MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190564

Evaluation of Similar Term Definitions in Medical Device Adverse Event Terminology

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

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

Abstract

The purpose of this study is to extract similar term definitions used in the terminology of Japanese medical device adverse events. We employed Levenshtein and Jaro–Winkler distances as edit distances and Skip-gram, continuous-bag of words, and fast text to produce distributed representations in Word2Vec. A comparison of the accuracies of the models showed that Levenshtein distance had higher specificity whereas Skip-gram had higher sensitivity as compared to the other models.

Keywords:

Machine Learning; Vocabulary, Controlled; Equipment and Supplies, Hospital

Introduction

In Japan, medical facilities and medical device manufacturing companies are required to submit medical device adverse event reports (MDAER) to the Ministry of Health, Labor, and Welfare when medical device adverse events (e.g., catheter breakage) occur during a medical procedure. Since detailed descriptions of adverse events and problems caused to patients by medical devices are documented through free-text in MDAER, they are hard to categorize and conduct statistical analysis.

Therefore, to standardize the terms in MDEAR, medical device adverse event terminology (terminology of Japan Federation of Medical Devices Associations: JFMDA terminology) was published in March 2015 [1]. This terminology consists of 89 medical device terminologies developed by 13 industry groups that are members of JFMDA. Each terminology has three categories: medical device problem, patient problem, and component. In addition, each term in "medical device problem" and "patient problem" categories have definition sentences, synonyms, and FDA code from the Center for Devices and Radiological Health (CDRH) Terminology [2] as shown in Fig. 1. Internationally, between 2009 and 2011, ISO TC 210 worked in conjunction with Global Harmonization Task Force (GHTF) and FDA to produce ISO/TS 19218 - 1 Medical devices-Hierarchical coding structure for medical device adverse events, including both event codes (part 1) and evaluation codes (part 2). In 2011, International Medical Device Regulators Forum (IMDRF) was conceived to accelerate international medical device regulatory harmonization and convergence [3].

The 13 industry groups independently constructed each terminology using a bottom-up approach by gathering the terms used regularly in medical facilities to facilitate communication between medical staff and medical device manufacturers. We are now trying to map these terminologies to ensure consistency. One problem associated with mapping these terminologies includes cases where the notation of terms is considered to be the same concept but differed depending on the terminologies. Because there are about 3,500 terms related to medical device problems, manual verification requires great effort.

Thus, the purpose of this study was to detect synonyms automatically using definition sentences of the terms used in the JFMDA terminology.

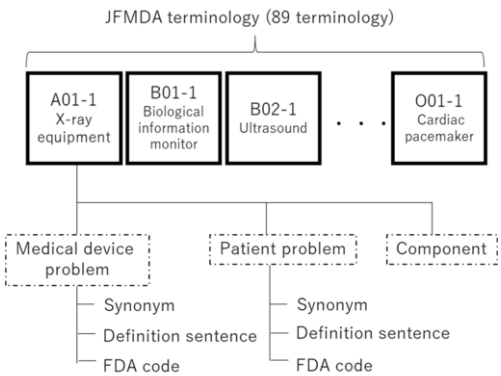

Figure 1–Overview of JFMDA terminology

Methods

Data

Terms used for describing medical device problems were the focus of this study. At first, we extracted the terms and their definition sentences. We made approximately 600,000 definition sentence pairs, of which 125 pairs were extracted arbitrarily.

Similarity Detection

We employed edit distance and Word2Vec as similarity detection methods. Edit distance is an algorithm for quantifying how two dissimilar strings are related to each other by counting the minimum number of operations required to transform one string into the other. Word2Vec is used to group vectors of similar words together into vector space to detect their similarities mathematically.

In this study, we employed Levenshtein [4] and Jaro–Winkler distances [5] as the edit distance and Skip-gram, continuous-bag of words [6], and fast text [7] to produce distributed representations in Word2Vec.

Japanese Wikipedia was used to create a distributed representation model of Word2Vec. In each model, the number of dimensions of the vector was set to 300.

Evaluation

As a gold standard, 50 similar definition pairs and 75 other definitions from the 125 pairs were identified by three experts in medical device safety. Receiver operating characteristic (ROC) analysis was carried out to evaluate the extraction accuracy of similar definition sentences, and area under curve (AUC) was calculated. The cutoff value was identified from the ROC curve using the Youden Index. Sensitivity and specificity were also calculated. ROC analysis was conducted using JMP 13.2.1.

Results

The AUC obtained from the ROC curve analysis and the sensitivity and specificity obtained from the cutoff value are shown in Table 1.

In comparing the models, we observed that both editing distance algorithms tend to have higher AUC and specificity compared to the three Word2Vec models. In particular, the AUC and specificity of the Levenshtein distance algorithm were highest. In Word2Vec, the sensitivity of Skip-gram had the highest score.

For the Levenshtein distance algorithm, the difference in the number of characters of definition sentence pairs was small, and the more common characters they had, the higher was the similarity between them. However, even though the concepts are the same, if the difference in the number of characters of a definition sentence pair is large, the similarity becomes low. Conversely, the accuracy of Skip-gram in Word2Vec did not depend on the number of characters. Therefore, to further improve accuracy, we believe that devising a method that is a combination of both Skip-gram and Levenshtein distance methods is necessary.

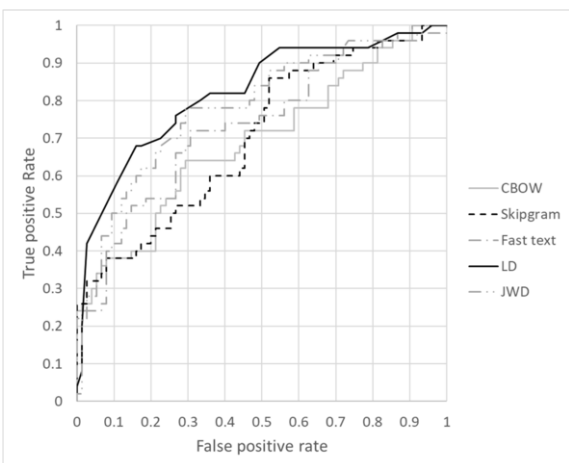

Figure 2–ROC curve

Conclusions

As a result of the experiments performed in this study to extract similar words using definition sentences, we drew the following conclusions: Skip-gram was the most sensitive and Levenshtein distance had the highest specificity and high AUC. To further improve accuracy, we suggest that devising a method that is a combination of both Skip-gram and Levenshtein distance methods is necessary.

Table 1–Comparison of the accuracy of each algorithm

Algorithm		Sensitivity	Specificity	AUC
Edit distance	Levenshtein distance	0.680	0.840	0.821
	Jaro–Winkler distance	0.780	0.707	0.782
Word2Vec	CBOW	0.640	0.707	0.690
	skip-gram	0.860	0.480	0.703
	fast text	0.720	0.693	0.734

Acknowledgements

This research is supported by the Research on Regulatory Science of Pharmaceuticals and Medical Devices from Japan Agency for Medical Research and development, AMED.

References

[1] Ministry of Health, Labour, and Welfare, Publication and Utilization of Medical Device Adverse Event Terminology (in Japanese), (n.d.). http://www.pmda.go.jp/files/000204082.pdf (accessed November 12, 2018).

[2] National Cancer Institute, Centers for Devices and Radiological Health (CDRH) Terminology Files, (n.d.). https://evs.nci.nih.gov/ftp1/FDA/CDRH/About.html (accessed November 12, 2018).

[3] International Medical Device Regulators Forum, About IMDRF, (n.d.). http://www.imdrf.org/about/about.asp (accessed November 12, 2018).

[4] V.I. Levenshtein, Binary Codes Capable of Correcting Deletions, Insertions and Reversals, *Soviet Physics Doklady*. **10** (1966) 707–710.

[5] W. Winkler, String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, *Proceedings of the Section on Survey Research Methods*. (1990) 354–359.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient Estimation of Word Representations in Vector Space, in: Proceedings of ICLR Workshops Track, 2013.

[7] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, FastText.zip: Compressing Text Classification Models, *ArXiv:1612.03651 [cs]*. (2016).

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