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. (2013b). 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. I primarily use the CBOW model, but it is worth presenting both. I standardized the dataset so that the mean and the variance of the representations are 0 and 1, respectively. 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 RNNLM 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.

Continuous Bag of Words (CBOW) :

CBOW is similar to the NNLM

The first proposed architecture is similar to the feedforward NNLM, where the non-linear hidden layer is removed and the projection layer is shared for all words (not just the projection matrix); thus, all words get projected into the same position (their vectors are averaged). We call this architecture a bag-of-words model as the order of words in the history does not influence the projection. Furthermore, we also use words from the future; we have obtained the best performance on the task introduced in the next section by building a log-linear classifier with four future and four history words at the input, where the training criterion is to correctly classify the current (middle) word. Training complexity is then Q = N D + D log2(V). (4) We denote this model further as CBOW, as unlike standard bag-of-words model, it uses continuous distributed representation of the context. The model architecture is shown at Figure 1. Note that the weight matrix between the input and the projection layer is shared for all word positions in the same way as in the NNLM.

Skip-Gram: The second architecture is 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. More precisely, we use each current word as an input to a log-linear classifier with continuous projection layer, and predict words within a certain range before and after the current word. We found that increasing the range improves quality of the resulting word vectors, but it also increases the computational complexity. Since the more distant words are usually less related to the current word than those close to it, we give less weight to the distant words by sampling less from those words in our training examples. The training complexity of this architecture is proportional to Q = C (D + D log2(V)), (5) where C is the maximum distance of the words. Thus, if we choose C = 5, for each training word we will select randomly a number R in range ¡ 1; C ¿, and then use R words from history and

R words from the future of the current word as correct labels. This will require us to do R 2 word classifications, with the current word as input, and each of the R + R words as output. In the following experiments, we use C = 10.

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 88 ”gold-standard” half adjectival scales. 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

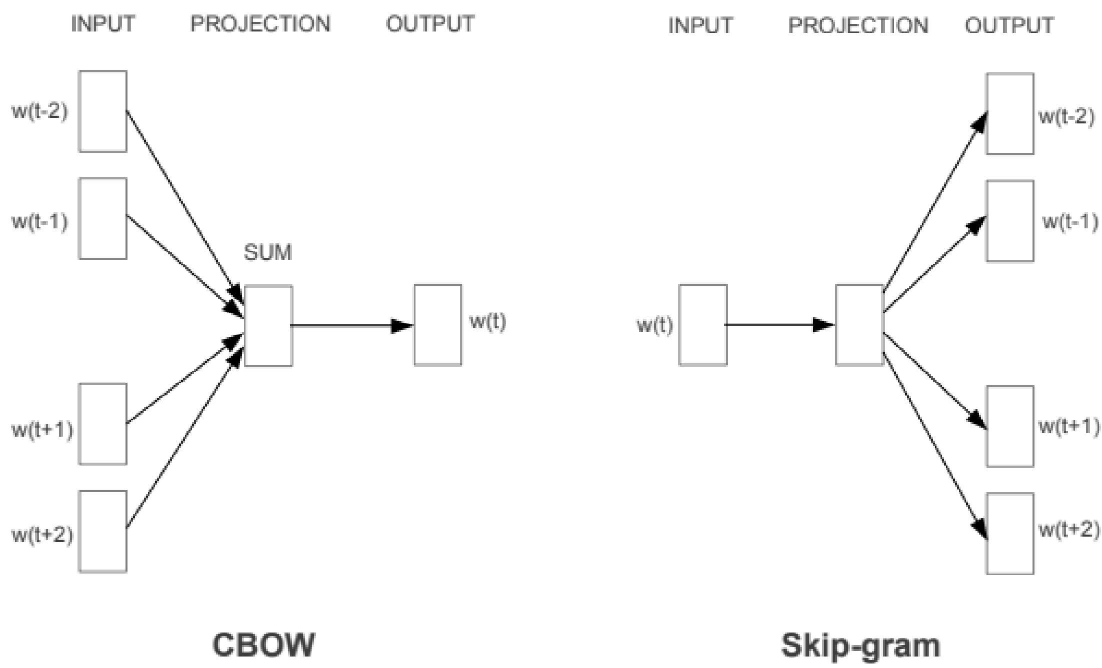

Figure 1: CBOW vs Skip-gram (via Mikolov et al. (2013a))

partition this data into four segments: the initial 88 half scale clusters, the extended XXX half scale development set, a recreation of the original full scales, and then a recreation of the extended full scales. 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 XX "initial" full scales and XX 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 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. (2013b) and Mikolov et al. (2013a). That is to say: there exists some relationship between these continuous word representations

Following Kim and de Marneffe (2013), we assume there exists linear relationship between adjectives.

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.

7.1 The Ruler

The NAACL-HLT 2019 style defines a printed ruler which should be presented in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document without the provided style files, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler. (LaTeX users may uncomment the `\aclfinalcopy` command in the document preamble.)

Reviewers: note that the ruler measurements do not align well with lines in the paper – this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. In most cases one would expect that the approximate location will be adequate, although you can also use fractional references (*e.g.*, the first paragraph on this page ends at mark 108.5).

7.2 Electronically-available resources

NAACL-HLT provides this description in LaTeX2e (`naaclhlt2019.tex`) and PDF format (`naaclhlt2019.pdf`), along with the LaTeX2e style file used to format it (`naaclhlt2019.sty`) and an ACL bibliography style (`acl_natbib.bst`) and example bibliography (`naaclhlt2019.bib`). These files are all available at `http://naacl2019.org/downloads/naaclhlt2019-latex.zip`. We strongly recommend the use of these style files, which have been appropriately tailored for the NAACL-HLT 2019 proceedings.

7.3 Format of Electronic Manuscript

For the production of the electronic manuscript you must use Adobe's Portable Document Format (PDF). PDF files are usually produced from LaTeX using the *pdflatex* command. If your version of LaTeX produces Postscript files, you can convert these into PDF using *ps2pdf* or *dvipdf*. On Windows, you can also use Adobe Distiller to generate PDF.

Please make sure that your PDF file includes all the necessary fonts (especially tree diagrams, symbols, and fonts with Asian characters). When you print or create the PDF file, there is usually an option in your printer setup to include none, all or just non-standard fonts. Please make sure that you select the option of including ALL the fonts. **Before sending it, test your PDF by printing it from a computer different from the one where it was created.** Moreover, some word processors may generate very large PDF files, where each page is rendered as an image. Such images may reproduce poorly. In this case, try alternative ways to obtain the PDF. One way on some systems is to install a driver for a postscript printer, send your document to the printer specifying "Output to a file", then convert the file to PDF.

It is of utmost importance to specify the **A4 format** (21 cm x 29.7 cm) when formatting the paper. When working with `dvips`, for instance, one should specify `-t a4`. Or using the command `\special{papersize=210mm,297mm}` in the latex preamble (directly below the `\usepackage` commands). Then using `dvipdf` and/or `pdflatex` which would make it easier for some.

Print-outs of the PDF file on A4 paper should be identical to the hardcopy version. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs as soon as possible.

7.4 Layout

Format manuscripts two columns to a page, in the manner these instructions are formatted. The exact dimensions for a page on A4 paper are:

- Left and right margins: 2.5 cm
- Top margin: 2.5 cm
- Bottom margin: 2.5 cm
- Column width: 7.7 cm
- Column height: 24.7 cm
- Gap between columns: 0.6 cm

Papers should not be submitted on any other paper size. If you cannot meet the above requirements about the production of your electronic submission, please contact the publication chairs above as soon as possible.

7.5 Fonts

For reasons of uniformity, Adobe's **Times Roman** font should be used. In LaTeX2e this is accomplished by putting

```
\usepackage{times}
\usepackage{latexsym}
```

in the preamble. If Times Roman is unavailable, use **Computer Modern Roman** (LaTeX2e's default). Note that the latter is about 10% less dense than Adobe's Times Roman font.

7.6 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.

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.

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 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

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.

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 (Goodman et al., 2016) to show you how papers with a DOI will appear in the bibliography. We cite (Harper, 2014) 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**.

7.7 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]

[1] This is how a footnote should appear.

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

7.8 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.

Preparing References:

Include your own bib file like this:

```
\bibliographystyle{acl_natbib}
\bibliography{naaclhlt2019}
```

where `naaclhlt2019` corresponds to a naaclhlt2019.bib file.

References

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for abstract meaning representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11. Association for Computational Linguistics.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, page 1. Dublin City University and Association for Computational Linguistics.

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, Wen-tau Yih, and Geoffrey Zweig. 2013b. 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)*.