Deriving adjectival scales from continuous space word representations

Joo-Kyung Kim
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210, USA
kimjook@cse.ohio-state.edu

Marie-Catherine de Marneffe
Department of Linguistics
The Ohio State University
Columbus, OH 43210, USA
mcdm@ling.ohio-state.edu

Abstract

Continuous space word representations extracted from neural network language models have been used effectively for natural language processing, but until recently it was not clear whether the spatial relationships of such representations were interpretable. Mikolov et al. (2013) show that these representations do capture syntactic and semantic regularities. Here, we push the interpretation of continuous space word representations further by demonstrating that vector offsets can be used to derive adjectival scales (e.g., *okay* $<$ *good* $<$ *excellent*). We evaluate the scales on the indirect answers to *yes/no* questions corpus (de Marneffe et al., 2010). We obtain 72.8% accuracy, which outperforms previous results ($\sim$60%) on this corpus and highlights the quality of the scales extracted, providing further support that the continuous space word representations are meaningful.

1 Introduction

There has recently been a surge of interest for deep learning in natural language processing. In particular, neural network language models (NNLMs) have been used to learn distributional word vectors (Bengio et al., 2003; Schwenk, 2007; Mikolov et al., 2010): the models jointly learn an embedding of words into an n-dimensional feature space. One of the advantages put forth for such distributed representations compared to traditional n-gram models is that similar words are likely to have similar vector representations in a continuous space model, whereas the discrete units of an n-gram model do not exhibit any inherent relation with one another. It has been shown that the continuous space representations improve performance in a variety of NLP tasks, such as POS tagging, semantic role labeling, named entity resolution, parsing (Collobert and Weston, 2008; Turian et al., 2010; Huang et al., 2012).

Mikolov et al. (2013) show that there are some syntactic and semantic regularities in the word representations learned, such as the singular/plural relation (the difference of singular and plural word vectors are equivalent: *apple* $-$ *apples* $\approx$ *car* $-$ *cars* $\approx$ *family* $-$ *families*) or the gender relation (a masculine noun can be transformed into the feminine form: *king* $-$ *man* $+$ *woman* $\approx$ *queen*).

We extend Mikolov et al. (2013)'s approach and explore further the interpretation of the vector space. We show that the word vectors learned by NNLMs are meaningful: we can extract scalar relationships between adjectives (e.g., *bad* $<$ *okay* $<$ *good* $<$ *excellent*), which can not only serve to build a sentiment lexicon but also be used for inference. To evaluate the quality of the scalar relationships learned by NNLMs, we use the indirect *yes/no* question answer pairs (IQAP) from (de Marneffe et al., 2010), where scales between adjectives are needed to infer a *yes/no* answer from a reply without explicit *yes* or *no* such as *Was the movie good? It was excellent.* Our method reaches 72.8% accuracy, which is the best result reported so far when scales are used.

2 Previous work

We use the continuous word representations from (Mikolov et al., 2011), extracted from a recurrent neural network language model (RNNLM), whose

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1625–1630,
Seattle, Washington, USA, 18-21 October 2013.

three-layer architecture is represented in Figure 1.

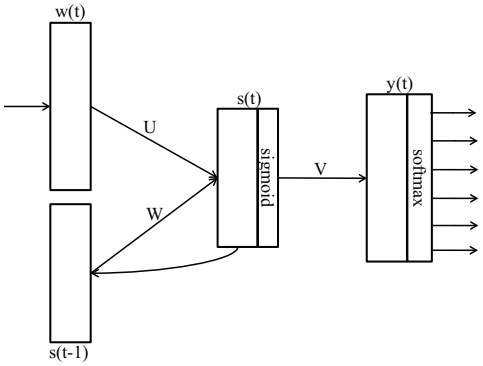

Figure 1: The architecture of the RNNLM.

In the input layer, $w(t)$ is the input word represented by 1-of-N coding at time t when the vocabulary size is N. When there are M nodes in the hidden layer, the number of connections between the input layer and the hidden layer is NM and the connections can be represented by a matrix U.

The hidden layer is also connected recurrently to the context $s(t-1)$ at time $t-1$ ($s(0)$ is initialized with small values like 0.1). The connections between the previous context and the hidden layer are represented by a matrix W. The dimensionality of the word representations is controlled by the size of W. The output of the hidden layer is $s(t) = f(Uw(t) + Ws(t-1))$, where f is a sigmoid function.

Because the inputs of the hidden layer consist of the word $w(t)$ and the previous hidden layer output $s(t-1)$, the current context of the RNN is influenced by the current word and the previous context. Therefore, we can regard that the continuous representations from the RNNLM exploit the context implicitly considering the word sequence information (Mikolov et al., 2010).

V is a N by M matrix representing the connections between the hidden layer and the output layer. The final output is $y(t) = g(Vs(t))$, where g is a softmax function to represent the probability distribution over all the words in the vocabulary.

When the RNN is trained by the back propagation algorithm, we can regard the ith column vector of U as the continuous representation of the ith word in the vocabulary since the column was adjusted correspondingly to the ith element of $w(t)$. Because the $s(t)$ outputs of two input words will be similar when they have similar $s(t-1)$ values, the corresponding column vectors of the words will also be similar.

Mikolov et al. (2013) showed that constant vector offsets of word pairs can represent linguistic regularities. Let w_a and w_b denote the vectors for the words a and b, respectively. Then the vector offset of the word pair is $w_a - w_b$. If a and b are syntactically or semantically related, the vector offset can be interpreted as a transformation of the syntactic form or the meaning. The offset can also be added to another word vector c. The word vector nearest to $w_a - w_b + w_c$ would be related to word c with the syntactic or semantic difference as the difference between a and b, as it is the case for the *king*, *man*, and *woman* example, where *king* $-$ *man* $+$ *woman* would approximately represent *king* with feminine gender (i.e., *queen*). They also tried to use the continuous representations generated by Latent Semantic Analysis (LSA) (Landauer et al., 1998). However, the results using LSA were worse because LSA is a bag-of-words model, in which it is difficult to exploit word sequence information as the context.

For all the experiments in this paper, we use the precomputed word representations generated by the RNNLM from (Mikolov et al., 2013). Their RNN is trained with 320M words from the Broadcast News data (the vocabulary size is 82,390 words), and we used word vectors with a dimensionality of 1,600 (the highest dimensionality provided).[1] We standardized the dataset so that the mean and the variance of the representations are 0 and 1, respectively.[2]

3 Deriving adjectival scales

Here we explore further the interpretation of word vectors. Assuming that the transformation of form or meaning represented by the vector offset is linear, an intermediate vector between two word vectors would represent some "middle" form or meaning. For example, given the positive and superlative forms of an adjective (e.g., *good* and *best*), we expect that the word representation in the middle of

[1] We also experimented with smaller dimensions, but consistent with the analyses in (Mikolov et al., 2013), the highest dimensionality gave better results.

[2] http://www.fit.vutbr.cz/~imikolov/rnnlm/word_projections-1600.txt.gz

Input words	Words with highest cosine similarities to the mean vector			
good:best	**better: 0.738**	strong: 0.644	normal: 0.619	less: 0.609
bad:worst	terrible: 0.726	great: 0.678	horrible: 0.674	**worse: 0.665**
slow:slowest	**slower: 0.637**	sluggish: 0.614	steady: 0.558	brisk: 0.543
fast:fastest	**faster: 0.645**	slower: 0.602	quicker: 0.542	harder: 0.518

Table 1: Words with corresponding vectors closest to the mean of positive:superlative word vectors.

First word (-)		1st quarter		Half		3rd quarter		Second word (+)	
furious	1	angry	0.632	unhappy	0.640	pleased	0.516	**happy**	1
furious	1	angry	0.615	tense	0.465	quiet	0.560	**calm**	1
terrible	1	horrible	0.783	incredible	0.714	wonderful	0.772	**terrific**	1
cold	1	mild	0.348	warm	0.517	sticky	0.424	**hot**	1
ugly	1	nasty	0.672	wacky	0.645	lovely	0.715	**gorgeous**	1

Table 2: Adjectival scales extracted from the RNN: each row represent a scale, and for each intermediate point the closest word in term of cosine similarity is given.

them will correspond to the comparative form (i.e., *better*). To extract the "middle" word between two word vectors w_a and w_b, we take the vector offset $w_a - w_b$ divided by 2, and add w_b: $w_b + (w_a - w_b)/2$. The result corresponds to the midpoint between the two words. Then, we find the word whose cosine similarity to the midpoint is the highest.

Table 1 gives some positive:superlative pairs and the top four closest words to the mean vectors, where the distance metric is the cosine similarity. The correct comparative forms (in bold) are quite close to the mean vector of the positive and superlative form vectors, highlighting the fact that there is some meaningful interpretation of the vector space: the word vectors are constituting a scale.

We can extend this idea of extracting an ordering between two words. For any two semantically related adjectives, intermediate vectors extracted along the line connecting the first and second word vectors should exhibit scalar properties, as seen above for the positive-comparative-superlative triplets. If we take two antonyms (*furious* and *happy*), words extracted at the intermediate points x_1, x_2 and x_3 should correspond to words lying on a scale of happiness (from "less furious" to "more happy"), as illustrated in Figure 2. Table 2 gives some adjectival scales that we extracted from the continuous word space, using antonym pairs. We picked three points with equal intervals on the line from the first to the second word (1st quarter, half and 3rd quarter). The extracted scales look quite reasonable: the words form a continuum from more negative to more positive meanings.

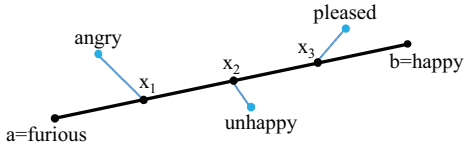

Figure 2: An example of vectors with the highest cosine similarity to intermediate points on the line between *furious* and *happy*.

Tables 1 and 2 demonstrate that the word vector space is interpretable: intermediate vectors between two word vectors represent a semantic continuum.

4 Evaluation: Indirect answers to *yes/no* questions

To evaluate the quality of the adjective scales learned by the neural network approach, we use the corpus of indirect answers to *yes/no* questions created by (de Marneffe et al., 2010), which consists of question-answer pairs involving gradable modifiers to test scalar implicatures. We focus on the 125 pairs in the corpus where both the question and answer contain an adjective: e.g., *Is Obama qualified? I think he's young.*[3] Each question-answer pair has

[3]These 125 pairs correspond to the 'Other adjective' category in (de Marneffe et al., 2010).

been annotated via Mechanical Turk for whether the answer conveys *yes*, *no* or *uncertain*.

4.1 Method

The previous section showed that we can draw a line passing through an adjective and its antonym and that the words extracted along the line are roughly semantically ordered. To infer a *yes* or *no* answer in the case of the IQAP corpus, we use the following approach illustrated with the *Obama* example above (Figure 3). Using WordNet 3.1 (Fellbaum, 1998), we look for an antonym of the adjective in the question *qualified*: *unqualified* is retrieved. Since the scales extracted are only roughly ordered, to infer *yes* when the question and answer words are very close, we set the decision boundary perpendicular to the line connecting the two words and passing through the midpoint of the line.

Since the answer word is *young*, we check whether *young* is in the area including *qualified* or in the other area. We infer a *yes* answer in the former case, and a *no* answer in the latter case. If *young* is on the boundary, we infer *uncertain*. If a sentence contains a negation (e.g., *Are you stressed? I am not peaceful.*), we compute the scale for *stressed-peaceful* and then reverse the answer obtained, similarly to what is done in (de Marneffe et al., 2010).

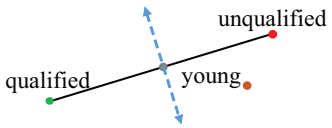

Figure 3: An example of the decision boundary given *qualified* as the question and *young* as the answer.

Since a word can have multiple senses and different antonyms for the senses, it is important to select the most appropriate antonym to build a more accurate decision boundary. We consider all antonyms across senses[4] and select the antonym that is most collinear with the question and the answer. For the word vectors of the question w_q, the ith antonym w_{ant_i}, and the answer w_a, we select ant_i where $argmax_{ant_i}|cos(w_q - w_a, w_q - w_{ant_i})|$. Figure 4 schematically shows antonym selection when the question is *good* and the answer is *excellent*: *bad* and *evil* are the antonym candidates of *good*. Because the absolute cosine similarity of *good-excellent* to *good-bad* is higher than to *good-evil*, we choose *bad* as the antonym in this case.

[4]Antonyms in WordNet can be directly opposed to a given word or indirectly opposed via other words. When there are direct antonyms for the question word, we only consider those.

	Acc	Macro P	Macro R	Macro F1
de Marneffe (2010)	60.00	59.72	59.40	59.56
Mohtarami (2011)	–	62.23	60.88	61.55
RNN model	**72.80**	**69.78**	**71.39**	**70.58**

Table 3: Score (%) comparison on the 125 scalar adjective pairs in the IQAP corpus.

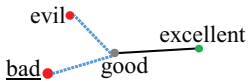

Figure 4: An example of antonym selection.

4.2 Results and discussion

Table 3 compares our results with previous ones where adjectival scales are considered: de Marneffe et al. (2010) propose an unsupervised approach where scales are learned from distributional information in a Web corpus; Mohtarami et al. (2011)'s model is similar to ours but uses word representations obtained by LSA and a word sense disambiguation system (Zhong and Ng, 2010) to choose antonyms. To compare with Mohtarami et al. (2011), we use macro-averaged precision and recall for *yes* and *no*. For the given metrics, our model significantly outperforms the previous ones ($p < 0.05$, McNemar's test).

Mohtarami et al. (2011) present higher numbers obtained by replacing the answer words with their synonyms in WordNet. However, that approach fails to capture orderings. Two words of different degree are often regarded as synonyms: even though *furious* means extremely angry, *furious* and *angry* are synonyms in WordNet. Therefore using synonyms, the system will output the same answer irrespective of the order in the pair. Mohtarami et al. (2012) also presented results on the interpretation of indirect questions on the IQAP corpus, but their method

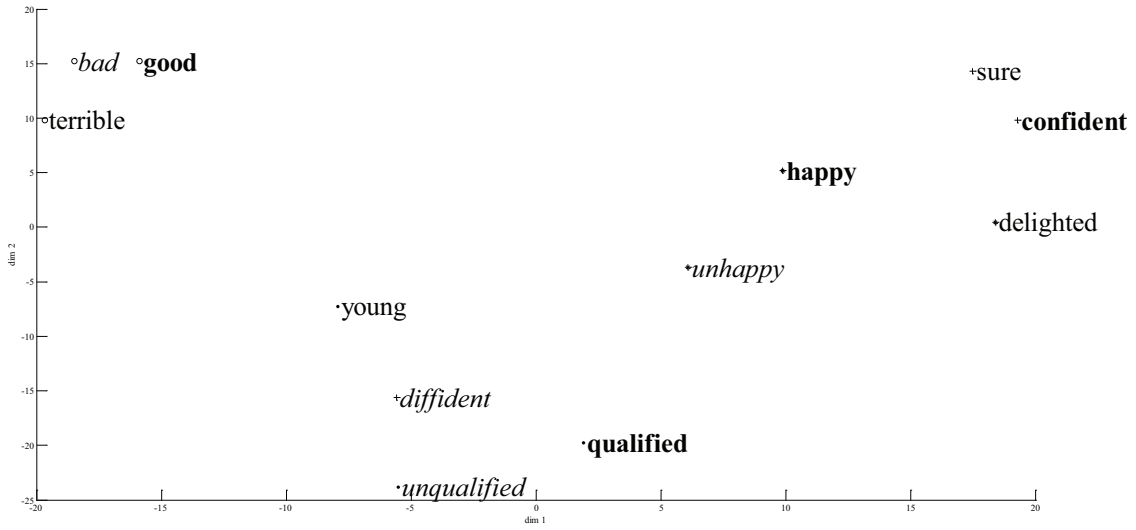

Figure 5: Question words (bold), their antonyms (italic), and answer words (normal) of four pairs from the IQAP dataset. The words are visualized by MDS.

did not involve learning or using scalar implicatures.

Figure 5 gives a qualitative picture: the question words, antonyms and answer words for four of the IQAP pairs are visualized in 2D space by multi-dimensional scaling (MDS). Note that MDS introduces some distortion in the lower dimensions. Bullet markers correspond to words in the same pair. Question words, antonyms, and answer words are displayed by bold, italic, and normal fonts, respectively. In the *Obama* example previously mentioned (*Is Obama qualified? I think he's young.*), the question word is *qualified* and the answer word is *young*. In Figure 5, *qualified* is around (2,-20) while its antonym *unqualified* is around (-6,-24). Since *young* is around (-7,-8), we infer that *young* is semantically closer to *unqualified* which corroborates with the Turkers' intuitions in this case. (1), (2) and (3) give the other examples displayed in Figure 5.

(1) A: Do you think she'd be *happy* with this book?
B: I think she'd be *delighted* by it.

(2) A: Do you think that's a *good* idea?
B: It's a *terrible* idea.

(3) A: The president is promising support for Americans who have suffered from this hurricane. Are you *confident* you are going to be getting that?
B: I'm not so *sure* about my insurance company.

In (1), *delighted* is stronger than *happy*, leading to a *yes* answer, whereas in (2), *terrible* is weaker than *good* leading to a *no* answer. In (3), the presence of a negation will reverse the answer inferred, leading to *no*.

5 Conclusion

In this paper we give further evidence that the relationships in the continuous vector space learned by recurrent neural network models are interpretable. We show that using vector offsets, we can successfully learn adjectival scales, which are useful for scalar implicatures, as demonstrated by the high results we obtain on the IQAP corpus.

Acknowledgements

We thank Eric Fosler-Lussier and the anonymous reviewers for their helpful comments on previous versions of this paper.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2010. Was it good? It was provocative. Learning the meaning of scalar adjectives. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics*, pages 167–176.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. MIT Press.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 873–882. Association for Computational Linguistics.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Tomas Mikolov, Martin Karafiát, Lukáš Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of Interspeech*, pages 1045–1048.

Tomas Mikolov, Daniel Povey, Lukáš Burget, and Jan Cernocky. 2011. Strategies for training large scale neural network language models. In *Proceedings of ASRU*, pages 196–201.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pages 746–751.

Mitra Mohtarami, Hadi Amiri, Man Lan, and Chew Lim Tan. 2011. Predicting the uncertainty of sentiment adjectives in indirect answers. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2485–2488.

Mitra Mohtarami, Hadi Amiri, Man Lan, Thanh Phu Tran, and Chew Lim Tan. 2012. Sense sentiment similarity: an analysis. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 1706–1712.

Holger Schwenk. 2007. Continuous space language models. *Computer Speech & Language*, 21(3):492–518.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: a wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83.