Final Project Proposal

Jasmine Huang
Chuqing Zhao
CMSC 35300
Autumn 2021

Proposed Questions

The meaning of a word can vary across extralinguistic contexts. To capture these phenomena, various types of dynamic word embedding have been proposed. In this project, we will use Singular Value Decomposition (SVD) and Positive Pointwise Mutual Information (PPMI) to transform natural language (English) into dense vector representations and compare their performances with human judgements on MEN similarity tasks[1].

Additionally, we want to explore an orthogonal Procrustes method that aligns multiple learned low-dimensional embeddings to allow comparisons of word embeddings across time. This dynamic word embedding method factors in the temporal effect and is critical in understanding semantic shifts (Hamilton et. al, 2018).

We plan to compute embeddings for three sets of data. The first set of data includes over 18,000 lyrics from Spotify and Billboard Top 100 from 1950 to 2020[2]. The second source of data includes 581,684 historical patent data scraped from USPTO API[3]. The third dataset source data is news dataset from the Corpus of News on the Web (NOW)[4].

We purposefully choose these three corpora to represent the unique challenge when applying general pertained embeddings on domain-specific corpora. Compared to the NOW dataset, patent and lyric data contain industry jargons or slangs/expressions that require cultural knowledge to understand. We hope to contrast SVD and PPMI methods for word representations on both generic news corpus and domain-specific corpus, to draw recommendations on which method would be more appropriate in different thematic contexts.

Literature Review

1. Word embedding algorithms

There are two common methods to construct word embeddings: Positive Pointwise Mutual Information (PPMI) and SVD. These methods represent the word semantics in co-occurrence matrices.

PPMI represents the vector embedding for each word containing the positive point-wise mutual information values between the word and a large set of pre-specified context words (Hamilton et. al,

[1] MEN similarity task: https://staff.fnwi.uva.nl/e.bruni/MEN
[2] Billboard "The Hot 100" Songs data: https://www.kaggle.com/dhruvildave/billboard-the-hot-100-songs
[3] Patent USPTO data: https://developer.uspto.gov/api-catalog
[4] Corpus of News on the Web: https://digital.library.unt.edu/ark:/67531/metadc1234358/

2018). Since raw word frequency is quite skewed and captures less informative words such as “the” and “of”, PPMI evaluates whether a context word is particularly informative about the target word. Pointwise mutual information (PMI) ranges from −∞ to + ∞. Negative values represent the co-occurrence is less than we expected by chance, whereas the positive values are more commonly used to evaluate the relatedness between two words. Clipping the PPMI values above zero ensures they remain finite and has been shown to dramatically improve results. PPMI vectors tend to generate long and sparse vectors. However, long and sparse vectors might not be good at capturing synonymy.

$$\mathbf{M}_{i,j}^{\text{PPMI}} = \max\left\{\log\left(\frac{\hat{p}(w_i, c_j)}{\hat{p}(w)\hat{p}(c_j)}\right) - \alpha, 0\right\},$$

SVD embeddings perform singular value decomposition on PPMI embeddings to extract a low-dimensional approximation of the PPMI matrix. (Levy et al., 2015). Previous research has demonstrated the robustness of SVD representations, as the dimensionality reduction functions as a form of regularization (Hamilton et.al, 2018).

$$\mathbf{w}_i^{\text{SVD}} = (\mathbf{U}\boldsymbol{\Sigma}^{\gamma})_i\,,$$

2. Aligning historical embeddings

We can use orthogonal Procrustes to align the learned low-dimensional embeddings. Defining W(t) ∈ R d×|V| as the matrix of word embeddings learned at year t, we align across time-periods while preserving cosine similarities by optimizing:

$$\mathbf{R}^{(t)} = \arg\min_{\mathbf{Q}^{\top}\mathbf{Q}=\mathbf{I}} \|\mathbf{Q}\mathbf{W}^{(t)} - \mathbf{W}^{(t+1)}\|_F,$$

With R(t) ∈ R d×d . The solution corresponds to the best rotational alignment and can be obtained efficiently using an application of SVD (Hamilton et al., 2018).

References:

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word   co-occurrence statistics: A computational study. Behavior research methods, 39(3), 510-526.

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. Behavior research methods, 44(3), 890-907.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016b). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, August 2016, 1489-1501. https://doi.org/10.18653/v1/P16-1141

Levy, Omer & Goldberg, Yoav & Dagan, Ido. (2015). Improving Distributional Similarity with Lessons Learned from Word Embeddings. Transactions of the Association for Computational Linguistics. 3. 211-225. 10.1162/tacl_a_00134.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37, 141-188.

=