Visualizing Low-Dimensional Word Embeddings with Emoji Annotators

Yoshinari Fujinuma*
Department of Computer Science
University of Colorado Boulder

Shantanu Karnwal†
Department of Computer Science
University of Colorado Boulder

Abstract

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Index Terms: Human-centered computing—Visualization—Visualization techniques—Treemaps; Human-centered computing—Visualization—Visualization design and evaluation methods

1 Introduction

Word embeddings, especially cross-lingual embeddings, has been successful in multiple NLP applications such as machine translation [1, 4] and cross-lingual document classification [3]. One way of exploring such embeddings is to enable interaction between humans and visualizations. However, there are potential problems of naively displaying the word embeddings projected onto 2D space using t-SNE [7], which is mot commonly used in visualizing word embeddings, such as;

- Overlap of words when zoomed out.
- A counter-intuitive features of a t-SNE visualization (e.g., "cluster sizes mean nothing"[1])
- There are many other alternatives to visualize word embeddings than commonly used t-SNE (e.g., UMAP [5] or k-Nearest Neighbor graph), but no thorough comparison conducted.

Figure **??** shows an example of k-nearest neighbor graph and Figure **??** shows an example of visualization using t-SNE.

In this project, we would like to accomplish the followings:

- Visualize the change of word vectors while training skip-gram with negative sampling (SGNS) model [6] using t-SNE.
- Compare visualizations of word vectors between t-SNE, UMAP, k-nearest neighbor graph, or any other methods (any suggestions are welcome).

2 Description of the Project

Figure 1 shows the output of our visualization.

Our approach for constructing the visualization is as follows:

1. Train a word embedding

*e-mail: Yoshinari.Fujinuma@colorado.edu
†e-mail: Shantanu.Karnwal@colorado.edu

[1] https://distill.pub/2016/misread-tsne/

Figure 1: The visualizaiton of word embeddings using clustered network and emojis.

2. Run k-means and obtain clusters
3. Assign an emoji to each cluster
4. Visualize using D3.js

2.1 Annotation of Clusters with Emojis

When a human look at emoji, one connects with various possible concepts. Searching for a right word to represent the cluster requires external linguistic resources e.g., WordNet. However, images does not associate a single word. For example, when one looks at 🍅, the possible association of this words are "tomato", "vegatable", "food", or even "object". Therefore, we decide to use emojis to represent the clusters.

2.2 Interaction

Users can click on emojis to "drill-down" [2] the cluster and look into which words are in the cluster.

2.3 Force Layout

To solve the problem of overlapping texts, we also use force layout in D3.js to let the texts and emojis move and draggable.

3 Discussion

Outliers

Emojis are diverse

References

[1] M. Artetxe, G. Labaka, E. Agirre, and K. Cho. Unsupervised neural machine translation. In *Proceedings of the International Conference on Learning Representations*, 2018.

[2] N. Elmqvist and J.-D. Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE Transactions on Visualization and Computer Graphics*, 16(3):439–454, May 2010. doi: 10.1109/TVCG.2009.84

[3] A. Klementiev, I. Titov, and B. Bhattarai. Inducing crosslingual distributed representations of words. In *Proceedings of International Conference on Computational Linguistics*, 2012.

[4] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the International Conference on Learning Representations*, 2018.

[5] L. McInnes and J. Healy. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426, 2018.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems*, 2013.

[7] L. Van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 2008.