Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering

Clayton Greenberg, Asad Sayeed and Vera Demberg
Computational Linguistics and Phonetics / M^{2}CI Cluster of Excellence
Saarland University
66123 Saarbrücken, Germany
{claytong,asayeed,vera}@coli.uni-saarland.de

Abstract

Most recent unsupervised methods in vector space semantics for assessing thematic fit (e.g. Erk, 2007; Baroni and Lenci, 2010; Sayeed and Demberg, 2014) create prototypical role-fillers without performing word sense disambiguation. This leads to a kind of sparsity problem: candidate role-fillers for different senses of the verb end up being measured by the same "yardstick", the single prototypical role-filler.

In this work, we use three different feature spaces to construct robust unsupervised models of distributional semantics. We show that correlation with human judgements on thematic fit estimates can be improved consistently by clustering typical role-fillers and then calculating similarities of candidate role-fillers with these cluster centroids. The suggested methods can be used in any vector space model that constructs a prototype vector from a non-trivial set of typical vectors.

1 Introduction

Thematic fit estimations can be quite useful for many NLP applications and also for cognitive models of human language processing difficulty, since human processing difficulty is highly sensitive to semantic plausibilities (Ehrlich and Rayner, 1981). For example, we expect that after the word *mash*, *banana* would be easier to process because it fits well as the patient, or direct object, of *mash*, but *milk* would be harder to process because it does not fit well.

A common method for estimating the thematic fit between a verb and a proposed role filler involves computing a centroid, or vector average, over the most typical role fillers for that verb, and then calculating the cosine similarity between this centroid and the proposed role filler (Baroni and Lenci, 2010; Blacoe and Lapata, 2012; Erk, 2012). For instance, we use the cosine of the angle between the *banana* vector and a vector average of the 20 nouns that, according to training data, are most likely to be mashed as a score for how well *banana* fits as the patient of *mash*. Hopefully, the *banana* vector will be closer to the centroid than *milk*, so *banana* will have a higher cosine similarity to the centroid, and thus a higher thematic fit score, than *milk*.

This conceptualization assumes that the most typical fillers for a verb-role will all be variants of a single prototype, i.e. distributionally similar to each other. However, such an assumption may not be true for ambiguous verbs. A verb with many different senses may have typical fillers for each sense, which fit relatively equally well, but are distributionally very different from one another. This means that the calculated prototypical filler will be a mixture of the arguments that are typical role fillers for the main senses of the verb. For example, consider the verb *serve*, for which the 24 most typical prepositional arguments related via the preposition *with* fall into three different senses, as illustrated in Figure 1.

Supposing that the centroid occupies a part of the vector space between two typical role fillers, but is relatively far from any one of the typical role fillers from which it was composed, as in Figure 1, none of the original typical role fillers will achieve high the-

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 21–31,
Denver, Colorado, May 31 – June 5, 2015.

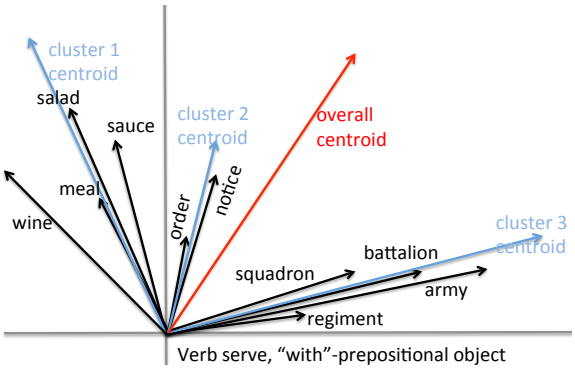

Figure 1: Illustration of *TypeDM* centroid for with-PP arguments of the verb *serve*.

matic fit scores. Also, verbs will be "penalized" for having many senses in that it will seem as though no role filler fits as well as they do with unambiguous verbs. This may produce inconsistent judgements when comparing one verb that is highly polysemous with a second, more restrictive verb whose meaning overlaps with the most dominant meanings of the first verb. For example, *cut* can be used in the sense of "cutting costs," which carries with it restrictions on instruments, locations, and so on that somewhat overlap with *eliminate* as in "eliminating costs." Things that are plausible to be eliminated are also plausible to be cut. But *cut* is also used in the sense of "cutting a cake" or "cutting (editing) a film." Without taking word sense into account, *costs* would be judged by the model as being less appropriate as a patient of *cut* than it should, and also its score for filling the patient role of *eliminate* would be infelicitously higher than its score for filling the patient role of *cut*.

One possible solution to this problem would be to do full word sense disambiguation on the resources from which these vector spaces are constructed. Then, there would be separate entries in the space for each meaning. This would however increase the overall size of the vector space by a significant factor and also cause an additional burden on corpus construction and annotation, even if automatic.

In this paper, we will approach the verb-role sense problem by clustering the most typical role-filler vectors and calculating the maximal cosine similarity for a candidate role filler with respect to each cluster prototype vector. So, to estimate the thematic fit of *salad* as an item with which something is served, in the vector space represented by Figure 1, we would use the cosine similarity with the nearest cluster centroid, the cluster 1 centroid. For a thematic fit task, the correlation between calculated estimates and human judgements can be expected to improve. In particular, good role fillers that are very different from one another and belong to different senses of a verb can all be assigned thematic fit scores as high as those of good role fillers of monosemous verbs.

We will evaluate our system using three distributional spaces: *TypeDM* (Baroni and Lenci, 2010), which is based on a syntactic dependency parser, *SDDM* (Sayeed and Demberg, 2014), which uses features obtained from the semantic role labeller SENNA (Collobert et al., 2011), and *SDDMX*, a novel extension of *SDDM*. This way, we can draw conclusions about feature space-specific and feature space-general trends.

The effects of clustering and choice of distributional space will be evaluated against the Padó (2007) and McRae et al. (1998) datasets of human judgements on thematic fit of agent and patient roles, and the Ferretti et al. (2001) datasets of human judgements on thematic fit of instrument and location roles. These different roles are conceptually interesting to compare, as instruments tend to be more strongly constrained by verbs than locations.

2 Background and related work

2.1 Thematic fit

The fit of a filler of a thematic role can be characterized as a semantic constraint on what can fill potentially available syntactic slots for a given predicate. For example, not every noun can satisfy the agent or patient roles of the typically transitive verb *eat*. There must be a valid "eater" for the agent and a valid "eatee" for the patient. Some nouns are simply more plausible than others in these positions: *lunch* is eaten, but rarely ever eats. But there can also be optional role assignments: there are certain utensils with which one is more or less likely to eat (i.e., appropriate instrument role-fillers) and even places where one is more or less likely to eat (i.e., location roles).

Verb	Noun	Semantic role	Score
advise	doctor	agent	6.8
advise	doctor	patient	4.0
confuse	baby	agent	3.7
confuse	baby	patient	6.0
eat	lunch	agent	1.1
eat	lunch	patient	6.9
kill	lion	agent	2.7
kill	lion	patient	4.9
kill	man	agent	3.4
kill	man	patient	5.4

Table 1: Sample of judgements from Padó (2007).

In order to model thematic roles, we use the insight that thematic fit correlates with human plausibility judgements (Padó et al., 2009; Vandekerckhove et al., 2009). Therefore, we can use datasets of human plausibility judgements to evaluate computational thematic fit estimates. One such dataset by Padó (2007) includes 18 verbs with up to 12 candidate nominal arguments and totals 414 verb-noun-role triples. The words were chosen based on their frequencies in the Penn Treebank and FrameNet. Human participants were asked to rate the appropriateness of given nouns as agents and as patients for given verbs on a scale from 1 to 7. The judgements were then averaged. We provide a small sample of these judgements in Table 1.

We use three other datasets as well. Ferretti et al. (2001) provide two datasets, one with 248 verb-instrument pairs and one with 274 verb-location pairs. Additionally, McRae et al. (1998) give a dataset of 1444 more agent/patient judgements. We write agent/patient as such because like Padó (2007), the agent plausibility and patient plausibility are given in the same dataset, albeit separately. Once again, human participants were asked to rate the appropriateness of given nouns as locations, instruments, and agents/patients, respectively, of the verbs in each dataset on a scale from 1 to 7. We will make use of these in our evaluation in order to see how well the models and algorithms we propose apply to various thematic roles, not just the most commonly tested and to-date most accurately estimated roles of agent and patient.

2.2 Distributional Semantics

2.2.1 Distributional Memory

Our semantic modeling technique comes from Baroni and Lenci (2010), who developed an explicitly multifunctional, i.e. not tightly bound to a particular task, framework for recording distributional information about linguistic co-occurrence. Distributional Memory (DM) records frequency information about links between words in a sentence as a third order tensor, in which words or lemmata are represented as two of the tensor axes and the syntactic or semantic link between them is the third axis.

The following corpora were used to construct the Baroni and Lenci (2010) version of DM:

- ukWaC, a corpus of about two billion words collected by crawling the `.uk` web domain (Ferraresi et al., 2008).
- WackyPedia, a snapshot selection of Wikipedia articles.
- The British National Corpus (BNC), a 100-million word corpus including documents such as books and periodicals.

The sentences from these sources were first run through MaltParser (Nivre et al., 2007). The dependency links (e.g. `SBJ`, `NMOD`) were run through a set of hand-crafted patterns to identify higher-level lexicalized links (e.g. `as-long-as`, `in-a-kind-of`). They then counted link type frequencies, so that links that involve the same lexical item (e.g. `long`, `kind`, as in the lexicalized links just mentioned) were collapsed into a single link, and the number of surface form realizations was used as the frequency count. All words were lemmatized and stored with basic part of speech information.

All these counts were then adjusted by Local Mutual Information (Baroni and Lenci, 2010), which is given by

$$LMI(i,j,k) = O_{ijk} \log \frac{O_{ijk}}{E_{ijk}} \qquad (1)$$

where i, j are words, k is the link between them, O is the observed frequency, and E is the expected frequency under independence. Tuples with nonpositive LMI values were removed. They called this tensor $TypeDM$.

2.2.2 DM Based on Semantic Role Labels

In order to create a competitor to the much less manually pruned cousin of $TypeDM$ named DepDM, Sayeed and Demberg (2014) based $SDDM$ (short for SENNA-DepDM) on similar corpora but used alternative features. Namely, this tensor was built from ukWaC and BNC, but the features came from a semantic role labelling (SRL) system called SENNA (Collobert and Weston, 2007; Collobert et al., 2011). SENNA uses a multi-layer neural network architecture that learns in a sliding window over token sequences working on raw text instead of syntactic parses, as other semantic role labellers do (Bohnet, 2010). SENNA extracts word features related to identity, capitalization, and suffix/tense (approximated by the last two characters of the word). From these features, in a process similar to decoding a conditional random field, the network derives features related to verb position, part of speech, and chunk membership.

SENNA was trained on PropBank and large amounts of unlabelled data. It achieves a role labelling F-score of 75.49% (in this case, tested on CoNNL 2005 data), which is slightly lower than state of the art SRL systems which use parse trees as input.

$SDDM$ was built by running the sentences from the input corpora through SENNA and using the role labels as links between predicates and role-fillers. Unlike $TypeDM$, $SDDM$ required almost no further processing; the raw frequency counts of triples were used in the LMI calculation.

In this paper, we present $SDDMX$, an extended version of the $SDDM$ model[1]. $SDDMX$ contains the same links as $SDDM$ and also contains links between nouns that belong to the same predicate instance, using the predicate as a link label. For example, supposing that during training the system encountered *the man eats a donut* with a role link between *man* and *eat* and another role link between *donut* and *eat*, then in $SDDMX$, a link was created between *man* and *donut*. This link was labelled with the verb lemma for the 400 most frequent verbs (`eat` in our example), and `vb` otherwise.

Sayeed and Demberg (2014) found that although

[1]We provide $SDDM$ and $SDDMX$ at `http://rollen.mmci.uni-saarland.de/`.

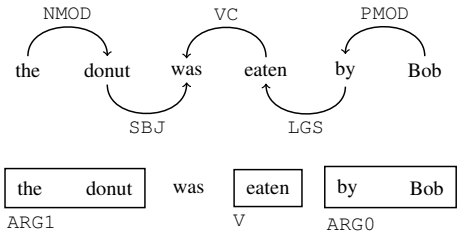

Figure 2: The same sentence with MaltParser (above) and SENNA (below) labels. Sayeed and Demberg (2014) used a simplified approach similar to the head percolation table of Magerman (1994) to find head nouns from SENNA annotation.

$SDDM$ is an arguably simpler DM model than $TypeDM$, it performs nearly as well as $TypeDM$ on a thematic fit estimation task using the Padó (2007) and McRae et al. (1998) agent/patient datasets. They also found that averaging the thematic fit scores of $SDDM$ with those of $TypeDM$ outperforms $TypeDM$ alone and nearly reaches the performance of a supervised model (Herdağdelen and Baroni, 2009). This suggests that $TypeDM$ and $SDDM$ cover different aspects of the corpora on which they were trained. Links generated by SENNA may directly access semantic role features that the MaltParser-based $TypeDM$ must infer through hand-crafted rules, such as tagging the subject as a patient instead of an agent in passive-voice contexts. Figure 2 illustrates the differences between the labelling approaches.

We make use of the $SDDM$, $SDDMX$, and $TypeDM$ tensors in our experiments to demonstrate how our techniques improve performance in thematic fit modelling across different feature spaces.

2.2.3 Centroid-based thematic fit calculation in DM

Investigating alternative ways to calculate thematic fit over the DM framework is a major goal of this work, so we now describe the baseline process.

Baroni and Lenci (2010) used the following approach to estimate thematic fit on the Padó (2007) agent/patient dataset: To assess the fit of a noun w_1 in a role r for a verb w_2, they construct a centroid from the 20 highest-ranked fillers for r with w_2 selected by LMI, using the relevant syntactic dependency links, such as subject and object, instead of

thematic roles. To illustrate, in order to determine how well *workshop* fits as a location for *eat*, they would construct a centroid of other locations for *eat* that appear in the DM, e.g. *kitchen, restaurant, cafeteria* up to 20.

Each of these top 20 represent a "slice" of the tensor along one of the word axes. One such slice, corresponding to w_1, is a matrix of links and words to which w_1 is connected. This tensor slice is collapsed into a vector whose components are word-link pairs. This is the vector of w_1.

All 20 such vectors are added up and the sum is the centroid that represents, e.g., the typical locations of *eat*. Then a vector is constructed from the slice of the tensor corresponding to *workshop*. The thematic fit score is the cosine of the location centroid of *eat* and the vector of *workshop*.

Accessing thematic roles in *SDDM* and *SDDMX* is straightforward, as the links in these models are PropBank roles. Agent is `ARG0`, patient is `ARG1`, location is `ARGM-LOC`, and we use a combination of `ARGM-MNR`, `ARG2`, and `ARG3` to represent instruments, based on a translation of the roles used by Ferretti et al. (2001). The role mapping for *TypeDM* involves a combination of `sbj_tr` and `subj_intr` (transitive and intransitive subjects) for agents, `obj` for patients, the prepositional links `in`, `at`, and `on` for locations, and `with` for instruments.

2.3 Word Sense Disambiguation in Distributional Models

While distributional models carry important information about the relative frequencies of word usages, and perhaps even phrase usages, they often must collapse such usages into one representation. For example, suppose within the domain of cooking recipes, *serve* occurs in its food sense (see cluster 1 in Figure 1) 97% of the time. The other senses will have negligible effect on the representation of *serve* because their frequencies are so much lower. But in a web crawl, the distribution is quite likely to be more uniform, which means the senses will "split the difference" in the representation and end up not being that similar to any instance of *serve*.

Many systems work to alleviate this problem by performing manipulations on words as they occur in training corpora (e.g., Thater et al., 2011). Namely, the base vector for the potentially ambiguous word is contextualized, as in scaled element-wise, by the vectors of the neighboring words for that instance. This is quite intuitive because if *serve* and *cake* occur next to each other, the chance that a non-food sense of the word *serve* was intended would be extremely small, in fact much smaller than a corpus-wide distribution would predict. These systems have been effective at improving correlation with human judgements for a verb-object composition model, i.e. approximating a vector for *serve cake* given a vector for *serve* and a vector for *cake* (Kartsaklis et al., 2014), and also reducing noise in similarity scores for a nearest neighbor-based prepositional phrase attachment disambiguator (Greenberg, 2014).

It remains a choice of the system whether to store explicit senses separately, and relatedly, whether to consult a knowledge base for the number of senses for each word, or even for meaning representations of those senses. Using a task-general knowledge base, in addition to the inherent cost of building one, is not particularly suited for our task because the items to be disambiguated are verb-role pairs, as opposed to just verbs, and usually such knowledge bases do not handle individual thematic roles separately. For instance, it may be optimal to analyze *serve* as having three senses with respect to instruments, two senses with respect to patients, and one sense with respect to agents.

Assigning semantic categories to the slots of a verb subcategorization frame harks back to work by Resnik (1996) and Rooth et al. (1999). Resnik's work presupposes predefined noun classes obtained from WordNet. Rooth et al. induced latent role-filler classes via expectation maximization. Erk et al. (2010) found that neither are good models of thematic fit. Padó et al. (2009) provided thematic fit scores that take into account verb class using a supervised model. In the vector space context, inducing different vectors for multiple verb senses has been investigated recently by Reisinger and Mooney (2010), Huang et al. (2012), and Neelakantan et al. (2014), although these were not focused on role-fillers for verbs. Our contribution is to make use of a large-scale, unsupervised vector space model to provide thematic fit scores after inducing implicit verb sense classes relative to thematic role.

3 Methods

We begin our discussion of sense disambiguation for thematic fit with the following insight: the baseline ($Centroid$) method takes as input a set of typical role-fillers, the highest-ranked ones according to the DM, and returns a single prototype vector. However, if we allow the system to return a *set* of prototype vectors, then the framework gains the capacity to handle multiple senses of the verb-role pair.

The first choice is how to handle the output. Now instead of one cosine similarity, we would have a set of cosines corresponding to the similarities between the test role-filler and each prototype vector in the set. But if we make the theoretical assumption that each prototype corresponds to a sense, then roughly only one should apply at a time. So, we choose to use the one that is most relevant, i.e. similar, to the test role-filler. Therefore, we use the maximum of the cosine similarities as the thematic fit score.

3.1 One best or nearest

In the extreme case, we can just use the unaltered set of highly-ranked role-fillers as our set of prototypes. For example, if we query $TypeDM$ for the top four instrument-fillers of *eat*, we would retrieve *spoon, hand, bread,* and *sauce*. Then, to assign a thematic fit score for *fork* as an instrument-filler, we compute the cosine similarities of (*fork*, *spoon*), (*fork*, *hand*), (*fork*, *bread*), and (*fork*, *sauce*). The cosine similarity of (*fork*, *spoon*) is the highest, so this cosine determines the score. We refer to this method as $OneBest$. Note that $OneBest$ requires the calculation of a large number of cosines, which is a relatively expensive operation given the sparse representations of words in DM spaces.

The number of retrieved top role-fillers (n) appears to be the only parameter for $OneBest$. Yet, this method poses a few theoretical questions. First, there most likely should be an upper bound on the number of role-fillers that the system can retrieve at once. Mathematically, allowing the system to retrieve the entire relevant cross-section of the tensor would be equivalent to reducing the thematic fit evaluation task to a binary decision, i.e. whether the verb-role has occurred with the test role-filler in the training data. So, we would not be able to model any graded effect on the fit of two seen role-fillers, even if one of them fits with the verb-role better than the other. Also, psycholinguistically, it seems implausible that one must remember all of the times that one has encountered a word in order to use it. Therefore, we impose 50 as an arbitrary upper bound on n. We also set a lower bound of 10 on n because values smaller than this generated quite erratic sets of top role-fillers.

Second, $OneBest$ might return a cosine of 1.0 if the DM retrieves the test role-filler itself as one of the top role-fillers. This could unfairly help the correlation between the cosines returned by the system and human judgements because the good role-fillers would all have the same cosine value, thus reducing the effect of the cosine ratings produced for the more distant (interesting) role-fillers. Therefore, we prohibit our system from returning any cosines of 1.0. The test role-filler thus achieves a high score by having a closely related role-filler in the prototype set, not by being present itself.

3.2 Clustering

In order to reduce noise from $OneBest$, we cluster similar top role-fillers together, calculate centroids for each cluster, and use these cluster centroids as the prototype set. This way, the presence of an anomalous vector in the centroid set has less effect. We use the group average agglomerative clustering package within NLTK (Bird et al., 2009). This algorithm works by initializing each top role-filler in its own cluster and iteratively combining the two most similar clusters.

For the stopping criterion, which determines the final number of clusters for the verb-role, we use the Variance Ratio Criterion (VRC) method (Caliński and Harabasz, 1974). Let c be the baseline centroid of all top role-fillers retrieved, f be a top role-filler, and c_f be the cluster centroid of the cluster to which f is assigned. Then, this method works by (a) calculating the VRC metric for each number of clusters (k), given by

$$VRC(k) = \frac{SS_B}{k-1} / \frac{SS_W}{n-k} \tag{2}$$

where we define

$$SS_B = \sum\nolimits_f (1 - \cos^2(c_f, c)) \tag{3}$$

and

$$SS_W = \sum\nolimits_f (1 - \cos^2(f, c_f)) \quad (4)$$

and then (2) choosing the final number of clusters such that

$$\omega_k = (VRC_{k+1} - VRC_k) - (VRC_k - VRC_{k-1}) \quad (5)$$

is minimized. Intuitively, this procedure is meant to find the number of clusters for which adding another cluster does not explain significantly more variance in the data. Also, note that the VRC metric is equivalent to the F-score in a one way ANOVA.

The main drawback of the VRC method is that it cannot evaluate fewer than three clusters, due to having both a VRC_{k+1} and a VRC_{k-1} term in Equation (5). However, as long as enough top role-fillers are retrieved, it should not hurt the system. Equivalently, we set VRC_0 and VRC_1 equal to VRC_2. To examine the effect of this choice, we evaluate two clustering methods: $2Clusters$, which chooses two clusters for every verb-role, and $kClusters$, which dynamically chooses a number of clusters between 3 and 10 based on the above criterion.

Once again, the system is prohibited from returning a cosine of 1.0. This means that if the DM retrieves the test role-filler itself as one of the top role-fillers, the system would skip comparing the test role-filler against itself if it were in a singleton cluster, but would not skip it if it were a member of a cluster of size two or greater. The alternative to this would have been removing the test role-filler before clustering, but we saw these role-filler-specific partitions as a form of supervision.

3.3 Evaluation procedure

The $Centroid$, $OneBest$, $2Clusters$, and $kClusters$ methods each determine their own prototype vector set for a verb-role, and then return the maximum cosine similarity value for each test role-filler. Prototype sets are stored in a dictionary so they can be reused. It is necessary to expand the sparse data structure of each vector in order to efficiently compute all of the necessary cosine similarities. Finally, we calculate Spearman's ρ values to measure the correlations between these sets of thematic fit scores and the four datasets of human judgements.

Dataset	$SDDM(X)$	$TypeDM$
Padó (2007)	98.6	100.0
McRae et al. (1998)	96.0	95.2
Ferretti et al. (2001) inst.	94.0	93.1
Ferretti et al. (2001) loc.	99.6	98.9

Table 2: Coverage (%) by dataset for each DM model.

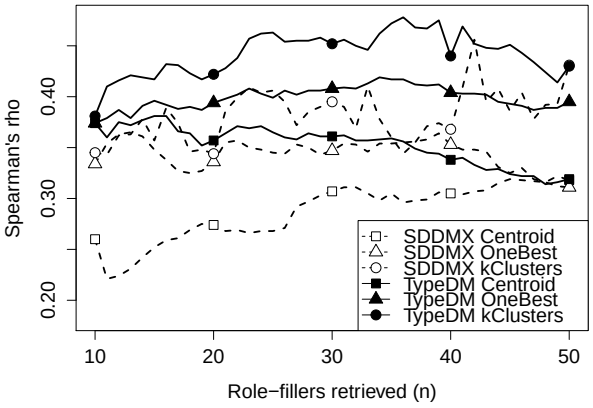

Figure 3: Spearman's ρ values for Ferretti et al. (2001) instruments vs. the number of vectors retrieved.

For our main experiment, we always retrieve the top 20 highest-ranked role-fillers for the verb-role pair to compute the prototype set. This allows our work to be more directly comparable with other implementations. Also, choosing a value of n that maximizes ρ would make this unsupervised system more supervised. However, it is useful to know how the number of top role-fillers retrieved affects the correlation with human judgements, so as a follow-up experiment, we evaluate versions of the $Centroid$, $OneBest$, and $kClusters$ methods, with the $SDDMX$ and $TypeDM$ models, retrieving from 10 to 50 top role-fillers, against the Ferretti et al. (2001) instruments dataset.

4 Results

In Table 2, we report the coverage percentages for the DM models on each of the thematic fit datasets. Note that since $SDDM$ and $SDDMX$ differ only in the additional links added between existing pairs of words, their coverages are the same.

Figure 3 shows the relationship between the number of vectors retrieved from the DM model and the correlation of the system with human judgements.

	Padó (2007) agents			**McRae et al. (1998) agents**			**Ferretti et al. (2001) instruments**		
	SDDM	*SDDMX*	*TypeDM*	*SDDM*	*SDDMX*	*TypeDM*	*SDDM*	*SDDMX*	*TypeDM*
Centroid	0.515	0.528	**0.535**	0.371	0.394	0.359	0.193	0.274	0.357
OneBest	0.321	0.324	0.464	0.375	0.376	**0.431**	0.274	0.336	0.394
2Clusters	0.489	0.412	0.522	0.367	0.373	0.370	0.252	0.331	0.388
kClusters	0.281	0.322	0.460	0.396	0.394	0.416	0.335	0.344	**0.422**
	Padó (2007) patients			**McRae et al. (1998) patients**			**Ferretti et al. (2001) locations**		
	SDDM	*SDDMX*	*TypeDM*	*SDDM*	*SDDMX*	*TypeDM*	*SDDM*	*SDDMX*	*TypeDM*
Centroid	0.511	0.505	0.525	0.133	0.131	0.343	0.187	0.248	0.230
OneBest	0.447	0.467	0.509	0.214	0.233	0.307	0.234	0.276	0.244
2Clusters	0.526	0.498	0.551	0.175	0.166	**0.353**	0.294	0.249	0.235
kClusters	0.401	0.428	**0.555**	0.212	0.227	0.350	0.293	**0.326**	0.289
	All from Padó (2007)			**All from McRae et al. (1998)**			**All datasets**		
	SDDM	*SDDMX*	*TypeDM*	*SDDM*	*SDDMX*	*TypeDM*	*SDDM*	*SDDMX*	*TypeDM*
Centroid	0.512	0.521	0.530	0.237	0.251	0.325	0.258	0.296	0.354
OneBest	0.385	0.395	0.482	0.273	0.287	0.345	0.275	0.304	0.359
2Clusters	0.508	0.458	**0.532**	0.252	0.256	0.336	0.287	0.289	0.366
kClusters	0.343	0.375	0.503	0.287	0.294	**0.359**	0.294	0.317	**0.385**

Table 3: Spearman's ρ for each method on each dataset and on all datasets together, using the 20 highest ranked words per verb-role.

The first six sections of Table 3 give the Spearman's ρ values for our four centroid set construction methods evaluated against the four datasets of human judgements, organized by thematic role, all using the 20 highest-ranked words per verb-role. We note that the ρ value for the Padó (2007) dataset using $TypeDM$ and the $Centroid$ method is slightly higher than the value reported in Baroni and Lenci (2010) due to correcting some transpositions in the original file. Finally, the last three sections of Table 3 give the performance of each method on the two whole agent/patient datasets (for comparison with previous work), as well as on all datasets merged together.

5 Discussion

While $SDDM$ and $SDDMX$ have marginally better coverage than $TypeDM$, we do not expect that this had an effect on our results. Figure 3 shows that for the various numbers of vectors retrieved from the DM models, $kClusters$ consistently outperforms $OneBest$, which consistently outperforms $Centroid$ on the Ferretti et al. (2001) instruments dataset. So, using just a single centroid that is a mixture of all possible good role-fillers for a verb leads to problems due to conflating different word meanings. But at the other extreme, we see how the ρ values for the $OneBest$ method peak (at $n = 13$ for $SDDMX$ and $n = 34$ for $TypeDM$) and then decrease instead of increasing monotonically. This is because we disallowed cosines of 1.0 and because as we increase the number of vectors retrieved, the easier it becomes to be close to one of the prototype vectors, regardless of thematic fit distinctions within the prototype set.

For the model comparison, we see that while $TypeDM$ generally performs better than $SDDMX$ on instruments, clustering reduces the gap considerably. Also $SDDMX$ outperforms $TypeDM$ for all methods on locations as shown in Table 3. This difference suggests that locations appear in sufficiently diverse syntactic configurations such that the handcrafted rules from $TypeDM$ do not work well.

From the All datasets section of Table 3, we see that both $OneBest$ and $kClusters$ improve the ρ values over the $Centroid$ baseline for all three DM models. This holds, too, for the individual instruments and locations datasets. Also, the two clustering methods perform better than $Centroid$ on Padó (2007) patients with all DM models and on McRae et al. (1998) patients with $TypeDM$. The fact that $Centroid$ performs best on Padó (2007) agents con-

firms previous analyses that have shown that the distribution of objects is more sensitive to verb sense than subjects. $kClusters$ outperforming $OneBest$ in a majority of cases suggests that clustering has successfully smoothed the top role-fillers, thus capturing sense-like patterns in the verb-roles.

As an example of the effect of the $kClusters$ method, we obtained the following top 20 instrument-fillers for the verb "eat" in 4 clusters using $TypeDM$:

- *gusto, relish*
- *family, friend*
- *chopstick, finger, fork, hand, knife, spoon*
- *appetite, bread, butter, cheese, food, meal, meat, mouth, rice, sauce*

The VRC method selected 7 to 9 clusters a little more often than 3 to 6, which is perhaps more clusters than the number of senses we could expect from a task general knowledge base. We can see from this example that the four clusters do not all correspond to separate senses, but instead, they rather nicely separate out noise from true instruments. Note that since these role-fillers came from $TypeDM$, they appeared as the object of "with," as a proxy for finding instruments. The true instruments ended up all in the third cluster, which created a cluster centroid that is less affected by noise and errors from the syntactic or semantic parse. So, the higher number of senses seems appropriate for this task and data.

We attribute the differences in results between the Padó (2007) and McRae et al. (1998) datasets to the differences in how these datasets were constructed. First, the Padó (2007) dataset contains only frequent verbs and most, but not all, of the verb-role pairs contain well-fitting and poorly-fitting role-fillers. The latter point is especially important because if the range of human judgements is small for a certain verb, then it is much more difficult to achieve a large ρ value regardless of the general performance level of the system. McRae et al. (1998), however, selected role-fillers much more automatically for their psycholinguistic study, so the data points do not necessarily reflect a typical sample of thematic role fitness decisions that occur in naturalistic language samples. So, it makes sense that the McRae et al. (1998) ρ values are systematically lower than those of Padó (2007). In fact, the Padó (2007) ρ values approach the ceiling of 0.6 as approximated by the supervised system.

Lastly, the effect of clustering was larger on instruments and locations than on agents and patients. A possible explanation is that instruments and locations are less-precisely defined thematic roles and better explained by several subclasses, i.e. clusters. In addition it could be that clustering helps to combat SRL inconsistencies.

6 Conclusions and future work

We show that clustering verb-roles into "senses" within a vector space framework achieves a higher correlation with human judgements on thematic fit over pure $Centroid$ and $OneBest$ methods. While we demonstrated this using the Distributional Memory technique by Baroni and Lenci (2010), the method will also be applicable to other vector space models.

This task has also been useful for comparing among DM models and the different thematic fit datasets. In particular, we can qualitatively evaluate how reliable syntax can be for determining the semantic notion of thematic fit, and the relative strength of human intuitions on verb-imposed restrictions on the various roles (agent, patient, instrument, and location).

In future work, we can investigate more sophisticated methods of vector clustering (such as expectation maximization and non-negative matrix factorization), interactions with verb and noun frequency, and interactions with number of word senses from a task-general knowledge-base such as WordNet. It would be especially useful to evaluate this system of a dataset of human judgements with verbs that systematically vary in polysemy, as this would more clearly expose the general trends we wish to model computationally.

Acknowledgments

This research was funded by the German Research Foundation (DFG) as part of SFB 1102: "Information Density and Linguistic Encoding." Also, the authors wish to thank the three anonymous reviewers whose valuable ideas contributed to this paper.

References

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Blacoe, W. and Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 546–556, Jeju Island, Korea. Association for Computational Linguistics.

Bohnet, B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 89–97, Stroudsburg, PA, USA. Association for Computational Linguistics.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, 3(1):1–27.

Collobert, R. and Weston, J. (2007). Fast semantic extraction using a novel neural network architecture. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 560–567, Prague, Czech Republic. Association for Computational Linguistics.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Ehrlich, S. F. and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, 20(6):641–655.

Erk, K. (2007). A simple, similarity-based model for selectional preferences. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 216–223, Prague, Czech Republic. Association for Computational Linguistics.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Erk, K., Padó, S., and Padó, U. (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Ferretti, T. R., McRae, K., and Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4):516–547.

Greenberg, C. (2014). Disambiguating prepositional phrase attachment sites with sense information captured in contextualized distributional data. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 71–77, Baltimore, Maryland, USA. Association for Computational Linguistics.

Herdağdelen, A. and Baroni, M. (2009). BagPack: A general framework to represent semantic relations. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 33–40, Athens, Greece. Association for Computational Linguistics.

Huang, E. H., Socher, R., Manning, C. D., and Ng, A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 873–882, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kartsaklis, D., Kalchbrenner, N., and Sadrzadeh, M. (2014). Resolving lexical ambiguity in tensor regression models of meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers)*, pages 212–217, Baltimore, USA. Association for Computational Linguistics.

Magerman, D. M. (1994). *Natural Lagnuage Parsing as Statistical Pattern Recognition*. PhD thesis, Stanford University.

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

Neelakantan, A., Shankar, J., Passos, A., and McCallum, A. (2014). Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069. Association for Computational Linguistics.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Padó, U. (2007). *The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing*. PhD thesis, Saarland University.

Padó, U., Crocker, M. W., and Keller, F. (2009). A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, 33(5):794–838.

Reisinger, J. and Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 109–117, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Rooth, M., Riezler, S., Prescher, D., Carroll, G., and Beil, F. (1999). Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 104–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sayeed, A. and Demberg, V. (2014). Combining unsupervised syntactic and semantic models of thematic fit. In *Proceedings of the first Italian Conference on Computational Linguistics (CLiC-it 2014)*.

Thater, S., Fürstenau, H., and Pinkal, M. (2011). Word meaning in context: A simple and effective vector model. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1134–1143, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Vandekerckhove, B., Sandra, D., and Daelemans, W. (2009). A robust and extensible exemplar-based model of thematic fit. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, pages 826–834.