Exploring Soft-Clustering for German (Particle) Verbs across Frequency Ranges

Moritz Wittmann
iteratec GmbH
Moritz.Wittmann@iteratec.de

Maximilian Köper
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
koepermn@ims.uni-stuttgart.de

Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
schulte@ims.uni-stuttgart.de

Abstract

In this paper we explore the role of verb frequencies and the number of clusters in soft-clustering approaches as a tool for automatic semantic classification. Relying on a large-scale setup including 4,871 base verb types and 3,173 complex verb types, and focusing on synonymy as a task-independent goal in semantic classification, we demonstrate that low-frequency German verbs are clustered significantly worse than mid- or high-frequency German verbs, and that German complex verbs are in general more difficult to cluster than German base verbs.

1 Introduction

Semantic classifications are of great interest to computational linguistics, specifically regarding the pervasive problem of data sparseness in the processing of natural language. Such classifications have been used in applications such as *word sense disambiguation* (Dorr and Jones, 1996; Kohomban and Lee, 2005; McCarthy et al., 2007), *parsing* (Carroll et al., 1998; Carroll and Fang, 2004), *machine translation* (Prescher et al., 2000; Koehn and Hoang, 2007; Weller et al., 2014), and *information extraction* (Surdeanu et al., 2003; Venturi et al., 2009), among many others.

Aiming for not only a hard assignment of word types to semantic classes but potentially distinguishing between various word senses, soft-clustering approaches have been exploited as the main tool for automatic semantic classification, e.g., Rooth et al. (1999); Schulte im Walde (2000); Korhonen et al. (2003); Iosif and Potamianos (2007); Köper and Schulte im Walde (2016). Most recently, sense-distinguishing classification approaches have also been defined for predict models by using multi-sense embeddings, e.g., Biemann (2006); Lau et al. (2012); Neelakantan et al. (2014); Li and Jurafsky (2015).

In general, clustering efforts are motivated by specific tasks or applications, so it is difficult to provide universal recommendations regarding the optimal clustering setup. This paper nevertheless addresses clustering parameters that are presumably of general importance on the meta level: Focusing on synonymy as a task-independent goal in semantic classification, we provide an extensive clustering setup to explore the role of verb frequency ranges across various numbers of clusters. The contributions of this paper are two-fold: We demonstrate that (1) low-frequency German verbs are clustered significantly worse than mid- or high-frequency German verbs, and that (2) German complex verbs are in general more difficult to cluster than German base verbs. While (1) the effect of clustering low-frequency target verbs has been investigated by a restricted number of earlier approaches, e.g. Schulte im Walde (2000); Korhonen et al. (2003); Schulte im Walde (2006); Scarton et al. (2014), (2) might be considered as general knowledge but has –as far as we are aware of– not explicitly been proven before.

2 Data and Algorithm

Using *DECOW* (Schäfer and Bildhauer, 2012; Schäfer, 2015) as one of the currently largest German web corpora, we extracted all base verbs and particle verbs from version *DECOW14*. The corpus sentences were morphologically annotated and parsed using *SMOR* (Faaß et al., 2010), *MarMoT* (Müller et al., 2013) and the MATE dependency parser (Bohnet, 2010). Relying on the morphological annotation, and after disregarding prefix verbs (i.e., non-separable complex verbs), we extracted a total of 4,871 base verb types and 3,173 particle verb types.

As vector spaces for the verbs, we relied on *word2vec* (Mikolov et al., 2013) using a symmetrical window of sizes 3 and 10. The underlying corpus was again *DECOW14*. We applied a min-frequency threshold of 50, the dimensionality was set to 400, and we used 10 corpus iterations and 15 negative samples. Other parameters were set to default.

For soft clustering, we used *Non-negative matrix factorization (NMF)*, a factorisation approach with an inherent (soft) clustering property (Ding et al., 2005). NMF has been applied successfully to other NLP tasks before, such as document clustering (Xu et al., 2003), topic number estimation (Yokoi, 2013), and preposition classification (Köper and Schulte im Walde, 2016). We applied the NMF algorithm from the *LAML* (Linear Algebra and Machine Learning) Java library, version 1.6.2 (Qian, 2016).

3 Clustering Experiments

3.1 Clustering Setup

In all clustering experiments, we clustered the German verbs using Non-negative Matrix Factorization with k-Means initialisation. We distinguished the following parameters.

- *Verb set*: We clustered (i) either the base verbs, or (ii) the particle verbs, or (iii) both base and particle verbs, to explore differences for simplex vs. complex verbs.
- *Frequency ranges*: The verbs were sorted by their corpus frequencies, and then split into three equally sized bins, to distinguish between low-, mid- and high-frequency verbs. We clustered only verbs from the same frequency range (LOW, MID, HIGH), or all verbs at the same time.
- *Verb vector spaces*: We applied two different vector spaces, relying on window sizes of 3 vs. 10.
- *Number of clusters*: We used 50, 100, and 250 clusters.
- *Number of iterations*: We let the clustering algorithm perform a maximum of 500 iterations (or less if it converged successfully).

Due to the combination of all parameters used, a total of 24 clusterings can be obtained for each of the three verb sets. For one parameter combination, the clustering algorithm failed to produce an output: base verbs, all frequencies, vectors relying on a window size of 3, and splitting into 250 clusters. The Java library used did not provide any reasons or explanations in the event of failure.

3.2 Clustering Evaluations

As mentioned in the introduction, clustering efforts are motivated by specific tasks or applications, so it is difficult to provide universal recommendations regarding the optimal clustering setup. However, we consider synonymy in cluster analyses as a meta-level goal for clustering approaches, because synonymy represents the strongest type of semantic relatedness. We therefore focus on the ability of the cluster analyses to detect synonymy as a task-independent goal in semantic classification, cf. Section 3.2.1. As a more task-specific evaluation for semantic classification we also assess the ability of the cluster analyses to predict the degree of compositionality of the particle verbs, cf. Section 3.2.2. Considering a strong compositionality of a particle verb regarding its base verb as a case of near-synonymy, the second evaluation targets a semantic relatedness between the complex and the simplex verbs that is not too different to the synonymy evaluation, yet more task-oriented.

3.2.1 Evaluation: Synonymy

We assess the cluster analyses on their ability to contain pairs of synonymous verbs in the same clusters. As basis for the evaluation, we use synonyms provided by the German online synonym dictionary *Duden*[1]. The dictionary contained 2,158 of our particle verbs (with an average of 19 synonyms), and 3,303 of our base verbs (with an average of 13 synonyms). Some examples are listed below:

aussehen *ausblicken, ausschauen, ausspähen, beobachten, entgegensehen, erwarten, spähen, umherblicken, ausgucken, luchsen, ähneln, anmuten, erscheinen, scheinen, vorkommen, wirken, sehen, suchen, umsehen*

zugestehen *akzeptieren, bewilligen, billigen, einwilligen, erlauben, genehmigen, gestatten, gewähren, zubilligen, zuerkennen, konzedieren, legitimieren, sanktionieren, tolerieren, zugutehalten, absegnen, unterschreiben, abnicken, stattgeben*

erklären *aufzeigen, auseinanderlegen, auseinandersetzen, ausführen, darlegen, definieren, entwickeln, erläutern, erörtern, konkretisieren, veranschaulichen, verdeutlichen, zeigen, exemplifizieren, explizieren, klarlegen, klarmachen, verdeutschen, verklickern, verkasematuckeln, auslegen, begründen, belegen, deuten, kommentieren, motivieren, rechtfertigen, fundieren, interpretieren, legitimieren, substanziieren, aufklären, einweihen, informieren, unterrichten, anbringen, anmelden, ausdrücken, äußern, aussprechen, bekennen, bekunden, eröffnen, formulieren, melden, mitteilen, sagen, verlautbaren, vorbringen, kundgeben, kundtun, offenbaren, unterbreiten, verkünden, verkündigen, artikulieren, dokumentieren, verbalisieren, angeben, ausweisen, bescheinigen, bezeichnen, deklarieren, kennzeichnen, einsetzen, einstehen, eintreten, zustimmen, starkmachen, enthüllen, offenbaren, outen*

siegen *bezwingen, gewinnen, schlagen, triumphieren*

Across the clusters within a cluster analysis, we check for all pairs of verbs whether they represent synonyms according to our gold standard or not, and compute precision, recall and the harmonic f-score.

As NMF clustering provides a membership score $x \geq 0$ for each verb and each cluster, we assume that the higher the membership score of a verb for a certain cluster, the more likely the verb is to be part of it. Before running the synonym evaluation, we thus apply an inclusion threshold in order to decide for each verb whether it is considered to be in a cluster or not. Since there is no maximum membership score, and since the values lie on different scales depending on the clustering parameters, determining the ideal membership threshold for each of the clusterings is not straightforward. We therefore employ a brute-force solution: after finding the largest membership score t_{max} for a specific cluster analysis, the synonym evaluation is applied to all non-negative thresholds in the set $t_{max} - k \cdot 0.001, k \in \mathbb{N}_0$. For example, if the largest membership value in a clustering is 0.8916, the synonym evaluation is applied to all thresholds in the set $\{0.8916, 0.8906, 0.8896, ..., 0.0036, 0.0026, 0.0016, 0.0006\}$.

For a given threshold value, the synonym evaluation counts all verb pairs given by the clustering. Two verbs are considered a pair if they share one or more clusters. Since verbs are included in more clusters as the threshold is lowered, we add an abort condition: as soon as 50% of all possible verb pairs are present in the clustering, the threshold is not lowered any further.

See Figure 1 for a small-scale example, listing all symmetric verb pairs for the gold standard and the clustering, marking the correct pairs among the clustering pairs, and calculating precision, recall and f-score. Since the clusterings in our experiments cover thousands of verbs, the actual number of verb pairs in our clusterings is large. This results in f-scores on a very low magnitude, which is not important for our evaluation, however, as the scores are used to compare clustering parameter variations, rather than providing impressive evaluation scores.

[1] www.duden.de

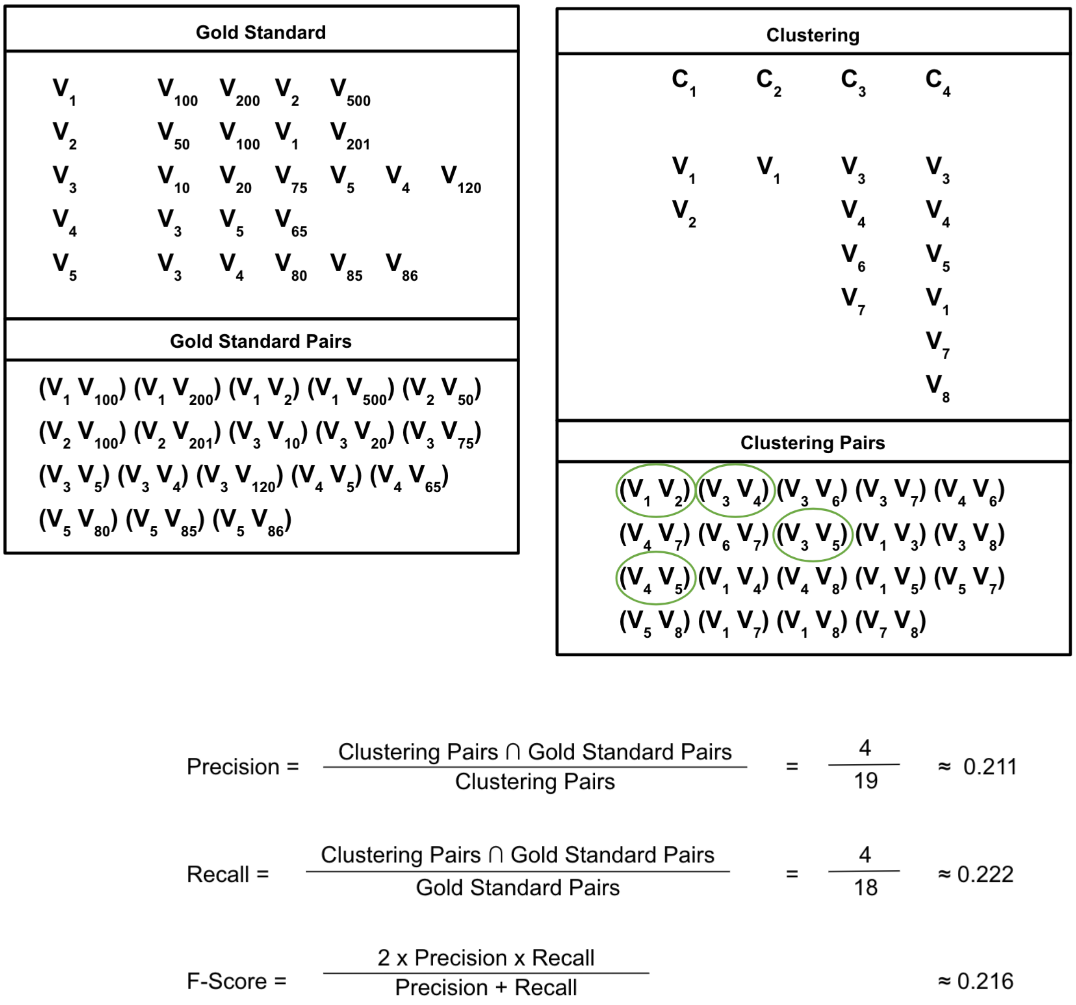

Gold Standard
V_1 V_{100} V_{200} V_2 V_{500}
V_2 V_{50} V_{100} V_1 V_{201}
V_3 V_{10} V_{20} V_{75} V_5 V_4 V_{120}
V_4 V_3 V_5 V_{65}
V_5 V_3 V_4 V_{80} V_{85} V_{86}
Gold Standard Pairs
$(V_1 V_{100})$ $(V_1 V_{200})$ $(V_1 V_2)$ $(V_1 V_{500})$ $(V_2 V_{50})$
$(V_2 V_{100})$ $(V_2 V_{201})$ $(V_3 V_{10})$ $(V_3 V_{20})$ $(V_3 V_{75})$
$(V_3 V_5)$ $(V_3 V_4)$ $(V_3 V_{120})$ $(V_4 V_5)$ $(V_4 V_{65})$
$(V_5 V_{80})$ $(V_5 V_{85})$ $(V_5 V_{86})$
Clustering
C_1 C_2 C_3 C_4
V_1 V_1 V_3 V_3
V_2 V_4 V_4
V_6 V_5
V_7 V_1
V_7
V_8
Clustering Pairs
$(V_1 V_2)$ $(V_3 V_4)$ $(V_3 V_6)$ $(V_3 V_7)$ $(V_4 V_6)$
$(V_4 V_7)$ $(V_6 V_7)$ $(V_3 V_5)$ $(V_1 V_3)$ $(V_3 V_8)$
$(V_4 V_5)$ $(V_1 V_4)$ $(V_4 V_8)$ $(V_1 V_5)$ $(V_5 V_7)$
$(V_5 V_8)$ $(V_1 V_7)$ $(V_1 V_8)$ $(V_7 V_8)$
Precision = $\frac{\text{Clustering Pairs} \cap \text{Gold Standard Pairs}}{\text{Clustering Pairs}}$ = $\frac{4}{19}$ ≈ 0.211
Recall = $\frac{\text{Clustering Pairs} \cap \text{Gold Standard Pairs}}{\text{Gold Standard Pairs}}$ = $\frac{4}{18}$ ≈ 0.222
F-Score = $\frac{2 \text{ x Precision x Recall}}{\text{Precision + Recall}}$ ≈ 0.216

Figure 1: Small-scale example of verb pair evaluation.

As an alternative to the brute-force search for the best inclusion threshold, we also apply a method for assigning verbs to their top n clusters, with $1 \leq n \leq \frac{N}{2}$ and N representing the total number of clusters. In this variant, verbs are added to the n clusters with the highest membership scores. For example, suppose that in a clustering of verbs into 6 clusters, verb v_1 has the membership values 0.7, 0.4, 0.45, 0.2, 0.5, and 0.8 for clusters 1 to 6 respectively. For $n = 1$, the verb will be included only in cluster 6, for $n = 2$, it will be considered part of clusters 6 and 1, and for $n = 3$, it belongs to clusters 6, 1, and 5. This variant is referred to as *top-n evaluation*, whereas the previously described method is referred to as *threshold evaluation*.

3.2.2 Evaluation: Compositionality

In this evaluation, we predict the degree of compositionality of the complex particle verbs, i.e., the degree of relatedness between the particle verbs and their corresponding base verbs (such as *abnehmen – nehmen* 'take over – take', and *anfangen – fangen* 'begin – catch'). We assume that if a particle verb and its base verb tend to co-occur in the same cluster within a cluster analysis, then the particle verb is semantically transparent, rather than opaque. The predictions are evaluated against an existing dataset of human ratings on German particle verb compositionality (Bott et al., 2016). The gold standard contains a total of 400 particle verbs across 11 particle types and 3 frequency bands.

Similarly to the evaluation metric described in the previous section, the compositionality evaluation is also applied to all thresholds in the set $t_{max} - k \cdot 0.001, k \in \mathbb{N}_0$, with t_{max} being the largest inclusion value found in the clustering, as well as to all top-n cluster assignments with $1 \leq n \leq \frac{N}{2}$. For each pair of particle verb and base verb, e.g., *abnehmen – nehmen*, we then compare the assignment of the two verbs to the same vs. different clusters in two different ways.

- *Pointwise Mutual Information (PMI)*:

  We calculate $\log \frac{p(PV,BV)}{p(PV)p(BV)}$, with $p(PV, BV)$ the proportion of clusters containing both the particle verb PV and the base verb BV, and $p(PV)$ and $p(BV)$ the proportions of clusters containing the particle and base verbs individually. The proportions are relative to the total number of clusters, so $p(PV, BV) = 0.2$ means that 20% of the clusters contain both PV and BV. A high PMI means that a pair tends to occur in the same clusters rather than in different ones.

- *Cosine similarity between average cluster centroid vectors*:

  For each cluster, we calculate the centroid vector as the average over all verb vectors in that cluster. In addition, we calculate average cluster centroid vectors for all verbs, as the average over all centroid vectors a verb has been assigned to. Then, each two verbs are compared by calculating the cosine of the angle between the respective average cluster centroid vectors. A high cosine similarity means that a pair tends to occur in the same clusters, or that the clusters in which the two verbs occur have similar centroids.

In the final evaluation step, we compute the correlation between the PV–BV similarity predictions relying on PMI/cosine in comparison to the gold standard ratings, using Spearman's Rank-Order Correlation Coefficient ρ (Siegel and Castellan, 1988).

4 Results

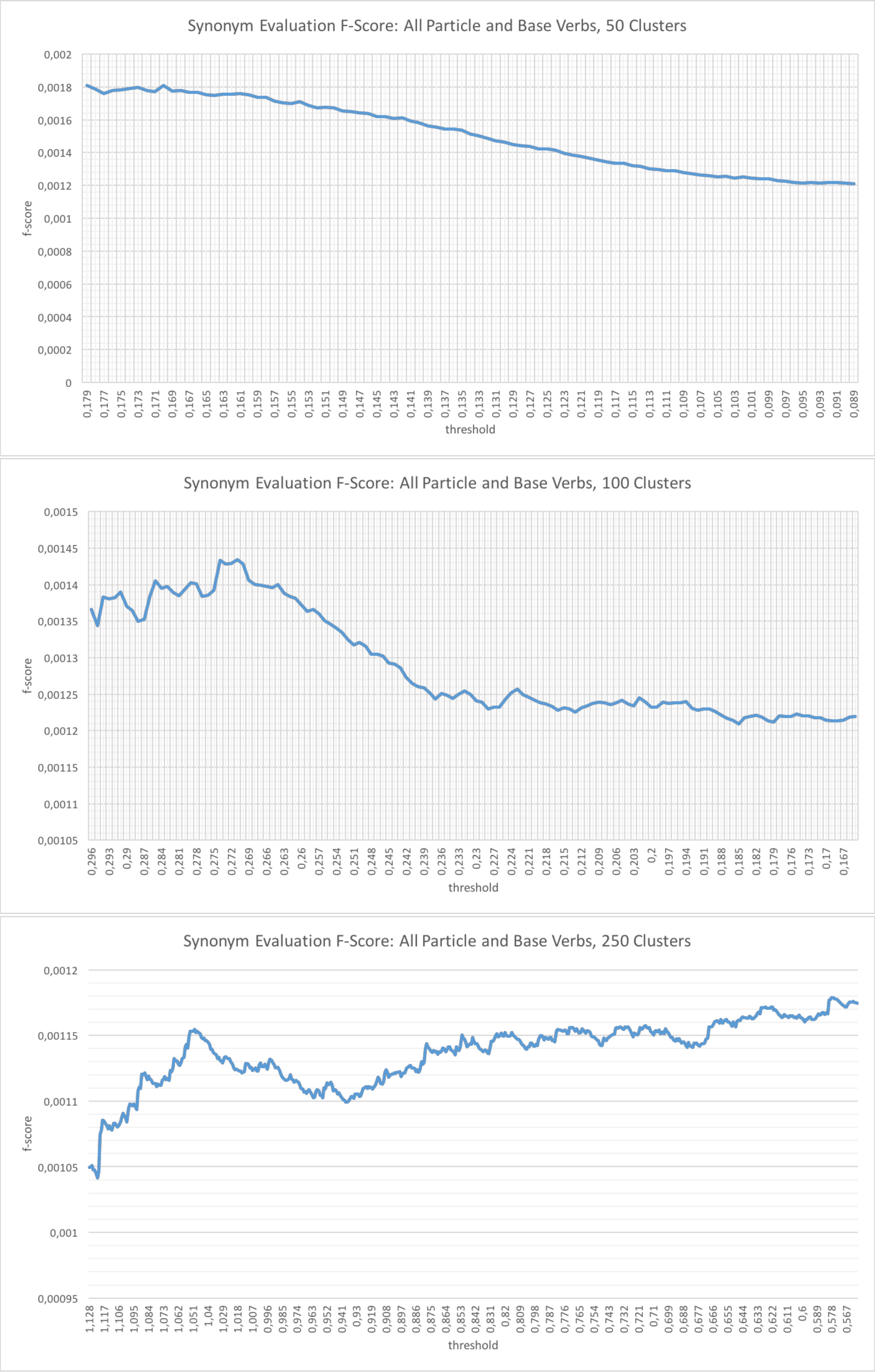
In the following, we present the results of our clustering experiments and evaluations. Please (a) remember that the f-score values for the synonym evaluation are in a very low range because they assess a comparably large number of verb pairs across 4,871 base verbs and 3,173 particle verbs within the cluster analyses; and (b) note that the compositionality evaluation is carried out on a subset of only 400 particle verbs for which the gold standard contains compositionality ratings.

Figure 2 presents the synonymy evaluation f-score values when clustering all particle and base verbs in 50, 100 and 250 clusters. With an increasing threshold (x-axis), a smaller number of verbs is included in the clusters. The resulting quality of the cluster analyses differs across the different numbers of clusters, as one would have expected. For 50 and 100 clusters, the correlation decreases with an increasing threshold along the x-axis, so a more general inclusion is better, but for 250 clusters, the clusters are better when they contain less verbs. As the different scales on the y-axis across the three plots show, overall a smaller number of clusters with generous assignment is best.

Table 1 zooms into the differences of clustering low-, mid, high-frequency or all verbs, regarding base verbs (BVs), particle verbs (PVs) and both BVs and PVs. For each cell, we show the best result across thresholds/top-n and vector spaces. For low- and mid-frequency verbs, we did not assess the compositionality evaluation because less than 10% of the particle verbs and corresponding base verbs from the gold standard were found in the clustering, regardless of the inclusion threshold or the top-n value used.

The results in the table demonstrate the following differences:

- The results for high-frequency verbs are generally better than for low- and mid-frequency verbs, demonstrating that target frequency (and, most probably, less sparse data) matters.



Synonym Evaluation F-Score: All Particle and Base Verbs, 50 Clusters

Synonym Evaluation F-Score: All Particle and Base Verbs, 100 Clusters

Synonym Evaluation F-Score: All Particle and Base Verbs, 250 Clusters

Figure 2: Synonymy f-score results for all verbs and 50/100/250 clusters.

Evaluation: synonymy (threshold)

Frequency	ALL			HIGH			MID			LOW		
Clusters	50	100	250	50	100	250	50	100	250	50	100	250
BVs	**.00640**	.00412	.00370	**.02337**	.01559	.01606	**.00955**	.00480	.00277	**.00212**	.00103	.00090
PVs	.00126	.00076	.00068	.01170	.00602	.00736	.00072	.00025	.00022	.00009	.00004	.00003
BVs+PVs	.00181	.00143	.00118	.01420	.00823	.00925	.00225	.00101	.00084	.00012	.00007	.00004

Evaluation: synonymy (top-n)

Frequency	ALL			HIGH			MID			LOW		
Clusters	50	100	250	50	100	250	50	100	250	50	100	250
BVs	**.01169**	.00736	.00428	**.03006**	.02271	.01999	**.01007**	.00514	.00324	**.00255**	.00144	.00099
PVs	.00217	.00124	.00119	.01335	.00616	.00788	.00088	.00026	.00018	.00004	.00003	.00003
BVs+PVs	.00368	.00351	.00214	.01935	.01206	.00917	.00239	.00152	.00101	.00012	.00007	.00004

Evaluation: compositionality (threshold)

Frequency	ALL			HIGH		
Clusters	50	100	250	50	100	250
BVs+PVs (PMI)	.274***	.183***	.248***	**.468****	.220*	.281**
BVs+PVs (Cos)	**.334*****	.264***	.287***	.439**	.301***	.283**

Evaluation: compositionality (top-n)

Frequency	ALL			HIGH		
Clusters	50	100	250	50	100	250
BVs+PVs (PMI)	.259***	.297***	**.377*****	**.421*****	.378***	.398***
BVs+PVs (Cos)	.197***	.186***	.203***	.311***	.257**	.207*

Table 1: Results across evaluations and clustering parameters (* = p≤0.05, ** = p≤0.01, *** = p≤0.001).

- The results for base verbs are generally better than for particle verbs (only applicable to the synonym evaluation), demonstrating that particle verbs are harder to assess semantically than base verbs, presumably because they are more ambiguous.

- Confirming insights from Figure 2, the results for clusterings with 50 clusters are generally better than for clusterings with 100 or 250 clusters.

- For predicting particle verb compositionality, PMI generally works better than the cosine.

- (not shown in the table:) There is no strong tendency for one of the vector spaces (i.e., using a window of 3 vs. 10 words) outperforming the other.

5 Conclusion

We provided an extensive clustering setup and focused on synonymy as a task-independent goal in semantic classification, in order to explore the role of verb frequency ranges across various numbers of clusters. We demonstrated that (1) low-frequency German verbs are clustered significantly worse than mid- or high-frequency German verbs, and that (2) German complex verbs are in general more difficult to cluster than German base verbs. While (1) the effect of clustering low-frequency target verbs has been investigated by a restricted number of earlier approaches, (2) might be considered as general knowledge but has –as far as we are aware of– not explicitly been proven before.

Acknowledgments

The research was supported by the DFG Collaborative Research Centre SFB 732/D12 *“Sense Discrimination and Regular Meaning Shifts of German Particle Verbs”* (Maximilian Köper, Moritz Wittmann), the DFG Research Grant SCHU-2580/2 *“Distributional Approaches to Semantic Relatedness”* (Moritz Wittmann), and the DFG Heisenberg Fellowship SCHU-2580/1 (Sabine Schulte im Walde).

References

Chris Biemann. 2006. Chinese Whispers: An Efficient Graph Clustering Algorithm and Its Application to Natural Language Processing Problems. In *Proceedings of the 1st Workshop on Graph-Based Methods for Natural Language Processing*. Stroudsburg, PA, USA, pages 73–80.

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

Stefan Bott, Nana Khvtisavrishvili, Max Kisselew, and Sabine Schulte im Walde. 2016. G$_{h}$ost-PV: A Representative Gold Standard of German Particle Verbs. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon*. Osaka, Japan, pages 125–133.

John Carroll and Alex C. Fang. 2004. The Automatic Acquisition of Verb Subcategorisations and their Impact on the Performance of an HPSG Parser. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*. Sanya City, China, pages 107–114.

John Carroll, Guido Minnen, and Ted Briscoe. 1998. Can Subcategorisation Probabilities Help a Statistical Parser? In *Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora*. Montréal, Canada, pages 118–126.

Chris Ding, Xiaofeng He, and Horst D. Simon. 2005. On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. In *Proceedings of the SIAM International Conference on Data Mining*. Newport Beach, CA, USA, pages 606–610.

Bonnie J. Dorr and Doug Jones. 1996. Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues. In *Proceedings of the 16th International Conference on Computational Linguistics*. Copenhagen, Denmark, pages 322–327.

Gertrud Faaß, Ulrich Heid, and Helmut Schmid. 2010. Design and Application of a Gold Standard for Morphological Analysis: SMOR in Validation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*. Valletta, Malta, pages 803–810.

Elias Iosif and Alexandros Potamianos. 2007. A Soft-Clustering Algorithm for Automatic Induction of Semantic Classes. In *Proceedings of the 8th Interspeech Conference*. Antwerp, Belgium, pages 1609–1612.

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic, pages 868–876.

Upali S. Kohomban and Wee Sun Lee. 2005. Learning Semantic Classes for Word Sense Disambiguation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Ann Arbor, MI, pages 34–41.

Maximilian Köper and Sabine Schulte im Walde. 2016. Automatic Semantic Classification of German Preposition Types: Comparing Hard and Soft Clustering Approaches across Features. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany, pages 256–263.

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*. Sapporo, Japan, pages 64–71.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word Sense Induction for Novel Sense Detection. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France, pages 591–601.

Jiwei Li and Dan Jurafsky. 2015. Do Multi-Sense Embeddings Improve Natural Language Understanding? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, pages 1722–1732.

Diana McCarthy, Sriram Venkatapathy, and Aravind K. Joshi. 2007. Detecting Compositionality of Verb-Object Combinations using Selectional Preferences. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, Czech Republic, pages 369–379.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*. Lake Tahoe, Nevada, USA, pages 3111–3119.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA, pages 322–332.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient Nonparametric Estimation of Multiple Embeddings per Word in Vector Space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, pages 1059–1069.

Detlef Prescher, Stefan Riezler, and Mats Rooth. 2000. Using a Probabilistic Class-Based Lexicon for Lexical Ambiguity Resolution. In *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken, Germany, pages 649–655.

Mingjie Qian. 2016. LAML. `https://github.com/MingjieQian/LAML`.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a Semantically Annotated Lexicon via EM-Based Clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Maryland, MD, pages 104–111.

Carolina Scarton, Lin Sun, Karin Kipper-Schuler, Magali Sanches Duran, Martha Palmer, and Anna Korhonen. 2014. Verb Clustering for Brazilian Portuguese. In Alexander Gelbukh, editor, *Proceedings of the 15th International Conference on Intelligent Text Processing and Computational Linguistics*. Kathmandu, Nepal, pages 25–39.

Roland Schäfer. 2015. Processing and Querying Large Web Corpora with the COW14 Architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*. Mannheim, Germany, pages 28–34.

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Istanbul, Turkey, pages 486–493.

Sabine Schulte im Walde. 2000. Clustering Verbs Semantically According to their Alternation Behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbrücken, Germany, pages 747–753.

Sabine Schulte im Walde. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics* 32(2):159–194.

Sidney Siegel and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Boston, MA.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using Predicate-Argument Structures for Information Extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan, pages 8–15.

Giulia Venturi, Simonetta Montemagni, Simone Marchi, Yutaka Sasaki, Paul Thompson, John McNaught, and Sophia Ananiadou. 2009. Bootstrapping a Verb Lexicon for Biomedical Information Extraction. In Alexander Gelbukh, editor, *Linguistics and Intelligent Text Processing*, Springer, Heidelberg, pages 137–148.

Marion Weller, Sabine Schulte im Walde, and Alexander Fraser. 2014. Using Noun Class Information to model Selectional Preferences for Translating Prepositions in SMT. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*. Vancouver, Canada, pages 275–287.

Wei Xu, Xin Liu, and Yihong Gong. 2003. Document Clustering Based on Non-negative Matrix Factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Toronto, Canada, pages 267–273.

Takeru Yokoi. 2013. Topic Number Estimation by Consensus Soft Clustering with NMF. In Tai-Hoon Kim, Young-Hoon Lee, Byeong Ho Kang, and Dominik Slezak, editors, *Future Generation Information Technology*, Springer, volume 8105 of *Lecture Notes in Computer Science*, pages 63–73.