







# Development of Kazakh Named Entity Recognition Models

Darkhan Akhmed-Zaki<sup>1,2</sup> , Madina Mansurova<sup>1</sup> , Vladimir Barakhnin<sup>3,4</sup> ,  
Marek Kubis<sup>5</sup> , Darya Chikibayeva<sup>1</sup>, and Marzhan Kyrgyzbayeva<sup>1</sup>

<sup>1</sup> Al-Farabi Kazakh National University, Almaty, Kazakhstan  
darhan\_a@mail.ru, mansurova.madina@gmail.com,  
dashachikibaeva@gmail.com, marzhan.kyrgyzbaeva@gmail.com

<sup>2</sup> University of International Business, Almaty, Kazakhstan

<sup>3</sup> Institute of Computational Technologies, Siberian Branch of the Russian  
Academy of Sciences, Novosibirsk, Russian Federation  
bar@ict.nsc.ru

<sup>4</sup> Novosibirsk State University, Novosibirsk, Russian Federation

<sup>5</sup> Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland  
mkubis@amu.edu.pl

**Abstract.** Named entity recognition is one of the important tasks in natural language processing. Its practical application can be found in various areas such as speech recognition, information retrieval, filtering, etc. Nowadays there are a variety of available methods for implementing named entity recognition. In this work we experimented with three models and compared the performances of machine learning based models and probabilistic sequence modeling method on the task of Kazakh language named entity recognition. We considered three models based on BERT, Bi-LSTM and CRF baseline. In the future these models can be parts of an ensemble learning system for name entity recognition in order to achieve better performance results.

**Keywords:** Named entity recognition · Conditional random fields · BERT · Bi-LSTM

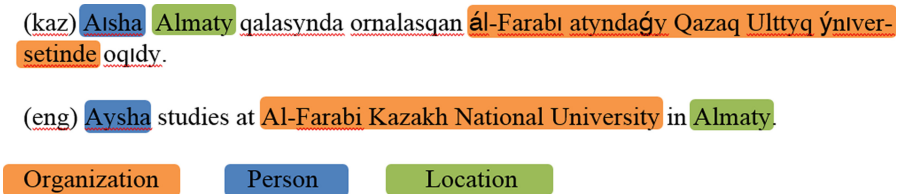
## 1 Introduction

In the information age with the increasing amount of digital data the need for automatic information extraction tools is bigger than ever. While there is a large number of information extraction tools available now for such languages as English or Russian, the situations with Kazakh differs. Kazakh is one of the low-resourced languages and it belongs to the group of agglutinative languages. In this paper we experiment on Kazakh data using different named entity recognition methods.

Currently, there are various approaches for extracting information. They are diverse and it is difficult to say that one is better than the other, since one or another shows good results in different situations. Information retrieval approaches can be classified into the following categories:

- rule-based approaches. The experts manually create the rule sets needed to extract certain data.
- knowledge-based approaches. These include models based on ontologies [1], models based on thesauri [2].
- statistical approaches. They include hidden Markov models [3–5], conditional Markov models [6], conditional random fields [7].
- machine learning based approaches [8].

One of the foundational tasks in the process of information extraction is the recognition of named entities, i.e. spans of text that are proper names of people, organizations, locations and other objects<sup>1</sup>. The task consists of identifying the location of names in text and recognizing their type, as illustrated in Fig. 1.



**Fig. 1.** A sentence with the proper names and their types denoted by square brackets.

Named entity recognition is an important preliminary step in a significant number of extraction tasks such as entity linking, relation extraction, event extraction and template filling (see [9]). Thus, having an accurate model for detecting and classifying proper names is not only significant on its own, but it also contributes to the performance of all the downstream tasks in the process of information extraction.

The existing methods for recognizing named entities can be divided into two categories:

- Rule-based methods. These are the earliest systems for recognizing named entities. Rules are based on lexico-syntactic patterns specific to a particular language.
- Supervised methods. These techniques require training data, manually labeled by experts. Then, on the basis of the annotated data, the system learns the rules for recognizing named entities.

In recent years state-of-the-art models for the named entity recognition task are based on pre-trained language models. They include ELMo [10] and BERT [11]. BERT is a language representation model developed and pre-trained by Google. It is based on transformers and unlike other language representation models designed to “pre-train

<sup>1</sup> In a broad sense it can also encompass recognition of temporal and numerical expressions and identification of terms specific to a particular subject area, such as names of chemical compounds in the biological domain.

deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers” [11]. In this paper we compare three different models. One is based on our previous work [12], the second one is based on BERT and the third is based on CRF. Furthermore, we juxtapose results of neural network models with the output of a non-neural named entity recognizer in order to verify, if the amount of training data that we have collected so far enables neural models to achieve state-of-the-art results in named entity recognition for Kazakh.

## 2 Related Work

Classical systems for extracting named entities use manually selected properties [13]. Some early systems used manual rules [14, 15], but the vast majority of modern systems rely on machine learning models [16], such as conditional random field (CRF) [17, 18], Hidden Markov model (HMM) [19], the support vector method (SVM). Although traditional machine learning models are not based on manual rules, they require a manual function development process, which is quite expensive and depends on the domain and language. Recently, many works using neural networks have surpassed classical systems [20–22]. In recent years, models with a recurrent neural network (RNN) such as Long-Short-Term-Memory (LSTM) [23], Gated Recurrent Unit (GRU) [24] have been very successful in sequence modeling problems, for example, Language Modeling [25, 26], machine translation [27], Dialog Act classification [28, 29]. One of the strengths of RNN models is their ability to learn from the main components of the text (i.e. words and symbols). This generalization feature facilitates the construction of language-independent NER models [30, 31], which are based on an uncontrolled study of properties and a small annotated case.

The first use of neural models in the task of marking a sequence was proposed by Collobert et al. [32]. However, there are some limitations to this model. Firstly, a simple neural network with direct connection is used here, which limits the range of the considered context around words. The model forgets the useful relationships between words over a long distance. Secondly, due to the dependence solely on the vectorization of words, it is impossible to define and use properties represented at the symbol level, such as suffixes and prefixes.

Later, modified models using bidirectional LSTM or Stacked LSTM were proposed [33, 34]. For example, in [33], architecture based on bi-LSTM and CRF is used. The authors of [35] use the bi-LSTM-CNNs architecture. To vectorize characters, they suggest the use of convolutional neural networks. New approaches have been found that use CNN or LSTM to extract subword information from input characters, the results of which are superior to other models [33]. Rei et al. [36] proposed a model, in which words and symbols are fed as input.

## 3 Models

### 3.1 Bi-LSTM

The model is based on a bi-LSTM block using vectorization of characters and words (see Fig. 2).

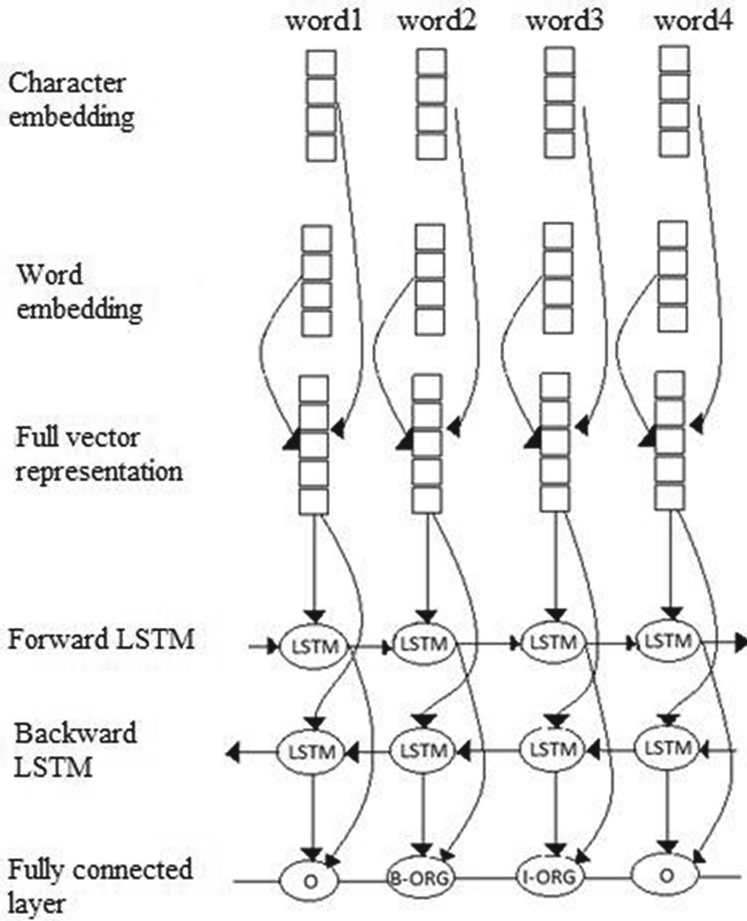


Fig. 2. Bi-LSTM based model architecture.

LSTM (Long Short-Term Memory) is a type of recurrent neural network. Recursive neural networks have the ability to remember the results of past iterations, but they are not able to remember them long-term. The problem of the disappearance of the gradient appears [37]. LSTM networks [35] are designed to combat this problem. They contain three main blocks that control which information will be forgotten and which will be transmitted to subsequent iterations.

### 3.2 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a natural language processing method based on the use of new architecture neural networks for working with sequences known as “transformers” [11]. BERT features consist in the fact that the technology is trained based on the entire set of words in a sentence or request. Previously, neural networks were trained on an ordered sequence of words (from left to right or from

right to left). BERT allows the language model to examine the context of a word based on all the words surrounding it. For example, the word “бет” (which can be translated from Kazakh as “face” or “surface”) will have the same context-free representation in “адамның беті” (person’s face) and “ыстелдің беті” (surface of a table). At the same time, BERT considers word context and represents “бет” using both the previous and the following surrounding word sequences.

In this paper BERT is used for the single sentence tagging task (see Fig. 3). In our work the model architecture consists of the BERT model followed by classifier (see Fig. 4). In this work BERT used twice. First time we use it to represent sentence as tokens and then BERT is used to get encoded representations of spans. The sentence is represented as sequence of words ( $w_1, w_2, \dots, w_n$ ). BERT takes as input sequences of up to N tokens. The output is the last hidden state of the sequence with dimension H. Classifier constitutes of linear layer which takes as input that last hidden state.

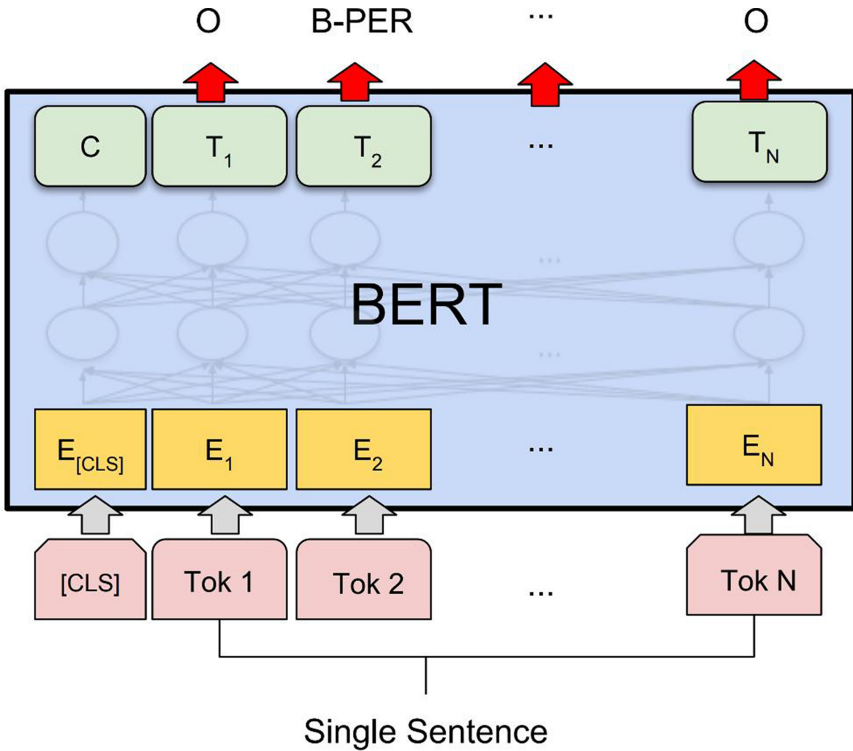


Fig. 3. BERT for single sentence tagging task [11].

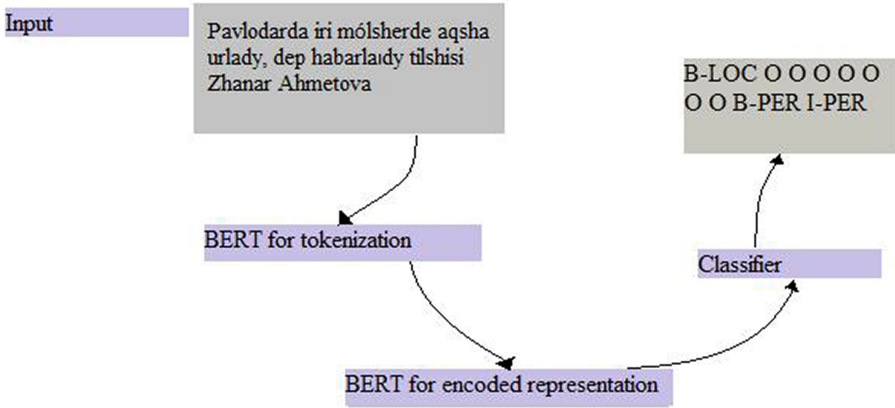


Fig. 4. BERT-classifier model architecture.

### 3.3 Conditional Random Fields

Neural network based models are data intensive, therefore we decided to confront their performance with a method that attains satisfactory results even with respect to the modest amount of data. We choose for this purpose Conditional Random Fields, a probabilistic sequence modeling framework [38] that was used for constructing named entity recognizers in the past [17].

As defined by Laferty et al. [38], for a sequence of data points  $X$  in conjunction with labels  $Y$  and a graph  $G = (V, E)$  such that  $Y$  is indexed by vertices from  $V$  (i.e.  $Y = (Y_v), v \in V$ ), the  $(X, Y)$  pair is a conditional random field, if the random variables  $Y_v$  obey the Markov property with respect to  $G$ , i.e.

$$p(Y_v \vee X, Y_w, w \neq v) = p(Y_v \vee X, Y_w, w \sim v) \tag{1}$$

where  $w \sim v$  means that  $(w, v) \in E$ . In our case  $X$  represents a sentence and  $Y$  consists of named entity labels for the words within the sentence.

CRF models require feature engineering in order to attain satisfactory results (see Table 1).

Table 1. CRF baseline feature set.

Feature	Examples
Current word	word[0] = Tailandta
Prefixes and suffixes of the current word	p[1] = T p[2] = Ta p[3] = Tai s[1] = a s[2] = ta s[3] = dta
Word shapes	shape[0] = ul shape [1] = l
Predecessors and successors of the current word	word[1] = su

The model inherits the set of features from the Named Entity Recognizer developed for [39] with exception of lemmata and part-of-speech based features that are not available in our corpus.

## 4 Training and Evaluation

The experiments were conducted on data collected from Kazakh online news sources. The volume of the data was 7153 sentences. The dataset was labeled manually with 4 entity classes (Location, Organization, Person and Other). IOB scheme was used to denote boundaries of entities. The data were partitioned into training, validation and test sets with [6507, 2531 and 3015 sentences, respectively. The distribution of entities among classes is shown in Table 2.

**Table 2.** Entity class distribution

Entiy class	Sample count
Location	4763
Organization	1650
Person	4352
Other	99650

For the purpose of training we used BERT<sub>BASE</sub> model with the following parameters: H = 768, S = 512, Total Parameters = 110 M. The pre-trained “bert-base-multilingual-cased” model that covers 104 languages was used for initialization. Parameters of the CRF model were estimated using the Passive Aggressive algorithm [39] with the maximum number of iterations set to 25. The CRFsuite library [40] was used for training the model.

We followed the established practice to compute precision, recall and F<sub>1</sub> scores with respect to the explicitly defined testset for the models under evaluation [41]. The metrics are computed according to the following formulas.

$$Precision = \frac{tp}{tp + fp}, \quad (2)$$

$$Recall = \frac{tp}{tp + fn}, \quad (3)$$

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

where *tp* means true positives (number of correctly recognized entities), *fp* means false positives (number of tokens that were mistakenly recognized as named entities) and *fn* means false negatives (number of named entities that were not recognized).

## 5 Results

The main results of the paper are presented in Table 3. As it can be seen from the table the BERT-classifier model outperforms simple Bi-LSTM system and shows better results in precision, recall and  $F_1$  metrics. However, it does not surpass the CRF baseline.

**Table 3.** Comparison of the results on the test set.

Architecture	Precision	Recall	$F_1$
Bi-LSTM	86.81	83.94	85.31
BERT-classifier	98.93	97.07	97.99
CRF baseline	97.73	91.32	94.27

Tables 4 and 5 present the detailed results of the experiments. In Table 4 we report the precision, recall and  $F_1$  scores obtained for the entity classes separately. Table 5 shows per tag results according to the IOB scheme that was used for training.

**Table 4.** Per class results on the test set.

Architecture	Entity class	Precision	Recall	$F_1$
Bi-LSTM	LOC	0.8573	0.8464	0.8525
	ORG	0.8735	0.8253	0.8456
	PER	0.8693	0.8577	0.8618
BERT-classifier	LOC	0.9791	0.9858	0.9872
	ORG	0.9539	0.8465	0.9406
	PER	0.9680	0.9642	0.9627
CRF baseline	LOC	0.9776	0.9839	0.9807
	ORG	0.9488	0.8268	0.8836
	PER	0.9676	0.9555	0.9615



**Table 5.** Per tag results on the test set.

Architecture	Tag	Precision	Recall	F <sub>1</sub>
Bi-LSTM	B-LOC	0.8759	0.8654	0.8702
	B-ORG	0.9043	0.8783	0.8893
	B-PER	0.8614	0.8521	0.8582
	I-LOC	0.8675	0.8445	0.8579
	I-ORG	0.8564	0.8359	0.8431
	I-PER	0.8955	0.8852	0.8903
BERT-classifier	B-LOC	0.9973	0.9825	0.9899
	B-ORG	1.0000	0.9795	0.9896
	B-PER	0.9971	0.9663	0.9815
	I-LOC	1.0000	0.9545	0.9767
	I-ORG	0.9535	0.9535	0.9535
	I-PER	0.9881	0.9881	0.9881
CRF baseline	B-LOC	0.9899	0.9863	0.9881
	B-ORG	0.9840	0.8280	0.8993
	B-PER	0.9871	0.9656	0.9762
	I-LOC	0.9697	0.9412	0.9552
	I-ORG	0.9506	0.7857	0.8603
	I-PER	0.9825	0.9722	0.9773

## 6 Conclusion

In this paper we applied current state-of-the-art language representation model BERT to the Kazakh NER task and compared the results with the models based on Bi-LSTM module and CRF baseline. Despite the fact that the task of extracting named entities has many approaches to solving, the most popular are the approaches based on machine learning using contextual information. The Bi-LSTM algorithm uses the two-sided environment of the target word (before and after it) as contextual information, and as shown by the experiments, it demonstrates relatively high accuracy and completeness results, but it does not surpass the CRF baseline in our task. We found out that BERT-based model achieves significantly better results on the standard evaluation than both Bi-LSTM and CRF models. The results are especially notable in the case of ORG entities where neither Bi-LSTM model nor the CRF baseline get close to the performance of the BERT-based model.

**Acknowledgements.** This work was supported in part under grants of Foundation of Ministry of Education and Science of the Republic of Kazakhstan AP05132933 – “Development of a system for knowledge extraction from heterogeneous data sources to improve the quality of decision-making” (2018–2020) and O.0856 BR05236340 – «Creation of high-performance intelligent technologies

for analysis and decision making for the “logistics-agglomeration”; system in the framework of the digital economy of the Republic of Kazakhstan» (2018–2020).

## References

1. Embley, W.D., Campbell, M.D., Smith, D.R.: Ontology-based extraction and structuring of information from data-rich unstructured documents. In: *Information and Knowledge Management* (1998)
2. Cardie, C.: A case-based approach to knowledge acquisition for domain-specific sentence analysis. In: *Eleventh National Conference on Artificial Intelligence*, pp. 798–803. AAAI Press (1993)
3. Scheffer, T., Decomain, C., Wrobel, S.: Active hidden Markov models for information extraction. In: *International Symposium on Intelligent Data Analysis* (2001)
4. Scheffer, T., Wrobel, S., Popov, B., Ognianov, D., Decomain, C., Hoche, S.: Learning hidden Markov models for information extraction actively from partially labeled text. *Künstliche Intelligenz* (2) (2002)
5. Skounakis, M., Craven, M., Ray, S.: Hierarchical hidden Markov models for information extraction. In: *IJCAI* (2003)
6. McCallum, A.K., Freitag, D., Pereira, F.: Maximum entropy Markov models for information extraction and segmentation. In: *ICML* (2000)
7. McCallum, A.K., Jensen, D.: A note on the unification of information extraction and data mining using conditional-probability, relational models. In: *IJCAI'03 Workshop on Learning Statistical Models from Relational Data* (2003)
8. Mansurova, M., Barakhnin, V., Khibatkhanyuly, Y., Pastushkov, I.: Named entity extraction from semi-structured data using machine learning algorithms. In: Nguyen, N.T., Chbeir, R., Exposito, E., Anioirté, P., Trawiński, B. (eds.) *ICCCI 2019. LNCS (LNAI)*, vol. 11684, pp. 58–69. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-28374-2\\_6](https://doi.org/10.1007/978-3-030-28374-2_6)
9. Jurafsky, D., Martin, J.H.: *Speech and Language Processing: An Introduction to Natural Language processing, Computational Linguistics, and Speech Recognition*. Pearson Prentice Hall, Upper Saddle River (2009)
10. Peters, M., Neumann, M., Iyyer, M., et. al.: Deep contextualized word representations. In: *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1 (Long Papers), pp. 2227–2237 (2018)
11. Devlin, J., Chang, M., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository* (2018)
12. Chikibayeva D., Mansurova M., Nugumanova A., Kyrgyzbayeva M.: Named entity recognition from news sources based on BI-LSTM. In: *2019 IICT Conference*, pp. 519–525 (2019)
13. Luo, G., Huang, X., Lin, C., Nie, Z.: Joint entity recognition and dis-ambiguation. In: *2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp 879–888. Association for Computational Linguistics (2015)
14. Rau, L.F.: Extracting company names from text. In: *Seventh IEEE Conference on Artificial Intelligence Applications*, vol. 1, pp. 29–32. IEEE (1991)
15. Sekine, S., Nobata, C.: Definition, dictionaries and tagger for extended named entity hierarchy. In: *LREC*, pp. 1977–1980 (2004)
16. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. In: *Linguisticae Investigationes* (2007)

17. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: The seventh conference on Natural language learning at HLT-NAACL 2003, vol. 4, pp. 188–191. Association for Computational Linguistics (2003)
18. Kubis, M.: Quantitative analysis of character networks in polish XIX and XX century novels. In: Digital Humanities 2019 Conference, Utrecht, The Netherlands (2019)
19. Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R.: Nymble: a high-performance learning name-finder. In: Fifth Conference on Applied natural language processing, pp. 194–201. Association for Computational Linguistics (1997)
20. Nugumanova, A., Baiburin, Y., Apaev, K.: A new text representation model enriched with semantic relations. In: ICCAS 2015 – 2015 15th International Conference on Control, Automation and Systems, Proceedings (2015)
21. Nugumanova, A.B., Apayev, K.S., Baiburin, Y.M., Mansurova, M.Y.: A contrastive approach to term extraction: case-study for the information retrieval domain using BAWE corpus as an alternative collection. *Eurasian J. Math. Comput. Appl.* **5**, 73–86 (2017)
22. Asahara, M., Matsumoto, Y.: Japanese named entity extraction with redundant morphological analysis. In: 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1, pp. 8–15. Association for Computational Linguistics (2003)
23. Hochreiter, S., Schmidhuber, J.: Long short-term memory. In: *Neural Computation*, pp. 1735–1780 (1997)
24. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling (2014)
25. Mikolov, T., Karafi, M., Burget, L., Cernock'y, J., Khudanpur, S.: Recurrent neural network based language model. In: *Interspeech*, vol. 2, p. 3 (2010)
26. Sundermeyer, M., Schl'uter, R., Ney, H.: LSTM neural networks for language modeling. In: *Interspeech*, pp. 194–197 (2012)
27. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate (2014)
28. Kalchbrenner, N., Blunsom, P.: Recurrent convolutional neural networks for discourse compositionality (2013)
29. Tran, Q., Zukerman, I., Haffari, G.: A hierarchical neural model for learning sequences of dialogue acts. In: 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 1, Long Papers, pp. 428–437. Association for Computational Linguistics, Valencia (2017)
30. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF (2016)
31. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 260–270. Association for Computational Linguistics, San Diego (2016)
32. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. *Journal of Machine Learning Research* **12**, 2493–2537 (2011)
33. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural Architectures for Named Entity Recognition. *CoRR* (2016)
34. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging (2015)
35. Jason, P.C., Nichols, E.: Named Entity Recognition with Bidirectional LSTM-CNNs (2016)
36. Rei, M., Crichton, G., Pyysalo, S.: Attending to characters in neural sequence labeling models. In: 26th International Conference on Computational Linguistics, pp. 309–318 (2016)
37. Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks (2012)

38. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: 18th International Conference on Machine Learning, pp. 282–289 (2001)
39. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online passive-aggressive algorithms. *J. Mach. Learn. Res.* **7**(Mar), 551–585 (2006)
40. Okazaki, N.: CRFsuite: a fast implementation of Conditional Random Fields (CRFs) (2007). <http://www.chokkan.org/software/crfsuite/>. Accessed 14 Feb 2020
41. Tjong, E.F., Sang, K., Meulder, F.D.: Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: CoNLL-2003, Proceedings, Edmonton, Canada, pp. 142–147 (2003)