A Framework for Developing and Evaluating Word Embeddings of Drug-named Entity

Mengnan Zhao[1], Aaron J. Masino[2], Christopher C. Yang[1]
[1]College of Computing and Informatics, Drexel University, Philadelphia, PA, US
[2]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, PA, US
emails: mz438@drexel.edu, masinoA@email.chop.edu, ccy24@drexel.edu

Abstract

We investigate the quality of task specific word embeddings created with relatively small, targeted corpora. We present a comprehensive evaluation framework including both intrinsic and extrinsic evaluation that can be expanded to named entities beyond drug name. Intrinsic evaluation results tell that drug name embeddings created with a domain specific document corpus outperformed the previously published versions that derived from a very large general text corpus. Extrinsic evaluation uses word embedding for the task of drug name recognition with Bi-LSTM model and the results demonstrate the advantage of using domain-specific word embeddings as the only input feature for drug name recognition with F1-score achieving 0.91. This work suggests that it may be advantageous to derive domain specific embeddings for certain tasks even when the domain specific corpus is of limited size.

1 Introduction

The ability of word embeddings to capture latent, contextual information has proven useful to a variety of NLP tasks, such as named entity recognition (Santos & Guimarães, 2015), syntactic parsing (Levy & Goldberg, 2014), and question answering (Iyyer et al., 2014). Within biomedical research, word embeddings developed in most previous studies were generated from very large, generic corpora (e.g. news articles). This is appropriate for generalized language models. However, for specialized domains and tasks, it may be beneficial to generate word embeddings from a targeted corpus. We propose a biomedical domain-specific word embedding model and a novel evaluation framework, which mainly focus on representing drug names in the current stage. This framework can be expanded to other biomedical entities such as protein, gene, and chemical compound names in the future. We evaluate the developed word embeddings with a comprehensive intrinsic evaluation framework that includes relatedness, coherence, and outlier detection assessment, as well as an extrinsic evaluation that focuses on the task of drug name recognition and classification with a bidirectional long short-term memory (Bi-LSTM) RNN model.

2 Related Work

In the biomedical domain, word embeddings are primarily used for biomedical named entity recognition (BNER) with evaluations conducted on tasks such as JNLPBA (Kim et al., 2004), BioCreAtIvE (Hirschman et al., 2005), and BioNLP Shared Tasks. Tang et al. (2014) explored the impact of three different types of word representations (WR) on clustering-based representation, distributional representation and word embedding. Segura-Bedmar et al. (2015) generated word embeddings with *word2vec* and a combined Wikipedia and MedLine corpus. The results were evaluated on the SemEval-2013 Task 9.1 Drug Name Recognition dataset (Segura-Bedmar et al., 2013). Wang et al. (2015, November) used word embeddings for bio-event trigger detection. Li et al. (2015) incorporated word embedding features with bag-of-words (BOW) features for bio-event extraction and evaluated results on the BioNLP 2013 GENIA task (Nédellec et al., 2013).

Drug name recognition (DNR) in biomedical literature and clinical notes is essential for many medical information and relation extraction tasks (e.g. drug-drug interaction). Significant effort has been devoted to DNR and the common methods can be categorized as (Lu et al., 2015): (1) dictionary-based approaches (Rindflesch et al., 2000; Sanchez-Cisneros et al., 2013), (2) rule-based/ontology-based approaches (Hamon & Grabar, 2010; Coden et al., 2012), (3)

Proceedings of the BioNLP 2018 workshop, pages 156–160
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

machine learning-based approaches (Lamurias et al., 2013; Lu et al., 2015), and (4) hybrid approaches (Korkontzelos et al., 2015).

3 Word Embeddings Training

We extracted text from PubMed and DrugBank to construct our corpus. For PubMed, we used "*drug*" as the keyword of query to broadly select drug related abstracts, which yielded 474,273 abstracts. From DrugBank [1] Release Version 5.0.5 we extracted the fields: "description" "indication" "pharmacodynamics" "mechanism-of-action" "toxicity" for 8,226 drugs.

We employed the skip-gram model in *word2vec* to generate word embeddings. Moreover, as studies have found that word embeddings have a consistent relationship with word frequencies, even after the interception of frequency-based effects by algorithms and vector length normalization (Schnabel et al., 2015), we employed correlation analysis between vectors and frequencies as the evaluation metric to tune the parameters for the word embedding model. For our final result, we trained the word embedding model in *word2vec* with parameters: *size* = 420, *window* = 5, *min_count* = 2.

4 Intrinsic Evaluation

4.1 Relatedness assessment

Relatedness evaluation is the most popular and direct intrinsic word embedding evaluation method. It is expected that high quality word embeddings will display significant correlation (e.g. Pearson's, Spearman's) between the cosine similarity of the embedding vectors for related word pairs and the human scores.

We evaluated the results on two biomedical domain inventories: UMNSRS-Rel and UMNSRS-Sim (Pakhomov et al., 2010). These datasets provide human-annotated scores of relatedness and similarity between clinical term pairs. We measured the correlation between the scores provided by the UMNSRS datasets and calculated by our model, using Spearman's correlation coefficient. We also compared our model to a publicly available word embedding set trained on about 100 billion words from Google News samples[2].

Corpora	PubMed+ DrugBank	Google News
drug-drug	**0.737**	0.430
drug-X	0.530	0.293
drug-nonDrug	0.492	0.245
whole dataset	0.555	0.345
nonDrug-nonDrug	0.565	0.368

Table 1: Relatedness assessment on UMNSRS-Rel dataset

Corpora	PubMed+ DrugBank	Google News
drug-drug	**0.764**	0.495
drug-X	0.529	0.435
drug-nonDrug	0.449	0.385
whole dataset	0.597	0.402
nonDrug-nonDrug	0.601	0.381

Table 2: Similarity assessment on UMNSRS-Sim dataset

As shown in Table 1 and 2, our model and UMNSRS show positive correlations in both relatedness and similarity assessment, with most of the correlation coefficients higher than 0.5, which means the relationship represented in vector space is consistent with human annotations. In particular, the highest consistency is achieved for the relationship of drug-drug pairs, where coefficients reach 0.737 and 0.764 for relatedness and similarity, respectively. In addition, the proposed model trained on PubMed+DrugBank shows significantly higher correlations with human scores than the model trained on a Google News corpus in all word pair types. This is important because the Google News based embeddings were trained on an extremely large dataset compared to our corpus.

4.2 Coherence assessment

Conceptually, we expect that a good word embedding should be surrounded by a coherent neighborhood of similar words. From this concept, we propose a novel intrinsic evaluation metric as a supplement to current relatedness analysis (Schnabel et al., 2015). In coherence assessment, we assess whether a given word embedding is mutually related to the word embeddings in its local neighborhood. Here we created a neighborhood for each drug name and explored the relation with the closest neighbor terms. We expect that other drug entities should be preferentially represented in the neighborhood. Setting the neighborhood size from 3 to 10, we calculated the percentage of

[1] www.drugbank.ca/releases/latest

[2] https://code.google.com/archive/p/word2vec/

drug names within the neighborhood of each drug, with selected results shown in Table 3.

Size of neighborhood	3	5	7	9	10
Percentage of drug/all_neighbors (%)	61.1	58.8	56.9	55.2	54.6

Table 3: Percentage of drug entities within a drug's neighborhood across all drugs.

From Table 3, we see that the percentage of drug entities declines with the expansion of neighborhood size. Noting that neighbors were arranged by the cosine similarity relative to the target word, such decline implies that drug entities tend to be the closest neighbors. Beyond that, drug entities still occupy more than half of the nearest 10 neighbors. These results suggest there is a strong coherence in the semantic space.

4.3 Outlier Detection

As a final intrinsic measure of word embedding quality, we consider a modification of a previously proposed outlier detection task. Given a group of words W, the compactness score of word $w_m \in W$ represents the compactness of the cluster $W\backslash\{w_m\}$. Performance on the outlier detection task can be evaluated by accuracy and outlier position percentage (OPP) (Camacho-Collados & Navigli, 2016). Ideally, if outliers in all the groups were identified and listed at the last position, accuracy and OPP should be 1 and 100% respectively.

In this study, the goal of outlier detection is to identify the non-drug words as outliers. We created two datasets each with 400 groups of words ($|D|$=400). Following the work of Camacho-Collado and Navigli, the first dataset, D-Manu, contains 4 to 8 drugs and 1 *manually* selected non-drug outlier ($|W| \in [5, 9]$). Additionally, we modify the previously presented work by forming a second dataset, D-Rand, in which each group contains 4 to 8 drugs and 1 *randomly* selected non-drug outlier ($|W| \in [5, 9]$). Tables 4 and 5 show the evaluation results of outlier detection on D-Rand and D-Manu. On D-Rand, outliers were identified in more than 40% of groups across different sizes, and OPP values indicate that the average outlier position was around 70% to the right end (100%) of the list arranged by compactness score. Meanwhile, for D-Manu, the accuracy values are all higher than 0.8 and the OPP values are all above 93%.

Group size-$|W|$	**5**	**6**	**7**	**8**	**9**
Accuracy	0.43	0.44	0.41	0.40	0.41
OPP(%)	69.2	72.0	73.6	70.3	72.4

Table 4: Accuracy and OPP of outlier detection on D-Rand

Group size-$|W|$	**5**	**6**	**7**	**8**	**9**
Accuracy	0.82	0.83	0.85	0.80	0.83
OPP	93.4	94.3	95.3	93.9	94.9

Table 5: Accuracy and OPP of outlier detection on D-Manu

To gain further insight on the potential correlation between the outlier task performance and the similarity distribution over the outlier term and the non-outlier terms, we calculated the average similarity between each pair of non-outlier terms and the average between non-outliers and the outlier for each group in D-Rand and D-Manu. We found that the average similarity between non-outliers was about 0.21. The average similarity between non-outliers and randomly selected outliers and manually selected outliers was about 0.16 and 0.12, respectively. This result confirmed that the greater distinction in word similarity is consistent with the better accuracies in outlier detection.

5 Extrinsic Evaluation - DNR

5.1 DNR with Bi-LSTM Model

We employ a bidirectional long short-term memory (Bi-LSTM) RNN model that is designed to process text input as a sequence of tokens (constituent parts, usually words) and predict the label for each token. The BLSTM-RNN model combines two RNNs: the forward RNN processes the sequence from left to right and the backward RNN processes it from right to left. We use a BIO scheme for the sequence labeling task. Specifically, each token is labeled as one B-X, I-X or O indicating it is at the beginning (B), inside (I), or outside (O) of the entity of type X (e.g. drug name).

In order to achieve the best results and compare the impact of the word embedding model in the labeling task, we introduced three BLSTM-RNN variants: (1) Fixed embedding (BLSTM-F): Word embedding values were provided by the pre-trained word embedding model and treated as fixed constants; (2) Varied embedding (BLSTM-V): Word embedding values were also provided by the pre-trained word embedding but treated as learnable parameters; (3) Randomly-initialized embedding (BLSTM-R): Word embedding values were initialized randomly and treated as learnable parameters.

5.2 Experiments on Drug Name Recognition

We evaluated our model on DDI-Extraction-2011 task (Segura-Bedmar et al., 2011) using two metrics: **Exact matching**-the predicted entity must have exactly the same boundary with the annotated entity and **Partial matching**-the predicted entity must have some overlap with the annotated entity. Table 6 shows the results of three BLSTM models. Regarding to the impact of pre-trained word embeddings, there is no obvious improvement when introducing the pre-trained embedding values instead of randomly initialized vector values. Moreover, the f1-score of BLSTM-V that sets embedding values as learnable parameters in RNN model is increased to 0.911 from 0.891 in BLSTM-F that treats them as fixed constants. Overall, our BLSTM models achieve very good results on DNR according to f1-scores, and treating embedding values as learnable parameters, regardless of pre-trained or randomly initialized, lead to better results than setting them fixed, indicating the great advantage of RNN models for drug name recognition task.

	Exact Matching			**Partial Matching**		
	P	**R**	**F1**	**P**	**R**	**F1**
BLSTM-F	0.89	0.90	0.89	0.91	0.92	0.91
BLSTM-V	**0.91**	0.91	**0.91**	**0.93**	0.94	**0.94**
BLSTM-R	0.90	**0.92**	0.91	0.93	**0.94**	0.93

*Bold indicates the highest score in the column.

Table 6: Evaluation results on DDI-Extraction-2011 test set.

5.3 Experiments on Drug Name Classification

In DDI-Extraction-2013 challenge (Segura-Bedmar et al., 2013), the drugs were annotated with four types instead of one type in 2011 task, including: *drug*, *brand*, *group*, and *drug_n*. Thus, it becomes a drug name recognition and classification task. We evaluated our results using four metrics provided by the organizers, with f1-scores shown in Table 7. Pre-trained word embeddings showed their advantages, for instance, f1 of strict matching were improved 16% in BLSTM-V than BLSTM-R. While updating the pre-trained embedding values did not show obvious improvement by comparing BLSTM-F and BLSTM-V.

DrugBank+MedLine	BLSTM-F	BLSTM-V	BLSTM-R
Strict matching	0.735	0.724	0.631
Type matching	0.753	0.737	0.654
Exact oundary matching	0.789	0.801	0.658
Partial boundary matching	0.816	0.823	0.688
drug	0.824	0.852	0.750
brand	0.722	0.588	0.344
group	0.722	0.702	0.697
drug_n	0.381	0.333	0

Table 7: Results on DDI-Extraction-2013 test set.

6 Conclusion

We presented biomedical domain-specific word embeddings formulated with the *word2vec* model using PubMed and DrugBank text sources and a comprehensive intrinsic and extrinsic evaluation framework for word embeddings that includes new and existing metrics. We found that our word embeddings demonstrated superior performance based on relatedness assessment, neighborhood coherence, and outlier detection. Moreover, we also found that these embeddings performed better than those generated from very large datasets such as Google News. This is significant because our training dataset is approximately two orders of magnitude smaller.
Since drug name recognition (DNR) is an important biomedical NLP task, we used DNR as the downstream task for extrinsic evaluation of the developed drug name embeddings. We utilized the pre-trained word embeddings in Bi-LSTM model for the task of drug name recognition and classification. For drug name recognition, setting embedding values as learnable parameters in RNN model has more impact on the performance than utilizing pre-trained word embeddings. For drug name classification, pre-trained word embeddings offer significant performance increases over randomly-initialized embeddings, while updating the pre-trained embedding values during the BLSTM model training has little improvement. This work provides a useful tool or framework for processing raw biomedical text and extracting drug entities, which could be helpful in processing other unstructured data and medical entities.

References

Camacho-Collados, J., & Navigli, R. (2016). Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations.

In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 43-50).

Chiu, B., Crichton, G., Korhonen, A., & Pyysalo, S. (2016). How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing* (pp. 166-174).

Coden, A., Gruhl, D., Lewis, N., Tanenblatt, M., & Terdiman, J. (2012, September). SPOT the drug! an unsupervised pattern matching method to extract drug names from very large clinical corpora. In *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on* (pp. 33-39). IEEE.

Hamon, T., & Grabar, N. (2010). Linguistic approach for identification of medication names and related information in clinical narratives. *Journal of the American Medical Informatics Association*, *17*(5), 549-554.

Hirschman, L., Yeh, A., Blaschke, C., & Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Kim, J. D., Ohta, T., Tsuruoka, Y., Tateisi, Y., & Collier, N. (2004, August). Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications* (pp. 70-75). Association for Computational Linguistics.

Korkontzelos, I., Piliouras, D., Dowsey, A. W., & Ananiadou, S. (2015). Boosting drug named entity recognition using an aggregate classifier. *Artificial intelligence in medicine*, *65*(2), 145-153.

Lamurias, A., Grego, T., & Couto, F. M. (2013, October). Chemical compound and drug name recognition using CRFs and semantic similarity based on ChEBI. In *BioCreative Challenge Evaluation Workshop* (Vol. 2, p. 75).

Levy, O., & Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *ACL (2)* (pp. 302-308).

Li, C., Song, R., Liakata, M., Vlachos, A., Seneff, S., & Zhang, X. (2015, July). Using word embedding for bio-event extraction. In *Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015). Stroudsburg, PA: Association for Computational Linguistics* (pp. 121-126).

Lu, Y., Ji, D., Yao, X., Wei, X., & Liang, X. (2015). CHEMDNER system with mixed conditional random fields and multi-scale word clustering. *Journal of cheminformatics*, *7*(1), S4.

Nédellec, C., Bossy, R., Kim, J. D., Kim, J. J., Ohta, T., Pyysalo, S., & Zweigenbaum, P. (2013, August). Overview of BioNLP shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop* (pp. 1-7).

Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., & Melton, G. B. (2010). Semantic similarity and relatedness between clinical terms: an experimental study. In *AMIA annual symposium proceedings* (Vol. 2010, p. 572). American Medical Informatics Association.

Rindflesch, T. C., Tanabe, L., Weinstein, J. N., & Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (p. 517). NIH Public Access.

Sanchez-Cisneros, D., Martínez, P., & Segura-Bedmar, I. (2013, November). Combining dictionaries and ontologies for drug name recognition in biomedical texts. In *Proceedings of the 7th international workshop on Data and text mining in biomedical informatics* (pp. 27-30). ACM.

Santos, C. N. D., & Guimarães, V. (2015). Boosting Named Entity Recognition with Neural Character Embeddings. *arXiv preprint arXiv:1505.05008*.

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 298-307).

Segura-Bedmar, I., Martinez, P., & Sánchez Cisneros, D. (2011). The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts.

Segura-Bedmar, I., Martínez, P., & Herrero Zazo, M. (2013). Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics.

Segura-Bedmar, I., Suárez-Paniagua, V., & Martınez, P. (2015, September). Exploring word embedding for drug name recognition. In *SIXTH INTERNATIONAL WORKSHOP ON HEALTH TEXT MINING AND INFORMATION ANALYSIS (LOUHI)* (p. 64).

Tang, B., Cao, H., Wang, X., Chen, Q., & Xu, H. (2014). Evaluating word representation features in biomedical named entity recognition tasks. *BioMed research international*, *2014*.

Wang, J., Zhang, J., An, Y., Lin, H., Yang, Z., Zhang, Y., & Sun, Y. (2015, November). Biomedical event trigger detection by dependency-based word embedding. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on* (pp. 429-432). IEEE.