Unsupervised Induction of Compositional Classes for English Adjective-Noun Pairs

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 examines how adjectival modification classes can be detected by applying unsupervised methods on adjective-noun co-occurrence data. It evaluates a k-means baseline, two graphical models, and a recently introduced bidirectional clustering algorithm against HeiPLAS, a manually annotated gold standard for hidden modificational classes. The paper shows that the bidirectional clustering algorithm performs best on this task, and discusses how the results of the unsupervised approaches can be employed for building a frame-based inventory of adjectival modification.

1 Introduction

Recent years have witnessed an increasing interest in computational models of compositionality that operate on vector-space representations (Baroni and Zamparelli, 2010; Guevara, 2010) or neural word embeddings (Socher et al., 2011; Mikolov et al., 2013). While many of these studies deal with the problem of predicting the distribution of compound phrases from the distributions of their elements, comparatively few of them examine which classes of compositional mechanisms are reflected by the learned models. The present paper focusses on adjectival modification and aims at uncovering adjective clusters with common modificational patterns.

The approach followed in this paper is frame-based. We assume that the meaning of a noun concept is represented by recursive attribute-value structures (Pustejovsky, 1995; Petersen, 2007; Löbner, 2014). Adjectives operate on these structures by (i) simply specifying the value of a single attribute or by (ii) enriching the concept structure by additional attributes and constraints. Examples of (i) are most property denoting adjectives with an intersective reading like 'red ball', where the adjective restricts the value range of the attribute COLOR to the value 'red'. Note that the noun concept may offer more than one adequate attribute like in 'red pen', where 'red' either specifies the color of the pen or the color of the writing of the pen. Examples of (ii) are, among many others, relational A+N phrases such as 'presidential speech' (not every speech given by the current president is presidential, i.e., a speech given at a private party[1]) and modal A+N phrases such as 'fake meat' (fake meat is no meat and lacks important attributes of meat).

Although linguistic research has proposed various classification schemes for adjectives, none of them reflects the modificational patterns of a frame-based approach, to our knowledge. They are either constructed from a lexicographic perspective, and aim at capturing fine-grained meaning distinctions (e.g., Hundsnurscher and Splett, 1982) that do not result in different modificational patterns from a frame perspective; or they focus on the semantics of adjective modification, but propose only a coarse distinction into major patterns like 'property' or 'relational' (e.g. Dixon, 2010). Furthermore, only a few of the

[1] See Anderson and Löbner (2017) for a frame based approach to relational adjectives.

proposed classification schemes are based on empirical evidence from larger corpora (e.g., Raskin and Nirenburg, 1995). As none of the proposed linguistic classifications is based on or suited for a frame perspective on adjectival modification, we intend to provide empirical evidence for adjectival classifications without resorting to linguistic theories, and test which quantitative methods are most suited for the unsupervised induction of compositional classes.

We hypothesize that the selectional restrictions of A+N pairs result from the attribute-value structures of the involved concepts, and that a clustering based on such A+N pairs may provide insights into these structures. We therefore focus on four unsupervised models that are able to cluster adjectives on the basis of the nouns they modify, and to induce latent compositional classes from A+N co-occurrence in this way. Apart from a baseline k-means clustering, we test two graphical models (Rooth et al., 1999; Séaghdha and Korhonen, 2014) and a recently proposed bidirectional clustering algorithm (Petersen and Hellwig, 2016). The graphical and bidirectional models were chosen, because they have proven to be effective tools in former studies on compositionality. The resulting clusters are compared with a gold data set that consists of attribute-adjective-noun triples and was originally designed for a semi-supervised attribute assignment task (HeiPLAS; Hartung, 2015). The data set is not fully appropriate for our task, as it only covers modificational patterns of type (i), i.e., adjectives specifying attribute values in noun frames. In spite of this restriction, it will turn out that bidirectional clustering slightly outperforms the k-means baseline and the two graphical models that were designed for the unsupervised induction of latent semantic classes.

The rest of the paper is organized as follows. Section 2 gives an overview of related research in NLP. Section 3 describes the training corpus, its preprocessing, and the structure of the evaluation data set. Section 4 sketches the applied models, and Section 5 presents an evaluation of the results.

2 Related Research

One important track of research in the field of computational approaches to semantic composition deals with predicting composed expressions on the basis of their atoms, either using vector space models (Mitchell and Lapata, 2008; Baroni and Zamparelli, 2010) or, more recently, relying on pretrained word embeddings (Socher et al., 2011; Dima, 2016). A group of studies with a stronger linguistic motivation examines which or how many compositional mechanisms are active in forming composed expressions (Tratz and Hovy, 2010; Hartung, 2015), and how the active mechanisms can be predicted using machine learning techniques (Hartung and Frank, 2011; Dima and Hinrichs, 2015). Most relevantly for our task, Hartung et al. (2017) aim at predicting hidden attributes of A+N phrases. The authors train a flat neural network that learns an embedded representation of an A+N phrase from the word embeddings of its components. The trained model is applied to the HeiPLAS data, and the hidden attribute is predicted using nearest neighbor search with the learned compositional embedding.

3 Data

A+N pairs are obtained by parsing the 2013 English news dump from `www.statmt.org` (Bojar et al., 2014) using the Stanford CoreNLP dependency parser (Manning et al., 2014), and extracting the lemmas of all sequences of the form JJ-[NN|NE|NNS]. By this method, we obtain 1,048,653 A+N pairs with 8,392 adjective lexemes, 17,560 noun lexemes, and 430,256 unique combinations of A+N; the density of the co-occurrence matrix is 0.002919691.

We use the HeiPLAS data set (Hartung, 2015) for evaluating the received adjective clusters. The design of HeiPLAS is linguistically motivated by a classification into basic, event- and object-related adjectives introduced by Boleda (2007) (refer to Hartung (2015, 57–59) for a short overview). Adjectives and the attributes they modify are extracted from the anchor structure of WordNet, and ambiguous cases were manually validated by a group of native speakers (Hartung, 2015, 98ff.). Therefore, the set of attributes available in HeiPLAS primarily reflects the linguistic intuition and the lexicographic aim underlying WordNet. The data set consists of 1,598 triples of the form (attribute, adjective, noun), such

as (VOLUME, big, voice) or (BEAUTY, repulsive, mask) with 849 distinct adjectives, 923 distinct nouns, and 253 distinct attributes. Due to its structure, the coverage of HeiPLAS is limited to basic adjectives, which denote attribute values of nouns ("blue car" for COLOR, "big house" for SIZE). The attributes in HeiPLAS are derived from WordNet attribute nouns, and were validated in a manual classification step. In order to use HeiPLAS as a gold standard for the evaluation of clusters of adjectives, we interpret the attributes as class labels. Thus, for each triple *(adj,attr,noun)* in HeiPLAS, we classify the adjective *adj* as belonging to the class *attr*.

4 Models

The paper compares four unsupervised models that can be used to cluster adjectives on the basis of the nouns they modifiy. **Rooth-LDA**[2] denotes a graphical model of compositionality that interprets each A+N pair as an independent observation (Rooth et al., 1999). During generation, the model samples a hidden topic variable z. Subsequently, adjective and noun distributions are sampled based on the value of z, resulting in the following joint probability of adjective a and noun n (Séaghdha and Korhonen, 2014, 602):[3]

$$p(a,n) = \sum_z p(a|z)p(n|z)p(z)$$

We use the Gibbs sampler described in Séaghdha and Korhonen (2014, 605–606) for training the model. The symmetric priors of the distributions are estimated using a grid search on held-out data, in order to maximize the Adjusted Rand Index (Hubert and Arabie, 1985) of the detected adjective clusters when compared with the gold standard provided by HeiPLAS (details in Section 5). The grid search produces the following values: $\alpha = 0.001, \beta = 0.01, \gamma = 0.1$. It is important to note that the chosen hyperparameter optimization method feeds in an element of supervision to the graphical model, because its parameters are tuned to reproduce the HeiPLAS classes. We will come back to this point in the evaluation.

Lex-LDA splits the generative process for an A+N pair into two separate subprocesses in order to deal with lexicalized and truly compositional A+N pairs (Séaghdha and Korhonen, 2014, 603-606). While the association between adjective and noun is modelled through their conditional probability in lexicalization mode, the model uses a topic model in non-lexicalized or compositional mode. The decision between lexicalized and compositional mode is directed by an adjective specific parameter σ_a that is learned from the data along with the model:

$$p(n|a) = \underbrace{\sigma_a p_{lex}(n|a)}_{\text{lexicalized}} + \underbrace{(1-\sigma_a)\sum_z p(n|z)p(z|a)}_{\text{compositional}}$$

We implement the Gibbs sampler described in Séaghdha and Korhonen (2014, 606). As noted in Séaghdha and Korhonen (2014, 605), this model poses the potential problem that hidden topics are not generated when the lexicalized subprocess is active. The symmetric priors are estimated in the same way as for Rooth LDA, above, resulting in $\alpha = 0.001, \beta = 0.1, \gamma = 0.01$. Using the gap statistics (Tibshirani et al., 2001) on pretrained GloVe vectors (Pennington et al., 2014) for English adjectives, we choose $k = 250$ hidden topics for training both graphical models.

Contrary to the two graphical models, the bidirectional clustering (**BidirClus**, Petersen and Hellwig (2016)) operates on a vector space matrix (VSM; Turney and Pantel, 2010) constructed from adjectives (rows) and the nouns they co-occur with (columns). In BidirClus, adjectives are first clustered on the basis of nouns they co-occur with. The centers of clusters detected in this way replace the adjective rows in the VSM, and the same process is repeated for the transposed VSM, such that nouns are clustered on the basis of adjectives in the second step. This process is repeated until no further clusters can be constructed from the semantic space.

[2]This model was first described in Rooth et al. (1999). We adhere to the naming proposed in Séaghdha (2010).

[3]As we are only interested in adjectives and nouns, we slightly adapted the notation found in Séaghdha and Korhonen (2014).

More specifically, the distributional distance between items in rows r_i and r_j, which are adjectives in the first step, is measured using their Jaccard distance:

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

In addition, an LDA topic model (Blei et al., 2003) with $k = 15$ hidden topics is built from the VSM. The topic similarity θ_{ij} of rows r_i and r_j is calculated from the Θ values of the LDA model:

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

The final similarity score d_{ij}^{lda} of rows r_i and r_j is calculated as the product of Jaccard distance and topic similarity:

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

In each iteration of BidirClus, new clusters are created from the 5% of rows that have the highest combined similarity scores d_{ij}^{lda}. New clusters R are constructed from transitive closures detected in the top 5%, and their distributional representation $\vec{R}$ is calculated using the following majority function:

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

The respective rows of the matrix are replaced by the new distributional representation $\vec{R}$, which means that the number of rows in the VSM is reduced by $|R| - 1$ in this step. It should be noted that BidirClus does not allow for multiple semantic readings of adjectives or nouns, because each item is assigned to a single cluster while building the transitive closure.[4]

As a baseline model for comparison, we use k-means clustering with the expected number of clusters set to $k = 250$.

5 Evaluation and Results

We evaluate the outcomes of the three models against HeiPLAS using Rand Index (RI; Rand, 1971) and Adjusted Rand Index (ARI; Hubert and Arabie, 1985). The two graphical models (Rooth-LDA, Lex-LDA) generate a hidden class variable z for each A+N pair, which denotes the (anonymous) compositional mechanism activated for this pair.[5] Because we are interested in clustering A's on the basis of their modificational properties, we operate directly with the sampled values of this class variable z. We average the counts of z for each A+N pair at every 20th iteration after a burn-in period of 100 iterations. Using z for labeling A+N pairs provides an elegant solution for clustering polysemous A's, because a single A can be labeled with different values of z, depending on which N it is combined with. However, the bidirectional clustering algorithm does not differentiate between semantic readings of the same A, but assigns each A to a single class. In order to provide a fair comparison of the three models, the evaluation of the two graphical models is broken down to the A level: Given the adjective set $\mathcal{A}$, noun set $\mathcal{N}$ and set of hidden topics $\mathcal{Z}$, let lda : $\mathcal{A} \times \mathcal{N} \rightarrow \mathcal{Z}$ be the partial function that models the topic assignment of LDA. We define $Z_A = \arg\max_{Z \in \mathcal{Z}} |\{N \in \mathcal{N} : \text{lda}(A, N) = Z\}|$ and assign each A to the class corresponding to Z_A.

[4] We are currently testing a new version of BidirClus that accounts for multiple readings by using a more flexible clustering strategy than transitive closure.

[5] For details on z see Figure 2 in Séaghdha and Korhonen (2014, 602) for Rooth-LDA, and Figure 3 in Séaghdha and Korhonen (2014, 603) for Lex-LDA. Note that for Lex-LDA the value of z is only defined if the model is in the class-based mode ($s_i = 0$).

Model	RI	ARI
k-means	0.9352	0.1778
Rooth-LDA (uninf.)	0.9668	0.3007
Rooth-LDA (optim.)	0.9669	0.3208
Lex-LDA (uninf.)	0.9719	0.3097
Lex-LDA (optim.)	0.9715	0.3150
BidirClus, 'leaf'	**0.9762**	**0.3472**
BidirClus, 'topmost'	0.9688	0.3215

Table 1: Evaluation of the four clustering models on the HeiPLAS dataset; uninf.: model uses an uninformed symmetric prior; optim: model uses a symmetric prior optimized on a held-out set of A+N pairs. RI: Rand Index; ARI: Adjusted Rand Index

BidirClus produces deeply nested, hierarchical clusterings that need to be transformed into hard, non-hierarchical clusterings for comparison with the other models. Following Petersen and Hellwig (2016), we report results for two brute force evaluation modes. In the mode 'topmost', each adjective obtains the label of the largest cluster containing it, and in the mode 'leaf' it obtains the label of the smallest non-singleton cluster containing it.

Table 1 reports Rand Index (RI) and Adjusted Rand Index (ARI) for the k-means baseline and the three more specialized methods.[6] Both graphical models and the bidirectional clustering clearly outperform k-means in terms of RI and ARI, while they generate similar coefficients when compared to each other. As noted above, however, the hyperparameters of Rooth-LDA and Lex-LDA were optimized using HeiPLAS data with the objective of maximizing their ARI on the HeiPLAS data set. To compare the influence of this optimization on model performance, we have repeated the training of Rooth-LDA and Lex-LDA with uninformed symmetric priors of 0.01 (settings marked with 'uninf.'). The respective results in Table 1 show that hyperparameter optimization gives both models a slight, but noticeable advantage over the uninformed versions. If one compares only the unsupervised models, the uninformed versions of Rooth-LDA and Lex-LDA are outperformed by BidirClus.

A closer look at the HeiPLAS gold standard reveals that the evaluated models may actually perform better than expressed by the values in Table 1. In several cases, the models capture frame semantic properties of adjectives that were not relevant for designing and building the HeiPLAS data set. The adjectives 'exceptional' and 'extraordinary', for example, are grouped into one cluster by BidirClus, while HeiPLAS labels 'exceptional' with the attribute COMMONNESS and 'extraordinary' with ORDINARINESS ('extraordinary beauty', 'extraordinary capacity') or MODERATION ('extraordinary desire'), as can be seen in the following list of HeiPLAS attributes and their associated adjectives:

COMMONNESS uncommon, common, special, exceptional, popular, rare, average

ORDINARINESS ordinary, remarkable, extraordinary, average, everyday, routine

MODERATION immoderate, reasonable, intermediate, extreme, moderate, modest, conservative, over-the-top, abnormal, average, exaggerated, excessive, exorbitant, extraordinary, outrageous, unreasonable

Looking into traditional lexicons, Merriam-Webster describes 'exceptional' by (a) forming an exception, (b) better than average, and (c) deviating from the norm and 'extraordinary' by (a) going beyond what is usual, regular, or customary and (b) exceptional to a very marked extent.[7] Although there is a subtle lexicographic distinction between 'extraordinary' and 'exceptional', they are often used synonymously in actual language use, and are even explicitly marked as synonyms by the Merriam-Webster.

[6]The Rand Index measures the probability that the automatic clustering agrees with the HeiPLAS classes on a randomly chosen pair of adjectives. The Adjusted Rand Index is a version of the Rand Index that is corrected for chance. As the values of the latter are less easy to interpret (they may take negative values), we have decided to present both indices.

[7]www.merriam-webster.com

A further example of an adjective pair that is not contained in one attribute cluster in HeiPLAS, but grouped together by BidirClus is 'current' and 'future'. HeiPLAS labels 'current' with the attribute CURRENTNESS and 'future' with TIMING. From a frame perspective, both adjectives specify the temporal relation of the modified noun and the utterance time ('current prize' versus 'future prize'). Related considerations apply to the pair 'impossible' and 'tricky', where HeiPLAS labels 'impossible' and its opposite 'possible' with POSSIBILITY, and 'tricky' with DIFFICULTY (along with adjectives such as 'troublesome' or 'challenging'). In a frame-based approach one reading of 'impossible' would place the adjective on the maximum on the difficulty scale.

On the whole, we could identify 32 pairs of adjectives that are, from a frame perspective, grouped correctly by the bidirectional clustering, but assigned to different groups in HeiPLAS. The main reason for the divergence between HeiPLAS and the unsupervised approach is the fact that HeiPLAS is derived from WordNet. WordNet focusses on modelling fine grained lexicographic distinctions, which may not necessarily be reflected in actual language use (Navigli, 2009). In the cases just described, BidirClus seems to succeed in grouping adjectives that belong to the same value ranges from a less fine grained, frame semantic perspective.

If one looks at the cases where pairs of adjectives are grouped in one class in HeiPLAS, but end up in different classes by BidirClus, the main weakness of the current implementation of BidirClus becomes obvious. Because BidirClus clusters into disjoint classes, meaning variants of adjectives are not captured. While, for example, 'right' is found in the same class APPROPRIATENESS as 'appropriate' and 'inappropriate' in HeiPLAS, BidirClus clusters 'right' together with 'left' in a group consisting of directional adjectives.

The results indicate two main lines of future research. First, BidirClus must be adapted in order to account for meaning variants, either by a preprocessing step that disambiguates adjectives and nouns, or by a less strict update procedure of the co-occurrence matrix. Second, a more dedicated gold standard is required. This gold standard should be based on a frame semantic perspective of A+N modification. It should provide data not only for attribute value specifying adjectives such as HeiPLAS, but for more complex frame modification patterns as triggered by relational or modal adjectives (see Section 1). Such a gold standard will serve for tuning and optimizing unsupervised models, as described in this paper. In addition, the gold standard itself and improved models derived from it will provide more fine grained empirical evidence for linguistic research in modificational structures of adjectives and nouns.

Acknowledgment

This work was supported by the DFG Collaborative Research Centre 991, "The Structure of Representations in Language, Cognition, and Science", University of Düsseldorf. In addition, we would like to thank the anonymous reviewers for their comments on the paper.

References

Anderson, C. and S. Löbner (2017). Roles and the lexical semantics of role-denoting relational adjectives. Abstract submitted to 12th International Tbilisi Symposium on Logic, Language, and Computation.

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

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993–1022.

Bojar, O., C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58.

Boleda, G. (2007). *Automatic Acquisition of Semantic Classes for Adjectives*. Ph. D. thesis, Pompeu Fabra University.

Dima, C. (2016). On the compositionality and semantic interpretation of English noun compounds. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 27–39.

Dima, C. and E. Hinrichs (2015). Automatic noun compound interpretation using deep neural networks and word embeddings. In *Proceedings of the 11th International Conference on Computational Semantics*, pp. 173–183.

Dixon, R. W. (2010). Where have all the adjectives gone? In *Where have All the Adjectives Gone? And Other Essays in Semantics and Syntax*, pp. 1–62. Berlin: De Gruyter Mouton.

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

Hartung, M. (2015). *Distributional Semantic Models of Attribute Meaning in Adjectives and Nouns*. Ph. D. thesis, University of Heidelberg.

Hartung, M. and A. 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*, pp. 540–551.

Hartung, M., F. Kaupmann, S. Jebbara, and P. Cimiano (2017). Learning compositionality functions on word embeddings for modelling attribute meaning in adjective-noun phrases. In *Proceedings of the 15th Meeting of the European Chapter of the ACL (EACL)*.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of classification 2*(1), 193–218.

Hundsnurscher, F. and J. Splett (1982). *Semantik der Adjektive im Deutschen: Analyse der semantischen Relationen*. Westdeutscher Verlag.

Löbner, S. (2014). Evidence for frames from human language. In T. Gamerschlag, D. Gerland, R. Osswald, and W. Petersen (Eds.), *Frames and Concept Types*, Volume 94 of *Studies in Linguistics and Philosophy*, pp. 23–67. Springer International Publishing.

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky (2014). The Stanford CoreNLP natural language processing toolkit. In *ACL System Demonstrations*, pp. 55–60.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Mitchell, J. and M. Lapata (2008). Vector-based models of semantic composition. In *Proceedings of 46th Annual Meeting of the ACL*, pp. 236–244.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Comput. Surv. 41*, 10:1–10:69.

Pennington, J., R. Socher, and C. D. Manning (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 EMNLP*, pp. 1532–1543.

Petersen, W. (2015/2007). Representation of concepts as frames. In T. Gamerschlag, D. Gerland, R. Osswald, and W. Petersen (Eds.), *Meaning, Frames, and Conceptual Representation*, Volume 2 of *Studies in Language and Cognition*, pp. 43 – 67. Düsseldorf University Press. Reprint with comments. Originally published 2007 in The Baltic International Yearbook of Cognition, Logic and Communication, Vol. 2.

Petersen, W. and O. Hellwig (2016). Exploring the value space of attributes: Unsupervised bidirectional clustering of adjectives in German. In *Proceedings of the COLING*, pp. 2839–2848.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

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

Raskin, V. and S. Nirenburg (1995). Lexical semantics of adjectives. *New Mexico State University, Computing Research Laboratory Technical Report, MCCS-95-288*.

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

Séaghdha, D. and A. Korhonen (2014). Probabilistic distributional semantics with latent variable models. *Computational Linguistics 40*(3), 587–631.

Séaghdha, D. O. (2010). Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the ACL*, pp. 435–444.

Socher, R., C. C. Lin, C. Manning, and A. Y. Ng (2011). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 129–136.

Tibshirani, R., 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.

Tratz, S. and E. Hovy (2010). A taxonomy, dataset, and classifier for automatic noun compound interpretation. In *Proceedings of the 48th Annual Meeting of the ACL*, pp. 678–687.

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