus, for a large number of verbs, it cannot provide any output. For the verbs that do occur in the training corpus, however, the model's predictions correlate very well with human plausibility ratings.

2.2 Padó et al. (2007)

The model of Padó et al. (2007) does not use semantically annotated resources, but approximates the agent and patient relations with the syntactic subject and object relations, respectively. The plausibility of a verb-role-argument triple (v, r, a) is found by calculating the weighted mean semantic similarity of the argument headword a to all headwords that have previously been seen together with the verb-role pair (v, r), as shown in Equation 2. The prediction is that high semantic similarity of a target headword a to seen headwords for a given (v, r) tuple corresponds to high thematic fit of the (v, r, a) tuple, while low similarity implies low thematic fit.

$$Plausibility_{v,r,a} = \sum_{a' \in Seen_r(v)} \frac{w(a') \times sim(a, a')}{|Seen_r(v)|} \qquad (2)$$

$w(a')$ is the weighting factor. Padó et al. (2007) used the frequency of the previously seen argument headwords as weights. Similarity between headwords was defined as the cosine between so-called 'dependency vector' representations of these headwords (Padó and Lapata, 2007). These vectors are constructed from the frequency counts with which the target items occur at one end of specific paths in a corpus of syntactic dependency trees. The argument headword vectors Padó et al. (2007) used in their experiments consisted of 2000 features, representing the most frequent $(head, subject)$ and $(head, object)$ pairs in the British National Corpus (BNC). The feature-values of the headword vectors were the log-likelihoods of the headwords occurring at the dependent end of these $(relation, head)$ pairs (so either as subjects or objects of the heads). The model's performance approaches that of the Padó et al. (2006) model on the correlation of its predictions with human ratings, and it attains higher cov-

erage (it can provide plausibility values for a larger proportion of the test items), since the model only requires that the verb occurs with subject and object arguments in the training corpus, and that the target argument headwords occur in the training data frequently enough to attain reliable dependency vectors.

3 Exemplar-based modeling of thematic fit

Exemplar-based models of cognition (also known as Memory-Based Learning or instance/case-based reasoning/learning models) (Fix and Hodges, 1951; Cover and Hart, 1967; Daelemans and van den Bosch, 2005) are classification models that extrapolate their behavior from stored representations of earlier experiences to new situations, based on the similarity of the old and the new situation. These models keep a database of stored exemplars and refer to that database to guide their behavior in new situations. Models can extrapolate from only one similar memory exemplar, a group of similar exemplars (a nearest neighbor set), or even the whole exemplar memory, using some decay function to give less weight to less similar exemplars.

Applied to our model of thematic fit, this means that the model should have a database in which semantic representations of verb-argument pairs are stored together with the semantic roles of the arguments. The plausibility of a semantic role given a new verb-argument pair is then determined by the support for that role among the verb-argument pairs in memory that are semantically most similar to the target pair.

An immediately obvious advantage of this approach should be its potential robustness for data sparsity, since similarity-based smoothing is an intrinsic part of the model. Even if neither the verb nor the argument of a verb-argument pair occur in the exemplar memory, role plausibilities can be predicted, as long as the similarity of the target exemplar's semantic representation with the semantic representations in the exemplar memory can be calculated. An additional advantage of similarity-based smoothing is that it does not involve the estimation of an exponential number of smoothing parameters, as is the case for backed-off smoothing methods (Zavrel and Daelemans, 1997).

For this study, we will implement three different kinds of exemplar-based models. The first model is a basic k-Nearest Neighbor (k-NN) model. In this model, the plausibility rating for a semantic role given a verb-argument pair is simply determined by the (relative) frequency with which that semantic role is assigned to the k verb-argument pairs that are nearest (i.e. most similar) to the target verb-argument pair (these exemplars constitute the nearest neighbor set). The second model adds a decay function to this simple k-NN model, so that not only the role frequency, but also the absolute semantic distance between the target item and the neighbors in the nearest neighbor set determine the plausibility rating. In the third model, a normalization factor ensures that distance of the exemplars in the nearest neighbor set to the target item determines their weight in the calculation of the plausibility rating while factoring out an effect of absolute distance.

The semantic distance between two verb-argument exemplars is determined by the semantic distance between the verbs and between the nouns. In all models described below, the distance between two exemplars i and j (d_{ij}) is given by the sum of the weighted distances (δ) between the semantic representations of the exemplars' nouns (n) and verbs (v):

$$d_{ij} = w_v \times \delta(v_i, v_j) + w_n \times \delta(n_i, n_j) \qquad (3)$$

We are not theoretically committed to any specific semantic representation or similarity metric for the computation of $\delta(v_i, v_j)$ and $\delta(n_i, n_j)$. The only requirement is that they should be able to distinguish nouns that typically occur in the same contexts, but in different roles (like *writer* and *book*), which probably excludes all vector-based approaches that do not take into account syntactic information (see also Padó et al. (2007)).

In the next three sections, each of the three exemplar-based models is discussed in more detail.

3.1 A basic k-NN model

The most basic of all exemplar-based models is a k-NN model in which the preference strength of a class upon presentation of a stimulus is simply the relative frequency of that class among the nearest neighbors of the stimulus. In the context of thematic fit, this means that the preference strength (PS) for a semantic role response J given a verb-argument stimulus i is found by summing the frequencies of all exemplars with semantic role J

Verb	Noun	Role	Rating
sentence	judge	agent	6.9
sentence	judge	patient	1.3
sentence	criminal	agent	1.3
sentence	criminal	patient	6.7

Table 1: Example mean thematic fit ratings from McRae et al. (1998)

among the k nearest neighbors of i (C_J^k) and dividing this by the total number of exemplars in the k-nearest neighbor set, with k (the number of nearest neighbors taken into consideration) being a free parameter:

$$PS(R_J|S_i) = \frac{\sum_{j \in C_J^k} f(j)}{\sum_{l \in C^k} f(l)} \quad (4)$$

We will call this model the k-NN frequency model (henceforth kNNf).

3.2 A distance decay model

The kNNf model uses the similarity between the target exemplar and the memory exemplars only to determine which items belong to the nearest neighbor set. Whether these nearest neighbors are very similar or only slightly similar to the target exemplar, or whether there are some very similar items but also some very dissimilar items among those neighbors does not have any influence on the class's preference strength; only relative frequency within the nearest neighbor set counts.

Only relying on the relative frequency of semantic roles within the nearest neighbor set to predict their plausibilities might indeed be a reasonable approach to modeling thematic fit in a lot of cases. Being a good agent for a given verb often entails being a bad patient for that same verb (or even in general), and the other way around. For example, *judge* is a very plausible agent of the verb *sentence*, while at the same time it is a rather unlikely patient of the same verb, while it is exactly the other way around for *criminal*, as the mean participant ratings (on a 7-point scale) in Table 1 show (these were taken from McRae et al. (1998)). The relative frequencies of the agent and patient roles in the nearest neighbor set could in theory perfectly explain these ratings: a high relative frequency of the agent role among the nearest neighbors of the verb-argument pair $(sentence, judge)$ should correspond to a high rating for the role, and implies low relative frequencies for other roles such as the patient role, which means the patient role should receive a low rating. For $(sentence, criminal)$ this works in exactly the opposite way.

Solely relying on the the relative semantic role frequencies in the nearest neighbor set might not always work, though, since it implies that plausibility ratings for different roles are always completely dependent on and therefore perfectly predictable from each other: high plausibility for a certain semantic role given a verb-argument pair always implies low plausibility for the other roles in the nearest neighbor set, and low plausibility for one semantic role invariably means higher plausibility for the other ones. However, nouns can also be more or less equally good as agents and patients for a given verb—one is hopefully as likely to be helped by a friend as to help a friend oneself—or equally bad—houses only kill in horror movies, and 'to kill a house' can only be made sense of in a metaphorical way. Therefore, we also implement a model that takes distance into account for its plausibility ratings. The basic idea is that a semantic role will receive a lower rating as the nearest neighbors supporting that role become less similar to the target item. The plausibility rating for a semantic role given a verb-argument pair in this model is a joint function of:

1. the frequency with which the role occurs in the set of memory exemplars that are semantically most similar to the target pair
2. the target pairs similarity to those exemplars

We will call this model the Distance Decay model (henceforth DD).

Formally, the preference strength (PS) for a semantic role J (R_J) given a verb-argument tuple i (S_i) is found by summing the distance-weighted frequency of all exemplars with semantic role J in the nearest neighbor set (C_J^k):

$$PS(R_J|S_i) = \sum_{j \in C_J^k} f(j) \times \eta_j \quad (5)$$

The weight of an exemplar j (η_j) is given by an exponential decay function, taken from Shepard (1987), over the distance between that exemplar and the target exemplar i (d_{ij}):

$$\eta_j = \mathrm{e}^{-\alpha \times d_{ij}} \quad (6)$$

In Equation 6, the free parameter α determines the rate of decay over d_{ij}. Higher values of α result in a faster drop in similarity as d_{ij} increases.

3.3 A normalized distance decay model

In Equation 5, we do not include a denominator that sums over the similarity strengths of all exemplars in the nearest neighbor set, because we want to keep the absolute effect of distance into the formula, so as to be able to accurately predict the bad fit of both the agent and patient roles for verb-argument pairs like $(kill, house)$ or the good fit of both agent and patient roles for a pair like $(help, friend)$. To find out whether a non-normalized model is indeed a better predictor of thematic fit than a normalized model, we also run experiments with a normalized version of the model presented in Section 3.2:

$$PS(R_J|T_i) = \frac{\sum_{j \in C_J^k} f(j) \times \eta_j}{\sum_{l \in C^k} f(l) \times \eta_l} \qquad (7)$$

Someone familiar with the literature on human categorization behavior might recognize Equation 7; this model is actually simply a Generalized Context Model (GCM) (Nosofsky, 1986), with the 'context' being restricted to the k nearest neighbors of the target item. Therefore, we will refer to this model using the shorthand kGCM.

4 Evaluation

4.1 The task: predicting human plausibility judgments

The model is evaluated by comparing its predictions to thematic fit or semantic role plausibility judgments from two rating experiments with human subjects. In these tasks, participants had to rate the plausibility of verb-role-argument triples on a scale from 1 to 7. They were asked questions like *How common is it for a judge to sentence someone?*, in which *judge* is the agent, or *How common is it for a judge to be sentenced?*, in which *judge* is the patient. The prediction is that model preference strengths of semantic roles given specific verb-argument pairs should correlate positively with participant ratings for the corresponding verb-role-argument triples.

4.2 Training the model

In exemplar-based models, training the model simply amounts to storing exemplars in memory. Our model uses an exemplar memory that consists of 133566 verb-role-noun triples extracted from the Wall Street Journal and Brown parts of the Penn Treebank (Marcus et al., 1993). These were first annotated with semantic roles using a state-of-the-art semantic role labeling system (Koomen et al., 2005).

Semantic roles are approximated by PropBank argument roles (Palmer et al., 2005). These consist of a limited set of numbered roles that are used for all verbs but are defined on a verb-by-verb basis. This contrasts with FrameNet roles, which are sense-specific. Hence PropBank roles provide a shallower level of semantic role annotation. They also do not refer consistently to the same semantic roles over different verbs, although the A0 and A1 roles in the majority of cases do correspond to the agent and patient roles, respectively. The A2 role refers to a third participant involved in the event, but the label can stand for several types of semantic roles, such as *beneficiary* or *recipient*. To create the exemplar memory, all lemmatized verb-noun-role triples that contained the A0, A1, or A2 roles were extracted.

4.3 Testing the model

To obtain the semantic distances between nouns and verbs for the calculation of the distance between exemplars (see Equation 3), we make use of a thesaurus compiled by Lin (1998), which lists the 200 nearest neighbors for a large number of English noun and verb lemmas, together with their similarity values. This resource was created by computing the similarity between word dependency vectors that are composed of frequency counts of $(head, relation, dependent)$ triples (dependency triples) in a 64-million word parsed corpus. To compute these similarities, an information-theoretic similarity metric was used. The basic idea of this metric is that the similarity between two words is the amount of information contained in the commonality between the two words, i.e. the frequency counts of the dependency triples that occur in the descriptions of both words, divided by the amount of information in the descriptions of the words, i.e. the frequency counts of the dependency triples that occur in either of the two words. See Lin (1998) for details. These similarity values were transformed into distances by subtracting them from the maximum similarity value 1.

Gain Ratio is used to determine the weights of

the nouns and verbs in the distance calculation. Gain Ratio is a normalization of Information Gain, an information-theoretic measure that quantifies how informative a feature is in the prediction of a class label; in this case how informative in general nouns or verbs are when one has to predict a semantic role. Based on our exemplar memory, the Gain Ratio values and so the feature weights are 0.0402 for the verbs, and 0.0333 for the nouns.

The model predictions are evaluated against two data sets of human semantic role plausibility ratings for verb-role-noun triples (McRae et al., 1998; Padó et al., 2006). These data sets were chosen because they are the same data sets that were originally used in the evaluation of the two other models discussed in sections 2.1 and 2.2.

The first data set, from McRae et al. (1998), consists of semantic role plausibility ratings for 40 verbs, each coupled with both a good agent and a good patient, which were presented to the raters in both roles. This means there are $40 \times 2 \times 2 = 160$ items in total. We divide this data set in the same 60-item development and 100-item test sets that were used by Padó et al. (2006) and Padó et al. (2007) for the evaluation of their models.

For most of the McRae items, being a good agent for a given verb also entails being a bad patient for that same verb, and the other way around. This leads us to predict that on this data set the kNNf model (see section 3.1) and the kGCM (see section 3.3) should perform no worse than the DD model (see section 3.2).

The second data set is taken from Padó et al. (2006) and consists of 414 verb-role-noun triples. Agent and patient ratings are more evenly distributed, so we predict that a model that exclusively relies on the relative role frequencies in the nearest neighbor sets of these items might not capture as much variability as a model that takes distance into account to weight the exemplars. Therefore, we expect the DD model to do better than the kNNf model on this data set. We randomly divide the data set in a 276-item development set, and a 138-items test set.

Because of the non-normal distribution of the test data, we use Spearman's rank correlation test to measure the correlation strength between the plausibility ratings predicted by the model and the human ratings. To estimate whether the strength with which the predictions of the different models correlate with the human judgments differs significantly between the models, we use an approximate test statistic described in Raghunathan (2003). This test statistic is robust for sample size differences, which is necessary in this case given the fact that the models differ in their coverage. We will refer to this statistic as the Q-statistic.

Experiments on the development sets are run to find optimal values per model for two parameters: k, the number of nearest neighbors that are taken into account for the construction of the nearest neighbor set, and α (for the DD and kGCM models), the rate of decay over distance (see Equation 6).

4.4 Results

4.4.1 McRae data

Results on the McRae test set are summarized in Table 2. The first three rows contain the results for the exemplar-based models. The last two rows show the results of the two previous models for comparison. The values for k and α that were found to be optimal in the experiments on the development set are specified where applicable.

The predictions of all three exemplar-based models correlate significantly with the human ratings, with the DD model doing somewhat better than the kNNf model and the kGCM model, although these differences are not significant ($Q(0.28) = 0.134, p = 2.8 \times 10^{-1}$ and $Q(0.28) = 0.116, p = 2.9 \times 10^{-1}$, respectively). Coverage of the exemplar-based models is very high.

When we compare the results of the exemplar-based models with those of the Padó models, we find that the predictions of the DD model correlate significantly stronger with the human ratings than the predictions of the Padó et al. (2007) model, $Q(0.98) = 4.398$, $p = 3.5 \times 10^{-2}$. The DD model also matches the high performance of the Padó et al. (2006) model. Actually, the correlation strength of the DD predictions with the human ratings is higher, but that difference is not significant, $Q(0.93) = 0.285, p = 5.6 \times 10^{-1}$. However, the DD model has a much higher coverage than the model of Padó et al. (2006), $\chi^2(1, N = 100) = 44.5, p = 2.5 \times 10^{-11}$.

4.4.2 Padó data

Table 3 summarizes the results for the Padó data set. We find that the predictions of all three exemplar-based models correlate significantly with the human ratings, and that there are

Model	k	α	Coverage	ρ	p
kNNf	9	-	96%	.407	$p = 3.9 \times 10^{-5}$
DD	11	5	96%	.488	$p = 4.6 \times 10^{-7}$
kGCM	9	21	96%	.397	$p = 6.2 \times 10^{-5}$
Padó et al. (2006)	-	-	56%	.415	$p = 1.5 \times 10^{-3}$
Padó et al. (2007)	-	-	91%	.218	$p = 3.8 \times 10^{-2}$

Table 2: Results for the McRae data.

Model	k	α	Coverage	ρ	p
kNNf	12	-	97%	.521	$p = 1.1 \times 10^{-10}$
DD	8	21	97%	.523	$p = 9.1 \times 10^{-11}$
kGCM	10	25	97%	.512	$p = 2.7 \times 10^{-10}$
Padó et al. (2006)	-	-	96%	.514	$p = 2.9 \times 10^{-10}$
Padó et al. (2007)	-	-	98%	.506	$p = 3.7 \times 10^{-10}$

Table 3: Results for the Padó data.

no significant differences between the three model instantiations. Coverage is again very high.

There are no significant performance differences between the exemplar-based models and the Padó models. Correlation strengths and coverage are more or less the same for all models.

4.5 Discussion

In general, we find that our exemplar-based, semantic role predicting approach attains a very good fit with the human semantic role plausibility ratings from both the McRae and the Padó data set. Moreover, because of the fact that generalization is determined by similarity-based extrapolation from verb-noun pairs, the high correlations of the model's predictions with the human ratings are accompanied by a very high coverage.

As concerns the comparison with the models of Padó et al. (2006) and Padó et al. (2007) on the Padó data, we can be brief: the exemplar-based models' performance matches that of the Padó models, and basically all models perform equally well, both on correlation strength and coverage.

However, there is a striking discrepancy between the performance of the Padó models and the DD model on the McRae data sets. We find that the DD model performs well for both correlation strength *and* coverage, as opposed to the Padó models, both of which score less well on one or the other of these two dimensions. Although the model of Padó et al. (2006) attains a good fit on the McRae data, its coverage is very low. This is especially problematic considering the fact that it is exactly this type of test items that is used in the kind of sentence comprehension experiments for which these thematic fit models should help explain the results. The model of Padó et al. (2007) succeeds in boosting coverage, but at the expense of correlation strength, which is reduced to approximately half the correlation strength attained by the Padó et al. (2006) model.

The model of Padó et al. (2006) requires the test verbs and their senses to be attested in the FrameNet corpus to be able to make its predictions. However, only 64 of the 100 test items in the McRae data set contain verbs that are attested in the FrameNet corpus, 8 of which involve an unattested verb sense. On the other hand, the only requirement for the exemplar-based model to be able to make its predictions is that the similarities between the verbs and the nouns in the target exemplars and the memory exemplars can be computed. In our case, this means that the verbs and nouns need to have entries in the thesaurus we use (see Section 4.3). In the McRae data set, this is the case for all verbs, and for 48 out of the 50 nouns. This explains the large difference in coverage between the DD model and the model of Padó et al. (2006).

Padó et al. (2007) attribute the poorer correla-

tion of their 2007 model with the human ratings in the McRae data set to the much lower frequencies of the nouns in that data set as compared to the frequencies of the nouns in the Padó data set. That is probably also the explanation for the difference in correlation strength between our model and the model of Padó et al. (2007). Both models use similarity-based smoothing to compensate for low-frequency target items, but the generalization problem caused by low frequency nouns is alleviated in our model by the fact that the model not only generalizes over nouns, but also over verbs. Since the model can base its generalizations on verb-noun pairs that contain the noun of the target pair coupled to a verb that is different from the verb in the target pair, the neighbor set that it generalizes from can contain a larger number of exemplars with nouns that are identical to the noun in the target pair. The model of Padó et al. (2007) has no access to nouns that are not coupled to the target verb in the training corpus.

In Section 3, we predicted that the kNNf and the kGCM should perform equally well as the DD model on the McRae data set, because of the balanced nature of that data set (all nouns are either good agents and bad patients, or the other way around), but that the DD model should do better on the less balanced Padó data set. This prediction is not borne out by the results, since the DD model does not perform significantly better on either of the data sets, although on both data sets it achieves the highest correlation strength of all three models. However, what we see is that the performance difference between the DD model on the one hand and the kNNf model and kGCM on the other hand is larger on the McRae data than on the Padó data, which is exactly the opposite of what we predicted. The fact that the differences are not significant makes us hesitant to draw any conclusions from this finding, though.

5 Conclusion

We presented an exemplar-based model of thematic fit that is founded on the idea that semantic role plausibility can be predicted by similarity-based generalization over verb-argument pairs. In contrast to previous models, this model does not implement semantic role plausibility as 'fit with verb selectional preferences', but directly captures the semantic role ambiguity problem comprehenders have to solve when confronted with sentences that contain structural ambiguities like the MV/RR ambiguity, namely deciding which semantic role a noun has in the event denoted by the verb. Therefore, the model should be easily extensible towards a complete model of any sentence-level ambiguity that revolves around a semantic role ambiguity.

We have shown that our model can account very well for human semantic role plausibility judgments, attaining both high correlations with human ratings and high coverage overall, and improving on two state-of-the-art models, the performance of which deteriorates when there is a small overlap between the verbs in the training corpus and in the test data, or when the test nouns have low frequencies in the training corpus. We suggest that this improvement is due to the fact that our model applies similarity-based smoothing over both nouns and verbs. Generally, one can say that the exemplar-based model's architecture makes it very robust for data sparsity.

We also found that a non-normalized version of our model that takes distance into account to weight the memory exemplars seems to perform somewhat better than a simple nearest neighbor model or a normalized distance decay model. However, these performance differences are not statistically significant, and we did not find the predicted advantage of the non-normalized distance decay model on the Padó data set.

In future work, we will test our claim of straightforward extensibility of the model by indeed extending our model to account for reading time patterns in the online processing of sentences exemplifying temporary semantic role ambiguities, more specifically the MV/RR ambiguity. Another avenue for future research is to see how our approach to thematic fit can be used to augment existing semantic role labeling systems.

Acknowledgments

This work was supported by a grant from the Research Foundation – Flanders (FWO). We are grateful to Ken McRae and Ulrike Padó for making their datasets available, Dekang Lin for the thesaurus, and the people of the Cognitive Computation Group at UIUC for their SRL system.

References

Thomas M. Cover and Peter E. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions*

on Information Theory, 13(1):21–27.

Walter Daelemans and Antal van den Bosch. 2005. *Memory-based language processing*. Cambridge University Press, Cambridge.

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

Evelyn Fix and Joseph L. Hodges. 1951. Discriminatory analysis—nonparametric discrimination: consistency properties. Technical Report Project 21-49-004, Report No. 4, USAF School of Aviation Medicine, Randolp Field, TX.

Lyn Frazier. 1987. Sentence processing: A tutorial review. In Max Coltheart, editor, *Attention and Performance XII: The Psychology of Reading*, pages 559–586. Erlbaum, Hillsdale, NJ.

Peter Koomen, Vasin Punyakanok, Dan Roth, and Wen-tau Yih. 2005. Generalized inference with multiple semantic role labeling systems. In Ido Dagan and Daniel Gildea, editors, *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 181–184. Association for Computational Linguistics, Morristown, NJ.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In Christian Boitet and Pete Whitelock, editors, *Proceedings of the 17th International Conference on Computational Linguistics*, pages 768–774. Association for Computational Linguistics, Morristown, NJ.

Maryellen C. MacDonald and Mark S. Seidenberg. 2006. Constraint satisfaction accounts of lexical and sentence comprehension. In Matthew J. Traxler and Morton A. Gernsbacher, editors, *Handbook of Psycholinguistics (Second Edition)*, pages 581–611. Academic Press, London.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Ken McRae, Michael J. Spivey-Knowlton, and Michael K. Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Robert M. Nosofsky. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology-General*, 115(1):39–57.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Ulrike Padó, Frank Keller, and Matthew Crocker. 2006. Combining syntax and thematic fit in a probabilistic model of sentence processing. In Ron Sun and Naomi Miyake, editors, *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 657–662. Cognitive Science Society, Austin, TX.

Sebastian Padó, Ulrike Padó, and Katrin Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In Jason Eisner, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 400–409. Association for Computational Linguistics, Morristown, NJ.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Trivellore Raghunathan. 2003. An approximate test for homogeneity of correlated correlation coefficients. *Quality and Quantity*, 4(1):99–110.

Philip Resnik. 1996. Selectional constraints: an information-theoretic model and its computational realization. *Cognition*, 61(1-2):127–159.

Roger N. Shepard. 1987. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323.

Jakub Zavrel and Walter Daelemans. 1997. Memory-based learning: Using similarity for smoothing. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 436–443. Association for Computational Linguistics, Morristown, NJ.