Revisiting Automatically-Generated Adjectival Scales with Continuous Space Word Representations

Gil Landau
glandau@seas.upenn.edu

Abstract

The goal of this study is to examine, replicate, and extend the model proposed by Kim and de Marneffe (2013). Their model uses the continuous space word representations described by Mikolov et al. (2010) to vectorize adjectives and discuss their relationships, with regards to their adjectival scale and relative intensity, in the new, high-dimensional, continuous space. The goal is discover what information and connections can be gleaned from these representations.

This study reviews and critiques a number of alternative approaches to generating an adjectival scale, and evaluates the performance of the original model on an expanded dataset and on the datasets of alternative models. The conclusion is clear: continuous space word representations are meaningful, but are inconsistent in determining adjectival scales.

1 Introduction

Continuous space word representations generated by neural networks capture syntactic and semantic meaning. The continuous model creates an n-dimensional space to represent a word, as compared to an n-gram model, which more directly bounds words to their discrete contexts. This makes them ideal to examine more complex relationships between words.

This paper attempts to use that meaning to construct a scale for adjective word representations. Using precomputed word representations, I map out the relationships between adjectives, under the assumption that the relationship is linear. There are a number of distance metrics one can use, depending on what attributes one wants to highlight.

Cosine similarity is one way to measure where a particular word fits on the scale (or what word fits at a particular point on the scale). Another is simple Euclidean distance. For example, to find the comparative adjective, one can find the word closest to the middle of the superlative and base adjective. Similarly, one can determine which adjective best fits a scale, when given a number of options, by measuring their similarity to words on the scale.

Our model trains the word2vec model developed by Mikolov et al. (2013a), on the Google News data set (6B words, 3M word vectors with 300 dimensions). Our test set includes adjectival scales introduced by Wilkinson and Oates (2016) and de Melo and Bansal (2013), as well as a more expansive dataset generated using the intensity scales introduced by Taboada et al. (2011).

We generate both *full* and *half* scales using the datasets and test the performance of our model on both. The difference between a *full* adjectival scale and a *half* adjectival scale is a matter of extremes. We define a *full* adjectival scale as an adjectival scale that goes from antonym to antonym, centering around a neutral or transitioning adjective. By contrast, I define a *half* adjectival scale as an adjectival scale that only has increasing intensity, centering around a comparative adjective. So, for example, *hot*, *lukewarm*, *cold* versus *tepid*, *warmer*, *hot* are full scale and half scale, respectively.

We compare our approach and results to those of Wilkinson and Oates (2016) and de Melo and Bansal (2013). Notably, I do not use the question-answer approach used by Kim and de Marneffe (2013) nor their IQAP data-set to determine accuracy, opting instead for a more explicit generation of an adjectival scale.

2 Model and related work

This paper is based on the observations and experiments of Kim and de Marneffe (2013), which use the continuous word representations described by

Mikolov et al. (2011) and expanded on in the recurrent neural network language model (RNNLM) discussed in Mikolov et al. (2013c). That paper trains the RNNLM on the Broadcast News dataset (320M words) with dimensionality 1,600.

I use a slightly different approach, word2vec, described in Mikolov et al. (2013a) (specifically the skip-gram and CBOW models) and trained on the Google News dataset (3M word vectors) with dimensionality 300.

To summarize these two models (from Mikolov et al. (2013a)):

First, continuous word vectors are learned using a simple model (will be explained later), and then the N-gram NNLM is trained on top of these word representations. The main difference between CBOW and skip-grams is that CBOW presents a word based on the surrounding contexts, and skip-grams presents contexts based on a word.

The probabilistic feedforward NNLM consists of input, projection, hidden layers, and output layers (though some models don't need any hidden layers!). At the input layer, N previous words are encoded using 1-of-V coding, where V is size of the vocabulary. The input layer is then projected to a projection layer P that has dimensionality N $\times$ D, using a shared projection matrix.

Continuous Bag of Words (CBOW) : CBOW is a variation of the feedforward NNLM. The weight matrix between the input and the projection layer is shared for all word positions in the same way as in the NNLM. All words get projected into the same position. This model is called "continuous bag of words" because it uses continuous word representations (unlike standard bag of words), that do not rely on word ordering (like standard bag of words).

Skip-gram: Skip-grams are similar to CBOW, but instead of predicting the current word based on the context, it tries to maximize classification of a word based on another word in the same sentence. Each current word as an input to a log-linear classifier with continuous projection layer and predicts words within a certain range before and after the current word. Increasing the range improves quality of the resulting word vectors. Since the more distant words are usually less related to the current word than those close to it, it gives less weight to the distant words by sampling less from those words in the training examples.

As shown in Mikolov et al. (2013b), the objective of the skip-gram model is to maximize the average log probability. Mikolov discusses a variety of approaches to optimize this calculation and to balance accuracy and training examples.

3 Data

As discussed above, I use the "gold-standard" adjectival scales (half and full) from both Wilkinson and Oates (2016) and de Melo and Bansal (2013), as well as generated adjectival scales using the intensity data provided by Taboada et al. (2011). I only include scales that have three or greater adjectives in the scale (since adjective pairs are not too useful to compare for our purposes). I run experiments on both the half scales and the full scales, but have separated the results into (Table 1) for half scales and (Table 2) for full scales.

Oates (Wilkinson and Oates (2016)) This dataset is simply 12 "gold-standard" full adjectival scales ranging in size (from four to seven adjectives) and complexity (defined loosely as a measure of how abstract the adjectives are). An example of a complex scale would be: *same, alike, similar, different*. A simple scale would be: *freezing, cold, warm, hot*. These scales were generated, cleaned, and sourced by crowd-sourcing answers via Mechanical Turk to determine which adjective was "higher" than the other.

Bansal (de Melo and Bansal (2013)) This dataset has an initial 76 "gold-standard" half adjectival scales that have greater than two values. They begin with full scale sets, which are are extracted from clustering WordNet dumbbell structures, extended with synonyms, and then split into two antonymous halves. I recreate the full scales by comparing the poles of different half scales and cross-listing them with WordNet to determine if they are antonyms. If they are, I join the two antonymous half scales. I effectively reverse the process discussed in de Melo and Bansal (2013). This results in 33 "initial" full scales and 1478 extended full scales. I partition this data into four segments: the initial 76 half scale clusters, the extended 261 half scale development set, a recreation of the original full scales, and then a recreation of the extended full scales.

Taboada (Taboada et al. (2011)) Unlike the other two dataset, this dataset has no "gold-standard" adjectival scale. Instead, I try to use this dataset to create my own. The model discussed in Taboada et al. (2011) is focused around analyzing

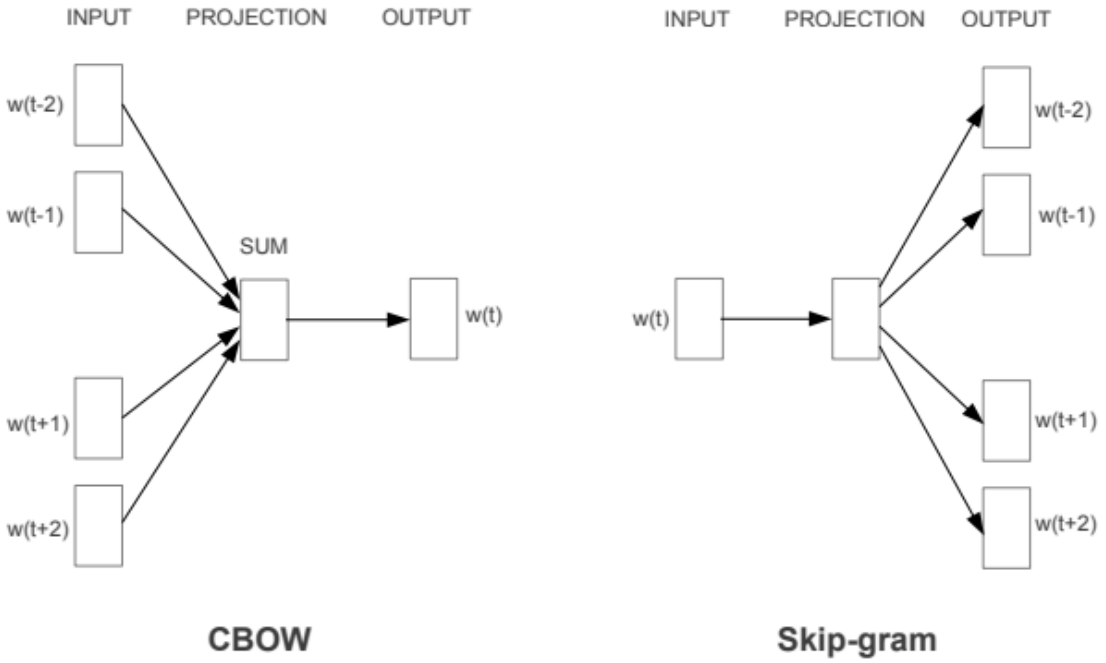

Figure 1: CBOW vs Skip-gram (via Mikolov et al. (2013a))

sentiment from text, and grades words based upon intensity (on a scale of 0 to 5) and sentiment (negative numbers indicate a negative opinion, positive numbers indicate a positive opinion). Here Dr. Marianna Apidianaki was indispensable. She cross-listed the words in the SO-CAL dictionaries with synonym and antonym sets in WordNet. She then created "intensity pairs," which are words in SO-CAL that are matched with their synonyms or antonyms that are also in SO-CAL. The end result was pairs of related words, with their intensity data. For example: *sinful unholy -2 -3*. Here, *unholy* is considered more negative than *sinful*. I was then able to use those pairings to create both full scales and half scales, based around the intensities of the words (as ranked by SO-CAL). This gives me 673 adjectival half scales and 3163 adjectival full scales of mixed quality.

4 Approach

The approach to this problem is similar to the one observed in Mikolov et al. (2013c) and Mikolov et al. (2013a). That is to say, there exists some relationship between these continuous word representations. Introducing the vector offset method: a method of determining the relationship between continuous word representations based around cosine distance. In this method, there is an assumption of that the relationship presented is a vector offset, so that in the embedding space, all pairs of words sharing a particular relation are related by the same constant offset. In this approach, one can approximate a word $y = x_b - x_a + x_c$. For example, Mikolov et al. (2013c) shows that the word closest (via cosine distance) to $king - man + woman$ is $queen$. Recall that cosine similarity is:

$$w^* = argmax_w \frac{x_w y}{||x_w||||y||}$$

This idea is expanded to adjectives in Kim and de Marneffe (2013). Moreover, the authors present the idea that an intermediate vector can be found between two adjectives, such that the middle vector is the intermediate intensity between the two adjectives. Under the assumption that there exists linear relationship between adjectives, they calculate the "middle" vector: $w_m = w_b + \frac{(w_a - w_b)}{2}$ and then calculate the vector most similar to that midpoint. Figure 2 shows their results (extending the idea to include quartiles) for antonymous adjectives.

My goal is to go just a bit farther, to adapt this approach to CBOW and skip-grams (versus an RNNLM) and to experiment with using this method to organize an adjectival scale (versus using the IQAP dataset).

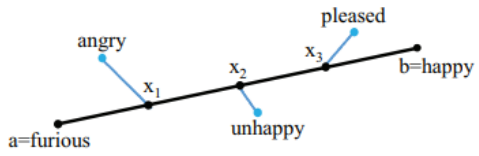

Figure 2: An example of vector with the highest cosine similarity to intermediate points between *furious* and *happy* (via Kim and de Marneffe (2013))

5 Evaluation

6 Discussion and Conclusion

7 General Instructions

Manuscripts must be in two-column format. Exceptions to the two-column format include the title, authors' names and complete addresses, which must be centered at the top of the first page, and any full-width figures or tables (see the guidelines in Subsection **??**). **Type single-spaced.** Start all pages directly under the top margin. See the guidelines later regarding formatting the first page. The manuscript should be printed single-sided and its length should not exceed the maximum page limit described in Section **??**. Pages are numbered for initial submission. However, **do not number the pages in the camera-ready version**.

By uncommenting `\aclfinalcopy` at the top of this document, it will compile to produce an example of the camera-ready formatting; by leaving it commented out, the document will be anonymized for initial submission. When you first create your submission on softconf, please fill in your submitted paper ID where `***` appears in the `\def\aclpaperid{***}` definition at the top.

The review process is double-blind, so do not include any author information (names, addresses) when submitting a paper for review. However, you should maintain space for names and addresses so that they will fit in the final (accepted) version. The NAACL-HLT 2019 LaTeX style will create a titlebox space of 2.5in for you when `\aclfinalcopy` is commented out.

The author list for submissions should include all (and only) individuals who made substantial contributions to the work presented. Each author listed on a submission to NAACL-HLT 2019 will be notified of submissions, revisions and the final decision. No authors may be added to or removed from submissions to NAACL-HLT 2019 after the submission deadline.

Type of Text	Font Size	Style
paper title	15 pt	bold
author names	12 pt	bold
author affiliation	12 pt	
the word "Abstract"	12 pt	bold
section titles	12 pt	bold
document text	11 pt	
captions	10 pt	
abstract text	10 pt	
bibliography	10 pt	
footnotes	9 pt	

Table 1: Font guide.

7.1 Sections

Headings: Type and label section and subsection headings in the style shown on the present document. Use numbered sections (Arabic numerals) in order to facilitate cross references. Number subsections with the section number and the subsection number separated by a dot, in Arabic numerals. Do not number subsubsections.

Citations: Citations within the text appear in parentheses as (**?**) or, if the author's name appears in the text itself, as Gusfield (**?**). Using the provided LaTeX style, the former is accomplished using `\cite` and the latter with `\shortcite` or `\newcite`. Collapse multiple citations as in (**??**); this is accomplished with the provided style using commas within the `\cite` command, *e.g.*, `\cite{Gusfield:97,Aho:72}`. Append lowercase letters to the year in cases of ambiguities. Treat double authors as in (**?**), but write as in (**?**) when more than two authors are involved. Collapse multiple citations as in (**??**). Also refrain from using full citations as sentence constituents.

We suggest that instead of

> "(**?**) showed that ..."

you use

> "Gusfield (**?**) showed that ..."

If you are using the provided LaTeX and BibTeX style files, you can use the command `\citet` (cite in text) to get "author (year)" citations.

If the BibTeX file contains DOI fields, the paper title in the references section will appear as a hyperlink to the DOI, using the hyperref LaTeX

output	natbib	previous ACL style files
(?)	\citep	\cite
?	\citet	\newcite
(?)	\citeyearpar	\shortcite

Table 2: Citation commands supported by the style file. The citation style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

package. To disable the hyperref package, load the style file with the `nohyperref` option:
`\usepackage[nohyperref]{naaclhlt2019}`

Digital Object Identifiers: As part of our work to make ACL materials more widely used and cited outside of our discipline, ACL has registered as a CrossRef member, as a registrant of Digital Object Identifiers (DOIs), the standard for registering permanent URNs for referencing scholarly materials. As of 2017, we are requiring all camera-ready references to contain the appropriate DOIs (or as a second resort, the hyperlinked ACL Anthology Identifier) to all cited works. Thus, please ensure that you use BibTeX records that contain DOI or URLs for any of the ACL materials that you reference. Appropriate records should be found for most materials in the current ACL Anthology at http://aclanthology.info/.

As examples, we cite (?) to show you how papers with a DOI will appear in the bibliography. We cite (?) to show how papers without a DOI but with an ACL Anthology Identifier will appear in the bibliography.

As reviewing will be double-blind, the submitted version of the papers should not include the authors' names and affiliations. Furthermore, self-references that reveal the author's identity, *e.g.*,

> "We previously showed (?) ..."

should be avoided. Instead, use citations such as

> "? (?) previously showed ... "

Any preliminary non-archival versions of submitted papers should be listed in the submission form but not in the review version of the paper. NAACL-HLT 2019 reviewers are generally aware that authors may present preliminary versions of their work in other venues, but will not be provided the list of previous presentations from the submission form.

Please do not use anonymous citations and do not include when submitting your papers. Papers that do not conform to these requirements may be rejected without review.

References: Gather the full set of references together under the heading **References**; place the section before any Appendices. Arrange the references alphabetically by first author, rather than by order of occurrence in the text. By using a .bib file, as in this template, this will be automatically handled for you. See the `\bibliography` commands near the end for more.

Provide as complete a citation as possible, using a consistent format, such as the one for *Computational Linguistics* or the one in the *Publication Manual of the American Psychological Association* (?). Use of full names for authors rather than initials is preferred. A list of abbreviations for common computer science journals can be found in the ACM *Computing Reviews* (?).

The LaTeX and BibTeX style files provided roughly fit the American Psychological Association format, allowing regular citations, short citations and multiple citations as described above.

- Example citing an arxiv paper: (?).
- Example article in journal citation: (?).
- Example article in proceedings, with location: (?).
- Example article in proceedings, without location: (?).

See corresponding .bib file for further details.

Submissions should accurately reference prior and related work, including code and data. If a piece of prior work appeared in multiple venues, the version that appeared in a refereed, archival venue should be referenced. If multiple versions of a piece of prior work exist, the one used by the authors should be referenced. Authors should not rely on automated citation indices to provide accurate references for prior and related work.

Appendices: Appendices, if any, directly follow the text and the references (but see above).

Letter them in sequence and provide an informative title: **Appendix A. Title of Appendix**.

7.2 Footnotes

Footnotes: Put footnotes at the bottom of the page and use 9 point font. They may be numbered or referred to by asterisks or other symbols.[1] Footnotes should be separated from the text by a line.[2]

7.3 Graphics

Illustrations: Place figures, tables, and photographs in the paper near where they are first discussed, rather than at the end, if possible. Wide illustrations may run across both columns. Color illustrations are discouraged, unless you have verified that they will be understandable when printed in black ink.

Captions: Provide a caption for every illustration; number each one sequentially in the form: "Figure 1. Caption of the Figure." "Table 1. Caption of the Table." Type the captions of the figures and tables below the body, using 10 point text. Captions should be placed below illustrations. Captions that are one line are centered (see Table 1). Captions longer than one line are left-aligned (see Table **??**). Do not overwrite the default caption sizes. The naaclhlt2019.sty file is compatible with the caption and subcaption packages; do not add optional arguments.

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

References

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1625–1630.

Gerard de Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, pages 1:279–290.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *ICLR Workshop*.

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

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

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

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stedel. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics 2011 Vol. 37*, pages 267–307.

Bryan Wilkinson and Tim Oates. 2016. A gold standard for scalar adjectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

[1] This is how a footnote should appear.

[2] Note the line separating the footnotes from the text.