MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220266

t-SNE Visualization of Vector Pairs of Similar and Dissimilar Definition Sentences Created by Word2vec and Doc2vec in Japanese Medical Device Adverse Event Terminology

Ayako Yagahara[a,b], Masahito Uesugi[c], Hideto Yokoi[d]

[a] *Department of Radiological Technology, Hokkaido University of Science, Sapporo, Hokkaido, Japan,*
[b] *Faculty of Health Sciences, Hokkaido University, Sapporo, Hokkaido, Japan,*
[c] *Department of Medical Management and Informatics, Hokkaido Information University, Sapporo, Hokkaido, Japan,*
[d] *Department of Medical Informatics, Kagawa University Hospital, kita-gun, Kagawa, Japan*

Abstract

The purpose of our study is to identify the patterns of vectors in similar/dissimilar pairs of definition sentence created by Word2vec and doec2vec for elaboration of the terminology for Japanese Medical Device Adverse Events. 2-dimension vector space created by t-SNE showed that the pair with true positive located closer in a vector space, especially Doc2vec had a strong tendency. Comparing with Word2vec, Similar vectors in Doc2vec were close and tended to form clusters.

Keywords:

Machine learning, Terminology, Equipment and Supplies.

Introduction

Medical facilities and medical device manufacturing companies in Japan are required to submit medical device adverse event reports (MDAERs) to the Ministry of Health, Labor and Welfare when a medical device malfunction occurs during medical procedure. To standardize the terms in MDEARs, Terminology of Japan Federation of Medical Devices Associations (JFMDA terminology) was published in March 2015 [1]. The 1st edition of the terminology consists of 89 medical devices terminology items developed by 13 industry groups independently. We have been trying to map these terminology items to establish consistency of JFMDA terminology.

In our previous study, we tried to identify similar definition sentences by machine learning methods to detect synonyms [2]. The results showed that the accuracy in edit distances were better than those in word embedding methods, and edit distances were useful for similar definition sentence pairs with common phrases. On the other hand, word embedding methods enabled detection of similar definition sentence pairs with different phrases occasionally. Detecting synonyms with different phrase automatically for work efficiency, it is important to identify the patterns. The t-distributed stochastic neighbor embedding (t-SNE), which is a dimension reduction technique, is useful for visualization to capture much more complex high-dimensional patterns.

The purpose of this study is to identify the patterns of the vector distributions in similar/dissimilar definition sentence pairs using t-SNE.

Methods

We used the training models created by Word2vec and Doc2vec with 300 dimension using Japanese Wikipedia in our previous study [2]. There were 125 pairs of definition sentences as the baseline in medical device problem, and each pair was assigned similar/dissimilar labels by specialists. Vectors of definition sentences in Word2vec were calculated an average vector using vectors allocated to each word in a sentence. Those in Doc2vec were inferred using the gensim package. The accuracy of synonym detection using four models from the previous study [2] are shown in Table 1. Reducing the vector dimensions by t-SNE, we used scikit-learn package and the parameters were as follows: perplexity was 40, iteration number was 2500. The sentence vectors were illustrated and identified the patterns of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) pairs. All experiments were run on a computer with the Ubuntu 18.04 operating system, Intel Core i7-9700K, and 64 GB RAM, with the Programming language Python 3.8.3.

Table 1–Accuracy of the algorithm [2]. AUC is area under curve in receiver operating characteristic analysis. CBOW is continuous bag of words, DBOW is distributed bag of words, DM is distributed memory model

Model	Sensitivity	Specificity	AUC
Word2vec with CBOW	0.600	0.733	0.707
Word2vec with skip-gram	0.560	0.787	0.723
Doc2vec with DBOW	**0.640**	0.853	**0.768**
Doc2vec with DM	0.480	**0.880**	0.681

Results

The results of t-SNE 2D maps using all sentence vectors in each model are illustrated in Figure 1. The distributions of sentence vectors in Word2vec tended to be widely. On the other hand, several clusters were found in Doc2vec. The result in comparison of the distribution in TP pairs, the pairs of the sentences tended to be closer to each other in Doc2vec with DBOW which has the highest sensitivity. This tendency was strongly found in t-SNE of Doc2vec with DBOW (Figure 2). This may be the reason why Doc2vec was the most sensitive. In the TN pairs of Word2vec, the vectors were distributed widely compared to those in TP. In Doc2vec, there were several clusters (a dotted circle in Figure 3). In the clusters, we found two features: existence of the quite same definition statements used in different definition statement pairs, and existence of definition statements with the same phrase other. The reason for the former is that the vectors of the same document are slightly different in the inference of the vector by Doc2vec or it may be due to the calculation error of t-SNE. In the latter, some cases were detected as FP. In FP, the tendency of the distributions of vectors in all four models was similar. The two vectors in each pair in Doc2vec with DM tended to be located closer than the others. In FN, the vectors tended to be distributed widely.

The vector distributions of two pairs with different phrase correctly identified as TP only Doc2vec with DBOW in previous study [2] were shown in Figure 6. One pair is "Puncturing the wrong part (A1) " and "Accidentally puncturing at a point that was not the target (A2)" and the other is "Puncturing the wrong part (B1)" and "Punctuation to a site that is not the intended area (B2)." Only the vectors in Doc2vec with DBOW were located closer than others, and in Word2vec with skip-gram, the vector positions in the pair were far apart but A2 and B2 were close.

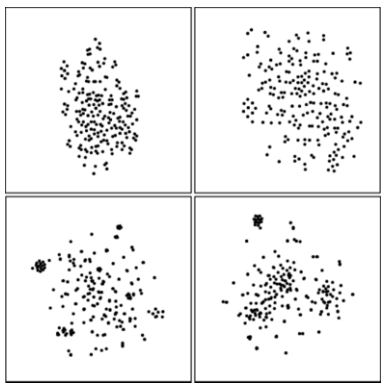

Figure 1 – 2-dimension visualization using t-SNE using all pairs in each model. Upper left: Word2vec with CBOW, upper right: Word2vec with skip-gram, lower left: Doc2vec with DBOW, lower right: Doc2vec with DM.

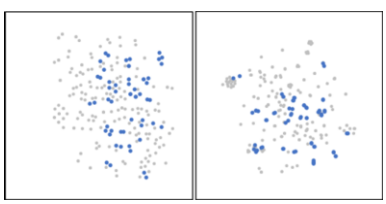

Figure 2 – Visualization of TP pairs (blue points). Left: Word2vec with skip-gram, right: Doc2vec with DBOW.

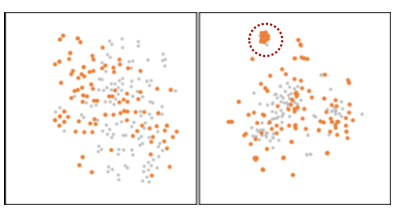

Figure 3 – Visualization of TN pairs (red points). Left: Word2vec with skip-gram, right: Doc2vec with DBOW. Dotted circles indicated the clusters.

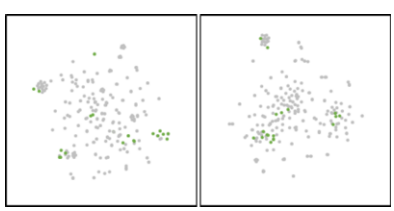

Figure 4 – Visualization of FP pairs (green points). Left: Doc2vec with DBOW, right: Doc2vec with DM.

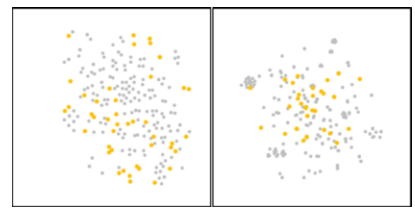

Figure 5 – Example of visualization of FN pairs (yellow points). Left: Word2vec with skip-gram, right: Doc2vec with DBOW.

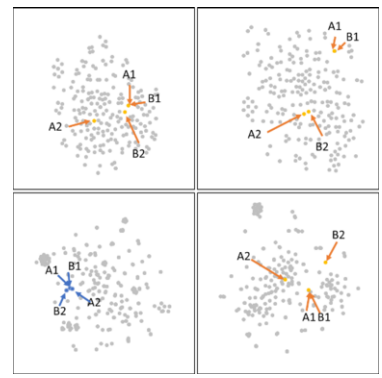

Figure 6 – Vector distributions of two pairs identified as TP only Doc2vec with DBOW. Upper left: Word2vec with CBOW, upper right: Word2vec with skip-gram, lower left: Doc2vec with DBOW, lower right: Doc2vec with DM. Indications refer to the main text.

Conclusions

In this study, we visualized the vectors of definition sentences using t-SNE. The vectors of definition sentences with true positive located closer in a vector space, especially Doc2vec had a strong tendency. In true negative and false negative, the pair of vectors tended to be far apart. Comparing with Word2vec, Similar vectors in Doc2vec were close and tended to form clusters. In future, we plan to conduct further examinations with other parameters such as cosine similarity.

Acknowledgements

This research is supported by the Research on Regulatory Science of Pharmaceuticals and Medical Devices from the Japan Agency for Medical Research and development (AMED). The authors have no conflict of interest to declare.

References

[1] Ministry of Health, Labour and Welfare. Publication and Utilization of Medical Device Adverse Event Terminology (in Japanese). [cited 2021 April 28]; Available from: https://www.pmda.go.jp/files/000204139.pdf.
[2] Yagahara A, Uesugi M, Yokoi H. Identification of Synonyms Using Definition Similarities in Japanese Medical Device Adverse Event Terminology. Appl. Sci. 2021, 11, 3659.

Address for correspondence

Corresponding author: Ayako Yagahara
Mailing address: 7-Jo 15-4-1 Maeda, Teine, Sapporo, Hokkaido, Japan 006-8585
Email: yagahara-a@hus.ac.jp
Phone: +81-11-676-8504