Visual Exploration of Neural Document Embedding in Information Retrieval: Semantics and Feature Selection

Xiaonan Ji, Han-Wei Shen, *Member, IEEE*, Alan Ritter, Raghu Machiraju, and Po-Yin Yen

Abstract—Neural embeddings are widely used in language modeling and feature generation with superior computational power. Particularly, neural document embedding - converting texts of variable-length to semantic vector representations - has shown to benefit widespread downstream applications, e.g., information retrieval (IR). However, the black-box nature makes it difficult to understand how the semantics are encoded and employed. We propose visual exploration of neural document embedding to gain insights into the underlying embedding space, and promote the utilization in prevalent IR applications. In this study, we take an IR application-driven view, which is further motivated by biomedical IR in healthcare decision-making, and collaborate with domain experts to design and develop a visual analytics system. This system visualizes neural document embeddings as a configurable document map and enables guidance and reasoning; facilitates to explore the neural embedding space and identify salient neural dimensions (semantic features) per task and domain interest; and supports advisable feature selection (semantic analysis) along with instant visual feedback to promote IR performance. We demonstrate the usefulness and effectiveness of this system and present inspiring findings in use cases. This work will help designers/developers of downstream applications gain insights and confidence in neural document embedding, and exploit that to achieve more favorable performance in application domains.

Index Terms—Neural document embedding, information retrieval, semantic analysis, feature selection

1 Introduction

Document representation is an important topic in text analytics and language modeling, aiming to encode essential features and underlying meanings of documents in a structured and machine understandable manner. Neural document embedding [1], [2], [3], as a successful extension to neural word embedding, has shown to leverage superior computational power of neural networks and generate effective document representations in concise feature vectors. In other words, neural embeddings not only mitigate the curse of dimensionality, but also demonstrate to capture text hidden semantics and outperform lexical and conventional embedding methods. The resulting semantic representations substantially benefit downstream applications [4], such as information retrieval (IR), sentiment analysis, sentence modeling, machine translation, etc. In particular, IR typically involves text similarity, clustering, or classification, and aims to identify relevant documents for an information need. It relies on effective document representations to capture document meanings and characterize document relevancy. The remarkable performance of neural embedding empowers IR to cope with the growing volume of information resources and fulfill critical needs.

Despite the superior performance in IR and other text analytics applications, neural document embedding is usually used as a black-box, and it is difficult to understand how the performance is achieved or how to tune the performance in different conditions. Many studies attempt to evaluate neural embedding models, but they generally rely on trial-and-error or benchmarks with limited characteristics [5]. Due to the presence of boundless analytic facets, noise, and redundancies in neural embedding, it is challenging to make thorough interpretations even with sophisticated mathematical tools. Thus, for prevalent IR applications, it remains unclear how underlying document meanings (e.g., semantics) are encoded in the hidden states (e.g., features or dimensions) in the embedding space, and how that contributes to IR performance, such as document semantics, similarity, classification, or clustering (of relevant documents). Moreover, we are motivated by IR applications in the biomedical and clinical domain, where advanced IR is needed to accelerate system review (SR) production in healthcare. In a real-world and critical SR, relevant biomedical documents (e.g., published studies or clinical trials) embracing high-quality research findings are retrieved to guide patient care and inform clinical decisions, thus supporting Evidence-based Practice (EBP). While the IR applications can benefit from

- *X. Ji is with the Institute for Informatics, Washington University, School of Medicine, St. Louis, MO 63108. E-mail: ji.62@osu.edu.*
- *H.-W. Shen, A. Ritter, and R. Machiraju are with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210. E-mail: {shen.94, ritter.1492, machiraju.1}@osu.edu.*
- *P.-Y. Yen is with the Institute for Informatics, Washington University, School of Medicine, St. Louis, MO 63108 and also with the Goldfarb School of Nursing, BJC Healthcare, St. Louis, MO 63108. E-mail: yenp@wustl.edu.*

Manuscript received 5 Oct. 2018; revised 21 Dec. 2018; accepted 4 Jan. 2019. Date of publication 15 Mar. 2019; date of current version 1 May 2019.
(Corresponding author: Xiaonan Ji.)
Recommended for acceptance by R. Maciejewski, J. Seo, and R. Westermann.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2019.2903946