Adapting Pre-trained Word Embeddings For Use In Medical Coding

Kevin Patel[1], Divya Patel[2], Mansi Golakiya[2], Pushpak Bhattacharyya[1], Nilesh Birari[3]
[1]Indian Institute of Technology Bombay, India
[2]Dharmsinh Desai University, India, [3]ezDI Inc, India
[1]{kevin.patel,pb}@cse.iitb.ac.in, [3]nilesh.b@ezdi.us
[2]{divya.patel.8796,golkiya.mansi}@gmail.com

Abstract

Word embeddings are a crucial component in modern NLP. Pre-trained embeddings released by different groups have been a major reason for their popularity. However, they are trained on generic corpora, which limits their direct use for domain specific tasks. In this paper, we propose a method to add task specific information to pre-trained word embeddings. Such information can improve their utility. We add information from medical coding data, as well as the first level from the hierarchy of ICD-10 medical code set to different pre-trained word embeddings. We adapt CBOW algorithm from the word2vec package for our purpose. We evaluated our approach on five different pre-trained word embeddings. Both the original word embeddings, and their modified versions (the ones with added information) were used for automated review of medical coding. The modified word embeddings give an improvement in f-score by 1% on the 5-fold evaluation on a private medical claims dataset. Our results show that adding extra information is possible and beneficial for the task at hand.

1 Introduction

Word embeddings are a recent addition to an NLP researcher's toolkit. They are dense, real-valued vector representations of words that capture interesting properties among them. Word embeddings are learned from raw corpora. Usually, the larger the corpora, the better is the quality of the embeddings learned. However, the larger the corpora, the larger is the amount of resources and time needed for their training. Thus, different groups release their learned embeddings publicly. Such pre-trained embeddings is a primary reason for the inclusion of word embeddings in mainstream NLP. However, such pre-trained embeddings are usually learned on generic corpora. Using such embeddings in a particular domain such as medical domain leads to following problems:

- No embeddings for domain-specific words. For example, *phenacetin* is not present in pre-trained vectors released by Google.
- Even those words that do have embeddings, may have a poor quality of the embedding, due to different senses of the words, some of which belonging to different domains.

It is difficult to obtain large amounts of domain-specific data. However, many NLP applications have benefited from the addition of information from small domain-specific corpus to that obtained from a large generic corpus (Ito et al., 1997). This raises the following questions:

- *Can we use additional domain-specific data to learn the missing embeddings?*
- *Can we use additional domain-specific data to improve the quality of already available embeddings?*

In this paper, we address the second question: Given pre-trained word embeddings, and domain specific data, we tune the pre-trained word embeddings such that they can achieve better performance. We tune the embeddings for and evaluate them on an automated review of medical coding.

The rest of the paper is organized as follows: Section 2 provides some background on different notions used later in the paper. Section 3 motivates our approach through examples. Section 4 explains our approach in detail. Section 5 enlists the experimental setup. Section 6 details the results and analysis, followed by conclusion and future work.

2 Background

2.1 Word Embeddings

Word embeddings are a crucial component of modern NLP. They are learned in an unsupervised manner from large amounts of raw corpora. Bengio et al. (2003) were the first to propose neural word embeddings. Many word embedding models have been proposed since then (Collobert and Weston, 2008; Huang et al., 2012; Mikolov et al., 2013; Levy and Goldberg, 2014). The central idea behind word embeddings is the distributional hypothesis, which states that *words which are similar in meaning occur in similar contexts* (Rubenstein and Goodenough, 1965). Consider the Continuous Bag of Words model by (Mikolov et al., 2013), where the following problem is poised to a neural network: given the context, predict the word that comes in between. The weights of the network are the word embeddings. Training the model over running text brings embeddings of words with similar meaning closer.

2.2 Medical Coding

Medical coding is the process of assigning predefined alphanumeric medical codes to information contained in patient medical records.

Babre et al. (2010) shows a typical medical coding pipeline. Note that the coding (both automatic and/or manual) is followed by a manual review. This is due to the critical nature of the coding process, and the high cost incurred due to any errors. However, any human involvement increases cost both in terms of time and money. Thus, in order to reduce human involvement in the review process, an automatic review component can be inserted just before the human review. Automated reviewing is a binary classification problem. Those instances that are rejected by the automated review component can be directly sent back for recoding, whereas those instances that are accepted by the automated review component should be sent to human reviewers for further checking. Such a modification decreases the load on the human reviewer, thereby reducing the cost of overall pipeline.

Given the textual nature of medical data, many natural language processing challenges manifest themselves while performing either automated medical coding or automated review of medical coding. Common challenges include, but are not limited to:

- Synonymy: Multiple words can have same meaning (Synonym). For instance, *High Blood Sugar* and *Diabetes* have the same meaning.

- Abbreviation: Medical staff, in their hurry, often abbreviate words and sentences. For instance, *hypertension* can be written as *HTN*. The automated system needs to understand that both these strings ultimately mean the same thing.

One can note that both in case of synonym and abbreviations, the context will be almost same. Thus, word embeddings are well suited to handle both these challenges.

3 Motivation

Consider the following medical terms (the abbreviations in parentheses will be used to refer to the terms later):
- High Blood Pressure (HBP)
- Low Blood Pressure (LBP)
- High Blood Sugar (HBS)
- Liver Failure (LF)
- Diabetes (D)
- Hypertension
- HTN

We would ideally like the embeddings of the terms to be learned such that the following constraints hold:

- Similarity (HBP, HBS) should be higher than Similarity (HBP, LBP), which in turn, should be higher than Similarity (HBP, LF) (as per medical knowledge).

- Similarity (HBS, D) should be high (as they are synonyms).

- Similarity (Hypertension, HTN) should be high (as HTN is abbreviation of hypertension).

Information about such relations might not be available in generic corpus on which most pretrained embeddings are trained. However, it might be available in domain specific corpora, or even labeled data, such as those used in medical claims. Approaches that can add that information to pretrained embeddings will definitely improve their utility.

4 Approach

We adapt the Continuous Bag Of Words (CBOW) approach (Mikolov et al., 2013) for our situation. Given labeled medical claims data, we consider the terms in the transcripts as context words, and the corresponding codes as target word. We have both positive and negative samples in our data. Thus we have both normal samples as well as negative samples needed for applying negative sampling.

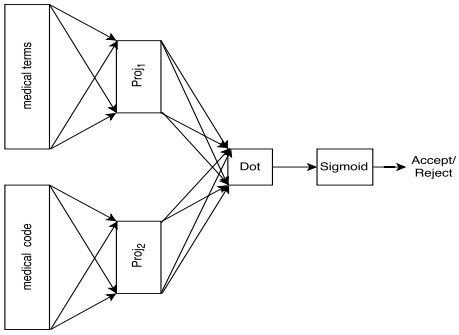

Figure 1: Network architecture of our approach

Figure 1 shows the network of our approach. The inputs to the network are a bag of words representation of medical terms, and a one-hot representation of the corresponding code. The output of the network is a binary value indicating whether the input code is accepted for the corresponding input medical terms.

Exploiting ICD10 Code hierarchy

Another information that can be included is the hierarchical nature of the ICD10 code set. Currently, the network considers the error of misclassifying codes in same subcategory, say F32.9 and F11.20, the same as the error of misclassifying codes belonging to different subcategories, say F32.9 and 30233N1. Ideally, error(F32.9, F11.20) should be less than error(F32.9, E87.1), which in turn should be less than error(F32.9, 30233N1). Such hierarchical information can be encoded by a network like the one in figure 2. Due to resource and time constraints, we have currently considered only the top level hierarchy, *i.e.* whether the code is ICD-10 Diagnosis or ICD-10 Procedural.

The learned weights between $\underline{\text{Proj}_1}$ and codes input in hierarchy network (figure 2) are used to initialize the weights between $\underline{\text{Proj}_2}$ and codes in the original network (figure 1). Then the original network is trained as usual. The weights between

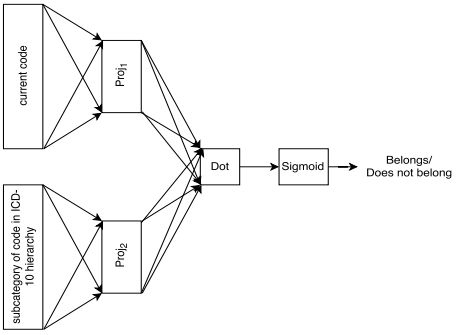

Figure 2: Encoding hierarchy information

$\underline{\text{Proj}_1}$ and medical terms in the original network are the modified word embeddings.

5 Experimental Setup

5.1 Dataset

We used a private medical claims review dataset, which we cannot release publicly due to privacy concerns. The dataset consists of 280k records, consisting of medical terms along with a code. Each entry is labeled as *accept* or *reject*, depending on whether the entry has correct code, or whether it was sent for recoding.

5.2 Pre-trained word embeddings

We used 5 different pre-trained word embeddings. The first one is the one released along with Google's word2vec toolkit. The remaining four are medical domain specific, and were released by (Pyysalo et al., 2013). They are as follows:

- PMC: Trained on 4 million PubMed Central's full articles
- PubMed: Trained on 26 million abstracts and citations in PubMed.
- PubMed_PMC: Trained on combination of previous two resources
- Wikipedia_PubMed_PMC: Trained on combination of Wikipedia, PubMed and PMC resources.

5.3 Classifiers

Once we tune the embeddings, we use them to learn a binary classifier. For our experiments, we report the results we got by using logistic regression..

		Medical Knowledge			Synonym	Abbreviation
		HBP,HBS	HBP,LBP	HBP,LF	HBS,Diabetes	Hypertension,HTN
Google	Orig	0.534	0.895	0.181	0.293	0
	Mod	0.549	0.640	0.089	0.350	-0.004
PMC	Orig	0.599	0.980	0.173	0.141	0.608
	Mod	0.638	0.477	-0.054	0.221	0.947
PubMed	Orig	0.529	0.970	0.006	0.091	0.465
	Mod	0.636	0.474	-0.090	0.188	0.952
PubMed_PMC	Orig	0.592	0.976	0.116	0.141	0.575
	Mod	0.641	0.450	-0.039	0.241	0.952
Wikipedia_ PubMed_PMC	Orig	0.595	0.976	0.158	0.156	0.617
	Mod	0.653	0.474	-0.061	0.190	0.950

Table 1: Cosine similarities of pairs of examples from Section 3

Pre-trained Embeddings	Original Embeddings	Modified Embeddings
Google	82.78	**83.37**
PMC	82.93	**83.96**
PubMed	83.18	**84.00**
PubMed_PMC	82.88	**83.92**
Wikipedia_ PubMed_PMC	83.12	**83.91**

Table 2: Average 5-fold cross validation F-score on automated review of medical coding

6 Results and Analysis

Table 2 shows the results of 5-fold evaluation on automated review of medical coding. Note that the modified embeddings consistently outperform the original ones for all pre-trained embeddings that we used. The reason behind this improvement is evident from the analysis table 1 where we show how the constraints are better modeled by the modified embeddings (Mod) as compared to the original embeddings (Orig).

7 Related Work

Word embeddings have proved to be useful for various tasks, such as Part of Speech Tagging (Collobert and Weston, 2008), Named Entity Recognition Sentence Classification (Kim, 2014), Sentiment Analysis (Liu et al., 2015), Sarcasm Detection (Joshi et al., 2016). Medical domain specific pre-trained word embeddings were released by different groups, such as Pyysalo et al. (2013), Brokos et al. (2016), *etc.* Wu et al. (2015) apply word embeddings for clinical abbreviation disambiguation.

8 Conclusion and Future Work

In this paper, we proposed a modification of the CBOW algorithm to add task and domain specific information to pre-trained word embeddings. We added information from a medical claims dataset and the ICD-10 code hierarchy to improve the utility of the pre-trained word embeddings. We obtained an improvement of approximately 1% using the modified word embeddings as compared to using the original word embeddings. Such improvement was achieved by including only the top level hierarchy. We hypothesize that using the full hierarchy will lead to better improvements, which we shall investigate in the future.

References

Deven Babre et al. 2010. Medical coding in clinical trials. Perspectives in clinical research 1(1):29.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. J. Mach. Learn. Res. 3:1137–1155.

Georgios-Ioannis Brokos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2016. Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. In Proceedings of 15th Workshop on Biomedical Natural Language Processing (BioNLP 2016), at the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016).

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, ICML. ACM, volume 307 of ACM International Conference Proceeding Series, pages 160–167.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In Annual Meeting of the Association for Computational Linguistics (ACL).

Akinori Ito, Hideyuki Saitoh, Masaharu Katoh, and Masaki Kohda. 1997. N-gram language model adaptation using small corpus for spoken dialog recognition. In ASJ. volume 3000, page 96779.

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. 2016. Are word embedding-based features useful for sarcasm detection? In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pages 1006–1011.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pages 1746–1751.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers. pages 302–308.

Pengfei Liu, Shafiq R Joty, and Helen M Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In EMNLP. pages 1433–1443.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems. pages 3111–3119.

S Pyysalo, F Ginter, H Moen, T Salakoski, and S Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In Proceedings of LBM 2013. pages 39–44.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. Commun. ACM 8(10):627–633. https://doi.org/10.1145/365628.365657.

Yonghui Wu, Jun Xu, Yaoyun Zhang, and Hua Xu. 2015. Clinical abbreviation disambiguation using neural word embeddings. In Proceedings of 14th Workshop on Biomedical Natural Language Processing (BioNLP 2016), at the 53th Annual Meeting of the Association for Computational Linguistics (ACL 2015). page 171.