Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings

Zenan Zhai[1], **Dat Quoc Nguyen**[1], **Saber A. Akhondi**[2], **Camilo Thorne**[2],
Christian Druckenbrodt[2], **Trevor Cohn**[1], **Michelle Gregory**[2], **Karin Verspoor**[1]

[1]The University of Melbourne, Australia; [2]Elsevier

[1]{zenan.zhai,dqnguyen,trevor.cohn,karin.verspoor}@unimelb.edu.au

[2]{s.akhondi,c.thorne.1,c.druckenbrodt,m.gregory}@elsevier.com

Abstract

Chemical patents are an important resource for chemical information. However, few chemical Named Entity Recognition (NER) systems have been evaluated on patent documents, due in part to their structural and linguistic complexity. In this paper, we explore the NER performance of a BiLSTM-CRF model utilising pre-trained word embeddings, character-level word representations and contextualized ELMo word representations for chemical patents. We compare word embeddings pre-trained on biomedical and chemical patent corpora. The effect of tokenizers optimized for the chemical domain on NER performance in chemical patents is also explored. The results on two patent corpora show that contextualized word representations generated from ELMo substantially improve chemical NER performance w.r.t. the current state-of-the-art. We also show that domain-specific resources such as word embeddings trained on chemical patents and chemical-specific tokenizers have a positive impact on NER performance.

1   Introduction

Chemical patents are an important starting point for understanding of chemical compound purpose, properties, and novelty. New chemical compounds are often initially disclosed in patent documents; however it may take 1-3 years for these chemicals to be mentioned in chemical literature (Senger et al., 2015), suggesting that patents are a valuable but underutilized resource. As the number of new chemical patent applications is drastically increasing every year (Muresan et al., 2011), it is becoming increasingly important to develop automatic natural language processing (NLP) approaches enabling information extraction from these patents (Akhondi et al., 2019). Chemical Named-Entity Recognition (NER) is a fundamental step for information extraction from chemical-related texts,

supporting relation extraction (Wei et al., 2016), reaction prediction (Schwaller et al., 2018) and retro-synthesis (Segler et al., 2018).

However, performing NER in chemical patents can be challenging (Akhondi et al., 2014). As legal documents, patents are written in a very different way compared to scientific literature. When writing scientific papers, authors strive to make their words as clear and straight-forward as possible, whereas patent authors often seek to protect their knowledge from being fully disclosed (Valentinuzzi, 2017).

In tension with this is the need to claim broad scope for intellectual property reasons, and hence patents typically contain more details and are more exhaustive than scientific papers (Lupu et al., 2011).

There are a number of characteristics of patent texts that create challenges for NLP in this context. Long sentences listing names of compounds in chemical patents are frequently used. The structure of sentences in patent claims is usually complex, and syntactic parsing in patents can be difficult (Hu et al., 2016). A quantitative analysis by Verberne et al. (2010) showed that the average sentence length in a patent corpus is much longer than in general language use. That work also showed that the lexicon used in patents usually includes domain-specific and novel terms that are difficult to understand. Some patent authorities use Optical Character Recognition (OCR) for digitizing patents, which can be problematic when applying automatic NLP approaches as the OCR errors introduces extra noise to the data (Akhondi et al., 2019).

Most NER systems for the chemical domain were developed, trained and tested on either chemical literature or only the title and abstract of chemical patents (Akhondi et al., 2019). There are substantial linguistic differences between ab-

stracts and the corresponding full text publications (Cohen et al., 2010). The performance of NER approaches on full patent documents has still not been fully explored (Krallinger et al., 2015).

Hence, this paper will focus on presenting the best NER performance achieved to date on full chemical patent corpus.

We use a combination of pre-trained word embeddings, a CNN-based character-level word representation and contextualized word representations generated from ELMo, trained on a patent corpus, as input to a BiLSTM-CRF model. The results show that contextualized word representations help improve chemical NER performance substantially. In addition, the impact of the choice of pre-trained word embeddings and tokenizers is assessed.

The results show that word embeddings that are pre-trained on chemical patents outperform embeddings pre-trained on biomedical datasets, and using tokenizers optimized for the chemical domain can improve NER performance in chemical patent corpora.

2 Related work

In this section, we summarize previous methods and empirical studies on NER in chemical patents.

Two existing Conditional Random Field (CRF)-based systems for chemical named entity recognition are tmChem (Leaman et al., 2015) and ChemSpot (Rocktäschel et al., 2012); each makes use of numerous hand-crafted features including word shape, prefix, suffix, part-of-speech and character N-grams in an algorithm based on modelling of tag sequences. A previous detailed empirical study explored the generalization performance of these systems and their ensembles (Habibi et al., 2016). The application of the tmChem model trained on chemical *literature* corpora of the BioCreative IV CHEMDNER task (Krallinger et al., 2015) and the ChemSpot model trained on a subset of the SCAI corpus (Klinger et al., 2008) resulted in a significant performance drop over chemical *patent* corpora.

Zhang et al. (2016) compared the performance of CRF- and Support Vector Machine (SVM)-based models on the CHEMDNER-patents corpus (Krallinger et al., 2015). The features constructed in that work included the binarized embedding (Guo et al., 2014), Brown clustering (Brown et al., 1992) and domain-specific features extracted by

detecting common prefixes/suffixes in chemical words. The obtained results show that the performance of CRF and SVM models can be significantly improved by incorporating unsupervised features (e.g. word embeddings, word clustering). The study also showed that the SVM model slightly outperformed the CRF model in the chemical NER task.

To perform chemical NER on the CHEMDNER patents corpus, Akhondi et al. (2016) proposed an ensemble approach combining a gazetteer-based method and a modified version of tmChem. Here, the gazetteer-based method utilized a wide range of chemical dictionaries, while additional features such as stems, prefixes/suffixes, chemical elements were added to the original feature set of tmChem. In the ensemble approach, tokens were predicted as chemical mentions if recognized as positive by either tmChem or the gazetteer-based method. The results showed that both gazetteer-based and ensemble approaches were outperformed by the modified tmChem version in terms of overall F_1 score, although these two approaches can obtain higher recall.

Huang et al. (2015) proposed a BiLSTM-CRF based on the use of a bidirectional long-short term memory network – BiLSTM (Schuster and Paliwal, 1997) – to extract (latent) features for a CRF classifier. The BiLSTM encodes the input in both forward and backward directions and passes the concatenation of outputs from both directions as input to a linear-chain CRF sequence tagging layer. In this approach, the BiLSTM selectively encodes information and long-distance dependencies observed while processing input sentences in both directions, while the CRF layer globally optimizes the model by using information from neighbor labels.

The morphological structures within words are also important clues for identifying named entities in biological domain. Such morphological structures are widely used in systematic chemical name formats (e.g. IUPAC names) and hence particularly informative for chemical NER (Klinger et al., 2008). Character-level word representations have been developed to leverage information from these structures by encoding the character sequences within tokens. Ma and Hovy (2016) uses Convolutional Neural Networks (CNNs) to encode character sequences while Lample et al. (2016)

developed a LSTM-based approach for encoding character level information.

Habibi et al. (2017) presented an empirical study comparing three NER models on a large collection of biomedical corpora including the BioSemantics patent corpus: (1) tmChem–the CRF-based model with hand-crafted features–used as the baseline; (2) a second CRF model based on CRFSuite (Okazaki, 2007) using pre-trained word embeddings; (3) and a BiLSTM-CRF model with additional LSTM-based character-level word embeddings (Lample et al., 2016). The performance of CRFSuite- and BiLSTM-CRF-based models with different sets of pre-trained biomedical word embeddings (Pyysalo et al., 2013) were also explored. The results showed that the BiLSTM-CRF model with the combination of domain-specific pre-trained word embedding and LSTM-based character-level word embeddings outperformed the two CRF-based models on chemical NER tasks in both chemical literature and chemical patent corpora. However, this work used only a general tokenizer (i.e. OpenNLP) and word embeddings pre-trained on biomedical corpora.

Corbett and Boyle (2018) presented word-level and character-level BiLSTM networks for chemical NER in literature domain. The word-level model employed word embeddings learned by GloVe (Pennington et al., 2014) on a corpus of patent titles and abstracts. The character-level model used two different transfer learning approaches to pre-train its character-level encoder. The first approach attempts to predict neighbor characters at each time step, while the other tries to predict whether a given character sequence is an entry in the chemical database ChEBI (Degtyarenko et al., 2007). Experimental results show that the character-level model can produce better NER performance than word-level model by leveraging transfer learning. In addition, for the word-level model, using pre-trained word embeddings learned from a patent corpus helps produce better performance than using the pre-trained ones learned from a general corpus.

3 Our empirical methodology

This section presents our empirical study of NER chemical on patent datasets. We first outline the experimental datasets (Section 3.1) and the tokenizers (Section 3.2) used to pre-process these datasets, and then we introduce the BiLSTM-CRF-based models (Section 3.3) with pre-trained word embeddings (Section 3.4), character-level word embeddings (Section 3.5), contextualized word embeddings (Section 3.6) and implementation details (Section 3.7).

3.1 Dataset

We conduct experiments on 2 patent corpora: the BioSemantics patent corpus (Akhondi et al., 2014) and Reaxys gold set (Akhondi et al., 2019).

The BioSemantics patent corpus (Akhondi et al., 2014) consists of 200 full chemical patent documents with 9 different entity classes. In particular, this corpus has 170K sentences and and 360K entity annotations, which is much larger than previously used datasets, e.g. the CHEMDNER patent abstract corpus (Krallinger et al., 2015). Therefore, this corpus can be considered as a more suitable resource for evaluating deep learning methods in which a large amount of training data is required (LeCun et al., 2015). A subset of 47 patents were annotated by multiple groups (at least 3) of annotators and evaluated through inner-annotator agreement. By harmonizing the annotations from different annotator groups, these 47 patents formed the "*harmonized*" set in the BioSemantics patent corpus. We use the harmonized set for both hyper-parameter tuning and error analysis as it has known high-quality annotations.

The Reaxys gold set (Akhondi et al., 2019) contains 131 patent snippets (parts of full chemical patent documents) from several different patent offices. The tagging scheme of this corpus includes 2 coarse-grained labels chemical class and chemical compounds, and 7 fine-grained labels of chemical compound (e.g. *mixture-part*, *prophetic*) and chemical class (e.g. *bio-molecule*, *Markush*, *mixture*, *mixture-part*). This corpus is relatively small in size, approximately 20,000 sentences in total, but very richly annotated. The relevancy score of each chemical entity and the relations between them were also annotated, which allows this corpus to be used in other tasks beyond named entity recognition.

In our experiments, each corpus is used separately. We follow Habibi et al. (2017) to use a ratio split of 60%/10%/30% for training/development/test. Note that on the BioSemantic patent corpus, our sampling of datasets may not be exactly the same as in Habibi et al. (2017).
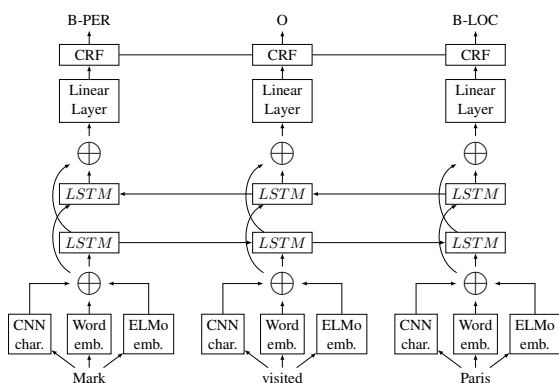
Figure 1: Architecture of EBC-CRF

3.2 Tokenizers

The morphological information captured by character-level word representations can be highly affected by tokenization quality. General-purpose tokenizers usually split tokens by spaces and punctuation. However, strict adherence to such boundaries may not be suitable for chemical texts as spaces and punctuation are commonly used in the IUPAC format for chemical names (e.g. *3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide*) (Jessop et al., 2011). Hence, the impact of different tokenizers on NER also needs to be explored.

A pre-processing step is applied to the patent corpora including sentence detection and tokenization. Following Habibi et al. (2017), we use the OpenNLP (Morton et al., 2005) English sentence detection model. To explore the relationship between tokenization quality and final NER performance, we apply different tokenizers and train/test models with each tokenizer individually. To investigate the effect of a general domain tokenizer, following Habibi et al. (2017), we also use the OpenNLP tokenizer. To investigate whether NER performance will be affected by tokenization quality, we employ three tokenizers optimized for chemical texts including ChemTok (Akkasi et al., 2016), OSCAR4 (Jessop et al., 2011) and NBIC UMLSGeneChemTokenizer.[1]

3.3 Models

We use the BiLSTM-CNN-CRF model (Ma and Hovy, 2016) as our baseline. We extend the baseline by adding the contextualized word representations generated from ELMo (Peters et al., 2018).

---

[1]NBIC UMLSGeneChemTokenizer is developed by the Netherlands Bioinformatics Center, available at https://trac.nbic.nl/data-mining/wiki.

For convenience, we call the extended version as EBC-CRF as illustrated in Figure 1. In particular, for EBC-CRF, we use a concatenation of pre-trained word embeddings, CNN-based character-level word embeddings and ELMo-based contextualized word embeddings as the input of a BiLSTM encoder. The BiLSTM encoder learns a latent feature vector for each word in the input. Then each latent feature vector is linearly transformed before being fed into a linear-chain CRF layer (Lafferty et al., 2001) for NER tag prediction. We assume binary potential between tags and unary potential between tags and words.

3.4 Pre-trained word embeddings

Dai et al. (2019) showed that NER performance is significantly affected by the overlap between pre-trained word embedding vocabulary and the target NER data. Therefore, we explore the effects of different sets of pre-trained word embeddings on the NER performance.

We use 200-dimensional pre-trained PubMed-PMC and Wiki-PubMed-PMC word embeddings (Pyysalo et al., 2013), which are widely used for NLP tasks in biomedical domain. Both the PubMed-PMC and Wiki-PubMed-PMC embeddings word embeddings were generated by training the Word2Vec skip-gram model (Mikolov et al., 2013) on a collection of PubMed abstracts and PubMed Central articles. Here, an additional Wikipedia dump was also used to learn the Wiki-PubMed-PMC word embeddings.

To explore whether word embeddings trained in the same domain can produce better performance in NER tasks, we learn another set of word embeddings, which we called *ChemPatent* embeddings, by applying the same model and hyperparameters from Pyysalo et al. (2013) on a collection of 84,076 full patent documents (1B tokens) across 7 patent offices (see Table 1 for details).

The pre-trained PubMed-PMC, Wiki-PubMed-PMC and ChemPatent word embeddings are fixed during training of the NER models. For a more concrete comparison, a set of 200-dimensional trainable word embeddings initialized from normal distribution is used as a baseline.

The 200-dimensional baseline word embeddings contain all words in the vocabulary of the dataset and are initialized from a normal distribution, the baseline word embeddings are learned during training process. The vocabulary of models

Patent Office	Document	Sentence	Tokens
AU	7,743	4,662,375	156,137,670
CA	1,962	463,123	16,109,776
EP	19,274	3,478,258	117,992,191
GB	918	182,627	6,038,837
IN	1,913	261,260	9,015,238
US	41,131	19,800,123	628,256,609
WO	11,135	4,830,708	159,286,325
Total	84,076	33,687,474	1,092,836,646

Table 1: Statistics of the unannotated patent corpus used for training ChemPatent embeddings and ELMo.

Hyper-para.	Value
Optimizer	Adam
Learning rate	0.001
Mini-batch size	16
Clip Norm(L2)	1
Dropout	[0.25, 0.25]

(a) BiLSTM-CRF

Hyper-para.	Value
charEmbedSize	50
filter length	3
# of filters	30
output size	30

(b) CNN-char

Table 2: Fixed hyper-parameter configurations.

using pre-trained word embeddings is built by taking the union of words in the pre-traied word embedding file and words with frequency more than 3 in training and development sets. We do not update weights for word embeddings if pre-trained word embeddings were used.

3.5 Character-level representation

The BiLSTM-CRF model with character-level word representations (Lample et al., 2016; Ma and Hovy, 2016) has been shown to have state-of-the-art performance in NER tasks on chemical patent datasets (Habibi et al., 2017). It has been shown that the choice of using LSTM-based or CNN-based character-level word representation has little effect on final NER performance in both general and biomedical domain while the CNN-based approach has the advantage of reduced training time (Reimers and Gurevych, 2017b; Zhai et al., 2018). Hence, we use the CNN-based approach with the same hyper-parameter settings of Reimers and Gurevych (2017b) for capturing character-level information (see Table 2 for details).

3.6 ELMo

ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) can be used to generate contextualized word representations by combining internal states of different layers in neural language models. Contextualized word representation can help to improve performance in various NLP tasks by incorporating contextual information, essentially allowing for the same word to have distinct context-dependent meanings. This could be particularly powerful for chemical NER since generic chemical names (e.g. *salts, acid*) may have different meanings in other domains. We therefore explore the impact of using contextualized word representations for chemical patents.

We train ELMo on the same corpus of 84K

patents (detailed in Table 1), which we use for training the ChemPatent embeddings (described in Section 3.4). We use the ELMo implementation provided by Peters et al. (2018) with default hyper-parameters.[2] Such neural language models require a large amount of computational resources to train. In ELMo, a maximum character sequence length of tokens is set to make training feasible. However, systematic chemical names in chemical patents are often longer than the typical maximum sequence length of these neural language models. As very long tokens tend to be systematic chemical names, we reduced the max length of word from 50 to 25 and replace tokens longer than 25 characters by a special token "Long_Token".

3.7 Implementation details

Our NER model implementation is based on the AllenNLP system (Gardner et al., 2017). We learn model parameters using the training set, and we use the overall F_1 score over development set as indicator for performance improvement. All models in this paper are trained with 50 epochs in maximum, and an early stopping is applied if there are no overall F_1 score improvement observed after 10 epochs.

In Reimers and Gurevych (2017a) and Zhai et al. (2018), optimal hyper-parameters of BiLSTM-CRF models in NER tasks were explored. Hence, we fix the hyper-parameters shown in Table 2 to the suggested values in our experiments, which means that only models with 2-stacked LSTM of size 250 are evaluated.

In this study, we also consider the choice of tokenizer and word embedding source as hyper-parameters. To compare the performance of different tokenizers, we tokenize the same split of datasets with different tokenizers and evaluate the overall F_1 score over development set. After the best tokenizer for pre-processing patent corpus is determined, we use datasets tokenized by the best

---

[2]https://github.com/allenai/bilm-tf

Tokenizer	BioSemantics	Reaxys	Avg.
OpenNLP	89.36	89.43	89.40
NBIC	**+0.86**	-0.13	+0.37
ChemTok	+0.04	+1.68	+0.86
OSCAR4	+0.08	**+1.86**	**+0.97**

Table 3: Best F_1 of EBC-CRF model with different tokenizations on development sets of BioSemantics patent (harmonized set) and Reaxys Gold with ChemPatent embeddings in use. Recall that the harmonized set of 47 patents is a subset of BioSemantics, which were annotated by multiple groups (i.e. better annotation quality than remaining patents).

Embeddings	BioSemantics	Reaxys	Avg.
Baseline	88.54	90.05	89.30
PubMed-PMC	+0.61	+1.03	+0.82
Wiki-PubMed-PMC	+1.24	+0.95	+1.10
ChemPatent	**+1.68**	**+1.24**	**+1.46**

Table 4: Best F_1 of EBC-CRF model with different word embeddings on development sets of BioSemantics patent (harmonized set) and Reaxys Gold (tokenized by NBIC and OSCAR 4 tokenizer respectively)

tokenizer to train models with different pre-trained word embeddings. The best set of pre-trained word embeddings for patent corpus is determined based on the overall F_1 score over development set. We then take the best performing tokenizer and pre-trained word embeddings by comparing the marco-average F_1 score improvement on both experimental datasets.

4 Results

4.1 Main Results

Effects of different tokenizers: Table 3 shows that all 3 tokenizers optimized for the chemical domain outperform the baseline general-purpose tokenizer (i.e. OpenNLP). The best performance on BioSemantics and Reaxys Gold are achieved by using the NBIC tokenizer (+1.86 F_1 score) and the OSCAR4 tokenizer (+0.86 F_1 score), respectively. The best overall tokenizer is OSCAR4 which obtains about 1.0 absolute macro-averaged F_1 improvement in comparison to the baseline.

Effects of different sets of word embeddings: Table 4 shows results obtained by training EBC-CRF with different sets of pre-trained word embeddings. On both BioSemantics and Reaxys Gold, it is not surprising that our ChemPatent word embeddings help produce the best performance on the development set, obtaining (on average) a higher F_1 score of 1.5 as compared to the

Model	P	R	F_1
tmChem	72.56	78.37	75.35
CRFSuite	**81.93**	78.38	80.12
BiLSTM-CRF + LSTM-char	79.72	**84.42**	**82.01**
BiLSTM-CNN-CRF	83.76	85.01	84.38
EBC-CRF	**84.30**	**87.11**	**85.68**

Table 5: NER scores on full BioSemantics test set (Akhondi et al., 2014). Results in the first 3 rows were reported in Habibi et al. (2017). BiLSTM-CRF + LSTM-char denotes the BiLSTM-CRF model with additional LSTM-based character-level word embeddings (Lample et al., 2016). Recall that our models use the OSCAR4 tokenizer and pre-trained ChemPatent word embeddings.

baseline embeddings. Specifically, ChemPatent does better than the second best Wiki-PubMed-PMC with about 0.4 improvement. In the rest of the Results section, obtained results are reported with the use of the OSCAR4 tokenizer and the ChemPatent embeddings on both experimental datasets.[3]

Final results: Table 5 compared results reported in Habibi et al. (2017) and our approach on the full BioSemantics test set. It is clear that all neural models outperform conventional CRF-based models tmChem and CRFSuite. Our EBC-CRF model outperforms the BiLSTM-CRF + LSTM-char model with a 3.7 F_1 score improvement. Compared to the baseline model BiLSTM-CNN-CRF, the ELMo-based contextualized word embeddings help to produce an F_1 improvement of 1.3 points.

Table 6 details our F_1 scores for BiLSTM-CNN-CRF and EBC-CRF with respect to each entity label on both the BioSemantics patent corpus and the Reaxys Gold set. The overall results show that ELMo-based contextualized word embeddings help improve the baseline by 1.3 and 4.8 absolute F_1 score on BioSemantics and Reaxys, respectively.

In BioSemantics patent corpus, we obtain 1+ F_1 score improvements on frequent entity labels (i.e. $> 3,000$ instances) except for the entity label *Formula*, which has 0.4 absolute improvement. Higher improvements can be observed on rare entity labels (e.g. 4 points on *Mode of Actions*, 6 points on *Registry numbers* and *Trademarks*). The highest improvement at 9 points is found for the most rare entity label *CAS Number*.

---

[3]OSCAR4 helped produce the highest "macro-averaged" improvement on both datasets.

Entity label	Count†		BiLSTM-CNN-CRF			+ELMo			Δ_{F_1}
	#	%	P	R	F_1	P	R	F_1	
B (Abbreviation)	6,558	5.78	85.90	87.02	86.46	85.78	89.98	87.83	+1.37
C (CAS Number)	13	0.01	54.55	46.15	50.00	57.14	61.54	59.26	+9.26
D (Trademark)	2,290	2.01	62.58	61.79	62.18	66.44	71.40	68.83	+6.65
F (Formula)	7,935	6.99	86.05	86.81	86.42	83.07	90.91	86.82	+0.40
G (Generic)	51,313	45.20	81.45	84.56	82.98	83.84	84.44	84.14	+1.16
M (IUPAC)	39,896	35.14	88.40	87.77	88.09	87.25	91.20	89.18	+1.09
MOA (Mode of Action)	1,137	1.00	68.97	63.32	66.02	67.62	72.74	70.08	+4.06
R (Registry #)	96	0.08	55.68	51.04	53.26	65.82	54.17	59.43	+6.17
T (Target)	4,290	3.78	77.77	77.32	77.55	77.21	82.68	79.85	+2.30
Micro Avg.	113,528	100.0	83.76	85.01	84.38	84.30	87.11	85.68	+1.30

(a) BioSemantics

Entity label	Count†		BiLSTM-CNN-CRF			+ELMo			Δ_{F_1}
	#	%	P	R	F_1	P	R	F_1	
1 (chemClass)	1,476	12.36	78.35	66.46	71.92	81.96	75.75	78.73	+6.81
2 (chemClass$_{biomolecule}$)	951	7.96	71.86	70.50	71.17	76.27	78.76	77.50	+6.33
3 (chemClass$_{markush}$)	38	0.32	42.86	47.37	45.00	42.86	47.37	45.00	+0.00
4 (chemClass$_{mixture}$)	387	3.24	76.49	59.69	67.05	74.18	64.60	69.06	+2.01
5 (chemClass$_{mixture-part}$)	161	1.35	71.00	44.10	54.41	78.10	50.93	61.65	+7.24
6 (chemClass$_{polymer}$)	609	5.10	81.40	72.82	76.87	89.20	84.07	86.56	+13.74
7 (chemCompound)	6,988	58.53	89.02	92.01	90.49	91.01	94.58	92.76	+2.27
8 (chemCompound$_{mixture-part}$)	904	7.57	90.02	81.86	85.75	90.63	85.62	88.05	+2.27
9 (chemCompound$_{prophetics}$)	426	3.57	18.52	2.35	4.17	77.75	79.58	78.65	+74.48
Micro Avg.	11,940	100.0	85.12	80.36	82.67	87.41	87.53	87.47	+4.80

(b) Reaxys Gold

Table 6: F_1 score with respect to each entity label. "Count†" denotes gold-entity counts in test sets. "+ELMo" denotes scores obtained by EBC-CRF.

In the Reaxys Gold set, with ELMo we obtain 2+ F_1 score improvements on entity labels *chemCompound*, *chemCompound-mixture part* and *chemClass-mixture*. Higher improvements (> 6 points) can be seen on some rare entity labels such as *chemClass*, *chemClass-biomolecule*, *chemclass-mixture-part* and *chemClass-polymer*. The improvements on entity label *chemClass-Markush* and *chemCompound-prophetics* are irregular compared to others. In particular, an absolute F_1 improvement of 74+ is achieved on entity label *chemCompound-prophetics*, while we do not find any improvement on *chemClass-Markush*.

4.2 Error Analysis

To perform error analysis on BioSemantics, we use its harmonized subset. Figure 2 (a) shows that most of the errors are confusions between non-chemical words and generic chemical names (e.g. *water, salt, acid*). For example, as illustrated in Figure 3 (a), the word "*salt*" which appears at the end of a systematic name should be identified as a part of the systematic name. However,

the same word is also widely used to describe a class of chemicals, e.g. "*pharmaceutically acceptable salt*" in Figure 3 (b). Disambiguation between chemical class and chemical compound is a challenging task even for human annotators, and is thus particularly difficult for a statistical model to learn. The confusion matrix of Reaxys Gold set in Figure 2 (b) also supports this point since most confusions are between non-chemical words, chemical classes and chemical compounds.

The Reaxys Gold set has a more complex tag set than the BioSemantics patent corpus, as it assigns separate fine-grained tags for subcategories of chemical classes (*chemClass*) and chemical compounds (*chemCompound*). As illustrated in Table 6, there is not sufficient training data for fine-grained sub-category labels. It is difficult for a high complexity neural model to learn characteristics of these sub-category labels and the key difference between the main categories and their subcategories. Figure 2 (b) shows that 50% the errors for "*chemical compound prophetics*" and 80% errors for "*chemical compound mixture part*" are
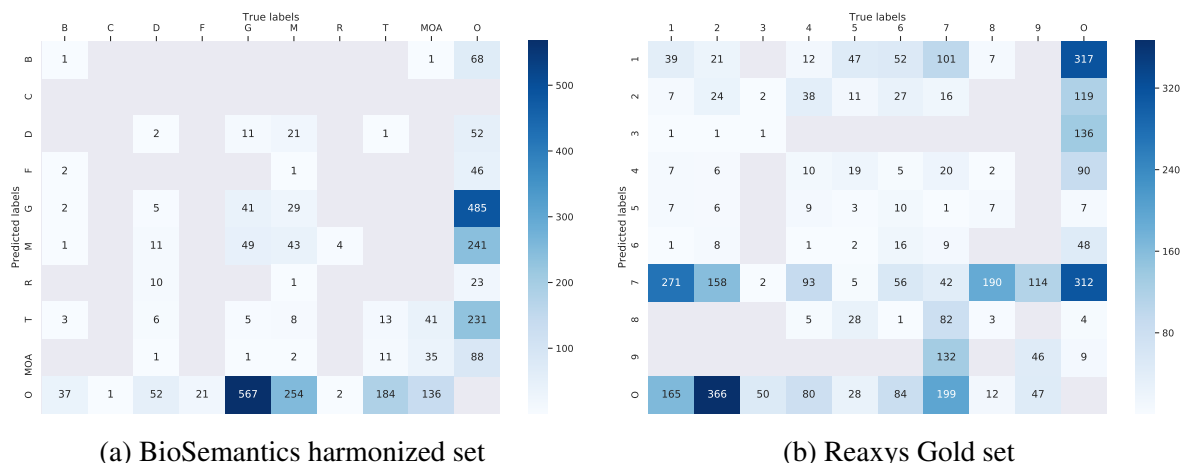
Figure 2: Confusion matrix of EBC-CRF model on BioSemantics (harmonized) and Reaxys Gold. x-axis: true labels; y-axis: predicted labels; numbers on cell where $x = y$ represent confusion between B (Begin) and I (Inside) tags. In (b) Labels 1–9 are detailed in Table 6 (b).

(a) *Salt* describing chemical compound was not detected.

(b) *Salt* describing chemical class being predicted as chemical compound.

Figure 3: Example of confusions caused by generic chemical names. (M: false negatives, G, M: false positives)

due to confusion with their parent category "*chemical compound*".

Another typical error observed frequently in BioSemantics and Reaxys is caused by participles. The most common example is word *'substituted'*. In "*substituted or un-substituted alkyl*", the token "*substituted*" refers to a specific chemical compound "*substituted alkyl*". Whereas in "*2-pyridinyl is optionally substituted with 1-3 substituents*", the token "*substituted*" refers to the substitution reaction.

We also observe that in both patent corpora, there are long sequences of systematic chemical names connected by comma only. Since there are no narrative words between the chemical names in such sequences, it is unlikely that the model can capture any contextual information when tagging them. This can potentially cause a "chain reaction" as shown in Figure 4, in which all chemical names fail to be recognized when the first chemical name is not tagged correctly.

4.3 Discussion

The results in Table 3 show that all chemical tokenizers outperform the OpenNLP general domain tokenizer. This is not surprising because tokenizers optimized for the chemical domain usually use either rule-based method or gazetteer-based methods to ensure that long systematic chemical names will be treated as a single token instead of being split into several tokens by symbols. This is reasonable as the character-level word representation will not be able to capture the morphological structures in a long chemical name if it is split into several tokens.

In the BioSemantics patent corpus, 80% of all entities are annotated as *Generic* or *IUPAC*. When adding ELMo-based word representations, we obtain smaller improvements in F_1 score for *Generic* and *IUPAC* than for remaining entity labels/types. This makes sense, as there are already enough training instances for these two labels in the dataset. By contrast, for rare entity labels with frequencies of less than 2 (e.g. *CAS Numbers, Trademarks, Mode of Actions, Registry numbers*), we obtain improvements of 4+ points when exploiting external information conveyed via ELMo.

The global F_1 score improvements on both experimental datasets confirm further this observation, viz., that score improvements due to ELMo

Figure 4: Example of "chain reaction" like errors. (⬛: false negatives, ⬛: false positives, ⬛: true positives)

decrease in inverse proportion to label frequency and training set size. Since the BioSemantics patent corpus contains 10 times more training instances than the Reaxys Gold set, we obtain an absolute improvement of 4.8 on Reaxys Gold set but of 1.3 points on the BioSemantics patent corpus.

Adding ELMo substantially improves the F_1 score on *chemCompound-prophetics*. This is because *chemCompound-prophetics* named entities are all long systematic chemical names which are arranged in lists. Since we replace all tokens longer than 25 characters with "Long_Token" when training ELMo, almost all sentences containing *chemCompound-prophetics* entities appear in the "Long_Token" style. This makes the ELMo-based representations of such long entities almost identical, and particularly easy to predict, thus resulting in an F_1 score improvement of 74 points for *chemCompound-prophetics*. We also observe no improvement for the *chemClass-Markush* label. The Markush structures are figures describing the structure of chemical compounds in which only a few parts/functional groups are labeled. When transforming to text, only the textual labels in the Markush structure are preserved. Thus, it is difficult for ELMo to learn any useful information from the broken Markush structures.

5 Conclusions

In this paper, we have made the following contributions towards improved chemical named entity recognition in chemical patents:

1. We improve on the current state-of-art for chemical NER in patents by +2.67 F_1 score.
2. We confirm that tokenizers optimized for chemical domain have a positive effect on NER performance by preserving informative morphological structures in systematic chemical names.
3. We demonstrate that word embeddings pre-trained on an in-domain chemical patent corpus help produce better performance than the word embeddings pre-trained on biomedical literature corpora.
4. We show that chemical NER performance can be improved by using contextualized word representations.
5. We release our ChemPatent word embeddings and an ELMo model trained from scratch on a newly collected corpus of 84K unannotated chemical patents, which can be utilized for downstream NLP tasks on chemical patents.[4]

Inspired by the patterns uncovered by our error analysis, our future work on chemical NER will focus on developing models which can be used to support disambiguation of general chemical words. In addition, it would be interesting to explore contextualized word embeddings learned by other neural models such as BERT (Devlin et al., 2019) or OpenAI GPT models (Radford et al., 2019) in future work.

Acknowledgments

---

[4] https://github.com/zenanz/ChemPatentEmbeddings

References

Saber A Akhondi, Alexander G Klenner, Christian Tyrchan, Anil K Manchala, Kiran Boppana, Daniel Lowe, Marc Zimmermann, Sarma ARP Jagarlapudi, Roger Sayle, Jan A Kors, et al. 2014. Annotated chemical patent corpus: a gold standard for text mining. *PloS one*, 9(9):e107477.

Saber A Akhondi, Ewoud Pons, Zubair Afzal, Herman van Haagen, Benedikt FH Becker, Kristina M Hettne, Erik M van Mulligen, and Jan A Kors. 2016. Chemical entity recognition in patents by combining dictionary-based and statistical approaches. *Database*, 2016.

Saber A. Akhondi, Hinnerk Rey, Markus Schwörer, Michael Maier, John P. Toomey, Heike Nau, Gabriele Ilchmann, Mark Sheehan, Matthias Irmer, Claudia Bobach, Marius A. Doornenbal, Michelle Gregory, and Jan A. Kors. 2019. Automatic identification of relevant chemical compounds from patents. *Database*, 2019:baz001.

Abbas Akkasi, Ekrem Varoğlu, and Nazife Dimililer. 2016. Chemtok: a new rule based tokenizer for chemical named entity recognition. *BioMed research international*, 2016.

Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

K. Bretonnel Cohen, Helen L. Johnson, Karin Verspoor, Christophe Roeder, and Lawrence E. Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11(1):492.

Peter Corbett and John Boyle. 2018. Chemlistem: chemical named entity recognition using recurrent neural networks. *Journal of cheminformatics*, 10(1):59.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2019. Using Similarity Measures to Select Pretraining Data for NER. In *Proceedings of NAACL-HLT 2019*, pages 1460–1470.

Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2007. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36(suppl_1):D344–D350.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform. In *arXiv:1803.07640*.

Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Revisiting embedding features for simple semi-supervised learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 110–120.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Maryam Habibi, David Luis Wiegandt, Florian Schmedding, and Ulf Leser. 2016. Recognizing chemicals in patents: a comparative analysis. *Journal of cheminformatics*, 8(1):59.

Mengke Hu, David Cinciruk, and John MacLaren Walsh. 2016. Improving automated patent claim parsing: Dataset, system, and experiments. *arXiv preprint arXiv:1605.01744*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

David M Jessop, Sam E Adams, Egon L Willighagen, Lezan Hawizy, and Peter Murray-Rust. 2011. OSCAR4: a flexible architecture for chemical textmining. *Journal of cheminformatics*, 3(1):41.

Roman Klinger, Corinna Kolářik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M Friedrich. 2008. Detection of iupac and iupac-like chemical names. *Bioinformatics*, 24(13):i268–i276.

Martin Krallinger, Obdulia Rabal, Analia Lourenço, Martin Perez Perez, Gael Perez Rodriguez, Miguel Vazquez, Florian Leitner, Julen Oyarzabal, and Alfonso Valencia. 2015. Overview of the CHEMDNER patents task. In *Proceedings of the Fifth BioCreative challenge evaluation workshop*, pages 63–75.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2015. tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, 7(1):S3.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436.

Mihai Lupu, Katja Mayer, John Tait, and Anthony J Trippe. 2011. *Current challenges in patent information retrieval*, volume 29. Springer.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.

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.

Thomas Morton, Joern Kottmann, Jason Baldridge, and Gann Bierner. 2005. Opennlp: A java-based nlp toolkit. In *Proceeding of the 10th Conference of the European Chapter of the Association of Computational Linguistics*.

Sorel Muresan, Plamen Petrov, Christopher Southan, Magnus J Kjellberg, Thierry Kogej, Christian Tyrchan, Peter Varkonyi, and Paul Hongxing Xie. 2011. Making every sar point count: the development of chemistry connect for the large-scale integration of structure and bioactivity data. *Drug Discovery Today*, 16(23-24):1019–1030.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan*, pages 39–43.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2017a. Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks. *arXiv preprint*, arXiv:1707.06799.

Nils Reimers and Iryna Gurevych. 2017b. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348.

Tim Rocktäschel, Michael Weidlich, and Ulf Leser. 2012. Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. 2018. "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science*, 9(28):6091–6098.

Marwin HS Segler, Mike Preuss, and Mark P Waller. 2018. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604.

Stefan Senger, Luca Bartek, George Papadatos, and Anna Gaulton. 2015. Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents. *Journal of cheminformatics*, 7(1):49.

Max E Valentinuzzi. 2017. Patents and scientific papers: Quite different concepts: The reward is found in giving, not in keeping [retrospectroscope]. *IEEE Pulse*, 8(1):49–53.

S Verberne, EKL D'hondt, NHJ Oostdijk, and CHA Koster. 2010. Quantifying the challenges in parsing patent claims. In *Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval at ECIR 2010*, pages 14–21.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database*, 2016.

Zenan Zhai, Dat Quoc Nguyen, and Karin Verspoor. 2018. Comparing cnn and lstm character-level embeddings in bilstm-crf models for chemical and disease named entity recognition. In *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis (LOUHI 2018)*, pages 38–43.

Yaoyun Zhang, Jun Xu, Hui Chen, Jingqi Wang, Yonghui Wu, Manu Prakasam, and Hua Xu. 2016. Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning. *Database*, 2016.