Visual Exploration and Comparison of Word Embeddings

Juntian Chen, Yubo Tao, Hai Lin

PII: S1045-926X(18)30124-1
DOI: https://doi.org/10.1016/j.jvlc.2018.08.008
Reference: YJVLC 856

To appear in: *Journal of Visual Languages and Computing*

Received date: 20 July 2018
Accepted date: 20 August 2018

Please cite this article as: Juntian Chen, Yubo Tao, Hai Lin, Visual Exploration and Comparison of Word Embeddings, *Journal of Visual Languages and Computing* (2018), doi: https://doi.org/10.1016/j.jvlc.2018.08.008

Visual Exploration and Comparison of Word Embeddings

Juntian Chen[a], Yubo Tao[a], Hai Lin[a]

[a]*State Key Lab of CAD&CG, Zhejiang University*

Abstract

Word embeddings are distributed representations for natural language words, and have been wildly used in many natural language processing tasks. The word embedding space contains local clusters with semantically similar words and meaningful directions, such as the analogy. However, there are different training algorithms and text corpora, which both have a different impact on the generated word embeddings. In this paper, we propose a visual analytics system to visually explore and compare word embeddings trained by different algorithms and corpora. The word embedding spaces are compared from three aspects, i.e., local clusters, semantic directions and diachronic changes, to understand the similarity and differences between word embeddings.

Keywords: Visual Comparison, Word Embeddings

1. Introduction

The word embedding is a kind of mathematical representation of vocabulary. Usually, there are two kinds of representations: one-hot vector representation and distributed representation. One-hot vector representation easily comes to our minds, which uses the index in the dictionary to represent word uniquely. However, this representation method only separates word and does not express the semantic meanings of the word. Distributed representation is a vector of real numbers, originally proposed by Hinton [1] in 1986, and it can encode semantic

Email addresses: chenjuntian@zju.edu.cn (Juntian Chen), taoyubo@cad.zju.edu.cn (Yubo Tao), lin@cad.zju.edu.cn (Hai Lin)