Development of a Deep Learning-Based Automated Mapping Tool in the Conversion Process of OMOP-CDM

Yourim Lee. MS[1], Ba Rom Kang. BS[2], Soondong Kim. MCS[1],
Soo-Yeon Cho. RN, MPH[1], Aelan Park, RN, CMD, PhD[1], Ha Young Kim. Ph.D[3], Byungkon Kang. Ph.D[4],
Hye Jin Kam. Ph.D[1]
[1]EvidNet, Seongnam, South Korea; [2]Dept. of Data Science, [3]Graduate School of Information, Yonsei University, Seoul, South Korea, [4]Dept. of Computer Science, SNYU Korea, Incheon, South Korea

Abstract

We proposed a mapping assistant tool, DEep Learning based AUtomated mapping tool (DEAU). DEAU reflects the meaning of words and can learn the information about the relationship among words. Unlike Usagi, which is based on TF-IDF, DEAU was created by applying a deep-learning methodology that reflects the meaning of words. The DEAU found condition domain concept matches 77.37% and Usagi did 65.87% to the target OMOP concepts. Both tools show a match rate of about 69% and a discrepancy of about 18% show, thus we confirmed the mutually complementary utilization possibility of the two tools.

Introduction

Vocabulary mapping from the original EMR terms to OMOP-CDM vocabularies is time consuming and labor-intensive process. For this reason, OHDSI consortium provides Usagi[1], an open-source software tool to help vocabulary mapping. Even though it is clear that Usagi is quite useful to map between English vocabularies, there are barriers in vocabulary mapping with the same meaning of different forms. Because the basic principle TF-IDF of Usagi based on the bag-of-words model, it cannot capture position in text, semantics, co-occurrences (e.g. as compared to topic models, word embeddings). We proposed a DEep learning based AUtomated mapping tool (DEAU) to reflect the information of the relationship between words and their semantic meanings.

Material and Methods

For model development and performance test, we utilized source codes from Korean diagnoses (n=83,113) that were mapped to condition domain concepts of OMOP. The vocabulary of condition domain we used is described in detail below. After preprocessing and filtering, we randomly split the dataset into train, validation, test in the ratio 7 : 1 : 2. But It took some time to test Usagi, so we only used 500 rows of the testset.

Applied code filtration condition was as follows:

- Korean diagnosis codes: standard concepts, SNOMED CT and condition domain

Detailed description of the network architecture of DEAU is provided Fig1. The DEAU was adopted two algorithms: fastText[2,3] for word embedding and Infersent[4] for sentence representation. Words that were not included in fastText were additionally pre-trained by getting context through google crawling. Vocabulary used dataset are converted into vectors, and the distance between similar words were made closer. After pre-training words with Google crawling, we trained DEAU model with 1:1 sampled (positive : negative) mapped data. To compare the performance of Usagi and our DEAU model, test sets of condition domain was tested with the two methodologies in parallel.

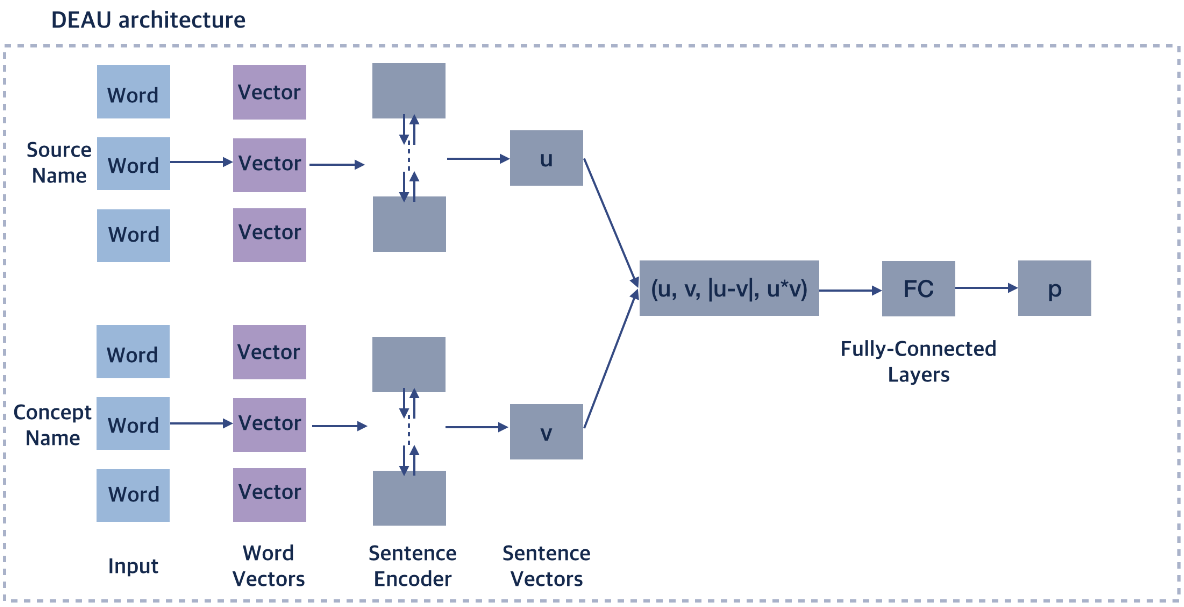

Figure 1. The architecture of DEAU

We adopted topk, precision@k, recall@k metrics which are often used as a recommendation system to compare the 1:N mapping performance of Usagi and DEAU. When the set of recommended sources is , the total number of concepts recommended by the algorithm is , the actual concepts are for the correct list of concepts, and the concept list which is actually correct among the items recommended by the algorithm is .

(1)

(2)

(3)

Results

The DEAU found concept matches of 77.37% of condition and Usagi did 65.87% of condition to the target OMOP concepts (Fig 2). At the venn diagrams of the concepts to be actually mapped in the top 100, we verified that the non-overlapping concepts were 18% of the whole test sets (overlap was 69%).

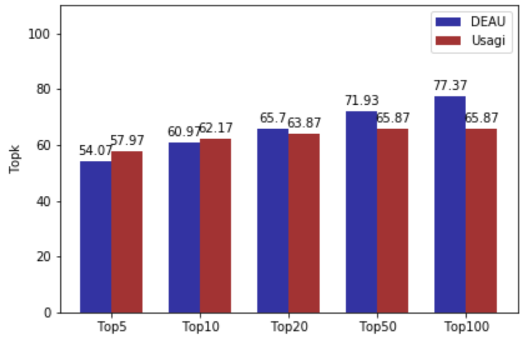

Figure 2. The number of concept sets

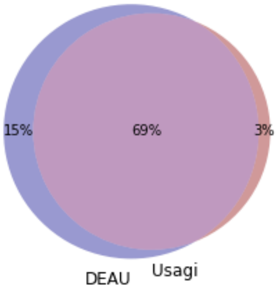

Figure 3. The venn diagram of concept sets in the top100

Table 1. The evaluation metrics between Usagi and DEAU.

Tool	Top5	Top10	Top20	Top50	Top100	Precision@100	Recall@100
Usagi	**57.97**	**62.17**	63.87	65.87	65.87	**0.0120**	0.5663
DEAU	54.07	60.97	**65.7**	**71.93**	**77.37**	0.0097	**0.7737**

Discussion

We developed an automated DEAU to aid users for map standard vocabulary of OMOP-CDM. By comparisons of the matching results between Usagi and DEAU, we confirmed that Usagi has a higher performance when the source name match with the words of the concept name. On the other hand, DEAU showed a high degree of similarity finding ability even the words of the concept name and the source name were formed in totally different forms. Table2,3 can found a part of source of matched real target by each tool. As shown in Fig 3, There were mapping concepts in the two tools that did not overlap each other. As working on mapping tool with OMOP-CDM, we expect that DEAU will have better results that the current mapping performance. We are to improve the performance of DEAU by (1) adding the context of source name and (2) increasing the volume of training dataset.

Table 2. A part of source of matched real target by only Usagi

Index	Source name	Matched concept name (contained synonym term)	Concept id
1	Exudative retinal detachments	Exudative retinal detachment	378414
2	Other specified arthritis, multiple sites	Allergic arthritis of multiple sites	74125
3	Cervicobrachial syndrome, cervical region	Cervicobrachial syndrome	77639
4	Macrostomia	Macrostomia	22426
5	Fracture of capitate bone, open	Fracture of capitate	4218884

Table 3. A part of source of matched real target by only DEAU

Index	Source name	Matched concept name (contained synonym term)	Concept id
1	Arthritis in other infectious and parasitic diseases classified elsewhere, lower leg	infective arthritis disorder	4262590
2	Spondylopathy, unspecified, cervicothoracic region	spondyloarthropathy	4157453
3	Other infection during labour	infection childbirth	45757688
4	Contact with agricultural machinery, school, other institution and public administrative area	accident caused agricultural machine event	436874
5	Striking against or struck by other objects, industrial and construction area	accidentally struck against objects persons	441192

Acknowledgement

This work was supported by the Bio Industrial Strategic Technology Development Program (20001234) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

References

1. Usagi. https://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:usagi. 2018.
2. Piotr B, Edouard G, Armand J, Tomas M. Enriching Word Vectors with Subword Information. arXiv. 2016;1607.04606.
3. Armand J, Edouard G, Piotr B, Tomas M. Bag of Tricks for Efficient Text Classification. arXiv. 2016;1607.01759.
4. Alexis C, Douse K, Holger S, Loic B, Antoine B. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. arXiv. 2017;1705.02364.