Learning Semantic Representations in a Bigram Language Model

Jeff Mitchell
mitchelljeff@hotmail.com

Abstract

This paper investigates the extraction of semantic representations from bigrams. The major obstacle to this objective is that while these word to word dependencies do contain a semantic component, other factors, e.g. syntax, play a much stronger role. An effective solution will therefore require some means of isolating semantic structure from the remainder. Here, the possibility of modelling semantic dependencies within the bigram in terms of the similarity of the two words is explored. A model based on this assumption of semantic coherence is contrasted and combined with a relaxed model lacking this assumption. The induced representations are evaluated in terms of the correlation of predicted similarities to a dataset of noun-verb similarity ratings gathered in an online experiment. The results show that the coherence assumption can be used to induce semantic representations, and that the combined model, which breaks the dependencies down into a semantic and a non-semantic component, achieves the best performance.

1 Introduction

Distributional semantics derives semantic representations from the way that words are distributed across contexts. The assumption behind this approach is that words that occur in similar contexts will tend to have similar meanings. Firth (1957) expressed this in a well known slogan - *you shall know a word by the company it keeps*. In application, these representations have proven successful in automatic thesaurus generation (Grefenstette, 1994), enhancing language models (Coccaro and Jurafsky, 1998) and modelling of reading times (Pynte et al., 2008) and the effects of priming (Landauer and Dumais, 1997).

However, the high level identification of meaning with distributional properties leaves the question of exactly which distributional properties are relevant to semantics a little vague. In practice, researchers evaluate various approaches and select those that produce the best performance. Moreover, other linguistic characteristics, such as syntax, are also analysed in terms of distributional properties. Bigram distributions, for example, are commonly used to induce POS classes (e.g Brown et al., 1992; Clark, 2003), but they have also been investigated as a basis for semantic representations (Bullinaria and Levy, 2007).

Here we examine the question of what statistical properties can be used to distinguish semantic factors from other dependencies in the distribution of words across bigram contexts. We carry this out in terms of class based bigram language models, and explore the possibility that semantic dependencies can be characterised in terms of coherence or similarity across the bigram. We then evaluate the induced representations in terms of their ability to predict human similarity ratings for noun-verb pairs. By evaluating the similarity predictions of our models across POS classes in this way, we assess the ability of the model to focus purely on the semantic content while ignoring other information, such as syntax.

2 Models

The intention is to induce semantic representations within a bigram model based on the assumption that semantic content is coherent across the bigram. Assume that semantic information can be captured in terms of a set, S, of semantic topics, with each word, w, having some independent probability of being used in a topic, $p(w|s)$. Then, if the probabilities of the topics are given by $p(s)$ and each bigram, w_1w_2, belongs to a single topic, then the joint probability, $p(w_1w_2)$, is given by:

$$p(w_1 w_2) = \sum_{s \in S} p(s) p(w_2|s) p(w_1|s) \tag{1}$$

Rewriting this in conditional form, with $p(s|w_1) = \frac{p(s)p(w_1|s)}{p(w_1)}$, gives:

$$p(w_2|w_1) = \sum_{s \in S} p(w_2|s) p(s|w_1) \tag{2}$$

This can also be expressed in a form that explicitly connects to the idea of a probability based on semantic similarity.

$$p(w_2|w_1) = p(w_2) \sum_{s \in S} \frac{p(s|w_2)}{p(s)} p(s) \frac{p(s|w_1)}{p(s)} \tag{3}$$

Equation 3 can be thought of as the unigram probability of w_2 modulated by its similarity to w_1, measured in terms of a weighted dot product between vectors representing the two words. In this case, the vector components are a ratio of probabilities measure, $\frac{p(s|w)}{p(s)}$, which has been widely used in distributional semantics (e.g. Bullinaria and Levy, 2007).

The key feature of this model is that the word probabilities in Equation 1 are independent of position in the bigram. It is this assumption that serves to ensure that the induced topics identify a characterisitic that is stable across the bigram, which, it is hoped, will relate to semantic content.

Relaxing this assumption produces a more general class based model, specifically the aggregate markov model of Saul and Pereira (1997). Using superscripts to indicate the position a word occurs in within the bigram, we write this model as:

$$p(w_2^r|w_1^l) = \sum_{z \in Z} p(w_2^r|z) p(z|w_1^l) \tag{4}$$

In contrast to Equation 1, this model makes no assumption of stability of content across the bigram, and instead allows the word distributions, $p(w|z)$ to be very different in the left and right positions. Thus, this model ought to more suited to handling the word order effects that the similarity based model cannot.

To construct a combined model, the bigram probabilities are expressed in terms of a sum over both S and Z.

$$p(w_2^r|w_1^l) = \sum_{s \in S, z \in Z} p(w_2^r|s,z) p(s,z|w_1^l) \tag{5}$$

These terms can be broken down further based on conditional independence of s and z. The rightmost probability, $p(s,z|w_1^l)$ separates straightforwardly.

$$p(s,z|w_1^l) = p(s|w_1) p(z|w_1^l) \tag{6}$$

On the other hand, s and z cannot in general also be conditionally independent given w_2. However, we can use this as an approximation and then normalise the final probabilities.

$$\hat{p}(w_2^r|s,z) = \frac{p(w_2^r) p(s|w_2) p(z|w_2^r)}{p(s,z)} \tag{7}$$

The final model then combines these components and divides through by a normalising constant $N(w_1)$.

$$p(w_2^r|w_1^l) = \sum_{s \in S, z \in Z} \frac{\hat{p}(w_2^r|s,z) p(s,z|w_1^l)}{N(w_1)} \tag{8}$$

$$N(w_1) = \sum_{w_2} \sum_{s \in S, z \in Z} \hat{p}(w_2^r|s,z) p(s,z|w_1^l) \tag{9}$$

	High	Medium	Low
Group 1	*anticipation - predict* *withdrawal - retire*	*analysis - derive* *invasion - merge*	*opinion - vanish* *disappearance - believe*
Group 2	*disappearance - vanish* *invasion - occupy*	*anticipation - believe* *opinion - predict*	*withdrawal - derive* *implication - retire*
Group 3	*opinion - believe* *implication - derive*	*disappearance - retire* *withdrawal - vanish*	*anticipation - succeed* *invasion - predict*

Table 1: Example items from the noun-verb similarity rating experiment.

2.1 Construction

Models were constructed based on three approaches: similarity based models, as defined by Equation 2, aggregate models, defined by Equation 4, and combined models, defined by Equation 8. The parameters of these bigram models were optimised over a set of sentences extracted from the BNC (BNC Consortium, 2001). 80,775,061 words from the written component of this corpus were used as a training set, with 9,759,769 words forming a development set and the final 9,777,665 words held back as a test set. Preprocessing included conversion to lowercase, addition of $\langle start \rangle$ and $\langle stop \rangle$ at the beginning and ends of sentences, and replacement of words that occurred fewer than 100 times in the training set with an $\langle unk \rangle$ token.

Optimisation of the parameters was based on the EM algorithm (Dempster et al., 1977), with training stopped when the log-likelihood over the development set began to increase. For the pure similarity and aggregate approaches, models were trained with numbers of induced classes ranging from 10 to 2,000. The numbers of classes, $|S|$ and $|Z|$, for the two components of the combined models, each ranged from 10 to 100. The ratio of probabilities measure from Equation 3 was used to construct the components of vectors which then formed the word representations, and similarity of these vectors was measured in terms of the cosine measure.

For comparison, a bigram language model with back-off and Kneser-Ney smoothing (Kneser and Ney, 1995) was also constructed using the SRILM toolkit (Stolcke, 2002).

3 Evaluation

The induced representations were evaluated in terms of their ability to predict semantic similarity ratings for a set of word pairs. We measured the cosine similarity of our word representations and correlated that with the human ratings to produce a measure of agreement. Because the strongest dependencies within the bigrams are likely to be syntactic effects based on the POS classes of the two words, measuring semantic similarity across POS classes is particularly relevant. That is, the semantic representations should contain as much information about the meaning of the words as possible, while containing as little part-of-speech information as possible, which should instead be shifted into the other part of the model. Predicting the similarity between nouns and verbs should therefore be an effective evaluation, as these two word classes contain the core of a sentence's semantic content while having substantially divergent distributional properties in regards of syntax. In this way, we can test whether the POS differences are genuinely being ignored to allow just the semantic similarity to be focussed on.

Thus, an experiment was run to collect similarity ratings for noun-verb pairs. Each participant rated one of three groups of 36 noun-verb pairs, giving a total of 108 items. Each group consisted of 12 high similarity pairs, 12 medium similarity pairs and 12 low similarity pairs.

Table 1 contains a small sample of these items, with rows corresponding to the three experimental groups of participants and columns corresponding to the high, medium and low similarity sets of items seen by each group. The items in the high similarity set (e.g. *anticipation-predict*) are related, via an intermediary word, by a combination of morphology (e.g. *anticipation-anticipate*) and synonymy (e.g. *anticipate-predict*), drawing on Catvar (Habash and Dorr, 2003) and WordNet (Miller, 1995) to identify these relationships. The medium and low sets are then recombinations of nouns and verbs from the high set, with the medium items being the most similar such pairings, as rated by WordNetSimilarity (Pedersen et al., 2004), and low being the least similar.

60 participants were paid $2 each to rate all 36 items from a single group, with equal numbers

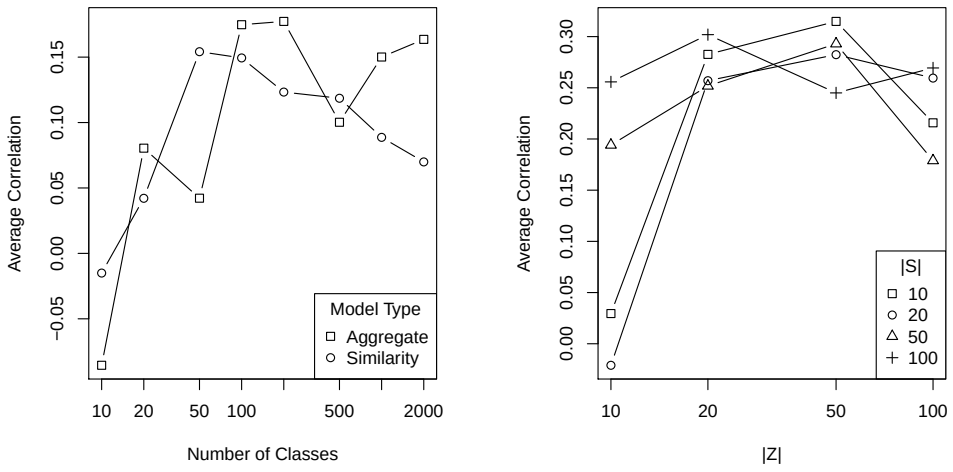
seeing each group. The experiments were conducted online, with participants instructed to rate the similarity in meaning of each pair of words on a scale of 1 to 5. They were initially presented with five practice items before the experimental materials were presented with randomisation of both within and between item orders.

Individual Spearman correlations were calculated for each participant's ratings against the predicted similarities, and the average of these values was used as the evaluation measure for the semantic representations induced by a model. A t-test on the z-transformed participant correlations was used to assess the significance of differences between these averages.

The performance of these models simply as language models was also evaluated, in terms of their perplexity over the test set, T, calculated in terms of the probability assigned to the test set, $p(T)$, and the number of words it contains, $|T|$.

$$perplexity = p(T)^{-\frac{1}{|T|}} \tag{10}$$

4 Results



(a) Average correlations by model size for the Similarity and Aggregate models.

(b) Average correlations by model size for the Combined models.

Figure 1: Correlations of model similarities with human ratings.

Figure 1(a) plots the average correlation between the model similarities and the human ratings for the similarity and aggregate representations. Both models show similar strengths of correlation and a similar pattern in relation to the size of the model, with a peak around the 50 - 200 range. The highest correlation is 0.18, achieved by the aggregate model with 200 classes, while the similarity model achieves a peak of 0.15 at $|S| = 50$. These values are not significantly different, $t(59) = 0.71$, $p = 0.24$. The equivalence in performance of the aggregate and similarity models is not entirely surprising, as both models, despite their differing forms, are directed at the problem of predicting the same bigram dependencies. It may therefore be expected that the weaker semantic factors play only a minor role within the representations generated.

In contrast, the combined models, which allow a separation of the dependencies into distinct components, are able to achieve higher correlations, as plotted in Figure 1(b). Among these models, the highest correlation of 0.31, which is significantly greater than the best aggregate model, $t(59) = 9.35$, $p < 0.001$, is achieved by a model having $|Z| = 50$ and $|S| = 10$. In fact, all the correlations over 0.2 in Figure 1(b) are significantly greater at the $p < 0.001$ level, except $|Z| = 100$, $|S| = 10$ and $|Z| = 20$, $|S| = 20$, which are only significant at the $p < 0.05$ and $p < 0.01$ levels respectively. This leaves only the four lowest performing combined models as not significantly outperforming the best aggregate model. Nonetheless, these values are substantially lower than the inter-subject correlations ($mean = 0.74$, $min = 0.64$), suggesting that the model could be improved further. In particular, extending the span of the model to longer ngrams ought to allow the induction of stronger

and more detailed semantic representations. The fact that the best performing model only contains 10 semantic classes underscores the limitations of extracting such representations from bigrams.

In addition to the ability of these models to induce semantic representations, their performance simply as language models was also evaluated. Figure 2 plots perplexity on the test set against number of parameters per word ($|S| + 2|Z|$) for the aggregate and combined models. In general lower perplexities are achieved by larger models for both approaches, as is to be expected. Within this trend, the combined model tends to have a lower perplexity than the aggregate model by about 5%. The single case in which the combined model is above the trend line of the aggregate model occurs for a model with in which a very small aggregate component, $|Z| = 10$, is dominated by a large similarity component, $|S| = 100$.

The performance of these models does not, however, rival that of a standard bigram model with back-off and Kneser and Ney (1995) smoothing, which achieves a perplexity of 185. On the other hand, neither the aggregate nor combined models are explicitly designed to address the issue of small or zero counts.

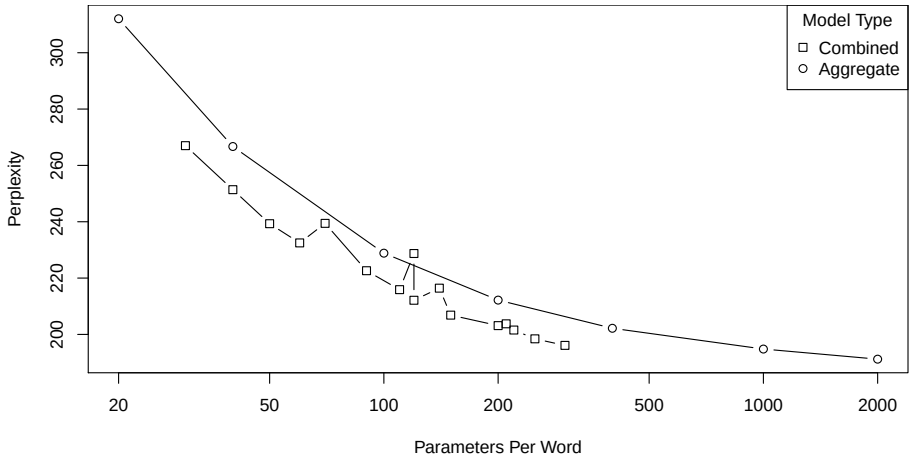


Figure 2: Perplexity by Number of Parameters for the Aggregate and Combined Models

5 Conclusions

Our experiments produced two novel results.

Firstly, we have shown that semantic representations can be induced from the dependencies within bigram word sequences, using an approach that derives word probabilities from similarity. This is similar in form to prior models (e.g Coccaro and Jurafsky, 1998; Bellegarda, 1997), but whereas they imported distributional representations from outside the model to enhance their performance, we use this model form to induce semantic representations within the model.

Secondly, we have shown that this approach is most effective when the model breaks these dependencies down into both a semantic and a non-semantic component. Typically, semantic classes have been induced in isolation (Landauer and Dumais, 1997; Bullinaria and Levy, 2007) or applied to long-range structure while short-range structure is handled by a separate component (Boyd-Graber and Blei, 2008; Griffiths et al., 2004; Wallach, 2006). Here, we have shown that even simple bigram dependencies can be conceived as breaking down into semantic and non-semantic components, as opposed to using those components to model two different types of dependency.

We also introduced a novel evaluation dataset for semantic representations, containing noun-verb similarity ratings. Correlation of these human ratings with the model similarities allows a quantification of the extent to which a model ignores POS information to focus on semantic content.

In future work, we hope to extend the span of our model and to characterise syntax, semantics and their interaction in a more sophisticated manner. Particularly interesting is the question of the extent to which the form of our model is specific to languages, such as English, in which syntax is identified with word order and how this might be adapted to free word order languages.

References

Bellegarda, J. R. (1997). A latent semantic analysis framework for large-span language modeling. In G. Kokkinakis, N. Fakotakis, and E. Dermatas (Eds.), *EUROSPEECH*. ISCA.

BNC Consortium (2001). The British National Corpus, Version 2. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Boyd-Graber, J. L. and D. M. Blei (2008). Syntactic topic models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (Eds.), *NIPS*, pp. 185–192. Curran Associates, Inc.

Brown, P. F., V. J. D. Pietra, P. V. de Souza, J. C. Lai, and R. L. Mercer (1992). Class-based n-gram models of natural language. *Computational Linguistics 18*(4), 467–479.

Bullinaria, J. and J. Levy (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods 39*(3), 510–526.

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In *EACL*, pp. 59–66. The Association for Computer Linguistics.

Coccaro, N. and D. Jurafsky (1998). Towards better integration of semantic predictors in statistical language modeling. In *ICSLP*. ISCA.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1–32. Oxford: Philological Society.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA: Kluwer Academic Publishers.

Griffiths, T. L., M. Steyvers, D. M. Blei, and J. B. Tenenbaum (2004). Integrating topics and syntax. In *NIPS*.

Habash, N. and B. J. Dorr (2003). A categorial variation database for english. In *HLT-NAACL*.

Kneser, R. and H. Ney (1995). Improved backing-off for M-gram language modeling. In *International Conference on Acoustics, Speech, and Signal Processing, 1995.*, Volume 1, pp. 181–184 vol.1.

Landauer, T. and S. Dumais (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review; Psychological Review 104*(2), 211.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM 38*(11), 39–41.

Pedersen, T., S. Patwardhan, and J. Michelizzi (2004). Wordnet: : Similarity - measuring the relatedness of concepts. In D. L. McGuinness and G. Ferguson (Eds.), *AAAI*, pp. 1024–1025. AAAI Press / The MIT Press.

Pynte, J., B. New, and A. Kennedy (2008). On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision Research 48*, 2172–2183.

Saul, L. K. and F. Pereira (1997). Aggregate and mixed-order markov models for statistical language processing. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, New York, NY, pp. 81–89. ACM Press.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In J. H. L. Hansen and B. L. Pellom (Eds.), *INTERSPEECH*. ISCA.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In W. W. Cohen and A. Moore (Eds.), *ICML*, Volume 148 of *ACM International Conference Proceeding Series*, pp. 977–984. ACM.