Learning Orthographic Features in Bi-directional LSTM for Biomedical Named Entity Recognition

Nut Limsopatham and **Nigel Collier**
Language Technology Lab
Department of Theoretical and Applied Linguistics
University of Cambridge
Cambridge, UK
{nl347,nhc30}@cam.ac.uk

Abstract

End-to-end neural network models for named entity recognition (NER) have shown to achieve effective performances on general domain datasets (e.g. newswire), without requiring additional hand-crafted features. However, in biomedical domain, recent studies have shown that hand-engineered features (e.g. orthographic features) should be used to attain effective performance, due to the complexity of biomedical terminology (e.g. the use of acronyms and complex gene names). In this work, we propose a novel approach that allows a neural network model based on a long short-term memory (LSTM) to automatically learn orthographic features and incorporate them into a model for biomedical NER. Importantly, our bi-directional LSTM model learns and leverages orthographic features on an end-to-end basis. We evaluate our approach by comparing against existing neural network models for NER using three well-established biomedical datasets. Our experimental results show that the proposed approach consistently outperforms these strong baselines across all of the three datasets.

1 Introduction

Named entity recognition (NER) is one of the first and important stages in a natural language processing (NLP) pipeline. In particular, an NER task is to identify mentions of entities (e.g. persons, locations and organisations) within unstructured text. In biomedical domain, NER tasks are particularly difficult, since the entities of interests are mainly genes, proteins, and chemical substances, which by nature (1) consist of millions of entities, (2) are created continuously, and (3) are non-standardised and can be referred to using different names (e.g. the use of acronyms and polysemy) (Kim et al., 2009; Kim et al., 2004; Smith et al., 2008a).

Traditionally, most of the effective NER approaches are based on machine learning techniques, such as conditional random field (CRF), support vector machine (SVM) and perceptrons (Lafferty et al., 2001; McCallum and Li, 2003; Settles, 2004; Luo et al., 2015; Ju et al., 2011; Ratinov and Roth, 2009; Segura-Bedmar et al., 2015). For instance, Ratinov and Roth (2009) effectively learned a perceptron model using features, including word classes induced using Brown clustering (Liang, 2005), and gazetteer extracted from Wikipedia. Campos et al. (2013) achieved effective performances for several biomedical NER tasks by learning a CRF model using multiple sets of features, including orthographic, morphological, linguistic-based, conjunctions and dictionary-based. However, these approaches rely heavily on feature engineering and domain knowledge (e.g. gazetteers), which are costly to develop. Consequently, they are difficult to be adapted to a new domain, since hand-engineered features are mostly specific to a target domain.

Recent advances in word vector representation (i.e. word embeddings) (Mikolov et al., 2013; Pennington et al., 2014), which represents a word in the form of a low-dimensional vector of real values, allow machine learning approaches to exploit semantic and syntactic information from word vectors, induced

Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016), pages 10–19, Osaka, Japan, December 12th 2016.

from a large dataset, for several NLP tasks, such as NER, part-of-speech (POS) tagging, sentiment analysis and concept normalisation (Collobert et al., 2011; Turian et al., 2010; Limsopatham and Collier, 2016a; Limsopatham and Collier, 2016b; Limsopatham and Collier, 2015). For example, Collobert et al. (2011) effectively used word embeddings as inputs of a feed-forward neural network for sequence labelling tasks, such as NER and POS tagging. Turian et al. (2010) learned a CRF model using word embeddings as input features for NER and chunking tasks. In the biomedical domain, Chiu et al. (2016) investigated the the use of different word embeddings in a feed-forward neural network for biomedical NER tasks. However, when using with word embedding features, traditional features (e.g. orthography and gazetteers) have shown to further improve the performance of an NER system (Segura-Bedmar et al., 2015; Turian et al., 2010; Huang et al., 2015).

In this work, we investigate a novel approach that allows an end-to-end neural network system for biomedical NER to explicitly learn and leverage orthographic features. Our approach is based on bi-directional long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) that learns to identify named entities in a sentence using both word and character embeddings as inputs. In particular, for each input sentence, we propose to generate and feed *an orthographic sentence* into a bi-directional LSTM to enable the model to explicitly learn orthographic features. We evaluate our proposed approach using three different well-established biomedical test collections, including the BioCreative II Gene Mention task corpus (BC2) (Smith et al., 2008b), the BioNLP 2009 shared task on event extraction (BioNLP09) (Kim et al., 2009) and the NCBI disease corpus (NCBI) (Doğan et al., 2014). Our experimental results show that the proposed approach consistently outperforms existing effective baselines in term of the f1-score measure.

The main contributions of this paper are three-folds:

1. We investigate the use of both word and character embeddings in bi-directional LSTM for biomedical NER tasks.

2. We propose a novel approach that enables bi-directional LSTM to automatically learn and leverage orthographic features without requiring feature engineering.

3. We thoroughly evaluate our proposed approach using three different standardised datasets for biomedical NER.

The remainder of this paper is organised as follows. In Section 2, we discuss related work and position our paper in the literature. In Section 3, we introduce our approach to learn and leverage orthographic features in bi-directional LSTM for biomedical NER. In Sections 4 and 5, we describe our experimental setup and empirically evaluate our approach, respectively. Section 6 provides concluding remarks.

2 Related Work

Biomedical NER, which aims to identify chunks of text mentioning specific entities of interest, is one of the fundamental biomedical text mining tasks. Due to the rapid growth of the number of biomedical documents, an automatic text mining system is needed to extract knowledge from the vast amount of data. Different from a general domain (e.g. newswire) where entities of interest are mainly places, persons and organisations (Tjong Kim Sang and De Meulder, 2003), entities that biomedical NER tasks focus on are, for example, genes, proteins, DNA and RNA. Existing studies (e.g. (Zhou et al., 2004; Fukuda et al., 1998; Liu et al., 2002)) showed that unique characteristics of biomedical text made NER a challenging task, such that existing NER approaches used in a general domain might not be effective. For example, Zhou et al. (2004) found that the names of many of biomedical entities were typically long (i.e. containing at least four words). In addition, the use of non-standardised naming conventions and abbreviation poses a significant challenge in biomedical NER (Smith et al., 2008a). For instance, 'cholesterol' can also be referred as '(3)-cholest-5-en-3-ol', '(3beta)-cholest-5-en-3-ol', '(3b)-cholest-5-en-3-ol', '5-Cholesten-3beta-ol' or '5-Cholesten-3b-ol'.

Machine learning-based approaches for NER have shown to achieve state-of-the-art performances for both general and biomedical domains. Conditional random field (CRF) is one of the most effective

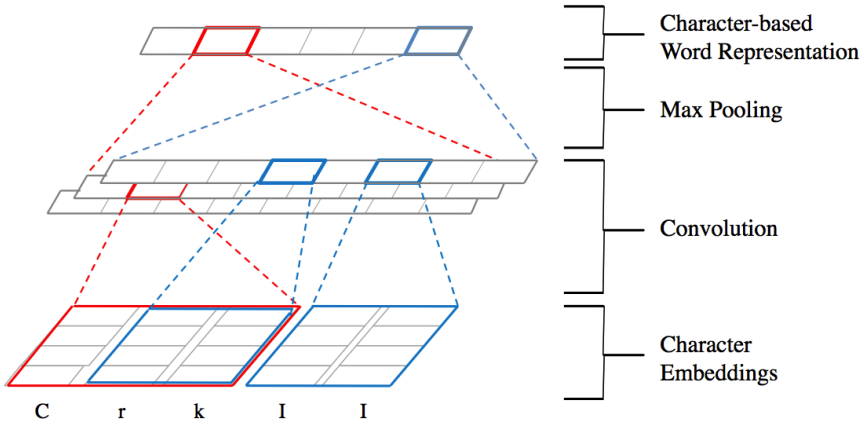

Figure 1: Our CNN architecture for learning word representation from character embeddings.

approaches used in NER tasks (Lafferty et al., 2001; McCallum and Li, 2003; Settles, 2004). Specifically, CRF is based on an undirected statistical graphical model that aims to learn a latent structure of an input sequence. Examples of effective biomedical NER tools that are based on CRF are ABNER (Settles, 2005), BANNER (Leaman et al., 2008) and Gimli (Campos et al., 2013). However, the performance of these CRF-based tools heavily depend on hand-crafted features, such as orthographic and contextual features (Bikel et al., 1999; Collier et al., 2000), which are task-specific and costly to develop. For example, Segura-Bedmar et al. (2015) manually created orthographic features, such as upperInitial (i.e. whether a given word begins with an upper-case character and then follows by any lower-case characters) and allCaps (i.e. whether all characters in a given word are upper-case), when learning a CRF model for drug name recognition. In this work, we investigate an automatic approach that could automatically induce orthographic features for biomedical named entity recognition.

Recently, neural network-based approaches have been effectively used for NER tasks. For example, Collobert et al. (2011) used a feed-forward neural network to effectively identify entities in a newswire corpus (Tjong Kim Sang and De Meulder, 2003) by classifying each word using contexts within a fixed number of surrounding words. Ma and Hovy (2016) and Lample et al. (2016) effectively used both character and word embeddings in a bi-directional LSTM for NER tasks, such as CoNLL03 (Tjong Kim Sang and De Meulder, 2003). Huang et al. (2015) combined hand-crafted features with bi-directional LSTM to further improve the performance. Chiu and Nichols (2016) achieved state-of-the-art performances by modelling both character and word embeddings before combining with hand-crafted features. Nevertheless, the studies of neural network models for biomedical NER tasks are limited. For instance, Chiu et al. (2016) investigated the use of the model of Collobert et al. (2011) with different word embeddings for the BioCreative II Gene Mention task (Smith et al., 2008b) and the JNLPBA task (Kim et al., 2004). In this work, we propose a novel end-to-end neural network model that can learn and leverage orthographic features, which are traditional domain-knowledge features widely used for NER tasks, without requiring any feature engineering.

3 Learning Orthographic Features in Bi-directional LSTM

In this section, we introduce our neural network architecture based on bi-directional LSTM for learning and leveraging orthographic features. In particular, our bi-directional LSTM model is composed of (1) character-based word representation, which induces a representation of a word from a character level using a convolutional neural network (CNN) (Section 3.1), (2) word representation, where any pre-trained word embeddings can be used (Section 3.2) and (3) bi-directional LSTM that learns to induce and leverage orthographic features when identifying named entities (Section 3.3).

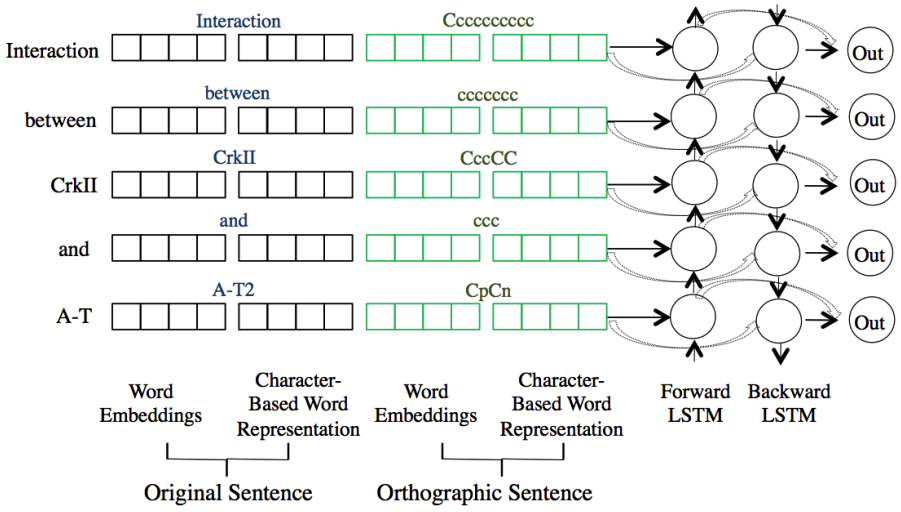

Figure 2: Our bi-directional LSTM for named entity recognition.

3.1 Character-based Word Representation

To learn a word representation from a character level, we use CNN to extract important features from character embeddings of a given word, as shown in Figure 1. In particular, we firstly represent a given word of length l characters (padded where necessary) using a word matrix $\mathbf{M} \in \mathbb{R}^{d \times l}$:

$$\mathbf{M} = \begin{bmatrix} | & | & | & | & \\ \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & ... & \mathbf{x}_l \\ | & | & | & | & \end{bmatrix} \tag{1}$$

where each column of $\mathbf{M}$ is the d-dimensional vector (i.e. character embedding) $\mathbf{x}_i \in \mathbb{R}^d$ of each character in the given word, which are initialised randomly.

Next, we apply a convolution operation using a filter $\mathbf{w} \in \mathbb{R}^{d \times h}$ to a window of h characters. The filter $\mathbf{w}$ is convolved over the sequence of characters in the word matrix $\mathbf{M}$ to create a feature matrix $\mathbf{C}$. Indeed, each feature c_i in $\mathbf{C}$ is extracted from a window of words $\mathbf{x}_{i:i+h-1}$, as follow:

$$c_i = f(\mathbf{w} \cdot \mathbf{x}_{i:i+h-1} + b) \tag{2}$$

where f is an activation function (such as tanh) and $b \in \mathbb{R}$ is a bias. Note that multiple filters can be used to extract multiple features. In this work, we use 200 filters, each of which has window size $h = 3$.

This convolution operation enables the learning of patterns of characters in words. In order to capture the most important features, max pooling (Collobert et al., 2011) is applied to take the maximum value of each row in the matrix $\mathbf{C}$:

$$\mathbf{c}_{max} = \begin{bmatrix} max(\mathbf{C}_{1,:}) \\ \vdots \\ max(\mathbf{C}_{d,:}) \end{bmatrix} \tag{3}$$

The $\mathbf{c}_{max}$ vector will later be used as a character-based word representation in bi-directional LSTM, since it captures important features of a given word.

3.2 Word Representation

We also use pre-trained word embeddings as inputs of bi-directional LSTM, since existing work (e.g. (Mikolov et al., 2013; Pyysalo et al., 2013; Pennington et al., 2014)) has shown that these embeddings could capture semantic and syntactic information of words.

Input Sentence	Orthographic Sentence
interaction between CrkII and A-T2	cccccccccc ccccccc CccCC ccc CpCn
Prognosis of asymptomatic multiple myeloma.	Ccccccccc cc ccccccccccc cccccccc cccccccp
activation of 3-hydroxy-3-methylglutaryl	ccccccccc cc npccccccpnpccccccccccc
Modification of dopamine D2 receptor activity	Cccccccccccc cc cccccccc Cn cccccccc cccccccc
G alpha i2 and G alpha i2	C ccccc cn ccc C ccccc cn
TPA induction of FGF-BP gene	CCC ccccccccc cc CCCpCC cccc
KAP-1 mediated repression in vivo	CCCpn cccccccc ccccccccc cc cccc

Table 1: Examples of biomedical sentences and their corresponding orthographic sentence.

	BC2	BioNLP09	NCBI
Target entities	Genes	Bio-molecular events	Diseases
Type of data	MEDLINE abstracts	MEDLINE abstracts	PubMed articles
Number of documents for training	201	1,436	8,662
Number of documents for development	488	995	2,872
Number of documents for testing	58	2,200	1,036

Table 2: The three datasets used to evaluate our proposed approach.

3.3 Bi-directional LSTM

We use bi-directional LSTM to learn to identify named entities in a sentence, because it can capture past (from the previous words) and future (from the next words) information effectively (Huang et al., 2015; Dyer et al., 2015). In addition, LSTM has shown to capture long-distance dependencies more effectively than a vanilla recurrent neural networks (RNNs), since it can cope with the gradient vanishing/exploding problems better (Dyer et al., 2015; Bengio et al., 1994).

To enable bi-directional LSTM to learn orthographic features, we create an orthographic pattern of the input sentence (denoted, *the orthographic sentence*). Specifically, given an input sentence (e.g. 'interaction between CrkII and A-T2'), we generate *an orthographic sentence* (e.g. 'cccccccccc ccccccc CccCC ccc CpCn') by using a set of simple rules, where each of the upper-case characters, lower-case characters, numbers and punctuations, are replaced with *C*, *c*, *n* and *p*, respectively. Examples of orthographic sentences are shown in Table 1. The orthographic sentence enables bi-directional LSTM to learn orthographic features automatically.

Next, as shown in Figure 2, given an input sentence and its orthographic sentence, we firstly extract both word embeddings (i.e. word representation) and character-based word representation corresponding to each word in the input sentence and the orthographic sentence, by using the approaches described in Sections 3.1 and 3.2[1]. Then, we concatenate word representations associated to the same words and sequentially feed them into bi-directional LSTM to model the contextual information of each word. Finally, at the output layer, we follow Huang et al. (2015) and optimise the CRF log-likelihood, which aims to maximise the likelihood of labelling the whole sentence correctly, by modelling the interactions between two successive labels using the Viterbi algorithm.

4 Experimental Setup

4.1 Datasets

To evaluate our proposed approach, we use three different well-estabished biomedical NER datasets, which are the BioCreative II Gene Mention task corpus (BC2) (Smith et al., 2008b), the BioNLP 2009 shared task on event extraction (BioNLP09) (Kim et al., 2009) and the NCBI disease corpus (NCBI) (Doğan et al., 2014), respectively. Table 2 shows the information of the three datasets. Firstly, the BC2 dataset consists of 20,000 sentences extracted from MEDLINE abstracts (15,000 sentences for

[1]Note that we use separated set of word and character embeddings for the input sentence and the orthographic sentence.

training and 5,000 sentences for testing), where the task is to annotate the mentions of genes. In order to create a development set, we randomly split the original 15,000 training sentences into 10,000 and 5,000 training and development sentences. Secondly, the BioNLP09 dataset is composed of 7,449, 1,450 and 2,447 sentences for training, development and testing, respectively. The target entities are bio-molecular events. Thirdly, the NCBI dataset contains more than 6,000 sentences from 793 PubMed articles (593, 100 and 100 articles for training, development and testing, respectively). The task aims to identify mentions of diseases in a given sentence.

4.2 Evaluation Measures

We evaluate the performance on the three biomedical NER tasks in terms of f1-score, precision and recall measures:

$$\textit{f1-score} = 2 \cdot \frac{precision \cdot recall}{precision + recall}, \tag{4}$$

$$precision = \frac{TP}{TP + FP}, \tag{5}$$

$$recall = \frac{TP}{TP + FN}, \tag{6}$$

where TP (true positive) is the number of named entity chunks that are correctly identified, FP (false positive) is the number of chunks that are mistakenly identified as entities, and FN (false negative) are the number of named entity chunks that are not identified.

4.3 Embeddings

4.3.1 Word Embeddings

As discussed in Section 3.2, our approach uses word embeddings as inputs when learning an NER model. We use pre-trained word embeddings of Moen et al. (2013), which are publicly available. In particular, the embeddings consists of 200-dimensional vectors of 5.4 million unique words, which are induced from a combined collection of PubMed, PMC and Wikipedia texts using the Skip-gram model from the word2vec tool (Mikolov et al., 2013). For the words that do not exist in the pre-trained embeddings, we use a vector of random values sampled from $[-\sqrt{\frac{3}{dim}}, +\sqrt{\frac{3}{dim}}]$ where dim is the dimension of embeddings as suggested by He et al. (2015).

We use a separated word embeddings for words in the orthographic sentences. In particular, for each word we use a 200-dimensional randomly generated vector, where each dimension is also uniformly sampled from $[-\sqrt{\frac{3}{dim}}, +\sqrt{\frac{3}{dim}}]$.

4.3.2 Character Embeddings

For both input sentence (i.e. original sentence) and orthographic sentence, we use 30-dimensional character embeddings for representing each character when inducing the character-based word representation (Equation (1) in Section 3.1). In particular, we initialise the character embeddings with uniform samples from $[-\sqrt{\frac{3}{dim}}, +\sqrt{\frac{3}{dim}}]$. Importantly, we have a separated embedding for each set of characters in the input and orthographic sentences.

4.4 Parameter Optimisation

Parameter optimisation is done by mini-batch stochastic gradient descent (SGD) with batch size 50. In particular, the stochastic gradient descent with back-propagation is performed using Adadelta update rule (Zeiler, 2012). Note that we also fine-tune both word and character embeddings by allowing their weights to be modified when performing gradient updates. To reduce the effects of gradient exploding, we follow Pascanu et al. (2013) and use a gradient clipping of 5.0.

To mitigate overfitting, we apply L_2 regularisation on the weight vectors, as well as applying dropout (Srivastava et al., 2014) with dropout rate 0.5 for all of the layers in our model. In addition, we use early stopping (Giles, 2001) based on the performance achieved on the development sets.

Approach	BC2			BioNLP09			NCBI		
	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall
FeedForward	66.13	76.43	58.28	76.83	77.25	76.42	73.55	72.05	75.12
BiLSTM	69.54	74.25	65.39	80.49	85.64	75.93	75.37	77.53	73.33
CNN-BiLSTM (Char-only)	79.98	81.85	78.20	85.11	87.54	82.81	82.70	83.00	82.40
CNN-BiLSTM	80.25	80.75	**79.76**	86.54	88.90	84.31	84.19	84.33	**84.06**
ORTH-CNN-BiLSTM	**80.58**	**83.01**	78.28	**87.06**	**88.91**	**85.29**	**84.26**	**86.67**	81.98

Table 3: Performances in terms of f1-score, precision and recall of our proposed approach and the baselines on the BC2, BioNLP09 and NCBI datasets.

4.5 Baselines

We compare our approach with four different baselines, which do not use any hand-engineered features:

1. *FeedForward*: A simple feed-forward neural network model similar to Collobert et al. (2011) with the context window size of 5 and the pre-trained word embeddings described in Section 4.3.1.

2. *BiLSTM*: A bi-directional LSTM model similar to the proposed model in Section 3, excepting that the orthographic sentence and the character-based word representation are discarded from the model. This baseline is similar to the model of Huang et al. (2015) when hand-crafted features are not taken into account.

3. *CNN-BiLSTM (Char-only)*: A bi-directional LSTM model similar to the proposed model in Section 3, excepting that the orthographic sentence and the word embeddings are discarded from the model.

4. *CNN-BiLSTM*: A bi-directional LSTM model similar to the model in Section 3, excepting that the orthographic sentence is not taken into account by the model.

5 Experimental Results

In this section, we compare the performance of our approach for learning and leveraging orthographic features in bi-directional LSTM for biomedical NER (denoted, *ORTH-CNN-BiLSTM*) against the four baselines introduced in Section 4.5. Table 3 compares the performances of our proposed approach with the baselines in terms of f1-score, precision and recall on the three datasets (i.e. BC2, BioNLP09 and NCBI).

From Table 3, we firstly observe that *FeedForward* is the weakest baseline, especially in terms of the f1-score. This is intuitive as feed-forward neural network is a simple model in comparison with bi-directional LSTM that could learn long-distance dependencies from sequences of words. Next, we compare the performance of *BiLSTM* and *CNN-BiLSTM (Char-only)*. Both *BiLSTM* and *CNN-BiLSTM (Char-only)* share a similar architecture for identifying named entities. The only difference is that *BiLSTM* uses pre-trained word embeddings for representing words in a sentence; meanwhile, *CNN-BiLSTM (Char-only)* learns word representation from character embeddings using a convolutional neural network. We observe that *CNN-BiLSTM (Char-only)* achieves better performances than *BiLSTM* in terms of all the three reported measures (i.e. f1-score, precision and recall), across the three datasets. This highlights the importance of the character-based word representation that could help to deal with non-standardised and continuously-growing biomedical vocabularies. Furthermore, we found that *CNN-BiLSTM*, which uses both pre-trained word embeddings and character-based word representation in a bi-directional LSTM model, further improves the f1-score and recall performances on all of the three datasets.

On the other hand, our approach, *ORTH-CNN-BiLSTM*, outperforms all of the baselines on the three datasets. In particular, *ORTH-CNN-BiLSTM* performs better than *CNN-BiLSTM*, which is the most effective baseline, in terms of f1-score and precision for all of the BC2, BioNLP09 and NCBI datasets. Importantly, we observe that our approach for automatically learning orthographic features could effectively boost the performance in term of precision. For example, for the BC2 and NCBI datasets,

ORTH-CNN-BiLSTM achieved 83.01% and 86.67% precision, while *CNN-BiLSTM* attains 80.75% and 84.33% precision, respectively.

When analysing the performance of *ORTH-CNN-BiLSTM*, we observe that the induced orthographic features could help to effectively identify complex biomedical entities, such as 'CrkII-23', 'ch-IAP1', 'HC-toxin', 'E.coli manX equivalent', 'cathepsin K', 'IL-2', and 'A-T', that do not appear in the training set by learning from the orthographic patterns of words. This shows the importance of orthographic features in biomedical NER tasks. Importantly, our approach shows a potential of enabling bi-directional LSTM to capture these patterns without resorting to hand-engineered features.

6 Conclusions

We have discussed recent advances in neural networks that could enable a machine learning-based NER system to performed effectively in a general domain, such as newswire, without requiring any hand-crafted features. However, the complexity and the continuous growth of biomedical vocabularies make biomedical NER a challenging task. Consequently, biomedical NER systems would require domain knowledge, in the forms of hand-crafted features, to achieve an effective performance. In this work, we investigate an approach that allows bi-directional LSTM to automatically learn and leverage orthographic features, which is one of the key features for biomedical NER. We evaluate our approach by comparing against existing effective end-to-end neural network models for NER. Our experimental results evaluated on three different well-established biomedical NER datasets showed that our approach consistently outperformed the baselines. Importantly, we found that our approach could help to identify named entities that did not appear in the training data by learning the orthographic patterns from similar entities. For future work, we aim to enable neural network models to automatically induce other hand-crafted features, such as gazetteers.

Acknowledgements

The authors wish to thank funding support from the EPSRC (grant number EP/M005089/1).

References

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what‘s in a name. *Mach. Learn.*, 34(1-3):211–231, February.

David Campos, Sérgio Matos, and José Luís Oliveira. 2013. Gimli: open source and high-performance biomedical name recognition. *BMC bioinformatics*, 14(1):1.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical nlp. In *Proceedings of BioNLP16*, page 166.

Nigel Collier, Chikashi Nobata, and Jun-ichi Tsujii. 2000. Extracting the names of genes and gene products with a hidden markov model. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, pages 201–207, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. *arXiv preprint arXiv:1505.08075*.

Ken-ichiro Fukuda, Tatsuhiko Tsunoda, Ayuchi Tamura, Toshihisa Takagi, et al. 1998. Toward information extraction: identifying protein names from biological papers. In *Pac symp biocomput*, volume 707, pages 707–718. Citeseer.

Rich Caruana Steve Lawrence Lee Giles. 2001. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, volume 13, page 402. MIT Press.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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

Zhenfei Ju, Jian Wang, and Fei Zhu. 2011. Named entity recognition from biomedical text using svm. In *Bioinformatics and Biomedical Engineering,(iCBBE) 2011 5th International Conference on*, pages 1–4. IEEE.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Nigel Collier. 2004. 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*, pages 70–75. Association for Computational Linguistics.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, pages 1–9. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, 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, San Diego, California, June. Association for Computational Linguistics.

Robert Leaman, Graciela Gonzalez, et al. 2008. Banner: an executable survey of advances in biomedical named entity recognition. In *Pacific symposium on biocomputing*, volume 13, pages 652–663.

Percy Liang. 2005. *Semi-supervised learning for natural language*. Ph.D. thesis, Massachusetts Institute of Technology.

Nut Limsopatham and Nigel Collier. 2015. Adapting phrase-based machine translation to normalise medical terms in social media messages. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1675–1680, Lisbon, Portugal, September. Association for Computational Linguistics.

Nut Limsopatham and Nigel Collier. 2016a. Modelling the combination of generic and target domain embeddings in a convolutional neural network for sentence classification. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics.

Nut Limsopatham and Nigel Collier. 2016b. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023, Berlin, Germany, August. Association for Computational Linguistics.

Hongfang Liu, Alan R Aronson, and Carol Friedman. 2002. A study of abbreviations in medline abstracts. In *Proceedings of the AMIA Symposium*, page 464. American Medical Informatics Association.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 879–888, Lisbon, Portugal, September. Association for Computational Linguistics.

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, Berlin, Germany, August. Association for Computational Linguistics.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. *ICML (3)*, 28:1310–1318.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

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.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pages 147–155. Association for Computational Linguistics.

Isabel Segura-Bedmar, Vıctor Suárez-Paniagua, and Paloma Martınez. 2015. Exploring word embedding for drug name recognition. In *SIXTH INTERNATIONAL WORKSHOP ON HEALTH TEXT MINING AND INFORMATION ANALYSIS (LOUHI)*, page 64.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107. Association for Computational Linguistics.

Burr Settles. 2005. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008a. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(2):1.

Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, et al. 2008b. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(2):1.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Guodong Zhou, Jie Zhang, Jian Su, Dan Shen, and Chewlim Tan. 2004. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190.