Exploring the value space of attributes: Unsupervised bidirectional clustering of adjectives in German

Wiebke Petersen
Düsseldorf University, SFB 991
petersen@
phil.uni-duesseldorf.de

Oliver Hellwig
Düsseldorf University, SFB 991
ohellwig@
phil-fak.uni-duesseldorf.de

Abstract

The paper presents an iterative bidirectional clustering of adjectives and nouns based on a co-occurrence matrix. The clustering method combines a Vector Space Models (VSM) and the results of a Latent Dirichlet Allocation (LDA), whose results are merged in each iterative step. The aim is to derive a clustering of German adjectives that reflects latent semantic classes of adjectives, and that can be used to induce frame-based representations of nouns in a later step. We are able to show that the method induces meaningful groups of adjectives, and that it outperforms a baseline k-means algorithm.

1 Introduction

The research presented in this paper is part of a larger project which aims at the semantic analysis of adjectival modification of nouns in German in a frame-based approach. Its approach is to model the conceptual mechanisms underlying adjectival modification by decomposing the meanings of adjectives (A) and nouns (N) in frames, i.e. in recursive attribute-value structures. The most common modification process is that the noun concept bears an attribute whose value is restricted by the adjective. The examples in (1) show that some adjectives are strongly associated with an attribute like the color attributes or the attributes of size or age. Some of these adjectives do not only require a special attribute but also a special noun concept (like the color adjective 'blond' that only applies to 'hair'). For other adjectives like *städtisch* 'urban' it is less clear the value of which attribute they specify.

(1) a. schwarzer Ball / Stift
black ball / pen
b. blondes Haar
blond hair
c. fröhlicher Junge / Abend
happy boy / evening
d. städtische Schule
city$_{adj}$ school
e. kindlicher Organismus
child$_{adj}$ organism
f. schneller Fahrer
fast driver

As adjectives often merely restrict the value of an attribute that is associated with the adjective, as 'black' modifying the attribute color in (1-a) such that $\text{COLOR}(ball) = black$, attribute-value structures are well-suited for the task of modeling adjectival modification (cf. Pustejovsky, 1995). However, flat attribute-value lists are not sufficient, because 'black pen' can denote a pen that is black, or a pen that writes in black, $\text{COLOR}(\textit{writing of the pen}) = black$. Similarly in (1-c), it cannot be the

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2839–2848, Osaka, Japan, December 11-17 2016.

same HAPPINESS-attribute applying to both 'boy' and 'evening' in the same way. Relational adjectives like 'städtisch' in (1-d) express a relation between the denotation of N and of the root of A, although the kind of relation is rather vague, a.o. OWNER(*school*) = *city*, OPERATOR(*school*) = *city*, LOCATION(*school*) = *city*. In cases like (1-e) where the noun is relational (cf. de Bruin and Scha, 1988; Löbner, 2011), the adjective fills in the argument slot, POSSESSOR(*organism*) = *child*. Finally, in an event-related A+N phrase as in (1-f) an event frame is activated and modified by A, SPEED(*driving event*) = *fast*.[1]

The examples in (1) show that the compositional mechanisms active in A+N phrases are based on an interplay of the adjective and the noun meaning. Our aim is to cluster adjectives and nouns by these mechanisms, that means by the attributes they exhibit or modify and by the structural position of the modified attribute in the A+N frame.

2 Related research

Distributional approaches to compositionality have been the subject of several recent publications (Blacoe and Lapata, 2012). Regarding A+N phrases, research concentrates on modeling their compositional meaning (Baroni and Zamparelli, 2010; Guevara, 2010), predicting the acceptability rates of novel A+N phrases (Vecchi et al., in press), and on connecting linguistic theory with computational models (Boleda et al., 2013). As an alternative to classical Vector Space Models (VSM) (Turney and Pantel, 2010; Erk, 2012), Latent Dirichlet Allocation (LDA) and related Dirichlet process mixture models are increasingly applied to questions in lexical semantics. Séaghdha and Korhonen (2014) model selectional preferences of verbs in a Bayesian framework, and learn semantic classes of arguments from the latent variables of the model.

Our approach is closely related to recent work in distributional semantics by Hartung and Frank (2010, 2011) on selecting attributes in A+N phrases. However, their task differs from ours in that they explicitly restrict themselves to property-denoting adjectives for which an associated attribute is explicitly assigned in WordNet (Fellbaum, 1998). Hartung and Frank (2011) apply supervised variants of LDA to the problem of attribute prediction for A+N phrases. The authors filter candidate phrases from synsets in WordNet, construct separate pseudo-documents for As and Ns, and compose the output vectors of the LDA with different arithmetic operations following Baroni and Zamparelli (2010), Guevara (2010), and Hartung and Frank (2010).

Unsupervised detection of semantic classes of adjectives has been the topic of several studies. Boleda et al. (2004) annotate Catalan adjectives with two types of coarse semantic labels (unary/binary, basic/object/event), and cluster morpho-syntactic sentence patterns in which the adjectives typically occur. When setting the number of expected clusters to 2 (unary/binary) and 3 (basic/object/event), the authors observe a high correlation between their semantic labelings and the detected clusters. Furthermore, our task is related to detecting synsets and gradability of adjectives using a corpus based approach (Schulam and Fellbaum, 2010). We are interested in detecting groups of adjectives that describe the full range of a (latent) attribute for a group of nouns. So, a group of nouns that denote motorized vehicles may be characterized by the adjectives *geparkt* 'parked', *gestartet* 'started', and *fahrend* 'driving', all of which describe an attribute 'motion state' in a vehicle frame, but don't necessarily belong to a single synset, or are part of a scale of values in a traditional definition.

3 Method

We aim at deriving an adjective and noun clustering that reflects all modificational mechanisms possible in A+N phrases. Due to the explorative nature of our approach, we need a full picture of the combinatorial possibilities of As and Ns, so that we need to investigate a huge corpus. Moreover, we cannot apply supervised methods, because we do not want to restrict ourselves to a subclass of adjectives. The basic idea is to extract as many attested A+N phrases as possible and to apply an iterative bidirectional clustering algorithm based on the co-occurrences of adjectives and nouns. First, the adjectives are clustered on the basis of co-occurring nouns, then the nouns are clustered on the basis of co-occurrences with

[1] For a recent overview on adjectival modification refer to Morzycki (2015).

	distinct A+N pairs	A types	N types	density
Google n-grams	894,743	4,460	10,996	0.0182
Wikipedia	82,022	12,616	19,849	0.0003
newspaper	261,327	4,507	7,955	0,0073

Table 1: Number of A+N pairs and A and N types extracted from the corpora

adjectives from the just gained clusters. The process is iterated, until it reaches a stable clustering or a point at which new clusters would become too diverse.

As detected clusters are reinserted into the vector space matrix, our clustering produces hierarchically structured representations of As and Ns. An example will be discussed in Section 4.

3.1 Data

3.1.1 Co-occurrence data

Data for the A+N co-occurrences are extracted from the Google n-gram corpus (Michel et al., 2011), from a corpus of German newspaper texts, and from a dump of the German Wikipedia, including the Wikisources.[2] The Google and Wikipedia corpora are chosen for their sizes, but also because they contain texts from literary domains that may display other patterns of A+N usage than newspapers. Note that our focus is on the binary question of whether an adjective can modify a noun or not. Thus, throughout the A+N pair extraction process, we prioritize precision above recall. Furthermore, we are not interested in the co-occurrence frequencies (as long as each A+N pair occurs at least 5 times in the corpus), which allows us to merge the data received from the two text corpora ('newspaper' and 'wikipedia') with the Google n-gram corpus.

German marks most A+N pairs in singular number and all definite A+N pairs in plural by using articles. In addition, all nouns are capitalized, while adjectives start with lowercase letters, such that A+N pairs can be extracted from the Google n-grams by applying a regular expression (remember that we do not aim at the extraction of all A+N pairs). Raw A+N pairs detected in this way are grouped, counted, lemmatized with a full form lexicon and normalized[3], resulting in a total of 894,743 distinct A+N pairs with 4,460 A and 10,996 N types. We process the Wikipedia data by removing Wiki specific formatting, and splitting the resulting text into sentences. After POS-tagging each sentence with MATE (Bohnet and Nivre, 2012), A+N pairs are extracted by searching for the POS tag sequence ADJA + [NN|NE]. After removing AN pairs that occur less than 5 times and applying the spelling normalization, this corpus contains 82,022 distinct A+N pairs with 12,616 A and 19,849 N types. The newspaper corpus is lemmatized and POS tagged in the same way as the Wikipedia corpus. This corpus yields 261,327 distinct lemmatized and counted A+N combinations with 4,507 A and 7,955 N types.

Table 1 summarizes the results of the A+N pair extraction process for the three corpora. The small number of distinct A+N pairs found in the Wikipedia corpus compared to the large number of A and N types can be explained by the characteristics of the texts in this corpus. Many Wikipedia entries deal with highly specified topics and introduce a professional terminology; especially the introduced adjectives are mainly used in fixed expressions. Furthermore, the table shows the relevance of the Google n-gram corpus, although it is built from slightly outdated texts and contains many OCR errors (Pechenick et al., 2015). Compared to the newspaper corpus the number of A and N types in the Google corpus is similar, while the number of distinct A+N combinations is much higher. This is due to the fact that the Google corpus is extracted from a huge corpus exhibiting more of the combinatorial possibilities of As and Ns. The density values illustrate the differences between the corpora (number of A+N pairs divided by the product of the number of noun types and number of adjective types).

[2]We use Version 20120701 of the German Google n-grams, the 2013 collection from `http://www.statmt.org/wmt14/training-monolingual-news-crawl/` and the Wikipedia dump from 1st of June 2016.

[3]Because the Google and Wikipedia corpora contain several texts in old German orthography, we apply a spelling normalization that transforms, for example, *th* into *t* (*Alterthum* 'antiquity' → *Altertum*), and corrects a few obvious OCR errors (*Iahr* 'year' → *Jahr*).

3.1.2 Gold data

A gold clustering is required for the evaluation of our clustering solutions. We tested the adjective classification given in GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) as a gold standard for our purpose. It turned out that the data was not ideal for our task for two reasons: First, GermaNet is hierarchically structured by the hyperonymy relation. That makes it difficult to automatically detect the appropriate granularity level which corresponds to the level of attribute value specifications. Second, GermaNet is constructed on the basis of binary semantic relations like 'synonymy', 'antinonymy' and 'hyperonymy' and not on the basis of attributes. For our purpose, a classification based on attributes and their potential values and bearers would be more suitable.

As an alternative, we used a dataset consisting of all simple adjectives from a German dictionary (Duden, 2004) that are neither loanwords nor derived from other words.[4] The 278 extracted adjectives have been manually classified by 45 attributes. Attributes that take characteristic values if applied to special noun classes are separated from each other. Hence, the attribute *Haarfarbe* 'hair color' is separated from the general attribute *Farbe* 'color' as it allows for hair specific values like *blond* 'blond'. Similarly, *Körpergestalt* 'shape of a human body' is separated from *Gestalt* 'shape', because it takes values like *hager* 'lean' or *mager* 'skinny' that cannot be used for non-human bodies. The main drawback of this data set is its restriction on non-derived adjectives, which are relatively rare in German. As a result, the data set is fairly small, and it contains several outdated and rarely used adjectives.

3.2 Algorithm

The proposed clustering is run in two configurations. The configuration **bin** operates with Jaccard distances calculated from the binarized vector space matrix (VSM) only. The configuration **lda** reweights these Jaccard distances with the output of an LDA topic model.

Both configurations start by constructing a vector space matrix V in which each A is described by the frequencies of the nouns with which it occurs in the corpus, such that rows represent adjectives, and columns represent nouns. This matrix is a structured distributional semantic model in the sense of Baroni and Lenci (2010), because the context words are defined by the syntactic relation between A and N. Sparsely populated rows (A) and columns (N) are removed by an iterative thinning step. After desparsification, a binary matrix V_B is derived from V by setting all non-zero values in V to 1.[5]

In the following, we describe the first iteration step in which the rows correspond to nouns and the columns to adjectives. Note, that after each iteration the matrix V_B is transposed, such that rows correspond to adjectives in even iterations and to nouns in odd iterations. Let $\vec{r_i}$ denote a logical row vector of V_B.

In the configuration **bin**, pairwise distance measures d_{ij} between all rows in V_B are calculated based on the Jaccard distances of the rows.

$$d_{ij} = 1 - \frac{|\vec{r_i} \wedge \vec{r_j}|}{|\vec{r_i} \vee \vec{r_j}|}$$

In order to avoid the clustering of nouns that share only a few adjectives, we set the distance measure to zero if the number of shared nouns is less than a given parameter t_J:

$$d_{ij}^{\text{bin}} = \begin{cases} d_{ij} & \text{if } |\vec{r_i} \wedge \vec{r_j}| \geq t_J \\ 0 & \text{else} \end{cases}$$

In the current configuration the parameter t_J is set to 3 to avoid the clustering of nouns sharing only two or less adjectives and vice versa.

In the configuration **lda**, LDA[6] is applied to the V_B, and another list of pairwise distance tuples

[4] These adjectives were collected and annotated by Sebastian Löbner and Thomas Gamerschlag in the project B02 "Dimensional Verbs", SFB 991, HHU Düsseldorf. Our special thanks to Thomas Gamerschlag for providing the data.

[5] The binarization step is performed, because we are only interested in whether a noun can be modified by an adjective, thus whether it bears an attribute the value of which can be restricted by the adjective. We could not rely on the frequency counts, as our corpus of A+N pairs results from an unbalanced merge of different corpora.

[6] LDA is performed with Gibbs Sampling using the library `GibbsLDA++`, `https://sourceforge.net/projects/gibbslda/`.

$(\vec{r_i}, \vec{r_j}, \theta_{ij})$ is created from the Θ values obtained from the LDA. For row vectors $\vec{r_i}$ and $\vec{r_j}$, θ_{ij} is given as

$$\theta_{ij} = \left(\sum_{k=1}^{K=15} (\Theta_{ik} - \Theta_{jk})^2 \right)^{\frac{1}{2}}$$

The number of latent topics $K = 15$ is intentionally kept low, because we don't derive semantic classes from the topic model of the LDA, but rather use the similarities between the Θ distributions for reweighting the Jaccard distances. In configuration **lda**, the Jaccard distances d_{ij} are reweighted with the distances from the LDA topic model, resulting in the final score d_{ij}^{lda} for each pair of words:

$$d_{ij}^{\text{lda}} = d_{ij}^{\text{bin}} \cdot \theta_{ij}$$

When the pairs of candidates $(\vec{r_i}, \vec{r_j})$ are reordered using this score, the top scoring pairs have high similarities in the binary vector space and in the topic model induced by the LDA.

Depending on the configuration mode, new word clusters are created either from the top 5% of the top scoring pairs with respect to the d^{bin} or to the d^{lda} distance measure. Clusters are built by constructing the transitive closures. Thus, if d_{ij} and d_{ik} belong to the 5% lowest distances measured, the words belonging to the rows $\vec{r_i}$, $\vec{r_j}$ and $\vec{r_k}$ are clustered together. Let R denote all rows that belong to one transitive closure and should thus be clustered together. The distributional representation of the new cluster, i.e. of the new row vector $\vec{R}$, is built by merging R with a majority based binary operator:

$$\vec{R_k} = \begin{cases} 1 & \text{if } \sum_{\vec{r} \in R} r_k \geq \frac{|R|}{2} \\ 0 & \text{else} \end{cases}$$

The rows in V which belong to R are deleted and the new row $\vec{R}$ is added, shrinking the matrix dimension in this way by $|R| - 1$.

The algorithm terminates, when no new classes are detected, because no pair of rows has more than t_J non-zero columns in common. Else, V_B is transposed, and the clustering method is applied to the other word class.

4 Evaluation and Results

We will restrict the evaluation of the obtained clusters to the adjective clusters, as we do not have access to an attribute based clustering of nouns that could be used as our gold data. We evaluate the results of the bidirectional clustering against the gold data described in Section 3.1, and against the results of k-means as a baseline clustering algorithm. As the adjective noun co-occurrence matrix is too sparse to run k-means successfully on it, we use information from context windows of the adjectives instead. For running the baseline k-means, we create neural embeddings of all words that occur at least 20 times in the merged corpus (newspaper, Wikipedia) using the `word2vec` tool (Mikolov et al., 2011),[7] and extract the embeddings of the adjectives that are processed in the bidirectional clustering. Remember that one of the main challenges in this task consists in determining the number of semantic classes from the raw data. In order to find this number, we calculate the gap statistic (Tibshirani et al., 2001) for cluster sizes between 40 (and thus just below the number of semantic classes in the gold data) and an – arbitrarily chosen – upper limit of 300, advancing in steps of 10 classes. Figure 1 shows that $\hat{k}$ as defined in Tibshirani et al. (2001, 415) changes from negative to positive values, when k-means is performed with approximately 160 classes, so that we use $k = 160$ for the baseline k-means clustering.

Since k-means and the gold standard provide hard non-hierarchical clusterings, we need to extract a hard non-hierarchical clustering from our bidirectional clustering as well, in order to compare them by standard measures as the Rand index (RI; Rand, 1971). The bidirectional clustering can produce deeply nested hierarchical clusterings, when later iterations of the method merge detected clusters and adjectives

[7] We cannot use the Google corpus for `word2vec` as it does not provide us with the necessary context of the A+N pairs. The settings for the `word2vec` tools are as follows: window size: 6 words, embedding size: 300, bow. All other parameters are set to their default values.

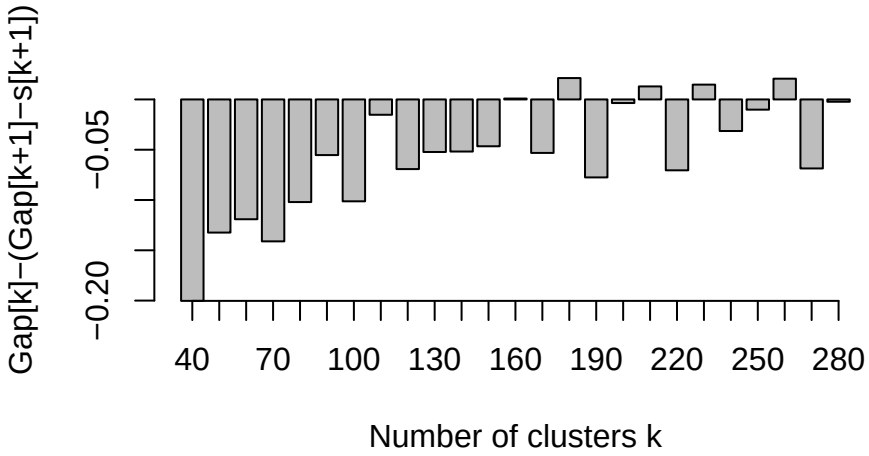

Figure 1: Gap statistic ($\hat{k}$) for neural embeddings of adjectives, showing a switch from negative to positive values at about 160 classes.

268 besonderer « 'topmost'

216 großartig außergewöhnlich

211 hervorragend

196 herausragend ausgezeichnet « 'leaf'

213 wunderschön wunderbar

Figure 2: Example demonstrating the effect of the two configurations 'leaf' and 'topmost'. Indentation indicates at which iteration a word cluster has been created. The configuration 'jaccard' cannot be displayed as its cannot be computed locally

into new classes. As such hierarchical representations provide no unambiguous hard clusters, we present three different modes of deriving hard clusterings from our representation. The following rules apply only to those adjectives that belong to at least one singleton cluster: In the configuration 'leaf' each such adjective is assigned to the smallest non-singleton cluster it belongs to; in the configuration 'topmost' it is assigned to the largest cluster it belongs to. Figure 2 illustrates the difference between these two configurations. The adjective *ausgezeichnet* 'excellent' is labeled with ID 196 in the 'leaf' and with ID 268 in the 'topmost' configuration. In addition, we derive a third hard clustering 'jaccard' that takes the homogeneity of the clusters into account. For this sake, we obtain pairwise Jaccard indexes between all adjectives that are subsumed under each node of the hierarchical output, and calculate their means for each node. To make this evaluation comparable with the result of k-means, the 160 nodes with the highest average Jaccard indexes are taken as hard cluster labels. All leaves that are not subsumed under one of these nodes are assigned to singleton clusters.

For the evaluation of the clustering we compute the Rand index (RI; Rand, 1971) and the Adjusted Rand index (Hubert and Arabie, 1985, ARI;). The Rand index of two clusterings C and C' is defined by

$$\mathcal{R}(C, C') = \frac{n_{11} + n_{00}}{\binom{n}{2}}$$

where n_{11} is the number of adjective pairs which belong to the same cluster in both clusterings, n_{00} is the number of adjective pairs that are assigned to different clusters in both clusterings, and n is the number of adjectives. Thus RI measures the proportion of the pairs which are clustered in the same

	RI	ARI
k-means	0.9499	0.0000

Table 2: (Adjusted) Rand index for the k-means clustering

configuration	'topmost'		'jaccard'		'leaf'	
	RI	ARI	RI	ARI	RI	ARI
newspaper (**bin**)	**0.9561**	0.0557	0.9568	0.0265	0.9583	0.0250
newspaper (**lda**)	0.9532	**0.0880**	0.9589	0.0263	0.9602	0.0269
google (**bin**)	0.9486	0.0687	0.9599	**0.0514**	0.9597	**0.0438**
google (**lda**)	0.9513	0.0759	**0.9607**	0.0442	**0.9607**	0.0370
merged (**bin**)	0.9382	0.0467	0.9566	0.0456	0.9571	**0.0454**
merged (**lda**)	**0.9534**	**0.0931**	**0.9576**	**0.0549**	**0.9581**	0.0386

Table 3: (Adjusted) Rand index for the bidirectional clustering

way in both clusterings. The adjusted Rand index corrects the Rand index by agreements that are solely due to chance. Note that while RI takes values between 0 and 1, ARI can take negative values as well (Meila, 2007). Table 2 shows the RI and ARI values for the k-means clusterings. Table 3 compares the clusterings resulting from the bidirectional clustering in different configurations and for different corpora.

While the RI values are high, ARI values are very low. This outcome is typical for sparse data and a large number of clusters. Remember that we have put each adjective that was not clustered by the algorithm into a singleton cluster. No clear favorite emerges when comparing the clusterings with and without LDA (Table 3, **bin** and **lda**). Only for the merged corpus which contains the data of all three corpora (newspaper, Google and Wikipedia), there is a tendency that **lda** outperforms **bin**. The different granularity levels ('topmost', 'jaccard' and 'leaf') by which the hierarchy is cut into a non-hierarchical clustering influence the RI and the ARI values as expected: a coarser clustering leads to higher ARI but slightly lower RI values. The most obvious result is that the bidirectional clustering outperforms the k-means clustering with respect to the more relevant ARI (independent of the chosen configuration). This result is remarkable, as the k-means clustering is based on textual context windows, while the bidirectional clustering only considers isolated A+N combinations. Thus, while for many tasks which aim at a semantic clustering it is better to look at the distribution of words in a larger context, the task of clustering adjectives by attributes seems to benefit from using a structured VSM.

To conclude this section, we will discuss some of the received clusters to give an impression of what kind of clusters can be expected. Many property-denoting adjectives turn out to be clustered very well, especially those which are derived from a numeral like *X-stellig* 'X place', *X-geschossig / X-stöckig* 'X floor', *X-malig* 'X times', *X-spurig* 'X lane', *X-jährig* 'X year', *X-tägig* 'X day' .... Although these clusters could have been easily identified by using morphological features, they are good candidates for a first proof of concept of our approach. Furthermore, we have gained satisfactory clusters of relational adjectives derived from concepts such as countries, languages, religions, cities, or territories.

A final example will be discussed in order to show the strengths and weaknesses of the approach. We get a cluster consisting of twenty adjectives describing properties of curves of some measure (temperature, price, ...), namely

5911 [*gleichbleibend* 'stable', *nachlassend* 'decreasing', *verringert* 'reduced', *gesunken* 'fallen']

5297 [*abnehmend* 'decreasing']

3163 [*vermindert* 'reduced', *vermehrt* 'increased']

2662 [*gesteigert* 'increased', *erhöht* 'raised']

1939 [*steigend* 'climbing']

1168 [*zunehmend* 'increasing / more and more', *wachsend* 'growing']

4541 [*höchstmöglich* 'highest possible']

3980 [*maximal* 'maximal', *größtmöglich* 'biggest possible']

5347 [*rückläufig* 'declining']

4687 [*sinkend* 'sinking', *gestiegen* 'climbed']

3331 [*stagnierend* 'stagnating', *schrumpfend* 'shrinking']

Most of the adjectives in this cluster denote some form of 'increasing', 'decreasing' or 'staying stable', thus a dynamic progression along a curve, and can be interpreted as values of the attribute 'progression' applied to curves. Others like *vermindert* 'decreased' or *vermehrt* 'increased' describe the general height of the curve. Only in subcluster 4541 one finds adjectives denoting extreme points of a curve like 'maximum' or 'highest possible'. Although the whole cluster 5911 is quite satisfactory, it could be improved in two respects. First, it obviously lacks many antinomies like *minimal* 'minimal' (to *maximal* 'maximal') or *fallend* 'decreasing' (to *steigend* 'increasing') which have been clustered elsewhere. Other adjectives describing the curve progression like *schwankend* 'floating' are missing as well. Second, from our frame-based perspective it would be desirable to receive clusters that reflect the different attributes by which a curve can be described like 'progression' and 'height' and which strictly separates adjectives that specify curves from others that specify points on a curve.

5 Conclusion

We have presented an iterative bidirectional clustering of adjectives and nouns by co-occurrences. The aim has been to derive adjective clusters which correspond to the value spaces of attributes. It has turned out that only some of the received clusters are perfect in that respect. Most miss some adjectives or mix adjectives belonging to different by familiar attributes. However, it has turned out that the clusters are a useful starting point for a manual analysis of the attribute value space.

In the quantitative evaluation the presented iterative bidirectional clustering outperformed the k-means clustering on word vectors. That indicates that for our task, the approach of only looking at individual adjective noun pairs instead of adjectives in bigger contexts is promising. By iteratively clustering adjectives on the basis of co-occurring nouns and vice versa, the hidden attributes connecting both can be crystallized out. However, bigger gold clusterings and more reliable evaluation measures are still missing.

The next steps to be taken are the following: The algorithm should be adapted to allow for overlapping clusters in order to account for polysemy. Better evaluation measures and more gold clusters which are specific to the task are needed. Ideally one would develop a frame-specific evaluation measure. One way could be to automatically induce frames from the derived adjective and noun clusters and to evaluate the resulting frames.

References

M. Baroni and R. Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*. Boston, pages 1183–1193.

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

William Blacoe and Mirella Lapata. 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.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *EMNLP-CoNLL*. pages 1455–1465.

Gemma Boleda, Toni Badia, and Eloi Batlle. 2004. Acquisition of semantic classes for adjectives from distributional evidence. In *Proceedings of the 20th international conference on Computational Linguistics*. page 1119.

Gemma Boleda, Marco Baroni, Louise McNally, and Nghia Pham. 2013. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Workshop on Computational Semantics*. pages 35–46.

Jos de Bruin and Remko Scha. 1988. The interpretation of relational nouns. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '88, pages 25–32.

Duden. 2004. *Duden - Die deutsche Rechtschreibung*. Bibliographisches Institut & F. A. Brockhaus AG Mannheim, 23rd edition. Electronic version.

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

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Emiliano Guevara. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics*. pages 33–37.

Birgit Hamp and Helmut Feldweg. 1997. Germanet - a lexical-semantic net for german. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

M. Hartung and A. Frank. 2010. A structured Vector Space Model for hidden attribute meaning in adjective-noun phrases. In *Proceedings of the 23rd International Conference on Computational Linguistics*. pages 430–438.

Matthias Hartung and Anette Frank. 2011. Exploring supervised LDA models for assigning attributes to adjective-noun phrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. pages 540–551.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT - the GermaNet editing tool. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. pages 2228–2235.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification* 2(1):193–218.

Sebastian Löbner. 2011. Concept types and determination. *Journal of Semantics* 28(3):279–333.

Marina Meila. 2007. Comparing clusterings – an information based distance. *Journal of Multivariate Analysis* 98:873–895.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, , Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014):176–182.

Tomáš Mikolov, Anoop Deoras, Daniel Povey, Lukáš Burget, and Jan Černockỳ. 2011. Strategies for training large scale neural network language models. In *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. pages 196–201.

Marcin Morzycki. 2015. *Modification*. Cambridge University Press.

Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the Google Books Corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE* 10(10).

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

W.M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66:846–850.

Peter F Schulam and Christiane Fellbaum. 2010. Automatically determining the semantic gradation of German adjectives. In *KONVENS*. pages 163–167.

Diarmuid Séaghdha and Anna Korhonen. 2014. Probabilistic distributional semantics with latent variable models. *Compututaional Linguistics* 40(3):587–631.

R. Tibshirani, G. Walther, and T. Hastie. 2001. Estimating the number of data clusters via the Gap statistic. *Journal of the Royal Statistical Society B* 63:411–423.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37(1):141–188.

E.M. Vecchi, M. Marelli, R. Zamparelli, and M. Baroni. in press. Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive Science.* .