Word Embeddings for Chemical Patent Natural Language Processing

Camilo Thorne **Saber Akhondi**

{s.akhondi,c.thorne.1}@elsevier.com

Latinx @ ICML 2020

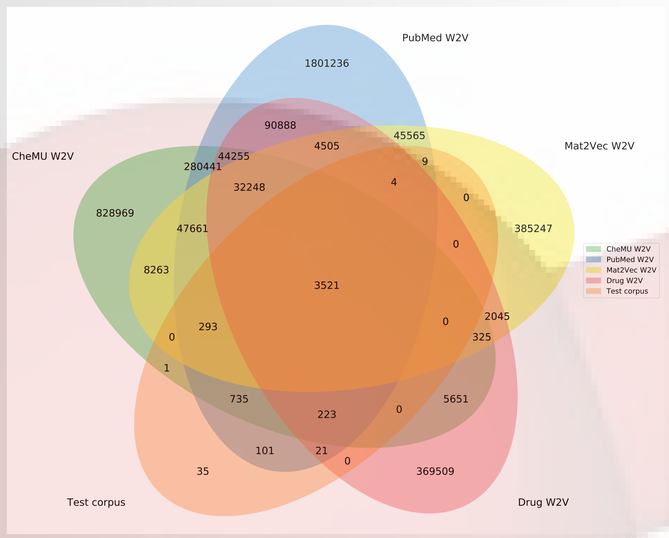
Problem

- ▷ NLP methods are key for processing chemistry patent, given the large quantities of patents filed each year
- ▷ Embeddings have been proposed/learnt over large biomedical and chemistry corpora:
  1. W2V / SkipGram: [4, 2, 3]
  2. Contextualized / ELMo: [1]
  3. Chemistry ELMo & W2V: [5]

Q Do chemistry embeddings outperform larger but generic embeddings on chemical text processing?

Datasets

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

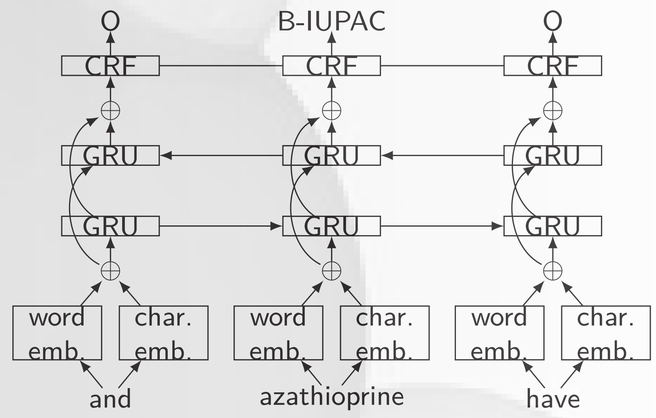


Embedding	Words	Dimensions
Mat2Vec W2V	529,686	200
PubMed W2V	2,351,706	200
Drug W2V	553,195	420
CheMU W2V	1,252,586	200
PubMed ELMo	—	1,204
CheMU ELMo	—	1,204

- ▷ NER: Sampled small chemical NER corpus from patents
- ▷ SIM: Restricted embeddings to test corpus and reduced dimensions

Extrinsic and Intrinsic Evaluation

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%



- ▷ Chemistry embeddings outperform biomedical embeddings on chemical texts
- ▷ PubMed ELMo gives reasonable F1-score on gold standard, using a simple biGRU-CRF model
- ▷ Contextualized embeddings invariably better ⇒ PubMed ELMo close in F1-score

"ibuprofen" query top 10 similarity rankings (cosine):

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

- ▷ Chemical embeddings return better similarity lists:
  1. All terms denote drugs and/or chemical compounds
  2. Terms closer in chemical (anti-inflammatory) properties ("tacrine", "atropine", "aspirin") to "ibuprofen"
- ▷ PubMed and CheMU ELMo again close
- ▷ Also: Better quality embeddings correlate highly (cor. >0.7, $p<0.05$) whereas lower quality embeddings correlate poorly (cor. <0.4, $p<0.05$)

Conclusions

- ▷ Studied the quality of embeddings trained over chemical patents against biomedical embeddings
- ▷ Patent-specific embeddings outperform larger scale but generic embeddings on NER and provide a better understanding of the chemistry domain
- ▷ But: Large scale generic ELMo embeddings provide OK performance

References

[1] Qiao Jin, Bhuwan Dhingra, William W Cohen, and Xinghua Lu. Probing biomedical embeddings from language models. *CoRR*, abs/1904.02181, 2019.
[2] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. Distributional semantics resources for biomedical text processing. In *Proceedings of LBM 2013*, 2013.
[3] Isabel Segura-Bedmar, Víctor Suárez-Paniagua, and Paloma Martínez. Exploring word embedding for drug name recognition. In *Proceedings of the Louhi @ EMNLP 2015 Workshop*, 2015.
[4] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.
[5] Zenan Zhai, Dat Quoc Nguyen, Saber Akhondi, Camilo Thorne, Christian Druckenbrodt, Trevor Cohn, Michelle Gregory, and Karin Verspoor. Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings. In *Proceedings of the BioNLP @ ACL 2018 Workshop*, 2019.