A Comparative Study of Sequential Tagging Methods for Domain Knowledge Entity Recognition in Biomedical Papers

Jian Wu*
Md Reshad Ul Hoque*
Old Dominion University
Norfolk, VA, USA
{j1wu,mhoqu001}@odu.edu

Gunnar W. Reiske
Michele C. Weigle
Computer Science
Old Dominion University
Norfolk, VA, USA

Brenda T. Bradshaw
School of Dental Hygiene
Old Dominion University
Norfolk, VA, USA

Holly D. Gaff
Biological Sciences
Old Dominion University
Norfolk, VA, USA

Jiang Li
Electrical & Computer Engineering
Old Dominion University
Norfolk, VA, USA

Chiman Kwan
Applied Research LLC.
Rockville, MD, USA

ABSTRACT

Named entity recognition has been extensively studied in the past decade. The state-of-the-art models, trained on general text such as Wikipedia articles and newsletters, have achieved $F_1 > 0.90$. Entity types are focused on people, location, organization, etc. However, entity recognition from domain-specific text, in particular research papers, is still challenging. In this paper, we perform a comparative study of sequential tagging methods on this task using a manually curated corpus from biomedical papers on Lyme disease. Each model we compare consists of a non-sequential classification and a sequential-tagging component. In this pilot study, we freeze the non-sequential component to study how the sequential tagging methods perform with different models including the conditional random field (CRF) and bidirectional long short-term memory (BiLSTM). The results shed light on the importance of pre-trained word embeddings such as ELMo for relatively small training samples, the roles of attention mechanism, and the residual unit. BiLSTM with attention+residual+ELMo achieves the best performance in recognizing entity strings. CRF with enriched features achieves the best performance in recognizing entity mentions and their positions.

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; *Data mining*; • **Computing methodologies** → **Information extraction**.

KEYWORDS

digital library, entity recognition, conditional random field, long short-term memory, natural language processing, domain knowledge entity

*Both authors contributed equally to this research.

JCDL '20, June 19–23, 2020, Wuhan, Hubei, China

ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

ACM Reference Format:
Jian Wu, Md Reshad Ul Hoque, Gunnar W. Reiske, Michele C. Weigle, Brenda T. Bradshaw, Holly D. Gaff, Jiang Li, and Chiman Kwan. 2020. A Comparative Study of Sequential Tagging Methods for Domain Knowledge Entity Recognition in Biomedical Papers. In *ACM/IEEE Joint Conference on Digital Libraries, June 19–23, 2020, Wuhan, Hubei, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Named entity recognition (NER) is a fundamental task in information extraction that seeks to extract named entities from unstructured text and classify them into predefined categories. The outcome can be utilized for many downstream applications such as constructing knowledge bases, data linking, and question answering. In the past decade, NER has been extensively studied based on models trained on general text, such as Wikipedia articles and newsletters, e.g., [14]. The most extracted entities fall into predefined categories including but not limited to people, organization, location, time expression, and monetary values. Although the general NER has achieved remarkable accuracy [14, 15], entity recognition from domain specific text, represented by scholarly papers published in research venues, is still challenging. Since the BioNLP shared task in 2004, much effort has been put on identifying DNA, RNA, cell line, cell type, and protein in biomedical papers. In this paper, we conduct a comparative study of sequential tagging models on domain knowledge entity extractions from biomedical papers on Lyme disease. Our research question is: *how do different sequential tagging approaches, with recently proposed boosting mechanisms, perform in extracting domain knowledge entities?*

We first define domain knowledge entities (DKEs), best described as noun phrases (NPs) representing domain knowledge of interest. DKEs are different from keyphrases [8], generally defined as important phrases or concepts in a paper. Keyphrases provide high-level description of a paper but DKEs can be at low-levels. For example, the article titled "Lyme disease: A rigorous review of diagnostic criteria and treatment" [3] has 4 keyphrases. However, the following sentence contains 6 DKEs, marked by underlines.

> "Spirochetes with similar morphology, protein profile and antigenic determinants were detected in Ixodes ricinus ticks from Switzerland and Ixodes pacificus ticks from Oregon and subsequently in Ixodes persulcatus ticks in Russia. "

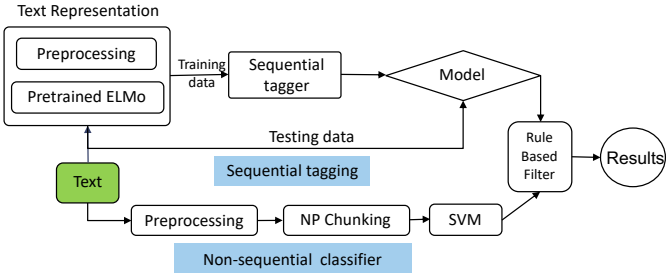

Figure 1: The general architecture of the entity extractors, including a non-sequential classifier and a sequential tagger. The results are merged and a rule-based filter is applied.

DKEs have been found useful in detection of fake scientific news [10], because they represent unique knowledge in a certain domain and a combination of them can be good identifiers of text snippet for research works. DKEs in medical science were extracted to generate knowledge triples and construct knowledge graphs from Electronic Medical Record (EMR) [5].

Sequential tagging can be used for recognizing DKEs from unstructured text. In this paper, our comparison is focused on conditional random field (CRF), bidirectional long-short term memory (BiLSTM), and their vaiants. We do not compare Hidden Markov Model (HMM) because CRF has already shown advantages over HMM in many sequence tagging problems, e.g., [16]. Our work is based on a relatively small dataset containing 100 documents in biomedical science. Each document consists of 1–3 paragraphs manually curated from journal articles on Lyme disease. The documents are manually annotated and validated by domain experts in biomedical and health sciences. To the best of our knowledge, there is no open access dataset on this particular domain. We demonstrate that despite of the relatively small size, the best model achieves a decent $F_1 = 0.55$ on extracting entity strings. The results also shed light on the critical roles of pre-trained WE and the attention mechanism in training with relatively small samples.

2 RELATED WORK

In a 2011 paper, key aspects of scientific papers were extracted by matching language patterns in dependency trees [7]. The authors extracted *focus*, *technique*, and *domain* from titles and abstracts in the ACL Anthology corpus using handwritten patterns. A similar work in 2017 used an unsupervised method [12] but only *application* and *technique* were extracted, and the evaluation was on computer science papers.

In 2015, SegPhrase was developed to extract quality phrases from text corpora using distant supervision [13]. The method rectifies the TF-IDF of segmented phrases in order to raise the rank of more informative phrases. A similar framework FacetGist extracts application, technique, evaluation metrics, and dataset from academic papers, using POS-tagging to segment phrases. The final selection of candidate concepts is made by solving a joint optimization problem. The experiments were based on ACL and DBLP titles and abstracts. Both farmeworks are best used on text corpora rather than individual documents.

The SemEval 2017 (SE17) had a task to extract "keyphrases" from scientific documents, which are essentially named entities in their context [2]. The winner of SE17 proposed a model [1] to stack CRF on top of BiLSTM. The model represents each word token using a vector $\mathbf{x}_k$ by concatenating a vector $\mathbf{c}_k$ (from character embedding) and a vector generated by word embedding (WE) $\mathbf{w}_k$, in which k denotes a token position. Next, the feature representations of words are learned using neural language models. The token representation $\mathbf{x}_k$ is fed through a BiLSTM to embed the history into a fixed dimensional vector. The bi-directional embeddings are concatenated and used for sequence tagging. The BiLSTM layers are followed by a CRF layer to predict the tag of each token. The model achieved an $F_1 = 0.54$ (the ensemble model achieved an $F_1 = 0.55$), however, the implementation was not open source.

3 MODELS

3.1 CRF and BiLSTM

Sequential tagging is a method to label individual tokens such as words in a sequence, a sentence for instance, in which order is important. One commonly used model is CRF, In CRF, the probability of tags for a token depends on its own features, and features *and tags* of the tokens surrounding it. CRF computes the joint probability distribution of the entire label sequence when an observation sequence intended for labeling is available. Recurrent neural network (RNN) is an nonlinear model for representation learning. The bidirectional long short-term memory (BiLSTM) has been proposed to in lieu of the vanilla RNN to overcome its vanishing gradient problem. In this model, the vector representation of the current token depends on the representations of context tokens. BiLSTM is usually followed by a fully connected layer or a CRF layer for sequence tagging, e.g., [18].

Recently, the attention mechanism was proposed to be incorporated in many BiLSTM-based models [21]. The idea is to apply attention weights of individual tokens, calculated using context vectors, when aggregating them to generate the output vector. The attention mechanism has been adopted in many NLP tasks such as machine translation [11] and question-answering [4]. In our work, we apply a special type called "self-attention", in which the weights are computed based on the correlation between a sentence itself. Self-attention has been used for semantic role labeling [20].

The residual unit structure was designed to solve the degraded performance of very deep neural networks. In the residual unit, the output of a shallow layer is directly added to the output of a deep layer, providing a clear path for gradients to back propagate to shallow layers, making the learning process smoother and faster. Residual networks have been applied to image classification and significantly boosted the performance and training time, e.g., [9].

Our comparative study also utilize pre-trained WE, which has shown advantageous to train an model when the dataset is relatively small [6]. In this paper, we use ELMo, a language model trained on 1 billion word benchmark [17]. The pre-trained model uses a multilayer BiLSTM and calculates the weighted sum of hidden states to represent each word.

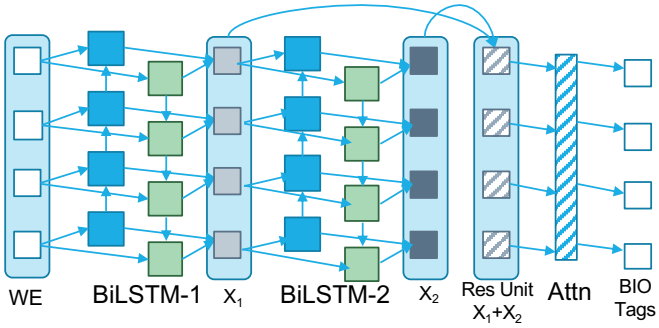

Figure 2: Architecture of Attn-Res-BiLSTM-ELMo model.

3.2 Baseline Models

In a previous work [22], we proposed HESDK, a hybrid method to extract DKEs (Figure 1). The method used an NP chunker followed by an SVM classifier to classify NPs based on TF-IDF calculated using 1M abstracts in the Medline 2016 database. The method also employs a CRF model to label word-level tokens in the Inside-Outside-Beginning (IOB) scheme [19]. The results of these two methods were merged and a rule-based filter was applied before the final results are obtained. In this pilot study, we freeze the non-sequential component and change the models in the sequential tagging component. The goal is to investigate *whether neural network models always achieve better performance than feature-based models and how the WE, attention, and residual unit affect the performance*. The sequential models are listed below.

(1) **CRF**. The sequential model used in HESDK [22]. The model extracts 9 features from the current word and the word before and after the current word (in total 27 features).
(2) **Enriched CRF**. Similar to (1) except that the number of features of each token increases to 16. New features include the first two characters of the POS-tags (e.g., VB from VBD), the type of the phrase the word is in (e.g., NP or VP), the first two characters of a phrase, the first two characters of the word, etc.
(3) **Res-BiLSTM**. This model contains two BiLSTMs (so four LSTM layers) and there is a residual unit after the second layer (Figure 2 without the attention layer, ELMo not used).
(4) **Attn-BiLSTM**. This model contains 1 BiLSTM followed by an attention layer (Figure 2 without $\mathbf{X_2}$ and the residual unit). A model with two BiLSTM layers was shown to underperform.
(5) **BiLSTM-CRF**. This is basically the model implemented by [1] with one BiLSTM followed by a CRF layer.
(6) **BiLSTM-ChE**. In this model, the BiLSTM is enhanced by character embedding. The motivation is that although WE is powerful to encode most words, rare or unseen words are usually embedded as dummy vectors. Character embedding is a solution to mitigate the out-of-vocabulary (OOV) problem. A word is modeled as a character sequence. An LSTM layer is first used to generate WE using character encoding. Another BiLSTM is used to further generate the encoding of each word. The word and character-WEs are concatenated to generate the final WEs.
(7) **Res-BiLSTM-ELMo**. This model is based on (3) except that the input to the Res-BiLSTM was initialized using the pre-trained ELMo WE [17] (Figure 2 without the attention layer).
(8) **Attn-Res-BiLSTM-ELMo**. This is the most complicated model we applied. The input was initialized using the pre-trained ELMo, followed by two BiLSTMs, a residual unit was used to add outputs of the first BiLSTM (i.e., $\mathbf{X_1}$) and the second BiLSTM (i.e., $\mathbf{X_2}$). An attention layer was applied after the residual layer and a `TimeDistributed` layer in Keras is applied to output IOB tags for each token (Figure 2).

4 EXPERIMENTS

4.1 Data Processing and Experimental Setups

The ground truth data are compiled by first searching a list of keywords, such as "Lyme disease", "tick-borne disease" on Google Scholar, resulting in 140 articles, from which 41 were randomly selected ranging from 1990 to 2018. We visually inspected them and extracted 100 documents to annotate, each consisting about 1–3 passages. Each document is manually cleansed such that (1) Each passage occupies only one line; (2) All characters (e.g., α) are encoded in UTF-8; (3) Superscripts and subscripts are expressed in the Latex way, e.g., "^{+}"; (4) Citation marks are preserved at the original places and canonicalized to Arabic numbers in square brackets, e.g., "[10]". Cleansing the data allows us to focus on information extraction, without affected by noise introduced when converting PDFs into text.

We use a web-based tool called *brat* for annotation. The annotator follows five general rules. (1) A DKE must be a noun or an NP; (2) Acronyms of DKEs are also DKEs (e.g., "LNB" for "Lyme Neuroborreliosis"); (3) Conjunction connected phrases are treated as a whole (e.g., "endemic and nonendemic areas"); (4) Try to label semantically meaningful superphrase when it contains a subphrase (e.g., "B. afzelii infection" instead of "B. afzelii"); (5) medicine names and body parts, even if commonly seen in daily life, are still labeled

Table 1: A comparison of models. All results are before the rule-based filters. Triplets are (Precision, Recall, F_1).

Sequential model	Hard			Soft		
Sequential model only						
CRF	0.13	0.17	0.15	0.41	0.39	0.40
Enriched CRF	**0.27**	**0.17**	**0.21**	0.43	0.25	0.31
BiLSTM	0.10	0.15	0.12	0.32	0.35	0.33
Res-BiLSTM	0.07	0.09	0.08	0.25	0.25	0.25
Attn-BiLSTM	0.06	0.11	0.07	0.21	0.29	0.24
BiLSTM-CRF	0.06	0.08	0.07	0.20	0.22	0.21
BiLSTM-ChE	0.08	0.12	0.10	0.27	0.31	0.29
Res-BiLSTM-ELMo	0.13	0.14	0.13	0.46	0.36	0.40
Att-Res-BiLSTM-ELMo	0.13	0.21	0.16	**0.44**	**0.45**	**0.46**
Sequential model + Non-sequential classification						
CRF	0.22	0.44	0.30	0.47	0.58	0.52
Enriched CRF	**0.35**	**0.43**	**0.39**	0.52	0.49	0.50
BiLSTM	0.20	0.42	0.27	0.41	0.55	0.47
Res-BiLSTM	0.21	0.39	0.27	0.35	0.47	0.40
Attn-BiLSTM	0.17	0.41	0.24	0.30	0.51	0.37
BiLSTM-CRF	0.20	0.40	0.26	0.30	0.51	0.37
BiLSTM-ChE	0.19	0.41	0.26	0.37	0.53	0.44
Res-BiLSTM-ELMo	0.24	0.41	0.30	**0.51**	**0.56**	**0.54**
Att-Res-BiLSTM-ELMo	0.21	0.47	0.29	**0.49**	**0.62**	**0.55**

as DKEs (e.g., "antibiotics" and "brain"). The final ground truth corpus contains 1952 DKEs.

We used Keras v2.1.6, Tensorflow v1.8.0, and Tensorflow-hub v0.3.0 for implementation. For BiLSTM, we set the number of memory units to 512. The input/output dropout rate and the recurrent dropout rates are both set to 0.20. The learning rate was set to 10^{-4}. The sparse categorical cross entropy was used as the loss function. The models were trained up to 15 epochs. The Adaptive Moment Estimation (Adam) optimizer was adopted in the training process. The pre-trained ELMo WE has a dimension size of 1024. We randomly chose 75% documents for training and the remaining 25% for testing.

4.2 Evaluation and Discussion

We consider two types of evaluations (Table 1). Under the hard criteria, an extracted entity is taken as a true positive (TP) if *both* phrase strings and positions are correctly identified. Under the soft criteria, only phrase strings are considered. It is intuitive that the performance is better under the soft criteria. For completeness, we present the performance with only the sequential tagging component and a combination with the non-sequential classifier. The following discussion pertains to the sequential tagging results unless otherwise noted.

(1) The neural models do not necessarily outperform the feature-based models (i.e., CRF and Enriched-CRF). Although the Attn-Res-BiLSTM-ELMo model achieves the best performance under the soft criteria. The Enriched-CRF achieves the best performance under the hard criteria.
(2) The pre-trained WE model (i.e., ELMo) plays an important role in the neural network models. The F_1 increases from 0.25 (Res-BiLSTM) to 0.40 (Res-BiLSTM-ELMo).
(3) The attention mechanism significantly increases the recall of Res-BiLSTM-ELMo (from 0.36 to 0.45) with a marginal loss of precision (from 0.46 to 0.44).
(4) The enriched features in the CRF model are helpful in predicting exact positions of entity mentions. A feature analysis indicates that the type of the phrase plays an important role.
(5) The character embedding model BiLSTM-ChE underperforms compared with the plain BiLSTM model. The residual unit also decreases the F_1. This is likely to be caused by the relatively small training sample size. There are not many OOV characters and the advantage of the residual unit is more prominent on tasks with a large amount of training data.
(6) The non-sequential classifier significantly boosts the overall performance. In the Attn-Res-BiLSTM-ELMo model, the F_1 increases by 9% (from 0.46 to 0.55) under the soft criteria. In the Enriched CRF model, the F_1 increases by 18% (from 0.21 to 0.39) under the hard criteria. This implies the benefit of combining sequential and non-sequential models.

5 CONCLUSION

We conducted a comparative study of sequential labeling methods on the task to recognize DKEs from biomedical papers on Lyme disease. The results indicate that the CRF models outperforms all variants of BiLSTM models under typical settings when predicting both entity strings and positions. The CRF underperforms BiLSTM with attention, residual unit, and ELMo when predicting only entity strings. When the training sample is relatively small, pre-trained WEs and the attention mechanism can significantly boost the performance. However, the overall performance of all sequential tagging methods on predicting the positions of DKEs still need to be improved. We will expand the ground truth size to at least 300. We will also fine-tune hyper-parameters and consider using Stochastic Gradient Descent (SGD) instead of the default Adam optimizer.

REFERENCES

[1] Waleed Ammar, Matthew E. Peters, Chandra Bhagavatula, and Russell Power. 2017. The AI2 system at SemEval-2017 Task 10 (ScienceIE): semi-supervised end-to-end entity and relation extraction.
[2] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017.*
[3] Andrea T. Borchers, Carl L. Keen, Arthur C. Huntley, and M. Eric Gershwin. 2015. Lyme disease: A rigorous review of diagnostic criteria and treatment. *Journal of Autoimmunity* 57 (2015), 82 – 115.
[4] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017.*
[5] Yang Deng, Yaliang Li, Ying Shen, Nan Du, Wei Fan, Min Yang, and Kai Lei. 2019. MedTruth: A Semi-supervised Approach to Discovering Knowledge Condition Information from Multi-Source Medical Data. In *Proceedings of CIKM 2019.*
[6] Ábel Elekes, Antonino Simone Di Stefano, Martin Schäler, Klemens Böhm, and Matthias Keller. 2019. Learning from Few Samples: Lexical Substitution with Word Embeddings for Short Text Classification. In *Proceedings of JCDL 2019.*
[7] Sonal Gupta and Christopher D. Manning. 2011. Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. In *Proceedings of the Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011.*
[8] Guoxiu He, Junwei Fang, Haoran Cui, Chuan Wu, and Wei Lu. 2018. Keyphrase Extraction Based on Prior Knowledge. In *Proceedings of JCDL 2018.*
[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.
[10] Md Reshad Ul Hoque, Dash Bradley, Chiman Kwan, Agnese Chiatti, Jiang Li, and Jian Wu. 2019. Searching for Evidence of Scientific News in Scholarly Big Data. In *Proceedings of the 10th International Conference on Knowledge Capture.*
[11] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *TACL* 5 (2017).
[12] Adit Krishnan, Aravind Sankar, Shi Zhi, and Jiawei Han. 2017. Unsupervised Concept Categorization and Extraction from Scientific Document Titles. In *Proceedings of the 2017 Conference on Information and Knowledge Management.*
[13] Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining Quality Phrases from Massive Text Corpora. In *Proceedings of SIGMOD 2015.*
[14] Xuezhe Ma and Eduard H. 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, ACL 2016.*
[15] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL System Demonstrations.*
[16] Manabu Ohta and Atsuhiro Takasu. 2008. Crf-based authors' name tagging for scanned documents. In *Proceedings of JCDL 2008.*
[17] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 NAACL-HLT.*
[18] Animesh Prasad, Manpreet Kaur, and Min-Yen Kan. 2018. Neural ParsCit: a deep learning-based reference string parser. *International Journal on Digital Libraries* 19, 4 (01 Nov 2018), 323–337.
[19] Lance A. Ramshaw and Mitch Marcus. 1995. Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora, 1995.*
[20] Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep Semantic Role Labeling With Self-Attention. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18).*
[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the NIPS 2017.*
[22] Jian Wu, Sagnik Ray Choudhury, Agnese Chiatti, Chen Liang, and C. Lee Giles. 2017. HESDK: A Hybrid Approach to Extracting Scientific Domain Knowledge Entities. In *Proceedings of JCDL 2017.*