Word Embeddings for Chemical Patent Natural Language Processing

Camilo Thorne [1] **Saber Akhondi** [1]

Abstract

We evaluate chemical patent word embeddings against known biomedical embeddings and show that they outperform the latter extrinsically and intrinsically. We also show that using contextualized embeddings can induce predictive models of reasonable performance for this domain over a relatively small gold standard.

1. Introduction

The chemical industry undoubtedly depends on the discovery of new chemical compounds. However, new chemical compounds are often disclosed first in patent documents (Senger et al., 2015), and only months or years later make it to scientific publications. Thus, most chemical compounds are only immediately available in patent documents (Bregonje, 2005). As the number of new chemical patent applications has been drastically increasing (Muresan et al., 2011), it is now crucial to develop natural language processing (NLP) approaches to automatically extract information from chemical patents (Akhondi et al., 2019).

A key tool in this endeavour are word embeddings (Baroni et al., 2014). Embeddings are crucial to derive word, sentence and text-level features in state-of-the-art neural models such as neural named entity recognition (NER) models. Large scale Word2Vec and contextualized embeddings have been developed for e.g., the biomedical (Pyysalo et al., 2013; Jin et al., 2019) and drug (Segura-Bedmar et al., 2015) domains. Smaller embeddings for analytic chemistry such as Mat2Vec (Tshitoyan et al., 2019) (covering materials science) have also been proposed. All such embeddings were trained on scientific papers (PubMed corpus). More recently, Zhai et al. (2019) learnt and successfully applied to chemical named entity recognition large scale Word2Vec and contextualized ELMo embeddings *learnt over chemistry patents* (a 1 billion word corpus), the so-called CheMU embeddings.

[1]Elsevier. Correspondence to: Camilo Thorne <c.thorne.1@elsevier.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020.

Table 1: Our training and test sets come from the SCAI corpus; the validation set from the Biosemantics corpus.

Split	Entities	Tokens
Train	731 IUPAC, 212 Modifier, 73 Partiupac	33,457
Validation	240 IUPAC	4,654
Test	48 IUPAC, 2 Modifier	28,240

In this paper we address the issue of *evaluating* the quality of these chemistry patent-specific embeddings against its predecessors. Two methods are usual in these kind of comparisons (Schnabel et al., 2015; Wang et al., 2018). On the one hand, *extrinsic* evaluation, in which the impact of each embedding on a prediction task – chemical NER in this paper – is reported. On the other hand, *intrinsic* evaluation, where we qualitatively analyze the quality of the distributed (semantic) representations each embedding assigns to chemical words. To this end it is customary to compare the top 10 most similar terms returned by each embedding to a fixed chemical word – "ibuprofen", a known anti-inflammatory drug, in this paper – or similarity query. We show that chemical patent embeddings outperform their predecessors both extrinsically and intrinsically.

2. Experiments

Data We use for our experiments a small gold standard sampled from two known chemical NER patent corpora, the SCAI corpus (Klinger et al., 2008) and the Biosemantics corpus (Akhondi et al., 2014). The SCAI corpus focuses on chemicals written using the UIPAC name standard[1]. In addition, we sampled IUPAC-annotated portions of the Biosemantics corpus (that covers a much wider variety of chemical entity types) to use as validation set. See Table 1.

We observed large vocabulary overlaps among the Word2Vec embeddings, and with the test set (3,521 words), see Figure 1. We exploit this fact to qualitatively compare the ELMo and Word2Vec embeddings over the test set vocabulary during the intrinsic evaluation.

[1]See: https://iupac.org

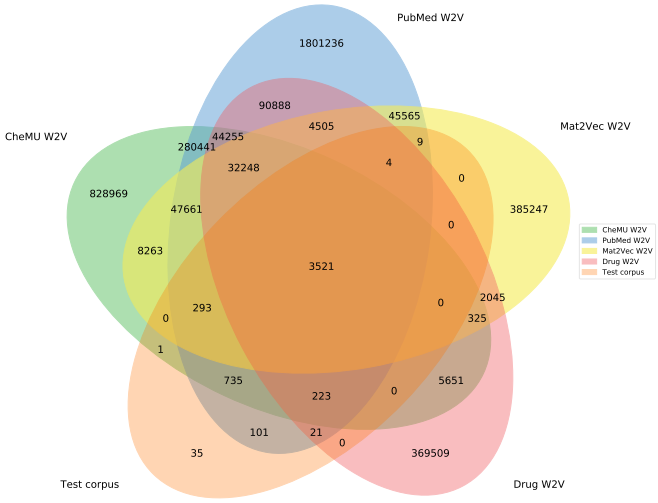

Figure 1: Vocabulary overlaps among Word2Vec embeddings and test corpus. Mat2Vec, PubMed, drug and CheMU embeddings share in total 35,769 words.

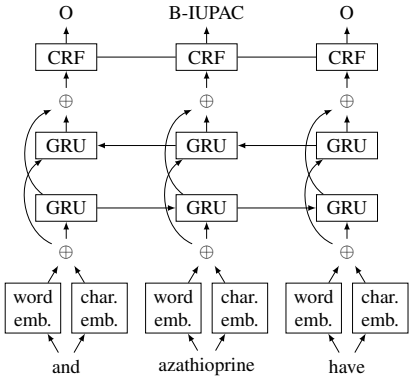

Figure 2: Architecture of the (neural) NER system used.

Extrinsic Evaluation For NER – viz., *extrinsic* evaluation – we use simplified variants of Lample et al. (2016)'s model by using GRU instead of LSTM (stacked) layers, to reduce parameters and prevent overfitting given the small size of our corpus. This is sufficient as we wanted to (a) test if the embeddings could induce reasonable (though not state of the art) performance, and (b) test if different embeddings give rise to different performances. We encoded words using the pre-trained chemical embeddings and a trainable character-level GRU encoder. See Figure 2. The models were trained for 50 epochs with early stopping (patience of 5 epochs), 80-dimensional bidirectional GRU character encodings with 0.25 dropout, and 300-dimensional bi-directional GRU token encoder with 0.5 dropout. We used a batch size of 16. As training algorithm, Adam was used, with a learning or decay rate of 0.01 and L2-regularization. We used AllenNLP as our main implementation[2]. The models were trained on a Tesla T4 GPU with 16GB of GPU RAM.

Table 2: Overview of the embeddings studied in this paper.

Embedding	Words	Dimensions
Mat2Vec W2V	529,686	200
PubMed W2V	2,351,706	200
Drug W2V	553,195	420
CheMU W2V	1,252,586	200
PubMed ELMo	—	1,204
CheMU ELMo	—	1,204

Table 3: Impact of the different chemical embeddings on chemical NER (sorted by F1 score).

Word Embedding	F1	Δ (F1)
Mat2Vec W2V	26.89%	—
PubMed W2V	27.23%	+ 0.3%
Drug W2V	48.48%	+21.3%
CheMU W2V	53.24%	+ 4.8%
PubMed ELMo	70.15%	+16.9%
CheMU ELMo	72.41%	+ 2.3%

As Table 3 shows, using large the scale PubMed Word2Vec embeddings increases only marginally F1 score w.r.t. Mat2Vec (baseline model). They are both beaten by a wide margin by the drug-specific Word2Vec embeddings. Using patent-specific chemical Word2Vec embeddings also yields the best F1 score for Word2Vec embeddings. The best results are obtained with the ELMo embeddings, which outperform the Word2Vec embeddings again by a wide margin, and interestingly allow the model to achieve a reasonable F1 score of 72,41%. It is also interesting to observe that PubMed ELMo embeddings contain sufficient domain knowledge as to reach a close (and also reasonable) F1 score of 70.15%.

Intrinsic Evaluation For qualitative or *intrinsic* evaluation, we adopted the following strategy to generate comparable embeddings. We restricted the vocabulary of Word2Vec embeddings to the vocabulary of our test corpus (we excluded stop and function words). To compare these restricted Word2Vec embeddings to the ELMo embeddings, we computed an ELMo embedding for each occurrence of a test corpus content word, and then averaged all such embeddings to derive *ad hoc* single corpus-level Word2Vec embeddings. While such approach potentially discards contextual information, it can still preserve syntactic and semantic word type information, given that contextualised embeddings tend to assign similar embeddings

[2] https://allennlp.org/

Table 4: Top 10 similarity lists ("ibuprofen" query).

CheMU ELMo	PubMed ELMo	CheMU W2V	PubMed W2V	Drug W2V	Mat2Vec W2V
tacrine	atropine	aspirin	aspirin	pronounced	drug
ondansetron	ondansetron	clopidogrel	ondansetron	ultrastructure	drugs
aspirin	sulfamethoxazole	prednisolone	clopidogrel	mimics	aspirin
clopidogrel	aspirin	azathioprine	propranolol	surgical	sulfamethoxazole
dipyridamole	tacrine	atropine	placebo	favorable	propranolol
atropine	trimethoprim	nifedipine	tacrine	intestine	trimethoprim
prednisolone	propranolol	sulfamethoxazole	nifedipine	trained	norfloxacin
propranolol	prednisolone	dipyridamole	prednisolone	extinct	estradiol
trimethoprim	clopidogrel	propranolol	mg	slightly	antibiotics
nifedipine	papaverine	papaverine	topical	combination	nifedipine

Table 5: Overlap of similarity lists ("ibuprofen" query).

	PubMed W2V	Drug W2V	Mat2Vec W2V	CheMU ELMo	PubMed ELMo
W2V CheMU	0.33	0.25	0.25	0.43	0.54
W2V PubMed	—	0.25	0.18	0.43	0.54
W2V Drug	—	—	0.18	0.25	0.18
W2V Mat2Vec	—	—	—	0.11	0.33
ELMo CheMU	—	—	—	—	0.54

to words that assume similar grammatical and semantic roles within a sentence. The dimensionality of the embeddings was also standardized, by reducing, using singular value decomposition (SVD), to 200 dimensions the 420-, and 1,024-dimensional embeddings. We carried out three analysis.

Similarity analysis. We chose a drug entity, "ibuprofen", mentioned in the test corpus and retrieved its topmost 10 most similar words using cosine – $sim(w, w') = (\vec{w} * \vec{w'})/||\vec{w} \cdot \vec{w'}||$ – similarity. Ideally, since ibuprofen is a drug, what we expect to see in such similarity lists are names of drugs or chemical compounds.

The results obtained align – with some caveats – with the results observed over the NER model. As Table 4 shows, chemical patent embeddings produce better rankings than their more generic or non-patent specific counterparts. Furthermore, ELMo embeddings again show better results. Mat2Vec, Drug and PubMed word embeddings return common nouns ("drugs"), abbreviations ("mg"), adjectives ("topical") or verbs ("mimics"), whereas chemical patent Word2Vec and ELMo embeddings return only drugs or chemicals. Interestingly as well, they return as their topmost most similar term names of substances ("aspirin", "tacrine", "atropine") with somewhat similar anti-inflammatory properties.

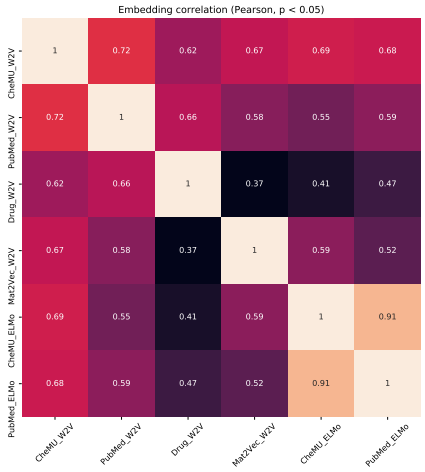

Figure 3: Embedding correlation (test corpus vocabulary).

Agreement analysis. These trends are largely confirmed when we check how much these lists semantically align. As chemical terms tend to be ambiguous, we used DrugBank[3] to normalise the terms w into their InChI $i(w)$ chemical identifiers[4]. Finally, we measured the set similarity – $sim(W, W') = |i(W) \cap i(W)'|/|i(W) \cup i(W)'|$ – of the ensuing normalised lists to measure how much the different embeddings "agree" on their understanding of "ibuprofen".

As Table 5 shows, chemical patent W2V and ELMo embeddings substantially align (0.43 and 0.54 set similarity resp.). On the other hand, the drug and Mat2Vec embeddings do not seem to align well to any other embedding. Interestingly, PubMed Word2Vec embeddings ex-

[3]DrugBank is a database of chemical substances, see: https://www.drugbank.ca

[4]An InChI identifier defines a unique representation of a molecule, see: https://www.inchi-trust.org.

hibit a higher than expected alignment with chemical patent embeddings.

Correlation analysis. Finally, we measured the degree of correlation between the embeddings. As Figure 3 shows, both ELMo embeddings correlate very highly (0.91)[5]. Chemical patent and PubMed Word2Vec embeddings moderately correlate with all embeddings, whereas Mat2Vec and drug embeddings give rise to lower correlation scores.

3. Conclusions

We have studied the quality of embeddings trained over chemical patents against those of embeddings of close biomedical domains. Our experiments show that patent specific embeddings outperform extrinsically other embeddings by giving rise to better F1 scores in chemical NER (72.41% on our test corpus). Correlation and similarity analysis indicate that they also outperform them intrinsically, and provide a better understanding of the chemistry domain. They also show that contextualized (ELMo) embeddings yield globally better results, and that generic but large scale PubMed ELMo embeddings (that cover the full life sciences domain) yield reasonable results for the chemistry domain.

References

Akhondi, S. A., Klenner, A. G., Tyrchan, C., Manchala, A. K., Boppana, K., Lowe, D., Zimmermann, M., Jagarlapudi, S. A., Sayle, R., Kors, J. A., et al. Annotated chemical patent corpus: a gold standard for text mining. *PloS one*, 9(9), 2014.

Akhondi, S. A., Rey, H., Schwörer, M., Maier, M., Toomey, J., Nau, H., Ilchmann, G., Sheehan, M., Irmer, M., Bobach, C., Doornenbal, M., Gregory, M., and Kors, J. A. Automatic identification of relevant chemical compounds from patents. *Database*, 2019:baz001, 2019.

Baroni, M., Dinu, G., and Kruszewski, G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL 2014*, 2014.

Bregonje, M. Patents: A unique source for scientific technical information in chemistry related industry? *World Patent Information*, 27(4):309–315, 2005.

Jin, Q., Dhingra, B., Cohen, W. W., and Lu, X. Probing biomedical embeddings from language models. *CoRR*, abs/1904.02181, 2019.

Klinger, R., Kolářik, C., Fluck, J., Hofmann-Apitius, M., and Friedrich, C. M. Detection of IUPAC and IUPAC-like chemical names. *Bioinformatics*, 24(13):i268–i276, 2008.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. Neural architectures for named entity recognition. *CoRR*, abs/1603.01360, 2016.

Muresan, S., Petrov, P., Southan, C., Kjellberg, M. J., Kogej, T., Tyrchan, C., Varkonyi, P., and Xie, P. H. Making every SAR point count: The development of Chemistry Connect for the large-scale integration of structure and bioactivity data. *Drug Discovery Today*, 16(23):1019–1030, 2011.

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

Schnabel, T., Labutov, I., Mimno, D. M., and Joachims, T. Evaluation methods for unsupervised word embeddings. In Márquez, L., Callison-Burch, C., Su, J., Pighin, D., and Marton, Y. (eds.), *Proceedings of EMNLP 2015*, 2015.

Segura-Bedmar, I., Suárez-Paniagua, V., and Martínez, P. Exploring word embedding for drug name recognition. In *Proceedings of the Louhi @ EMNLP 2015 Workshop*, 2015.

Senger, S., Bartek, L., Papadatos, G., and Gaulton, A. Managing expectations: Assessment of chemistry databases generated by automated extraction of chemical structures from patents. *Journal of Cheminformatics*, 7:49:1–49:12, 2015.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., and Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571 (7763):95–98, 2019.

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P. R., and Liu, H. A comparison of word embeddings for the biomedical natural language processing. *J. Biomed. Informatics*, 87:12–20, 2018.

Zhai, Z., Nguyen, D. Q., Akhondi, S., Thorne, C., Druckenbrodt, C., Cohn, T., Gregory, M., and Verspoor, K. Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings. In *Proceedings of the BioNLP @ ACL 2018 Workshop*, 2019.

[5] Which may explain their similar properties.