Bidirectional LSTM-CRF for Clinical Concept Extraction

Raghavendra Chalapathy
University of Sydney
Capital Markets CRC
rcha9612@uni.sydney.edu.au

Ehsan Zare Borzeshi
Capital Markets CRC
ezborzeshi@cmcrc.com

Massimo Piccardi
University of Technology Sydney
Massimo.Piccardi@uts.edu.au

Abstract

Automated extraction of concepts from patient clinical records is an essential facilitator of clinical research. For this reason, the 2010 i2b2/VA Natural Language Processing Challenges for Clinical Records introduced a concept extraction task aimed at identifying and classifying concepts into predefined categories (i.e., treatments, tests and problems). State-of-the-art concept extraction approaches heavily rely on handcrafted features and domain-specific resources which are hard to collect and define. For this reason, this paper proposes an alternative, streamlined approach: a recurrent neural network (the bidirectional LSTM with CRF decoding) initialized with general-purpose, off-the-shelf word embeddings. The experimental results achieved on the 2010 i2b2/VA reference corpora using the proposed framework outperform all recent methods and ranks closely to the best submission from the original 2010 i2b2/VA challenge.

1 Introduction

Patient clinical records typically contain longitudinal data about patients' health status, diseases, conducted tests and response to treatments. Analysing such information can prove of immense value not only for clinical practice, but also for the organisation and management of healthcare services. *Concept extraction* (CE) aims to identify mentions to medical concepts such as problems, test and treatments in clinical records (e.g., discharge summaries and progress reports) and classify them into predefined categories. The concepts in clinical records are often expressed with unstructured, "free" text, making their automatic extraction a challenging task for clinical Natural Language Processing (NLP) systems. Traditional approaches have extensively relied on rule-based systems and lexicons to recognise the concepts of interest. Typically, the concepts represent drug names, anatomical nomenclature and other specialized names and phrases which are not part of everyday vocabularies. For instance, "resp status" should be interpreted as "response status". Such use of abbreviated phrases and acronyms is very common within the medical community, with many abbreviations having a specific meaning that differ from that of other lexicons. Dictionary-based systems perform concept extraction by looking up terms on medical ontologies such as the Unified Medical Language System (UMLS) (Kipper-Schuler et al., 2008). Intrinsically, dictionary- and rule-based systems are laborious to implement and inflexible to new cases and misspellings (Liu et al., 2015). Although these systems can achieve high precision, they tend to suffer from low recall (i.e., they may miss a significant number of concepts). To overcome these limitations, various machine learning approaches have been proposed (e.g., conditional random fields (CRFs), maximum-entropy classifiers and support vector machines) to simultaneously exploit the textual and contextual information while reducing the reliance on lexicon lookup (Lafferty et al., 2001; Berger et al., 1996; Joachims, 1998). State-of-the-art machine learning approaches usually follow a two-step process of *feature engineering* and *classification*. The feature engineering task is, in its own right, very laborious and demanding on expert knowledge, and it can become the bottleneck of the overall approach. For this reason, this paper proposes a highly streamlined alternative: to employ a contemporary neural network - the bidirectional LSTM-CRF - initialized with general-purpose, off-the-shelf word embeddings such

Proceedings of the Clinical Natural Language Processing Workshop,
pages 7–12, Osaka, Japan, December 11-17 2016.

Sentence	*His*	*HCT*	*had*	*dropped*	*from*	*36.7*	*despite*	*2U*	*PRBC*	*and*
Concept class	*O*	*B-test*	*O*	*O*	*O*	*O*	*O*	*B-treatment*	*I-treatment*	*O*

Table 1: Example sentence in a concept extraction task. The concept classes are represented in the standard in/out/begin (IOB) format.

as GloVe (Pennington et al., 2014a) and Word2Vec (Mikolov et al., 2013b). The experimental results over the authoritative 2010 i2b2/VA benchmark show that the proposed approach outperforms all recent approaches and ranks closely to the best from the literature.

2 Related Work

Most of the research to date has framed CE as a specialized case of named-entity recognition (NER) and employed a number of supervised and semi-supervised machine learning algorithms with domain-dependent attributes and text features (Uzuner et al., 2011). Hybrid models obtained by cascading CRF and SVM classifiers along with several pattern-matching rules have shown to produce effective results (Boag et al., 2015). Moreover, (Fu and Ananiadou, 2014) have given evidence to the importance of including preprocessing steps such as truecasing and annotation combination. The system that has reported the highest accuracy on the 2010 i2b2/VA concept extraction benchmark is based on unsupervised feature representations obtained by Brown clustering and a hidden semi-Markov model as classifier (de-Bruijn et al., 2011). However, the use of a "hard" clustering technique such as Brown clustering is not suitable for capturing multiple relations between the words and the concepts. For this reason, Jonnalagadda et al. (Jonnalagadda et al., 2012) demonstrated that a random indexing model with distributed word representations can improve clinical concept extraction. Moreover, Wu et al. (Wu et al., 2015) have jointly used word embeddings derived from the entire English Wikipedia (Collobert et al., 2011) and binarized word embeddings derived from domain-specific corpora (e.g. the MIMIC-II corpus (Saeed et al., 2011)). In the broader field of machine learning, the recent years have witnessed a proliferation of deep neural networks, with outstanding 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 over traditional approaches is that they can learn the feature representations automatically from the data, thus avoiding the expensive feature-engineering stage. Given the promising performance of deep neural networks and the recent success of unsupervised word embeddings in general NLP tasks (Pennington et al., 2014a; Mikolov et al., 2013b; Lebret and Collobert, 2014), this paper sets to explore the use of a state-of-the-art deep sequential model initialized with general-purpose word embeddings for a task of clinical concept extraction.

3 The Proposed Approach

CE 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, *HCT* as the test and *2U PRBC* as the treatment. In this paper, we approach the CE task by the bidirectional LSTM-CRF framework where each word in the input sentence is first mapped to either a random vector or a vector from a word embedding. We therefore provide a brief description of both word embeddings and the model hereafter.

Word embeddings are vector representations of natural language words that aim to preserve the semantic and syntactic similarities between them. The vector representations can be generated by either count-based approaches such as Hellinger-PCA (Lebret and Collobert, 2014) or trained models such as Word2Vec (including skip-grams and continuous-bag-of-words) and GloVe trained over large, unsupervised corpora of general-nature documents. In its embedded representation, each word in a text is represented by a real-valued vector, x, of arbitrary dimensionality, d.

Recurrent neural networks (RNNs) are a family of neural networks that operate on sequential data. They take as input a sequence of vectors $(x_1, x_2, ..., x_n)$ and output a sequence of class posterior probabilities, $(y_1, y_2, ..., y_n)$. An intermediate layer of hidden nodes, $(h_1, h_2, ..., h_n)$, is also part of the model.

	Training set	Test set
notes	170	256
sentences	16315	27626
problem	7073	12592
test	4608	9225
treatment	4844	9344

Table 2: Statistics of the training and test data sets used for the 2010-i2b2/VA concept extraction.

In an RNN, the value of the hidden node at time t, h_t, depends on both the current input, x_t, and the previous hidden node, h_{t-1}. This recurrent connection from the past timeframe enables a form of short-term memory and makes the RNNs suitable for the prediction of sequences. Formally, the value of a hidden node is described as:

$$h_t = f(U \bullet x_t + V \bullet h_{t-1}) \tag{1}$$

where U and V are trained 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, $f(x) = 1/1 + e^{-x}$, that adds non-linearity to the layer. Eventually, $h(t)$ is input into the output layer and convolved with the output weight matrix, W:

$$y_t = g(W \bullet h_t), \text{ with } g(z_m) = \frac{e^{z_m}}{\Sigma_{k=1}^{K} e^{z_k}} \tag{2}$$

Eventually, the output is normalized by a multi-class logistic function, $g(\cdot)$, to become a proper probability over the class set. Therefore, the output dimensionality is equal to the number of concept classes. Although an RNN can, in theory, learn long-term dependencies, in practice it tends to be biased towards its most recent inputs. For this reason, the Long Short-Term Memory (LSTM) network incorporates an additional "gated" memory cell that can store long-range dependencies (Hochreiter and Schmidhuber, 1997). In its bidirectional version, the LSTM computes both a forward, $\overrightarrow{h_t}$, and a backward, $\overleftarrow{h_t}$, hidden representation at each timeframe t. The final representation is created by concatenating them as $h_t = [\overrightarrow{h_t};\overleftarrow{h_t}]$. In all these networks, the hidden layer can be regarded as an implicit, learned feature that enables concept prediction. A further improvement to this model is provided by performing joint decoding of the entire input sequence in a Viterbi-style manner using a CRF (Lafferty et al., 2001) as the final output layer. The resulting network is commonly referred to as the *bidirectional LSTM-CRF* (Lample et al., 2016).

4 Experiments

4.1 Dataset

The 2010 i2b2/VA Natural Language Processing Challenges for Clinical Records include a concept extraction task focused on the extraction of medical concepts from patient reports. For the challenge, a total of 394 concept-annotated reports for training, 477 for testing, and 877 unannotated reports were de-identified and released to the participants alongside a data use agreement (Uzuner et al., 2011). However, part of this data set is no longer being distributed due to restrictions later introduced by the Institutional Review Board (IRB). Thus, Table 2 summarizes the basic statistics of the training and test data sets which are currently publicly available and that we have used in our experiments.

4.2 Evaluation Methodology

Our models have been blindly evaluated on the 2010 i2b2/VA CE test data using a strict evaluation criterion requiring the predicted concepts to exactly match the annotated concepts in terms of both boundary

Methods	Precision	Recall	F_1 Score
Hidden semi-Markov Model (deBruijn et al., 2011)	86.88	83.64	85.23
Distributonal Semantics CRF (Jonnalagadda et al., 2012)	85.60	82.00	83.70
Binarized Neural Embedding CRF (Wu et al., 2015)	85.10	80.60	82.80
CliNER (Boag et al., 2015)	79.50	81.20	80.00
Truecasing CRFSuite (Fu and Ananiadou, 2014)	80.83	71.47	75.86
Random - Bidirectional LSTM-CRF	81.06	75.40	78.13
Word2Vec - Bidirectional LSTM-CRF	82.61	80.03	81.30
GloVe - Bidirectional LSTM-CRF	84.36	83.41	83.88

Table 3: Performance comparison between the bidirectional LSTM-CRF (bottom three lines) and state-of-the-art systems (top five lines) over the 2010 i2b2/VA concept extraction task.

and class. To facilitate the replication of our experimental results, we have used a publicly-available library for the implementation of the LSTM (i.e. the Theano neural network toolkit (Bergstra et al., 2010)) and we publicly release our code[1]. We have split the training set into two parts (sized at approximately 70% and 30%, respectively), using the first for training and the second for selection of the hyper-parameters ("validation") (Bergstra and Bengio, 2012).The hyper-parameters include the embedding dimension, d, chosen over $\{50, 100, 300, 500\}$, and two additional parameters, the learning and drop-out rates, that were sampled from a uniform distribution in the range $[0.05, 0.1]$. All weight matrices were randomly initialized from the uniform distribution within range $[-1, 1]$. The word embeddings were either initialized randomly in the same way or fetched from Word2Vec and GloVe (Mikolov et al., 2013a; Pennington et al., 2014b). Approximately 25% of the tokens were alphanumeric, abbreviated or domain-specific strings that were not available as pre-trained embeddings and were always randomly initialized. Early stopping of training was set to 50 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 state-of-the-art CE systems and the proposed bidirectional LSTM-CRF with different initialization strategies. As a first note, the bidirectional LSTM-CRF initialized with GloVe outperforms all recent approaches (2012-2015). On the other hand, the best submission from the 2010 i2b2/VA challenge (deBruijn et al., 2011) still outperforms our approach. However, based on the description provided in (Uzuner et al., 2011), these results are not directly comparable since the experiments in (deBruijn et al., 2011; Jonnalagadda et al., 2012) had used the original dataset which has a significantly larger number of training samples. Using general-purpose, pre-trained embeddings improves the F_1 score by over 5 percentage points over a random initialization. In general, the results achieved with the proposed approach are close and in many cases above the results achieved by systems based on hand-engineered features.

Conclusion

This paper has explored the effectiveness of the contemporary bidirectional LSTM-CRF for clinical concept extraction. The most appealing feature of this approach is its ability to provide end-to-end recognition using general-purpose, off-the-shelf word embeddings, thus sparing effort from time-consuming feature construction. The experimental results over the authoritative 2010 i2b2/VA reference corpora look promising, with the bidirectional LSTM-CRF outperforming all recent approaches and ranking closely to the best submission from the original 2010 i2b2/VA challenge. A potential way to further improve its performance would be to explore the use of unsupervised word embeddings trained from domain-specific resources such as the MIMIC-III corpora (Johnson et al., 2016).

[1] https://github.com/raghavchalapathy/Bidirectional-LSTM-CRF-for-Clinical-Concept-Extraction

References

Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research (JMLR)*, 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.

William Boag, Kevin Wacome, Tristan Naumann, and Anna Rumshisky. 2015. CliNER: A lightweight tool for clinical named entity recognition. In *AMIA Joint Summits on Clinical Research Informatics (poster)*.

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 (JMLR)*, 12:2493–2537.

Berry deBruijn, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Xiaodan Zhu. 2011. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *Journal of the American Medical Informatics Association*, 18(5):557–562.

Xiao Fu and Sophia Ananiadou. 2014. Improving the extraction of clinical concepts from clinical records. *Proceedings of the 4th Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM04)*.

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.

Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning (ECML)*, pages 137–142. Springer.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo A. Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.

Siddhartha Jonnalagadda, Trevor Cohen, Stephen Wu, and Graciela Gonzalez. 2012. Enhancing clinical concept extraction with distributional semantics. *Journal of biomedical informatics*, 45(1):129–140.

Karin Kipper-Schuler, Vinod Kaggal, James Masanz, Philip Ogren, and Guergana Savova. 2008. System evaluation on a named entity corpus from clinical notes. In *Language Resources and Evaluation Conference (LREC)*, pages 3001–3007.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems (NIPS)*.

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

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *The North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Rémi Lebret and Ronan Collobert. 2014. Word emdeddings through hellinger PCA. In *European Chapter of the Association for Computational Linguistics (EACL)*.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. GloVe: Global Vectors for Word Representation. `https://code.google.com/archive/p/word2vec/`. Accessed: 2016-08-30.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems (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. 2014a. GloVe: Global vectors for word representation. In *European conference on machine learning (ECML)*, pages 1532–1543.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014b. GloVe: Global vectors for word representation. `http://nlp.stanford.edu/projects/glove/`. Accessed: 2016-08-30.

Mohammed Saeed, Mauricio Villarroel, Andrew T. Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H. Kyaw, Benjamin Moody, and Roger G. Mark. 2011. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. *Critical Care Medicine*, 39:952–960.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Yonghui Wu, Jun Xu, Min Jiang, Yaoyun Zhang, and Hua Xu. 2015. A study of neural word embeddings for named entity recognition in clinical text. In *AMIA Annual Symposium Proceedings*, volume 2015.