Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining

Proceedings of the Workshop

December 11-16, 2016
Osaka, Japan
Copyright of each paper stays with the respective authors (or their employers).

ISBN978-4-87974-719-8
Preface

Biomedical natural language processing has grown from its roots in clinical language processing and bioinformatics into a thriving research field of its own. The search (“natural language processing”) OR (“text mining”) performed in PubMed today returns 5,056 hits, versus 1,903 at the turn of the century and 3,485 in 2010. The papers appearing in this volume reflect the diversity of trends in biomedical natural language processing today—movement from English-language texts to clinical texts in other languages; exploration of social media in addition to clinical documents and traditional scientific publications; and processing of full-text articles, versus abstracts only. In addition to reflecting the diversity of the field, the papers in this volume also reflect the homogenisation of approaches that has characterised some recent approaches, with 5 out of 16 papers involving some combination of neural networks and/or distributional semantics. The organisers thank the authors for sharing their science with this community, and the programme committee (listed elsewhere in this volume) for their contribution to maintaining the high standards of the BioTxtM series of meetings.

Keynote Talk by Dr. Makoto Miwa

Learning for Information Extraction in Biomedical and General Domains

Information extraction (IE) has been widely studied in various domains since IE is a key to bridge the gap between knowledge and texts. IE includes several core sub-problems, such as named entity recognition, relation extraction, and event extraction, and these sub-problems have been tackled using machine learning techniques. In this talk, I will give an overview of learning approaches for IE in biomedical and general domain, especially on corpus-based classification and structured learning approaches. I will then introduce recent deep learning approaches including our recent recurrent neural network (RNN)-based approach, and discuss the limitations and future directions.

Speaker Biography

Makoto Miwa is an associate professor of Toyota Technological Institute (TTI). He received his Ph.D. from the University of Tokyo in 2008. His research mainly focuses on information extraction from texts, deep learning, and representation learning. His projects include AkaneRE, EventMine, PathText and LSTM-ER.
Organisers

Sophia Ananiadou, National Centre for Text Mining, University of Manchester UK
Riza Batista-Navarro, National Centre for Text Mining, University of Manchester UK
Kevin Bretonnel Cohen, Computational Bioscience Program, University of Colorado School of Medicine, USA
Dina Demner-Fushman, National Library of Medicine, USA
Paul Thompson, National Centre for Text Mining, University of Manchester, UK

Programme Committee

Eiji Aramaki, Nara Institute of Science and Technology (NAIST), Japan
Hercules Dalianis, Stockholm University, Sweden
Graciela Gonzalez, Arizona State University, USA
Wen-Lian Hsu, Academia Sinica, Taipei, Taiwan
Rezarta Islamaj, NCBI/NLM/NIH, USA
Roman Klinger, University of Stuttgart, Germany
Robert Leaman NCBI/NLM/NIH, USA
Shervin Malmasi, Harvard Medical School, USA
Makoto Miwa, Toyota Technological Institute, Japan
Sung-Hyon Myaeng, Korea Advanced Institute of Science and Technology (KAIST), Korea
Claire Nedellec, French National Institute of Agronomy (INRA), France
Naoaki Okazaki, Tohoku University, Japan
Arzucan Özgür, Bogazici University, Turkey
Martha Palmer, University of Colorado at Boulder, USA
Stelios Piperidis, Institute for Language and Speech Processing, Greece
Guergana Savova, Boston Children’s Hospital and Harvard Medical School, USA
Hagit Shatkay, University of Delaware, USA
Mark Stevenson, University of Sheffield, UK
Yoshimasa Tsuruoka, University of Tokyo, Japan
Lucy Vanderwende, Microsoft, USA
Karin Verspoor, University of Melbourne, Australia
Stephen Wu, Oregon Health & Science University, USA
Yan Xu, Microsoft Research Asia, China
Pierre Zweigenbaum, Laboratoire d’Informatique pour la Mécanique et les Sciences de l’Ingénieur (LIMSI), France
# Table of Contents

*Cancer Hallmark Text Classification Using Convolutional Neural Networks*
Simon Baker, Anna Korhonen and Sampo Pyysalo ........................................ 1

*Learning Orthographic Features in Bi-directional LSTM for Biomedical Named Entity Recognition*
Nut Limspatham and Nigel Collier ............................................................. 10

*Entity Extraction in Biomedical Corpora: An Approach to Evaluate Word Embedding Features with PSO based Feature Selection*
Shweta Yadav, Asif Ekbal and Sriparna Saha ........................................... 20

*Building Content-driven Entity Networks for Scarce Scientific Literature using Content Information*
Reinald Kim Amplayo and Min Song ....................................................... 31

*Named Entity Recognition in Swedish Health Records with Character-Based Deep Bidirectional LSTMs*
Simon Almgren, Sean Pavlov and Olof Mogren ........................................ 41

*Entity-Supported Summarization of Biomedical Abstracts*
Schulze Frederik and Mariana Neves ....................................................... 51

*Fully unsupervised low-dimensional representation of adverse drug reaction events through distributional semantics*
Alicia Pérez, Arantza Casillas and Koldo Gojenola ................................... 61

*A Dataset for ICD-10 Coding of Death Certificates: Creation and Usage*
Thomas Lavergne, Aurelie Neveol, Aude Robert, Cyril Grouin, Grégoire Rey and Pierre Zweigenbaum ................................................................. 71

*A Corpus of Tables in Full-Text Biomedical Research Publications*
Tatyana Shmanina, Ingrid Zukerman, Ai Lee Cheam, Thomas Bochynek and Lawrence Cavedon 81

*Supervised classification of end-of-lines in clinical text with no manual annotation*
Pierre Zweigenbaum, Cyril Grouin and Thomas Lavergne .......................... 91

*BioDCA Identifier: A System for Automatic Identification of Discourse Connective and Arguments from Biomedical Text*
Sindhuja Gopalan and Sobha Lalitha Devi ................................................ 100

*Data, tools and resources for mining social media drug chatter*
Abeed Sarker and Graciela Gonzalez ...................................................... 110

*Detection of Text Reuse in French Medical Corpora*
Eva D’hondt, Cyril Grouin, Aurelie Neveol, Efstatios Stamatas and Pierre Zweigenbaum 119

*Negation Detection in Clinical Reports Written in German*
Viviana Cotik, Roland Roller, Feiyu Xu, Hans Uszkoreit, Klemens Budde and Danilo Schmidt 126

*Scoring Disease-Medication Associations using Advanced NLP, Machine Learning, and Multiple Content Sources*
Bharath Dandala, Murthy Devarakonda, Mihaela Bornea and Christopher Nielson 136

*Author Name Disambiguation in MEDLINE Based on Journal Descriptors and Semantic Types*
Dina Vishnyakova, Raul Rodriguez-Esteban, Khan Ozol and Fabio Rinaldi 145
Conference Program

12th December 2016

9:00–9:10  Welcome remarks

9:10–10:30  Session 1

9:10–9:30  Cancer Hallmark Text Classification Using Convolutional Neural Networks
Simon Baker, Anna Korhonen and Sampo Pyysalo

9:30–9:50  Learning Orthographic Features in Bi-directional LSTM for Biomedical Named Entity Recognition
Nut Limspopatham and Nigel Collier

9:50–10:00  Entity Extraction in Biomedical Corpora: An Approach to Evaluate Word Embedding Features with PSO based Feature Selection
Shweta Yadav, Asif Ekbal and Sriparna Saha

10:00–10:10  Building Content-driven Entity Networks for Scarce Scientific Literature using Content Information
Reinald Kim Amplayo and Min Song

10:10–10:20  Named Entity Recognition in Swedish Health Records with Character-Based Deep Bidirectional LSTMs
Simon Almgren, Sean Pavlov and Olof Mogren

10:20–10:30  Entity-Supported Summarization of Biomedical Abstracts
Schulze Frederik and Mariana Neves
12th December 2016 (continued)

10:30–10:50  Coffee break and Poster Session 1

10:50–12:00  Session 2

10:50–11:10  Fully unsupervised low-dimensional representation of adverse drug reaction events through distributional semantics
Alicia Pérez, Arantza Casillas and Koldo Gojenola

11:10–12:00  Keynote Talk: Learning for Information Extraction in Biomedical and General Domains
Dr. Makoto Miwa

12:00–14:00  Lunch break

14:00–15:20  Session 3

14:00–14:20  A Dataset for ICD-10 Coding of Death Certificates: Creation and Usage
Thomas Lavergne, Aurelie Neveol, Aude Robert, Cyril Grouin, Grégoire Rey and Pierre Zweigenbaum

14:20–14:40  A Corpus of Tables in Full-Text Biomedical Research Publications
Tatyana Shmanina, Ingrid Zukerman, Ai Lee Cheam, Thomas Bochynek and Lawrence Cavedon

14:40–14:50  Supervised classification of end-of-lines in clinical text with no manual annotation
Pierre Zweigenbaum, Cyril Grouin and Thomas Lavergne

14:50–15:00  BioDCA Identifier: A System for Automatic Identification of Discourse Connective and Arguments from Biomedical Text
Sindhuja Gopalan and Sobha Lalitha Devi

15:00–15:10  Data, tools and resources for mining social media drug chatter
Abeed Sarker and Graciela Gonzalez

15:10–15:20  Detection of Text Reuse in French Medical Corpora
Eva D’hondt, Cyril Grouin, Aurelie Neveol, Efstatios Stamatatos and Pierre Zweigenbaum
12th December 2016 (continued)

15:20–15:50  Coffee break and Poster Session 2

15:50–16:50  Session 4

15:50–16:10  Negation Detection in Clinical Reports Written in German  
Viviana Cotik, Roland Roller, Feiyu Xu, Hans Uszkoreit, Klemens Budde and Danilo Schmidt

16:10–16:30  Scoring Disease-Medication Associations using Advanced NLP, Machine Learning, and Multiple Content Sources  
Bharath Dandala, Murthy Devarakonda, Mihaela Bornea and Christopher Nielson

16:30–16:50  Author Name Disambiguation in MEDLINE Based on Journal Descriptors and Semantic Types  
Dina Vishnyakova, Raul Rodriguez-Esteban, Khan Ozol and Fabio Rinaldi

16:50–17:00  Closing remarks
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...
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 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) (Do˘gan 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
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 pretrained 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).
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 \( M \in \mathbb{R}^{d \times l} \):

\[
M = \begin{bmatrix}
    x_1 \\
    x_2 \\
    x_3 \\
    \vdots \\
    x_l
\end{bmatrix}
\]  

(1)

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

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

\[
c_i = f(w \cdot x_{i:i+h-1} + b)
\]  

(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 \( C \):

\[
c_{max} = \begin{bmatrix}
    \text{max}(C_{1,:}) \\
    \vdots \\
    \text{max}(C_{d,:})
\end{bmatrix}
\]  

(3)

The \( c_{max} \) 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.
Table 1: Examples of biomedical sentences and their corresponding orthographic sentence.

<table>
<thead>
<tr>
<th>Input Sentence</th>
<th>Orthographic Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>interaction between CrkII and A-T2</td>
<td>ccccccccc ccccccc CccCC ccc CpCn</td>
</tr>
<tr>
<td>Prognosis of asymptomatic multiple myeloma. activation of 3-hydroxy-3-methylglutaryl Modification of dopamine D2 receptor activity G alpha i2 and G alpha i2 TPA induction of FGF-BP gene KAP-1 mediated repression in vivo</td>
<td>Ccccccccc cc ccccccccccecccccecccccep ccccccccc cc npcccccccpnpccccccccccccc</td>
</tr>
</tbody>
</table>

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

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

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. ‘cccccccccccccccccccc CcCcCcC 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. 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-established 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...
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:

$$f1\text{-}score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

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

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

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

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 terms 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


Author Index

Almgren, Simon, 41
Amplayo, Reinald Kim, 31
Baker, Simon, 1
Bochynek, Thomas, 81
Bornea, Mihaela, 136
Budde, Klemens, 126
Casillas, Arantza, 61
Cavedon, Lawrence, 81
Cheam, Ai Lee, 81
Collier, Nigel, 10
Cotik, Viviana, 126
D’hondt, Eva, 119
Dandala, Bharath, 136
Devarakonda, Murthy, 136
Ekbal, Asif, 20
Frederik, Schulze, 51
Gojenola, Koldo, 61
Gonzalez, Graciela, 110
Gopalan, Sindhuja, 100
Grouin, Cyril, 71, 91, 119
Korhonen, Anna, 1
Lalitha Devi, Sobha, 100
Lavergne, Thomas, 71, 91
Limsopatham, Nut, 10
Mogren, Olof, 41
Neveol, Aurelie, 71, 119
Neves, Mariana, 51
Nielson, Christopher, 136
Ozol, Khan, 145
Pavlov, Sean, 41
Pérez, Alicia, 61
Pyysalo, Sampo, 1
Rey, Grégoire, 71
Rinaldi, Fabio, 145
Robert, Aude, 71
Rodriguez-Esteban, Raul, 145
Roller, Roland, 126
Saha, Sriparna, 20
Sarker, Abeed, 110
Schmidt, Danilo, 126
Shmanina, Tatyana, 81
Song, Min, 31
Stamatatos, Efstatios, 119
Uszkoreit, Hans, 126
Vishnyakova, Dina, 145
Xu, Feiyu, 126
Yadav, Shweta, 20
Zukerman, Ingrid, 81
Zweigenbaum, Pierre, 71, 91, 119