An Investigation of Recurrent Neural Architectures
for Drug Name Recognition

Anonymous EMNLP submission

Abstract

Drug name recognition (DNR) is an essential step in the Pharmacovigilance (PV) pipeline. DNR aims to find drug name mentions in unstructured biomedical texts and classify them into predefined categories. State-of-the-art DNR approaches heavily rely on hand-crafted features and domain-specific resources which are difficult to collect and tune. For this reason, this paper investigates the effectiveness of contemporary recurrent neural architectures - the Elman and Jordan networks and the bidirectional LSTM with CRF decoding - at performing DNR straight from the text. The experimental results achieved on the authoritative SemEval-2013 Task 9.1 benchmarks show that the bidirectional LSTM-CRF ranks closely to highly-dedicated, hand-crafted systems.

1 Introduction

Pharmacovigilance (PV) is defined by the World Health Organization as the science and activities concerned with the detection, assessment, understanding and prevention of adverse effects of drugs or any other drug-related problems. Drug name recognition (DNR) is a fundamental step in the PV pipeline, similarly to the well-studied Named Entity Recognition (NER) task for general natural language processing (NLP). DNR aims to find drug mentions in unstructured biomedical texts and classify them into predefined categories in order to link drug names with their effects and explore drug-drug interactions (DDIs). Conventional approaches to DNR sub-divide as rule-based, dictionary-based and machine learning-based. Intrinsically, rule-based systems are hard to scale, time-consuming to assemble and ineffective in the presence of informal sentences and abbreviated phrases. Dictionary-based systems identify drug names by matching text chunks against drug dictionaries. These systems typically achieve high precision, but suffer from low recall (i.e., they miss a significant number of mentions) due to spelling errors or drug name variants not present in the dictionaries (Liu et al., 2015a). Conversely, machine-learning approaches have the potential to overcome all these limitations since their foundations are intrinsically robust to variants. The current state-of-the-art machine learning approaches follow a two-step process of feature engineering and classification (Segura-Bedmar et al., 2015; Abacha et al., 2015; Rocktäschel et al., 2013). Feature engineering refers to the task of representing text by dedicated numeric vectors using domain knowledge. Similarly to the design of rule-based systems, this task requires much expert knowledge, is typically challenging and time-consuming, and has a major impact on the final accuracy. For this reason, this paper explores the performance of contemporary recurrent neural networks (RNNs) at providing end-to-end DNR straight from text, without any manual feature engineering stage. The tested RNNs include the popular Elman and Jordan networks and the bidirectional long short-term memory (LSTM) with decoding provided by a conditional random field (CRF) (Elman, 1990; Jordan, 1986; Lample et al., 2016; Collobert et al., 2011). The experimental results over the SemEval-2013 Task 9.1 benchmarks show an interesting accuracy from the LSTM-CRF

that exceeds that of various manually-engineered systems and approximates the best result in the literature.

2 Related Work

Most of the research on drug name recognition to date has focussed on domain-dependent aspects and specialized text features. The benefit of leveraging such tailored features was made evident by the results from the SemEval-2013 Task 9.1 (Recognition and classification of pharmacological substances, known as DNR task) challenge. The system that ranked first, WBI-NER (Rocktäschel et al., 2013), adopted very specialized features derived from an improved version of the ChemSpot tool (Rocktäschel et al., 2012), a collection of drug dictionaries and ontologies. Similarly, many other recent approaches (Abacha et al., 2015; Liu et al., 2015b; Segura-Bedmar et al., 2015) have been based on various combinations of general and domain-specific features. In the broader field of machine learning, the recent years have witnessed a rapid proliferation of deep neural networks, with unprecedented results in tasks as diverse as visual, speech and named-entity recognition (Hinton et al., 2012; Krizhevsky et al., 2012; Lample et al., 2016). One of the main advantages of neural networks is that they can learn the feature representations automatically from the data, thus avoiding the laborious feature engineering stage (Mesnil et al., 2015; Lample et al., 2016). Given these promising results, the main goal of this paper is to provide the first performance investigation of popular RNNs such as the Elman and Jordan networks and the bidirectional LSTM-CRF over DNR tasks.

3 The Proposed Approach

DNR can be formulated as a joint segmentation and classification task over a predefined set of classes. As an example, consider the input sentence provided in Table 1. The notation follows the widely adopted in/out/begin (IOB) entity representation with, in this instance, *Cimetidine* as the drug, *ALFENTA* as the brand, and words *volatile inhalation anesthetics* together as the group. In this paper, we approach the DNR task by recurrent neural networks and we therefore provide a brief description hereafter. In an RNN, each word in the input sentence is first mapped to a random real-valued vector of arbitrary dimension, d. Then, a measurement for the word, noted as $x(t)$, is formed by concatenating the word's own vector with a window of preceding and following vectors (the "context"). An example of input vector with a context window of size $s = 3$ is:

$$w_3(t) = [Cimetidine, \textbf{reduces}, effect],$$
$$'reduces' \rightarrow x_{reduces} \in \mathbb{R}^d,$$
$$'Cimetidine' \rightarrow x_{Cimetidine} \in \mathbb{R}^d, \quad (1)$$
$$'effect' \rightarrow x_{effect} \in \mathbb{R}^d,$$
$$x(t) = [x_{Cimetidine}, x_{\textbf{reduces}}, x_{effect}] \in \mathbb{R}^{3d}$$

where $w_3(t)$ is the context window centered around the t-th word, $'reduces'$, and x_{word} represents the numerical vector for $word$.

For the Elman network, both $x(t)$ and the output from the hidden layer at time $t - 1$, $h(t - 1)$, are input into the hidden layer for frame t. The recurrent connection from the past time frame enables a short-term memory, while hidden-to-hidden neuron connections make the network Turing-complete. This architecture, common in RNNs, is suitable for prediction of sequences. Formally, the hidden layer is described as:

$$h(t) = f(U \bullet x(t) + V \bullet h(t - 1)) \quad (2)$$

where U and V are randomly-initialized weight matrices between the input and the hidden layer, and between the past and current hidden layers, respectively. Function $f(\cdot)$ is the sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

that adds non-linearity to the layer. Eventually, $h(t)$ is input in the output layer:

$$y(t) = g(W \bullet h(t)), \text{ with } g(z_m) = \frac{e^{z_m}}{\Sigma_{k=1}^{K} e^{z_k}} \quad (4)$$

and convolved with the output weight matrix, W. The output is normalized by a multi-class logistic function, $g(\cdot)$, to become a proper probability over the class set. The output dimensionality is therefore

Sentence	Cimetidine	reduces	clearance	of	ALFENTA	and	volatile	inhalation	anesthetics
Entity class	B-drug	O	O	O	B-brand	O	B-group	I-group	I-group

Table 1: Example sentence in a DNR task with entity classes represented in IOB format.

	DDI-DrugBank		DDI-MedLine	
	Training+Test for DDI task	Test for DNR	Training+Test for DDI task	Test for DNR
documents	730	54	175	58
sentences	6577	145	1627	520
drug_n	124	6	520	115
group	3832	65	234	90
brand	1770	53	36	6
drug	9715	180	1574	171

Table 2: Statistics of training and test datasets used for SemEval-2013 Task 9.1.

determined by the number of entity classes (i.e., 4 for the DNR task).

The Jordan network is very similar to the Elman network, except that the feedback is sourced from the output layer rather than the previous hidden layer:

$$h(t) = f(U \bullet x(t) + V \bullet y(t-1)). \quad (5)$$

Although the Elman and Jordan networks can learn long-term dependencies, their exponential decay biases them toward their most recent inputs (Bengio et al., 1994). The LSTM was designed to overcome this limitation by incorporating a gated memory-cell to capture long-range dependencies within the data (Hochreiter and Schmidhuber, 1997). In the bidirectional LSTM, for any given sentence, the network computes both a left, $\overrightarrow{h}(t)$, and a right, $\overleftarrow{h}(t)$, representations of the sentence context at every input, $x(t)$. The final representation is created by concatenating them as $h(t) = [\overrightarrow{h}(t); \overleftarrow{h}(t)]$. All these networks utilize the $h(t)$ layer as an implicit feature for entity class prediction: although this model has proved effective in many cases, it is not able to provide joint decoding of the outputs in a Viterbi-style manner (e.g., an I-group cannot follow a B-brand; etc). Thus, another modification to the bidirectional LSTM is the addition of a conditional random field (CRF) (Lafferty et al., 2001) as the output layer to provide optimal sequential decoding. The resulting network is commonly referred to as the bidirectional LSTM-CRF (Lample et al., 2016).

4 Experiments

4.1 Datasets

The DDIExtraction 2013 shared task challenge from SemEval-2013 Task 9.1 (Segura-Bedmar et al., 2013) has provided a benchmark corpus for DNR and DDI extraction. The corpus contains manually-annotated pharmacological substances and drug-drug interactions (DDIs) for a total of $18,502$ pharmacological substances and $5,028$ DDIs. It collates two distinct datasets: DDI-DrugBank and DDI-MedLine (Herrero-Zazo et al., 2013). Table 2 summarizes the basic statistics of the training and test datasets used in our experiments. For proper comparison, we follow the same settings as (Segura-Bedmar et al., 2015), using the training data of the DNR task along with the test data for the DDI task for training and validation of DNR. We split this joint dataset into a training and validation sets with approximately 70% of sentences for training and the remaining for validation.

4.2 Evaluation Methodology

Our models have been blindly evaluated on unseen DNR test data using the *strict* evaluation metrics. With this evaluation, the predicted entities have to match the ground-truth entities exactly, both in boundary and class. To facilitate the replication of our experimental results, we have used a publicly-available library for the implementation (i.e., the Theano neural network toolkit (Bergstra et al., 2010)) and we publicly release our code. The experiments have been run over a range of values for the hyper-parameters, using the validation set for selection (Bergstra and Bengio, 2012). The hyper-parameters include the number of hidden-

3

Methods	DDI-DrugBank			DDI-MedLine		
	Precision	Recall	F_1 Score	Precision	Recall	F_1 Score
WBI-NER (Rocktäschel et al., 2013)	88.00	87.00	87.80	61.00	56.00	58.10
Hybrid-DDI (Abacha et al., 2015)	93.00	70.00	80.00	74.00	25.00	37.00
Word2Vec+DINTO (Segura-Bedmar et al., 2015)	69.00	82.00	75.00	65.00	51.00	57.00
Elman RNN	79.91	60.91	69.13	43.23	33.56	37.78
Jordan RNN	77.59	60.91	68.25	59.47	30.20	40.06
Bidirectional LSTM-CRF	87.07	83.39	85.19	52.93	52.57	52.75

Table 3: Performance comparison between the recurrent neural networks (bottom three lines) and state-of-the-art systems (top three lines) over the SemEval-2013 Task 9.1.

	Entities	DDI-DrugBank			DDI-MedLine		
		Precision	Recall	F_1 Score	Precision	Recall	F_1 Score
Bidirectional LSTM-CRF	group	76.92	90.91	83.33	59.52	53.76	56.50
	drug	90.59	84.62	87.50	65.22	61.05	63.06
	brand	91.30	79.25	84.85	0.0	0.0	0.0
	drug_n	0.0	0.0	0.0	40.20	45.45	42.67

Table 4: SemEval-2013 Task 9.1 results by entity for the bidirectional LSTM-CRF.

layer nodes, $H \in \{25, 50, 100\}$, the context window size, $s \in \{1, 3, 5\}$, and the embedding dimension, $d \in \{50, 100, 300, 500, 1000\}$. Two additional parameters, the learning and drop-out rates, were sampled from a uniform distribution in the range $[0.05, 0.1]$. The embedding and initial weight matrices were all sampled from the uniform distribution within range $[-1, 1]$. Early training stopping was set to 100 epochs to mollify over-fitting, and the model that gave the best performance on the validation set was retained. The accuracy is reported in terms of micro-average F_1 score computed using the CoNLL score function (Nadeau and Sekine, 2007).

4.3 Results and Analysis

Table 3 shows the performance comparison between the explored RNNs and state-of-the-art DNR systems. As an overall note, the RNNs have not reached the same accuracy as the top system, WBI-NER (Rocktäschel et al., 2013). However, the bidirectional LSTM-CRF has achieved the second-best score on DDI-DrugBank and the third-best on DDI-MedLine. These results seem interesting on the ground that the RNNs provide DNR straight from text rather than from manually-engineered features. Given that the RNNs learn entirely from the data, the better performance over the DDI-DrugBank dataset is very likely due to its larger size. Accordingly, it is reasonable to expect higher relative performance should larger corpora become available in the fu-

ture. Table 4 also breaks down the results by entity class for the bidirectional LSTM-CRF. The low score on the $brand$ class for DDI-MedLine and on the $drug\_n$ class (i.e., active substances not approved for human use) for DDI-DrugBank are likely attributable to the very small sample size (Table 2). This issue is also shared by the state-of-the-art DNR systems.

5 Conclusion

This paper has investigated the effectiveness of recurrent neural architectures, namely the Elman and Jordan networks and the bidirectional LSTM-CRF, for drug name recognition. The most appealing feature of these architectures is their ability to provide end-to-end recognition straight from text, sparing effort from laborious feature construction. To the best of our knowledge, ours is the first paper to explore RNNs for entity recognition from pharmacological text. The experimental results over the SemEval-2013 Task 9.1 benchmarks look promising, with the bidirectional LSTM-CRF ranking closely to the state of the art. A potential way to further improve its performance would be to initialize its training with unsupervised word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). This approach has proved effective in many other domains and still dispenses with expert annotation effort; we plan this exploration for the near future.

References

Asma Ben Abacha, Md Faisal Mahbub Chowdhury, Aikaterini Karanasiou, Yassine Mrabet, Alberto Lavelli, and Pierre Zweigenbaum. 2015. Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug–drug interaction extraction and classification. *Journal of Biomedical Informatics*, 58:122–132.

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.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305.

James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. 2010. Theano: A CPU and GPU math compiler in Python. In *The 9th Python in Science Conference*, pages 1–7.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics*, 46(5):914–920.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

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

Michael I. Jordan. 1986. Serial order: A parallel distributed processing approach. Technical report, San Diego: University of California, Institute for Cognitive Science.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL-HLT*.

Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. 2015a. Drug name recognition: Approaches and resources. *Information*, 6(4):790–810.

Shengyu Liu, Buzhou Tang, Qingcai Chen, Xiaolong Wang, and Xiaoming Fan. 2015b. Feature engineering for drug name recognition in biomedical texts: Feature conjunction and feature selection. *Computational and Mathematical Methods in Medicine*, 2015:1–9.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, et al. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):530–539.

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.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26.

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

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

Tim Rocktäschel, Torsten Huber, Michael Weidlich, and Ulf Leser. 2013. WBI-NER: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs. In *The 7th International Workshop on Semantic Evaluation*, pages 356–363.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *The 7th International Workshop on Semantic Evaluation*.

Isabel Segura-Bedmar, Víctor Suárez-Paniagua, and Paloma Martínez. 2015. Exploring word embedding for drug name recognition. In *The 6th International Workshop on Health Text Mining and Information Analysis*, page 64.