Learning Semantics and Selectional Preference of Adjective-Noun Pairs

Karl Moritz Hermann
Department of Computer Science
University of Oxford
Oxford OX1 3QD, UK
karl.moritz.hermann@cs.ox.ac.uk

Chris Dyer
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
cdyer@cs.cmu.edu

Phil Blunsom
Department of Computer Science
University of Oxford
Oxford OX1 3QD, UK
phil.blunsom@cs.ox.ac.uk

Stephen Pulman
Department of Computer Science
University of Oxford
Oxford OX1 3QD, UK
stephen.pulman@cs.ox.ac.uk

Abstract

We investigate the semantic relationship between a noun and its adjectival modifiers. We introduce a class of probabilistic models that enable us to to simultaneously capture both the semantic similarity of nouns and modifiers, and adjective-noun selectional preference. Through a combination of novel and existing evaluations we test the degree to which adjective-noun relationships can be categorised. We analyse the effect of lexical context on these relationships, and the efficacy of the latent semantic representation for disambiguating word meaning.

1 Introduction

Developing models of the meanings of words and phrases is a key challenge for computational linguistics. Distributed representations are useful in capturing such meaning for individual words (Sato et al., 2008; Maas and Ng, 2010; Curran, 2005). However, finding a compelling account of semantic *compositionality* that utilises such representations has proven more difficult and is an active research topic (Mitchell and Lapata, 2008; Baroni and Zamparelli, 2010; Grefenstette and Sadrzadeh, 2011). It is in this area that our paper makes its contribution.

The dominant approaches to distributional semantics have relied on relatively simple frequency counting techniques. However, such approaches fail to generalise to the much sparser distributions encountered when modeling compositional processes and provide no account of selectional preference. We propose a probabilistic model of the semantic representations for nouns and modifiers. The foundation of this model is a latent variable representation of noun and adjective semantics together with their compositional probabilities. We employ this formulation to give a dual view of noun-modifier semantics: the induced latent variables provide an explicit account of selectional preference while the marginal distributions of the latent variables for each word implicitly produce a distributed representation.

Most related work on selectional preference uses class-based probabilities to approximate (sparse) individual probabilities. Relevant papers include Ó Séaghdha (2010), who evaluates several topic models adapted to learning selectional preference using co-occurence and Baroni and Zamparelli (2010), who represent nouns as vectors and adjectives as matrices, thus treating them as functions over noun meaning. Again, inference is achieved using co-occurrence and dimensionality reduction.

2 Adjective-Noun Model

We hypothesize that *semantic classes* determine the semantic characteristics of nouns and adjectives, and that the distribution of either with respect to other components of the sentences they occur in is also mediated by these classes (i.e., not by the words themselves). We assume that in general nouns select for adjectives,[1] and that this selection is dependent on both their latent semantic classes. In the next section, we describe a model encoding our hypotheses.

2.1 Generative Process

We model a corpus $\mathcal{D}$ of tuples of the form $(n, m, c_1 \dots c_k)$ consisting of a noun *n*, an adjective *m* (modifier), and k words of context. The context variables $(c_1 \dots c_k)$ are treated as a bag of words and

[1] We evaluate this hypothesis as well as its inverse.

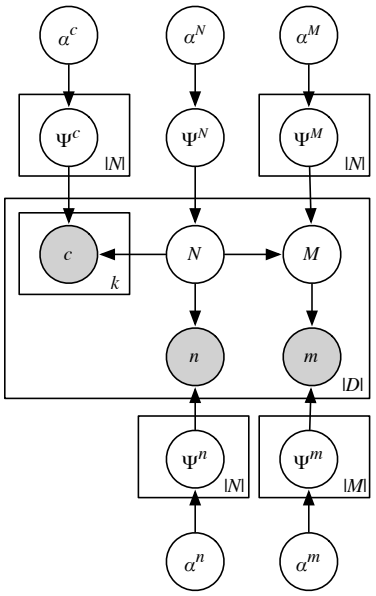

Figure 1: Plate diagram illustrating our model of noun and modifier semantic classes (designated N and M, respectively), a modifier-noun pair (m,n), and its context.

include the words to the left and right of the noun, its siblings and governing verbs. We designate the vocabulary V_n for nouns, V_m for modifiers and V_c for context. We use z_i to refer to the i^{th} tuple in $\mathcal{D}$ and refer to variables within that tuple by subscripting them with i, e.g., n_i and $c_{3,i}$ are the noun and the third context variable of z_i. The latent noun and adjective class variables are designated N_i and M_i.

The corpus $\mathcal{D}$ is generated according to the plate diagram in figure 1. First, a set of parameters is drawn. A multinomial Ψ^{N} representing the distribution of noun semantic classes in the corpus is drawn from a Dirichlet distribution with parameter α^{N}. For each noun class i we have distributions Ψ_i^{M} over adjective classes, Ψ_i^{n} over V_n and Ψ_i^{c} over V_c, also drawn from Dirichlet distributions. Finally, for each adjective class j, we have distributions Ψ_j^{m} over V_m.

Next, the contents of the corpus are generated by first drawing the length of the corpus (we do not parametrise this since we never generate from this model). Then, for each i, we generate noun class N_i, adjective class M_i, and the tuple z_i as follows:

$$
\begin{aligned}
N_i \mid \Psi^{\text{N}} &\sim \text{Multi}(\Psi^{\text{N}}) \\
M_i \mid \Psi^{\text{M}}_{N_i} &\sim \text{Multi}(\Psi^{\text{M}}_{N_i}) \\
n_i \mid \Psi^{\text{n}}_{N_i} &\sim \text{Multi}(\Psi^{\text{n}}_{N_i}) \\
m_i \mid \Psi^{\text{m}}_{M_i} &\sim \text{Multi}(\Psi^{\text{m}}_{M_i}) \\
\forall k\text{: } c_{k,i} \mid \Psi^{\text{c}}_{N_i} &\sim \text{Multi}(\Psi^{\text{c}}_{N_i})
\end{aligned}
$$

2.2 Parameterization and Inference

We use Gibbs sampling to estimate the distributions of N and M, integrating out the multinomial parameters Ψ^x (Griffiths and Steyvers, 2004). The Dirichlet parameters $\boldsymbol{\alpha}$ are drawn independently from a $\Gamma(1,1)$ distribution, and are resampled using slice sampling at frequent intervals throughout the sampling process (Johnson and Goldwater, 2009). This "vague" prior encourages sparse draws from the Dirichlet distribution. The number of noun and adjective classes $\mathcal{N}$ and $\mathcal{M}$ was set to 50 each; other sizes (100,150) did not significantly alter results.

3 Experiments

As our model was developed on the basis of several hypotheses, we design the experiments and evaluation so that these hypotheses can be examined on their individual merit. We test the first hypothesis, that nouns and adjectives can be represented by semantic classes, recoverable using co-occurence, using a sense clustering evaluation by Ciaramita and Johnson (2003). The second hypothesis, that the distribution with respect to context and to each other is governed by these semantic classes is evaluated using pseudo-disambiguation (Clark and Weir, 2002; Pereira et al., 1993; Rooth et al., 1999) and bigram plausibility (Keller and Lapata, 2003) tests.

To test whether noun classes indeed select for adjective classes, we also evaluate an inverse model (Mod_i), where the adjective class is drawn first, in turn generating both context and the noun class. In addition, we evaluate copies of both models ignoring context (Mod_{nc} and Mod_{inc}).

We use the British National Corpus (BNC), training on 90 percent and testing on 10 percent of the corpus. Results are reported after 2,000 iterations including a burn-in period of 200 iterations. Classes are marginalised over every 10th iteration.

4 Evaluation

4.1 Supersense Tagging

Supersense tagging (Ciaramita and Johnson, 2003; Curran, 2005) evaluates a model's ability to cluster words by their semantics. The task of this evaluation is to determine the WORDNET supersenses of a given list of nouns. We report results on the WN1.6 test set as defined by Ciaramita and Johnson (2003), who used 755 randomly selected nouns with a unique supersense from the WORDNET 1.6

corpus. As their test set was random, results weren't exactly replicable. For a fair comparison, we select all suitable nouns from the corpus that also appeared in the training corpus. We report results on type and token level (52314 tokens with 1119 types). The baseline[2] chooses the most common supersense.

	k	Token	Type
Baseline		.241	.210
Ciaramita & Johnson		.523	.534
Curran		-	**.680**
Mod	10	**.592**	.517
Mod_{nc}	10	.473	.410

Table 1: Supersense evaluation results. Values are the percentage of correctly assigned supersenses. k indicates the number of nearest neighbours considered.

We use cosine-similarity on the marginal noun class vectors to measure distance between nouns. Each noun in the test set is then assigned a supersense by performing a distance-weighted voting among its k nearest neighbours. Results of this evaluation are shown in Table 1, with Figure 2 showing scores for model Mod across different values for k.

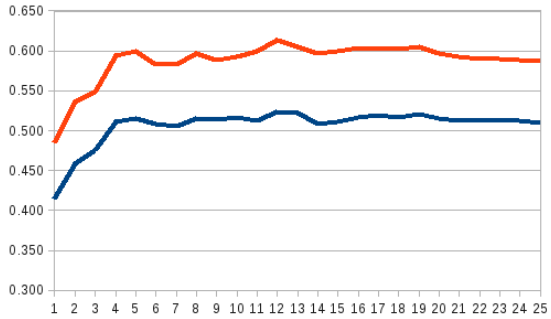

Figure 2: Scores of Mod on the supersense task. The upper line denotes token-, the lower type-level scores. The y-axis is the percentage of correct assignments, the x-axis denotes the number of neighbours included in the vote.

The results demonstrate that nouns can semantically be represented as members of latent classes, while the superiority of Mod over Mod_{nc} supports our hypothesis that context co-occurence is a key feature for learning these classes.

4.2 Pseudo-Disambiguation

Pseudo-disambiguation was introduced by Clark and Weir (2002) to evaluate models of selectional preference. The task is to select the more probable of two candidate arguments to associate with a given predicate. For us, this is to decide which adjective, a_1 or a_2, is more likely to modify a noun n.

We follow the approach by Clark and Weir (2002) to create the test data. To improve the quality of the data, we filtered using bigram counts from the Web1T corpus, setting a lower bound on the probable bigram (a_1, n) and chosing a_2 from five candidates, picking the lowest count for bigram (a_2, n).

We report results for all variants of our model in Table 2. As baseline we use unigram counts in our training data, chosing the more frequent adjective.

L-bound	0	100	500	1000
Size	5714	5253	3741	2789
Baseline	.543	.543	.539	.550
Mod	**.783**	**.792**	**.810**	**.816**
Mod_i	.781	.787	.800	.810
Mod_{nc}	.720	.728	.746	.750
Mod_{inc}	.722	.730	.747	.752

Table 2: Pseudo-disambiguation: Percentage of correct choices made. L-bound denotes the Web1T lower bound on the (a_1, n) bigram, size the number of decisions made.

While all models decisively beat the baseline, the models using context strongly outperform those that do not. This supports our hypothesis regarding the importance of context in semantic clustering.

The similarity between the normal and inverse models implies that the direction of the noun-adjective relationship has negligible impact for this evaluation.

4.3 Bigram Plausibility

Bigram *plausibility* (Keller and Lapata, 2003) is a second evaluation for selectional preference. Unlike the frequency-based pseudo-disambiguation task, it evaluates how well a model matches human judgement of the plausibility of adjective-noun pairs. Keller and Lapata (2003) demonstrated a correlation between frequencies and plausibility, but this does not sufficiently explain human judgement. An example taken from their *unseen* data set illustrates the dissociation between frequency and plausibility:

- Frequent, implausible: "educational water"
- Infrequent, plausible: "difficult foreigner"[3]

The plausibility evaluation has two data sets of 90 adjective-noun pairs each. The first set (*seen*) contains random bigrams from the BNC. The second set (*unseen*) are bigrams not contained in the BNC.

[2]The baseline results are from Ciaramita and Johnson (2003). Using the majority baseline on the full test set, we only get .176 and .160 for token and type respectively.

[3]At the time of writing, Google estimates 56,900 hits for "educational water" and 575 hits for "difficult foreigner". "Educational water" ranks bottom in the gold standard of the *unseen* set, "difficult foreigner" ranks in the top ten.

Recent work (Ó Séaghdha, 2010; Erk et al., 2010) approximated plausibility with joint probability (JP). We believe that for semantic *plausibility* (not *probability*!) mutual information (MI), which factors out acutal frequencies, is a better metric.[4] We report results using JP, MI and MI^2.

	Seen		Unseen	
	r	ρ	r	ρ
AltaVista	.650	—	.480	—
BNC (Rasp)	.543	.622	.135	.102
Padó et al.	.479	.570	.120	.138
LDA	.594	.558	.468	.459
ROOTH-LDA	.575	.599	**.501**	**.469**
DUAL-LDA	.460	.400	.334	.278
Mod (JP)	.495	.413	.286	.276
Mod (MI)	.394	.425	.471	.457
Mod (MI^2)	.575	.501	.430	.408
Mod_{nc} (JP)	.626	.505	.357	.369
Mod_{nc} (MI)	.628	.574	.427	.385
Mod_{nc} (MI^2)	**.701**	**.623**	.423	.394

Table 3: Results (Pearson r and Spearman ρ correlations) on the Keller and Lapata (2003) plausibility data. Bold indicates best scores, underlining our best scores. High values indicate high correlation with the gold standard.

Table 3 shows the performance of our models compared to results reported in Ó Séaghdha (2010). As before, results between the normal and the inverse model (omitted due to space) are very similar. Surprisingly, the no-context models consistently outperform the models using context on the *seen* data set. This suggests that the *seen* data set can quite precisely be ranked using frequency estimates, which the no-context models might be better at capturing without the 'noise' introduced by context.

	Standard		Inverse (i)	
	r	ρ	r	ρ
Mod (JP)	.286	.276	.243	.245
Mod (MI)	.471	.457	.409	.383
Mod (MI^2)	.430	.408	.362	.347
Mod_{nc} (JP)	.357	.369	.181	.161
Mod_{nc} (MI)	.427	.385	.220	.209
Mod_{nc} (MI^2)	.423	.394	.218	.185

Table 4: Results on the *unseen* plausibility dataset.

The results on the *unseen* data set (Table 4) prove interesting as well. The inverse no-context model is performing significantly poorer than any of the other models. To understand this result we must investigate the differences between the *unseen* data set and the *seen* data set and to the pseudo-disambiguation evaluation. The key difference to pseudo-disambiguation is that we measure a human plausibility judgement, which — as we have demonstrated — only partially correlates with bigram frequencies. Our models were trained on the BNC, hence they could only learn frequency estimates for the *seen* data set, but not for the *unseen* data.

Based on our hypothesis about the role of context, we expect Mod and Mod_i to learn semantic classes based on the distribution of context. Without the access to that context, we argued that Mod_{nc} and Mod_{inc} would instead learn frequency estimates.[5] The hypothesis that nouns generally select for adjectives rather than vice versa further suggests that Mod and Mod_{nc} would learn semantic properties that Mod_i and Mod_{inc} could not learn so well.

In summary, we hence expected Mod to perform best on the *unseen* data, learning semantics from both context and noun-adjective selection. Also, as supported by the results, we expected Mod_{inc} to performs poorly, as it is the model least capable of learning semantics according to our hypotheses.

5 Conclusion

We have presented a class of probabilistic models which successfully learn semantic clusterings of nouns and a representation of adjective-noun selectional preference. These models encoded our beliefs about how adjective-noun pairs relate to each other and to the other words in the sentence. The performance of our models on estimating selectional preference strongly supported these initial hypotheses.

We discussed plausibility judgements from a theoretical perspective and argued that frequency estimates and JP are imperfect approximations for plausibility. While models can perform well on some evaluations by using either frequency estimates or semantic knowledge, we explained why this does not apply to the *unseen* plausibility test. The performance on that task demonstrates both the success of our model and the shortcomings of frequency-based approaches to human plausibility judgements.

Finally, this paper demonstrated that it is feasible to learn semantic representations of words while concurrently learning how they relate to one another.

Future work will explore learning words from broader classes of semantic relations and the role of context in greater detail. Also, we will evaluate the system applied to higher level tasks.

[4] See (Evert, 2005) for a discussion of these metrics.

[5] This could also explain their weaker performance on pseudo-disambiguation in the previous section, where the negative examples had zero frequency in the training corpus.

References

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1183–1193, Stroudsburg, PA, USA. Association for Computational Linguistics.

Massimiliano Ciaramita and Mark Johnson. 2003. Supersense tagging of unknown nouns in wordnet. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 168–175, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stephen Clark and David Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Comput. Linguist.*, 28:187–206, June.

James R. Curran. 2005. Supersense tagging of unknown nouns using semantic similarity. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 26–33, Stroudsburg, PA, USA. Association for Computational Linguistics.

Katrin Erk, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36:723–763.

Stefan Evert. 2005. *The statistics of word cooccurrences: word pairs and collocations*. Ph.D. thesis, Universität Stuttgart, Holzgartenstr. 16, 70174 Stuttgart.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011. Experimental support for a categorical compositional distributional model of meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1394–1404, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 317–325, Stroudsburg, PA, USA. Association for Computational Linguistics.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, pages 459–484.

Andrew L. Maas and Andrew Y. Ng. 2010. A probabilistic model for semantic word vectors. In *Workshop on Deep Learning and Unsupervised Feature Learning*, NIPS '10.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *ACL-HLT'08*, pages 236 – 244.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 435–444, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, pages 183–190, Stroudsburg, PA, USA. Association for Computational Linguistics.

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 on Computational Linguistics*, ACL '99, pages 104–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Issei Sato, Minoru Yoshida, and Hiroshi Nakagawa. 2008. Knowledge discovery of semantic relationships between words using nonparametric bayesian graph model. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 587–595, New York, NY, USA. ACM.