Detecting Polysemy in Hard and Soft Cluster Analyses of German Preposition Vector Spaces

Sylvia Springorum, Sabine Schulte im Walde and Jason Utt
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
Pfaffenwaldring 5b, 70569 Stuttgart, Germany
{riestesa,schulte,uttjn}@ims.uni-stuttgart.de

Abstract

This paper presents a methodology to identify polysemous German prepositions by exploring their vector spatial properties. We apply two cluster evaluation metrics (the *Silhouette Value* (Kaufman and Rousseeuw, 1990) and a fuzzy version of the *V-Measure* (Rosenberg and Hirschberg, 2007)) as well as various correlations, to exploit hard vs. soft cluster analyses based on Self-Organising Maps. Our main hypothesis is that polysemous prepositions are outliers, and thus represent either (i) singletons or (ii) marginals of the clusters within a cluster analysis. Our analyses demonstrate that (a) in a subset of the clusterings, singletons have a tendency to contain polysemous prepositions; and (b) misclassification and cluster membership rate exhibit a moderate correlation with ambiguity rate.

1 Introduction

Vector space models have become a steadily increasing, integral part of data-intensive lexical semantics over the past 20 years (cf. Turney and Pantel (2010) and Erk (2012) for two recent surveys). They have been exploited in psycholinguistic (Lund and Burgess, 1996) and computational linguistic research (Schütze, 1998), to explore distributional properties of target objects and the notion of "similarity" within a geometric setting.

While individual vector space approaches have been concerned with sense discrimination, it is still largely unknown how to identify polysemous objects within a vector space model, and which geometric properties characterise the polysemous objects. For example, Schütze (1998) performed sense discrimination of ambiguous word tokens, based on their second-order co-occurrence distributions; Erk (2009) presented two variants of defining regions of word meaning in vector spaces; Erk and Padó (2010) defined a model where polysemous words activated several word vectors; Boleda et al. (2012b) compared two models of representing regular polysemy, one with multiple class assignments for multiple senses, and one incorporating classes with polysemy properties. Our work is different from all these approaches, since we aim to investigate prototypical spatial properties of polysemous objects.

More specifically, this paper is part of a larger framework that systematically explores the vector spatial properties of German prepositions, a notoriously polysemous closed word class. Relying on Self-Organising Maps (SOMs, cf. Kohonen (2001)) and preposition-dependent nouns as vector-space features, we present a methodology to identify the degree of polysemy of the prepositions. For this task, the methodology applies two cluster evaluation metrics, the *Silhouette Value* (Kaufman and Rousseeuw, 1990) and the *V-Measure* (Rosenberg and Hirschberg, 2007), to hard vs. soft cluster analyses based on the Self-Organising Maps. Since we start out with a hard clustering, a sub-task is concerned with transferring the SOM hard clusters to soft clusters. Similarly, the original V-Measure applies to hard clusters only, so a second sub-task is concerned with defining a Fuzzy V-Measure that applies to soft clusters. Our main hypothesis is that polysemous prepositions are outliers, and thus represent either (i) singletons or (ii) marginals of the clusters within a cluster analysis.

The paper is organised as follows. After introducing our preposition data in Section 2, Section 3 describes the preposition vector-space features, and the hard and soft clusterings. Section 4 is devoted to the evaluations, and Section 5 relies on the cluster analyses and the evaluations, to detect and discriminate polysemous prepositions.

International Joint Conference on Natural Language Processing, pages 632–640,
Nagoya, Japan, 14-18 October 2013.

2 Preposition Data

Although prepositions contribute a considerable portion to the meaning of texts, comparably little effort in computational semantics has gone beyond a specific choice of prepositions (such as spatial prepositions), towards a systematic classification of preposition senses. In recent years, computational research on prepositions has been enforced, mainly driven by the ACL Special Interest Group on Semantics (ACL-SIGSEM). The SIG has organised a series of workshops on prepositions, and a special issue in the *Computational Linguistics* journal (Baldwin et al., 2009).

Related work across languages includes *The Preposition Project* for English prepositions (Litkowski and Hargraves, 2005), *PrepNet* for French prepositions (Saint-Dizier, 2006), and a German project on the role of preposition senses in determiner omission in prepositional phrases (Kiss et al., 2010). The latter is most closely related to the present work, as it is also aimed at German. Their focus however is on manual classifications and corpus annotation, in contrast to our automatic classification approach.

As in many other languages, German prepositions are notoriously ambiguous, e.g. note the quite distinct senses of the German preposition *nach* in *nach drei Stunden/Berlin/Meinung* 'after three hours/to Berlin/according to', referring to a temporal, directional, and accordance meaning. Our gold standard in terms of preposition senses is the German grammar book by Helbig and Buscha (1998). Starting with their class hierarchy, we selected the classes of prepositions that contained more than one preposition. We deleted those prepositions from the classes that appeared less often than 10,000 times in our web corpus containing 880 million words (cf. Section 3.1). This selection process resulted in 12 semantic classes covering between 2 and 27 prepositions each (cf. Table 1). The included prepositions exhibit ambiguity rates of 1 (monosemous) up to 6 (cf. Table 2). Out of the 47 prepositions, 24 are polysemous (51%).

Class		Size
lokal	'local'	27
modal	'modal'	24
temporal	'temporal'	21
kausal	'causal'	5
distributiv	'distributive'	6
final	'final'	4
urheber	'creator'	3
konditional	'conditional'	3
ersatz	'replacement'	2
restriktiv	'restrictive'	2
partitiv	'partitive'	2
kopulativ	'copulative'	2

Table 1: Preposition classes.

#Senses	#Prepositions
6	1
5	3
4	3
3	11
2	6
1	23

Table 2: Degrees of preposition ambiguity.

3 Cluster Analyses

The pipeline in our framework is as follows.

1. The prepositions are associated with a distributional feature set.
2. The vector space of prepositions is hard-clustered using Self-Organising Maps.
3. The hard clustering is transferred to a soft clustering.
4. The cluster analyses are evaluated.

The following subsections describe these steps in more detail. While the larger framework plans to perform this pipeline for various cluster algorithms and many feature sets, the current setup of experiments focuses rather on the methodology towards polysemy detection, and is thus restricted to one algorithm (SOMs) and one feature set (nouns).

3.1 Preposition Corpus Features

The distributional features for the German prepositions were induced from the *sdeWaC* corpus (Faaß and Eckart, 2013), a cleaned version of the German web corpus *deWaC* created by the *WaCky* group. The corpus cleaning had focused mainly on removing duplicates from the *deWaC*, and on disregarding sentences that were syntactically ill-formed (relying on a parsability index provided by a standard dependency parser (Schiehlen, 2003)). The sdeWaC contains approx. 880 million words.

In this paper, we focus on one specific feature set that is expected to provide salient properties towards preposition meaning, i.e., the nouns that are subcategorised by the prepositions. This dependency information was extracted from a parsed version of the sdeWaC using Bohnet's MATE dependency parser (Bohnet, 2010). So each preposition was associated with a feature vector over its

subcategorised nouns. The overall set of noun features was restricted to the 10,000 nouns from the corpus which co-occurred with the largest number of prepositions.

3.2 Hard Clustering

For hard-clustering the German prepositions, we relied on the Self-Organising Maps (SOMs) artificial neural networks provided by the `kohonen` library of the *R Project for Statistical Computing*[1]. We expected SOMs to be especially useful for this task, as they create typology-preserving maps, and should thus provide a suitable model to look into the spatial properties of polysemous vectors. Furthermore, SOMs have successfully been applied to semantic classification before (Ontrup and Ritter, 2001; Kanzaki et al., 2002; Guida, 2007).

We created SOM maps with k clusters, for $2 \leq k \leq 47$, where 47 represents the total number of prepositions. For each k, we initiated two-dimensional spacings for all possible hexagonal grids. For example, we trained four SOM maps with 30 clusters, using a 30×1 grid, a 15×2 grid, a 10×3 grid, and a 6×5 grid. The distance measure used in the maps was *Euclidean Distance*, which is the only option for SOMs in *R*.

3.3 Soft Clustering

The soft clustering of the German prepositions was based on the various hard cluster analyses. We performed the *hard* $\rightarrow$ *soft* clustering transfer in two alternative ways, providing two different types of soft cluster analyses.

(1) Centroid-based softening: For each cluster c within a hard cluster analysis C, we calculated the mean distance $prep2cluster(c)$ over all prepositions p to the cluster centroid z_c, ignoring any hard assignments in the hard clustering, cf. Equation 1. The individual distances between a preposition p and a cluster centroid z_c are denoted as $d(p, z_c)$.

$$prep2cluster(c) = \frac{\sum^{p} d(p, z_c)}{|p|} \qquad (1)$$

For the corresponding soft cluster analysis $S_t(C)$ of a hard cluster analysis C, a preposition p was assigned to a cluster c if the distance $d(p, z_c)$ was below a threshold $t \times prep2cluster(c)$, with $t = 0.05, 0.1, 0.15, \ldots, 0.95$. For example, if a distance of a preposition p to a cluster c was 5, and the mean distance $prep2cluster(c)$ was 10, then p would *not* be assigned to c for $t = 0.05, 0.1 \ldots, 0.5$ but for $t = 0.6, \ldots, 0.95$. In this way, we created 19 different soft cluster analyses $S_t(C)$ for each hard clustering C, one for each t. With low values of t, few prepositions (i.e., only those that were very close to the respective cluster centroids) were assigned to the clusters, and the resulting cluster analyses were likely to contain not all of our prepositions, and a low ambiguity rate; with high values of t, more prepositions were assigned to each of the clusters, and the resulting cluster analyses were likely to contain many of the 47 prepositions, and a high ambiguity rate.

(2) Preposition-based softening: For each preposition p within a hard cluster analysis C, we calculated the mean distance $cluster2prep(p)$ over all cluster centroids z_c to the preposition p, ignoring any hard assignments in the hard clustering, cf. Equation 2. Again, the individual distances between a preposition p and a cluster centroid z_c are denoted as $d(p, z_c)$.

$$cluster2prep(p) = \frac{\sum^{c} d(p, z_c)}{|c|} \qquad (2)$$

Similarly to the centroid-based softening, for the corresponding soft cluster analysis $S_t(C)$ of a hard cluster analysis C, a preposition p was assigned to a cluster c if the distance $d(p, z_c)$ was below a threshold $t \times cluster2prep(p)$, with $t = 0.05, 0.1, 0.15, \ldots, 0.95$. By relying on the threshold, we again created 19 different soft cluster analyses $S_t(C)$ for each hard clustering C, one for each t. In this case, however, we compared the mean distances of an individual preposition to all cluster centroids, and only performed soft cluster assignments if the preposition was close to a cluster centroid in comparison to its distance to other cluster centroids. With low values of t, the prepositions were assigned to none or few clusters, and the resulting cluster analyses were likely to contain not all of our prepositions, and a low ambiguity rate; with high values of t, the prepositions were assigned to many clusters, and the resulting cluster analyses were likely to contain many of the 47 prepositions, and a high ambiguity rate.

4 Evaluation

The evaluation metrics play an important role in our work. On the one hand, we created a large number of hard clustering SOMs (i.e., 96 cluster

[1] `http://www.r-project.org/`

analyses since we took all possible grids for each $2 \leq k \leq 47$ into account), and for each hard cluster analysis we created 38 soft cluster analyses (19 centroid-based versions, and 19 preposition-based versions). We thus needed evaluation measures to decide about the quality of a cluster analysis. On the other hand, our methodology relies on evaluation metrics to identify polysemous prepositions, so the measures are crucial to perform this work.

There is a large body of research regarding the question of how to compare and evaluate two cluster analyses. For example, with respect to the specific task of semantic classification, Schulte im Walde (2003), compared a range of evaluation measures. Related work in this area partly adopted the suggested measures, and in addition relied on *Purity* or *Accuracy* (Korhonen et al., 2003; Stevenson and Joanis, 2003). In more general terms, there is an ongoing discussion about cluster comparison, mainly in the field of Machine Learning, but also elsewhere. Recent examples include Meila (2007), Rosenberg and Hirschberg (2007), and Vinh and Bailey (2010). These approaches all concentrate on evaluations relying on the entropy between two cluster analyses, in order to compare them. Entropy is an information-theoretic measure of uncertainty; in our context, entropy measures how uncertain a clustering is, given the information provided by a gold standard, and vice versa.

We decided to make use of two evaluation measures, in order to (i) evaluate and compare our hard and soft cluster analyses, and (ii) detect polysemy. The two measures were expected to provide complementary perspectives on the properties of our cluster analyses, and on the properties of ambiguous prepositions. The following paragraphs describe these measures, and how they were applied.

(1) With the ***Silhouette Value*** (Kaufman and Rousseeuw, 1990), each cluster is represented by a silhouette displaying which objects lie well within a cluster and which objects are marginal to a cluster. The evaluation appeared specifically suited to our task, as according to our hypotheses, ambiguous prepositions were expected to represent marginals in a cluster analysis, i.e., to be comparably far away from all cluster centroids.

To obtain the silhouette value sil for an object o_i within a cluster c_A, we compared the average distance a between o_i and all other objects in c_A with the average distance b between o_i and all objects in the neighbouring cluster c_B, cf. Equations 3 to 5. For each object o_i, $-1 \leq sil(o_i) \leq 1$. If $sil(o_i)$ is large, the average object distance within the cluster is smaller than the average distance to the objects in the neighbour cluster, so o_i is well classified. If $sil(o_i)$ is small, the average object distance within the cluster is larger than the average distance to the objects in the neighbour cluster, so o_i has been misclassified. The silhouette value was only calculated if cluster C_A has at least two members, i.e. if it is not a singleton.

$$a(o_i) = \frac{1}{|c_A| - 1} \sum_{o_j \in c_A, o_j \neq o_i} d(o_i, o_j) \quad (3)$$

$$b(o_i) = min_{c_B \neq c_A} \frac{1}{|c_B|} \sum_{o_j \in C_B} d(o_i, o_j) \quad (4)$$

$$sil(o_i) = \frac{b(o_i) - a(o_i)}{max\{a(o_i), b(o_i)\}} \quad (5)$$

In addition to providing information about the quality of classification of a single object, the silhouette value can be extended to evaluate the individual clusters and the entire clustering. The *average silhouette width* $sil(c)$ of a cluster c is defined as the average silhouette value for all objects within cluster c, cf. Equation 6, and the *average silhouette width for the clustering* C with k clusters $sil(C_k)$ is defined as the average silhouette value for the individual clusters, cf. Equation 7.

$$sil(c) = \frac{1}{|c|} \sum_{o_i \in c} sil(o_i) \quad (6)$$

$$sil(C_k) = \frac{1}{k} \sum_{i=1}^{k} sil(c) \quad (7)$$

(2) The ***V-Measure*** (Rosenberg and Hirschberg, 2007) is an entropy-based cluster evaluation measure. We chose this measure over other entropy-based measures (e.g., *Variance of Information (VI)* (Meila, 2007), and variants suggested by Vinh and Bailey (2010)) because the V-Measure $v(C)$ balances two desirable properties for a clustering C of a given dataset: homogeneity (hom) and completeness (com), cf. Equations 8 to 10.[2]

[2]Note that Equations 8 and 9 differ from those in Rosenberg and Hirschberg (2007) in the denominators of the *else* condition because there were typos in the definitions (personal communication with Andrew Rosenberg).

Homogeneity is similar to *purity*, and measures how well the clusters within a cluster analysis map to the classes within a gold standard. If each cluster contains only objects from one gold-standard class, then the entropy is at its minimum, $H(C|G) = 0$. This represents a maximally homogeneous clustering. *Completeness* measures how well the classes within a gold-standard map to the clusters within a cluster analysis. If each gold-standard class contains only objects from one cluster, then the entropy is at its minimum, $H(G|C)$. This represents a maximally complete clustering, because each gold-standard class is completely contained in a cluster.

$$hom(C) = 1 \; if H(C,G) = 0; \; else \; 1 - \frac{H(C|G)}{H(C,G)} \quad (8)$$

$$com(C) = 1 \; if H(G,C) = 0; \; else \; 1 - \frac{H(G|C)}{H(G,C)} \quad (9)$$

$$v(C) = \frac{2 \times hom(C) \times com(C)}{hom(C) + com(C)} \quad (10)$$

There is however a limitation to the V-Measure because it can only be applied to hard classifications which represent an $N : 1$ relationship between data points and gold-standard classes. This means a given object only belongs to a single class. In our data, this is clearly not the case due to the inherent ambiguity of the prepositions. We thus extended the V-Measure to a fuzzy version *Fuzzy V-Measure (fuzzy v)* that applies to $N : M$ classifications, where a data point can belong to any number of classes.[3]

As for the original calculation of the entropy values, we must define the joint and conditional probabilities across clusters and gold-standard classes. In Rosenberg and Hirschberg (2007), the joint probability of a cluster c and a gold-standard class g was estimated as

$$\hat{p}(c,g) = \frac{a_{cg}}{N}, \quad (11)$$

where a_{cg} is the number of prepositions shared by c and g and N is the total number of prepositions. Due to the polysemy of prepositions, we must assume that a preposition occurs in multiple classes. Calculating the probability as above would however give too much weight to highly ambiguous prepositions. Our approach is to give each preposition a total mass of 1 and then equally divide its mass across the classes of which it is a member. Thus, Equation 11 becomes:

$$\hat{p}(c,g) = \frac{\mu(c \cap g)}{M}, \quad (12)$$

where $\mu(c \cap g)$ is the total mass of the prepositions shared by c and g, and M is the total mass of the clustering. Note that M will only be equal to N if each preposition belongs to exactly as many clusters as classes.

Example: The prepositions p_1, p_3 and p_4 each belong to two classes, while preposition p_2 belongs to three classes (cf. Table 3). Assuming cluster c_1 contains p_1, and p_2, and c_2 contains p_3 and p_4, the contingency table for the clusters c_1 and c_2 is given as in Table 4. Thus, while both c_1 and c_2 each share two prepositions with the gold-standard classes g_1 and g_2 respectively, the higher ambiguity of p_2 in the first case means there is less evidence for c_1 given g_1 than c_2 given g_2, namely: $\hat{p}(c_1|g_1) = .83/2 \; < \; 1/2 = \hat{p}(c_2|g_2)$.

In addition to being applicable to ambiguous data on the side of the classes themselves, our adaptation of the V-Measure also allows for the application to soft clusterings. In this case, the data points may be present in multiple clusters and simply add their respective mass to the cells in the contingency table.

	g_1	g_2	g_3	g_4
p_1	0.5	0.5	0	0
p_2	0.33	0	0.33	0.33
p_3	0	0.5	0.5	0
p_4	0	0.5	0	0.5

Table 3: Prepositions in gold standard.

	g_1	g_2	g_3	g_4	$\sum$
c_1	0.83	0.5	0.33	0.33	$= 2$
c_2	0	1	0.5	0.5	$= 2$

Table 4: Evidence for clusters.

5 Detecting Polysemy

This section applies the evaluation measures to our cluster analyses, in order to detect polysemous prepositions, and to identify their spatial properties. Our hypothesis is that polysemous prepositions are outliers, and thus represent either (i) singletons or (ii) marginals of the clusters within a

[3]Thanks to Andrew Rosenberg for valuable discussions.

cluster analysis. We present a series of assumptions regarding this main hypothesis, and check them according to our hard and soft clusterings.

Singletons represent polysemy. Our first analysis applies to the hard cluster analyses. The assumption here is that clusters that represent singletons contain polysemous prepositions, because singletons contain objects that do not belong to any of the other clusters. Figure 1 plots the number of polysemous singletons (i.e., those singletons whose only cluster member is a polysemous preposition) against the total number of singletons, for each SOM map. The baseline is provided by 51% of the total number of singletons, as 24 out of our 47 preposition types (51%) are polysemous, so the baseline corresponds to a random assignment of preposition types to singletons.

For SOM maps with up to $k = 13$ clusters, there is maximally one singleton in the cluster analyses (except for $k = 4$ and a grid of 2×2, which contains two singletons), so it is difficult to judge about the correctness of our prediction. For $14 \leq k \leq 26$, in most cases the number of polysemous singletons clearly outperforms the baseline. For $k = 22$ with a grid of 22×1 and $k = 26$ with a grid of 13×2, the difference to the baseline is even significant (χ^2, p<0.1). For $k > 27$, the number of polysemous singletons outperforms the baseline in fewer cases than for smaller k. In sum, our prediction that singletons represent polysemy holds for a restricted subset of our SOM maps, most strongly for $22 \leq k \leq 26$.

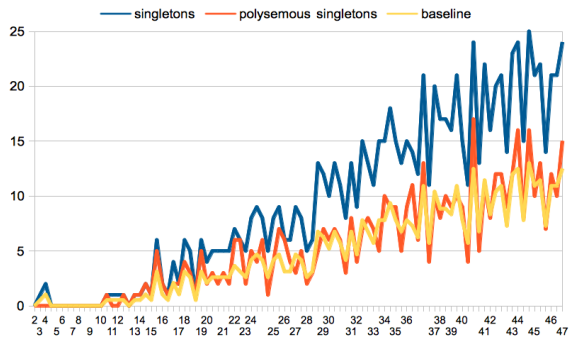

Figure 1: Number of (ambiguous) singletons.

Polysemous prepositions are misclassified. Our second analysis also applies to the hard cluster analyses. Figure 2 exploits the Silhouette Value to predict polysemous prepositions. Since prepositions with several senses are exptected to represent marginals in a cluster analysis, they should be comparably far away from all cluster centroids, and thus their silhouette value sil should be low, i.e., misclassify them. Figure 2 plots the correlation values of *Kendall's tau-b*[4] between the silhouette value $sil(p)$ and the ambiguity rate $amb(p)$ as defined by the gold standard, across all hard cluster analyses. According to our hypothesis, *tau* should be negative: the higher the ambiguity rate, the lower the silhouette value.

The plot demonstrates that our assumption is only partly correct: There are cluster analyses where we find a weak negative correlation, but most clusterings do not exhibit a noticeable correlation, and some clusterings even have a moderate positive correlation. For $k = 24$ with a grid of 24×1 and $k = 27$ with a grid of 27×1, we however find cluster analyses with a moderate negative correlation, $tau = -0.30$ and $tau = -0.32$.

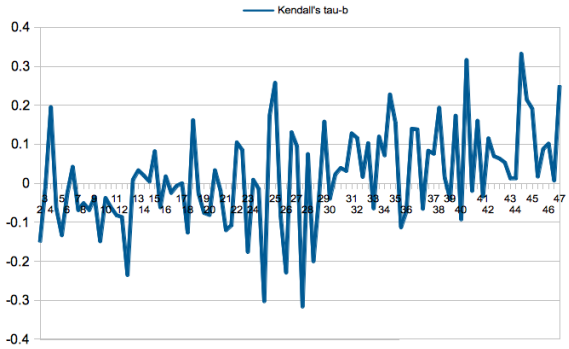

Figure 2: Correlation between $sil(p)$ and $amb(p)$.

General evaluation of soft clusterings. Before we move on to exploring a further hypothesis regarding polysemous prepositions, we present a general evaluation of our two types of softening approaches. Figures 3 and 4 plot the *homogeneity*, *completeness* and *fuzzy v* scores after applying centroid-based and preposition-based softening to k hard clusters, respectively. The soft cluster analyses depend on the threshold t that controls the assignment of prepositions to clusters. We chose $t = 0.7$ as a medium threshold for the two figures. Since the various k cause strong differences in the coverage of the preposition types in the soft cluster analyses, we also plot the *coverage*, and the harmonic mean of *fuzzy v* and *coverage*.

The best *fuzzy v* scores for the centroid-based soft clusters were obtained with $k = 16$ and a 8×2 grid (0.380), $k = 12$ with a 6×2 grid (0.379) and $k = 10$ with a 10×1 grid (0.377).

[4]Kendall's tau-b is a measure of association based on concordant and discordant pairs, adjusted for the number of ties.

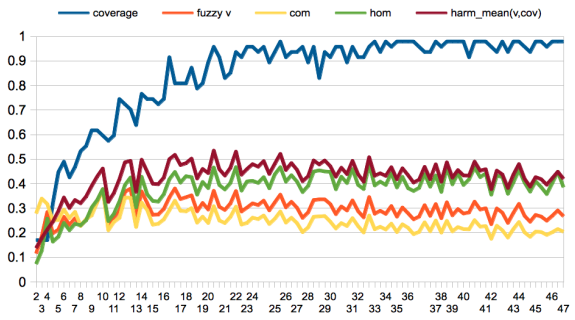

Figure 3: Centroid-based softening: evaluation.

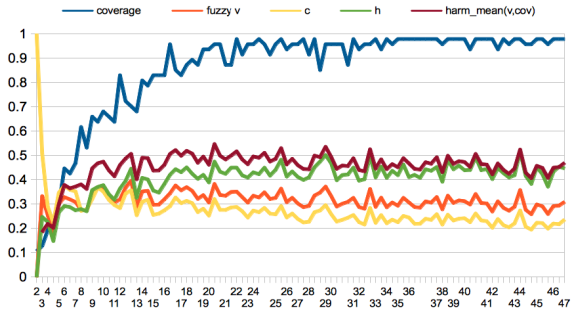

Figure 4: Preposition-based softening: evaluation.

If we take the coverage into account, the best results were obtained with $k = 20$ with a 20×1 grid (0.534), $k = 22$ with a 11×2 grid (0.530) and $k = 25$ with a 5×5 grid (0.521). For the preposition-based soft clusters the respective *fuzzy v* scores were $k = 12$ with a 6×2 grid (0.396), $k = 16$ with a 8×2 grid (0.376) and $k = 29$ with a 29×1 grid (0.372); taking coverage into account, the respective scores were $k = 20$ with a 20×1 grid (0.547), $k = 29$ with a 29×1 grid (0.536) and $k = 25$ with a 5×5 grid (0.530). In sum, the best *fuzzy v* scores for both types of soft cluster analyses were in most cases obtained for k being similar to the number of gold standard classes. Taking coverage into account, the best results were obtained for cluster analyses with $20 \leq k \leq 29$.

A threshold of $t = 0.7$ seemed appropriate for our descriptions, since lower and also higher values of t resulted in less clear preferences for k, and the threshold appeared like a useful compromise between low coverage in assigning prepositions to clusters, and highly ambiguous clusters.

Correlation of cluster membership rate with ambiguity rate. This final analysis investigates the relationship between the cluster membership rate of a preposition and its ambiguity rate. Our assumption is that the more clusters a specific preposition is assigned to, the more ambiguous it is. As basis for this analysis we used both the centroid-based and the preposition-based soft clusters, with varying t. Figures 5 and 6 present the correlation results, again relying on *Kendall's tau-b*. For presentation reasons, we restrict the plots to $10 \leq k \leq 30$ with grid shapes $k \times 1$ only, and $t = 0.6, 0.7, 0.8, 0.9$.

Both plots demonstrate that the highest threshold $t = 0.9$ corresponding to highly ambiguous cluster analyses exhibits the best correlations with the ambiguity rates of the prepositions. For the centroid-based softening, this is true for $12 \leq k \leq 20$, for the preposition-based softening, this is true for all but two values of k. For lower thresholds, it seems that $t = 0.8 > t = 0.7 > t = 0.6$, but the differences are not at all clear but rather vary depending on k. Overall, we reached moderate correlation values, the best correlation being $tau = 0.45$. Interestingly, the best correlation values in the two types of softening approaches were obtained for similar values of k, and with k being very similar to the number of gold standard classes (12): the prediction of the centroid-based softening was best with $k = 13$ and $k = 12$ ($tau = 0.453$ and $tau = 0.449$, respectively), and the prediction of the preposition-based softening was best with $k = 12$ and $k = 14$ ($tau = 0.439$ and $tau = 0.368$).

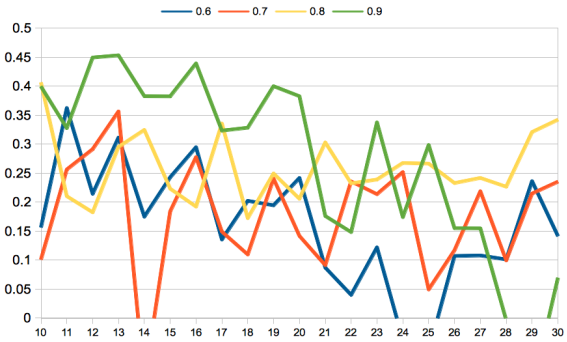

Figure 5: Centroid-based softening: ambiguity.

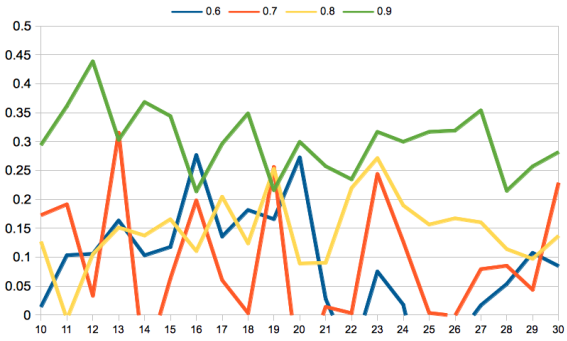

Figure 6: Preposition-based softening: ambiguity.

6 Discussion

In the previous section, we performed a series of analyses to investigate the spatial properties of polysemous prepositions in vector space models. Our main hypothesis is that polysemous prepositions are outliers, and thus represent either (i) singletons or (ii) marginals of the clusters within a cluster analysis. Concerning option (i), we showed that for specific values of k, there were significantly more polysemous prepositions in the singletons of the hard clusterings than there would be by chance. The relationship did not hold across k, however. Concerning option (ii), we performed two analyses. First, we checked whether the silhouette value of a preposition in a hard clustering correlated with its ambiguity rate, based on the assumption that the silhouette value identifies cluster marginals. Again, we found a strong correlation for specific values of k, but not across k. Second, relying on the soft clusterings we checked whether the cluster membership rate of a preposition correlated with its ambiguity rate: Especially in highly ambiguous cluster analyses there were strong correlations in both types of soft clusterings, for k similar to the number of gold standard classes.

In sum, our analyses confirmed our hypothesis, but (a) with regard to specific k only, and (b) the k varied across the analyses. This might partly be due to our clustering approaches (SOMs for hard clustering, and our two versions of softening approaches), so we are currently experimenting with alternatives. Furthermore, the *fuzzy v* measure that we developed in order to evaluate soft clusterings still seems to provide sub-optimal evidence of clustering quality: The magnitude of the score depends on the threshold, so it is difficult to decide which threshold performed best.

On the other hand, several of our analyses pointed towards similar numbers for an optimal k, and these optimal k values were reasonable, as they were close to the number of gold standard classes. Last but not least, we looked into a range of clusterings that performed well according to our *fuzzy v*, and it turned out that within a certain magnitude of k, the clusterings were very similar to each other, with similar strengths and weaknesses. We thus conclude this paper with a qualitative analysis of the centroid-based soft clustering with $k = 16$ and a 8×2 grid, the best clustering according to the general evaluation.

The clustering actually contained only 15 clusters (so one cluster was an empty cluster). Three of the clusters were singletons, one with a 3-way ambiguous preposition (*nach*: local, modal, temporal), one with a 2-way ambiguous preposition (*unter*: local, modal), and one with a monosemous preposition (*samt*: modal). From the remaining 12 clusters, 8 could unambiguously be assigned a major sense according to the gold standard classes, and 4 clusters contained prepositions from various gold standard classes.

Overall, we found 27 local preposition senses, 24 modal senses, 21 temporal senses, 5 causal and 3 replacement senses. The minor senses (according to the sizes of the gold standard classes), i.e., final, creator, distributive, partitive, conditional, copulative and restrictive, were not found in the clustering. So there was a clear bias towards the assignment of majority senses. This bias might well be due to the very different sizes of the gold standard classes, so in future work we will experiment with sub-classifications of the large classes.

7 Conclusion

In this paper, we presented a methodology to identify polysemous German prepositions by exploring their vector spatial properties in hard and soft clusterings. The analyses demonstrated that – when looking at clusterings with a similar or slightly larger number of clusters than the gold standard – (a) singletons have a tendency to contain polysemous prepositions; and (b) misclassification and cluster membership rate exhibit a moderate correlation with ambiguity rate.

Acknowledgements

The research presented in this paper was funded by the DFG Collaborative Research Centre SFB 732 (Sylvia Springorum, Jason Utt), and the DFG Heisenberg Fellowship SCHU-2580/1-1 (Sabine Schulte im Walde).

References

Timothy Baldwin, Valia Kordoni, and Aline Villavicencio, editors. 2009. *Computational Linguistics, Volume 35, Number 2, June 2009 - Special Issue on Prepositions*, volume 35. MIT Press.

Bernd Bohnet. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97, Beijing, China.

Gemma Boleda, Sebastian Padó, and Jason Utt. 2012a. Regular Polysemy: A Distributional Model. In

Proceedings of the 1st Joint Conference on Lexical and Computational Semantics, pages 151–160, Montréal, Canada.

Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. 2012b. Modelling Regular Polysemy: A Study on the Semantic Classification of Catalan Adjectives. *Computational Linguistics*, 38(3):575–616.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based Models for Word Meaning in Context. In *Proceedings of the ACL Conference Short Papers*, pages 92–97, Uppsala, Sweden.

Katrin Erk. 2009. Representing Words in Regions in Vector Space. In *Proceedings of the 13th Conference on Computational Natural Language Learning*, pages 57–65, Boulder, Colorado.

Katrin Erk. 2012. Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass*, 6(10):635–653.

Gertrud Faaß and Kerstin Eckart. 2013. SdeWaC – a Corpus of Parsable Sentences from the Web. In *Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology*, Darmstadt, Germany. To appear.

Annamaria Guida. 2007. The Representation of Verb Meaning within Lexical Semantic Memory: Evidence from Word Associations. Master's thesis, Universit degli studi di Pisa.

Gerhard Helbig and Joachim Buscha. 1998. *Deutsche Grammatik*. Langenscheidt – Verlag Enzyklopädie, 18th edition.

Kyoko Kanzaki, Qing Ma, Masaki Murata, and Hitoshi Isahara. 2002. Classification of Adjectival and Non-Adjectival Nouns based on their Semantic Behaviour by using a Self-Organizing Semantic Map. In *Proceedings of the COLING Workshop SEMANET: Building and Using Semantic Networks*, Taipei, Taiwan.

Leonard Kaufman and Peter J. Rousseeuw. 1990. *Finding Groups in Data – An Introduction to Cluster Analysis*. Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.

Tibor Kiss, Katja Keßelmeier, Antje Müller, Claudia Roch, Tobias Stadtfeld, and Jan Strunk. 2010. A Logistic Regression Model of Determiner Omission in PPs. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 561–569, Beijing, China.

Teuvo Kohonen. 2001. *Self-Organizing Maps*. Springer, Berlin, 3rd edition.

Anna Korhonen, Yuval Krymolowski, and Zvika Marx. 2003. Clustering Polysemic Subcategorization Frame Distributions Semantically. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 64–71, Sapporo, Japan.

Kenneth C. Litkowski and Orin Hargraves. 2005. The Preposition Project. In *Proceedings of the 2nd ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 171–179, Colchester, England.

Kevin Lund and Curt Burgess. 1996. Producing High-Dimensional Semantic Spaces from Lexical Co-Occurrence. *Behavior Research Methods, Instruments, and Computers*, 28(2):203–208.

Marina Meila. 2007. Comparing Clusterings - An Information-based Distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Jörg Ontrup and Helge J. Ritter. 2001. Hyperbolic Self-Organizing Maps for Semantic Navigation. *Advances in Neural Information Processing Systems*.

Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-based External Cluster Evaluation Measure. In *Proceedings of the joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 410–420, Prague, Czech Republic.

Patrick Saint-Dizier. 2006. PrepNet: a Multilingual Lexical Description of Prepositions. In *Proceedings of the 5th Conference on Language Resources and Evaluation*, pages 1021–1026, Genoa, Italy.

Michael Schiehlen. 2003. A Cascaded Finite-State Parser for German. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 163–166, Budapest, Hungary.

Sabine Schulte im Walde. 2003. *Experiments on the Automatic Induction of German Semantic Verb Classes*. Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. Published as AIMS Report 9(2).

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123. Special Issue on Word Sense Disambiguation.

Suzanne Stevenson and Eric Joanis. 2003. Semi-supervised Verb Class Discovery Using Noisy Features. In *Proceedings of the 7th Conference on Natural Language Learning*, pages 71–78, Edmonton, Canada.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Nguyen Xuan Vinh and James Bailey. 2010. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11:2837–2854.