ASSESSMENT OF DEEP LEARNING METHODS FOR SEQUENCE LABELING

V Raviteja Kanakala[1],

Dept of CSE, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, Pin:522501,

raviteja.kanakala@gmail.com[1]

K.Jagan Mohan[2],

Dept of Information Technology, Annamalai University, Tamilnadu, Pin:508002,

aucsejagan@gmail.com[2]

V.Krishna Reddy[3],

Dept of CSE, Gandhi Institute of Technology and Management, Andhra Pradesh,
Pin:530045, kvuyyuru@gitam.edu[3]

Y Jnapika[4]

Dept of Computer Science, Smt.Nps Govt. Degree College, Chittoor, Pin:517002, Andhra
Pradesh, jnapikagdl@gmail.com[4]

Abstract

Now a day's sequence labeling has become most interesting topic in the current technical era. Sequence labeling is a type of pattern recognition task that involves the algorithmic assignment of a categorical label to each member of sequence of observed values, and also, it is treated as an independent task. By using traditional methods such as HMM and CRF we can implement sequence labeling. Both the methods take the sequence of input and learn to predict an optimal sequence of labels. These are very powerful methods, but they have not experienced the great success due to some drawbacks like lack of semantic awareness and can't handle longer sequential dependencies. So by using deep learning techniques such as recurrent neural networks, they can capture the local dependencies and find longer patterns. Real world applications where the sequence labeling can be applied are Google search Engine. Where in the search box if we type some words automatically Google will suggest some sentences or words which makes our work easier.

1. Introduction

Sequence labeling is one of the main tasks that are focusing on Natural Language Processing over the past decade. The main aim of NLP is to convert a human language into a formal representation so that it is easy for the computers to manipulate. Some of the applications are search, information extraction, and machine translation [1]. POS, Word Sense Disambiguation, NER, Word Segmentation are some of the subtasks of sequence labeling. Among these NER is the important task of NLP. The most traditional sequence labeling models which have shown high performance are linear statistical models like HMM and CRF [2] but these models highly rely on specific task resources and on hand crafted features. By using high performance approaches we can get the better results when compared to the linear statistical Models. Accuracy what we get through these linear statistical models is not accurate when compared to the deep learning techniques. So to overcome the drawbacks in the existing techniques and to get the good results we are training and testing the datasets by using Deep learning techniques so that we can improve the accuracy [3-5]. Deep Learning is the key Technology that is used in many exciting novel applications like Google translators like Siri and Alexa [6]. The difference between the traditional methods and the deep learning methods is both the methods take the sequence of input and learn the optimal sequence and predict the labels for the sequence, But the traditional methods like HMM simply works on the words (type of tokens), and the CRF works on the set of some features like input token or phrases. These are the very powerful methods but they have not been experienced great success due to some of the drawbacks. By using the deep neural nets, we can overcome some of the drawbacks and they have shown great power to learn latent features.

2. Literature Survey

Sequence labeling is a type of machine learning technique which is used for pattern recognition and allows categorizing labels for the observed values. Some of the techniques of Sequence Labeling are Parts Of Speech Tagging (POS), Named Entity Recognition (NER), and Word Segmentation [13]. Initially linear statistical models used for sequence labeling, the models are HMM CRF and SVM.

2.1 HMM

HMM is a generative model which assigns the joint probability for the observations and the label sequence. Then the parameters are trained so that to maximize the joint likelihood of training sets [7]. HMM is called hidden because only the symbols emitted by the system we can see, but not the process which is undergoing between the layers or the states [10]. In HMM there will be Transition Probabilities and Emission Probabilities.

*Transition Probability***:** It is the probability of the state‘s’ appearing after observing sequences *u* and *v* in the sequence of observations.

$$Q(s|u, v)$$

*Emission Probability***:** It is the Probability of an observation *x* given that the state was *s*

$$E(x|s)$$

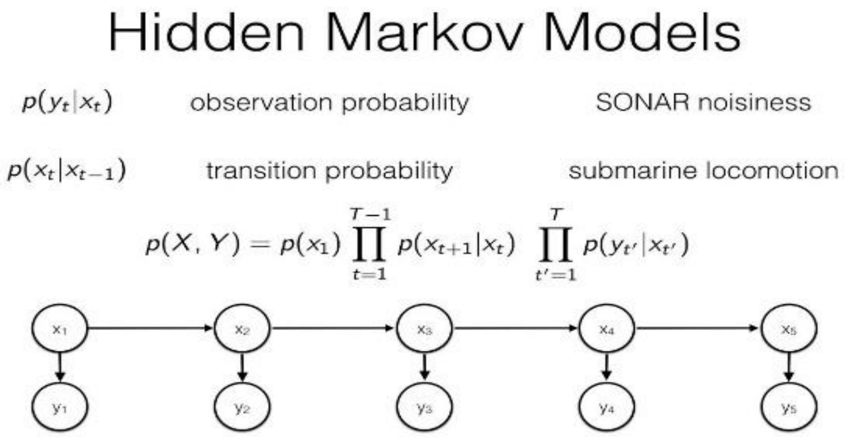


Fig 2.1: HMM

There are some drawbacks like here there are many unstructured parameters and they cannot express the dependencies among the hidden states and are unable to capture the higher order information. To overcome these drawbacks CRF came into existence.

2.2 CRF

CRF means Conditional Random Field is a statistical modeling method. This method is widely used in the pattern recognition and also for structure recognition. CRF comes under the category of sequence modeling. CRF is a type of discriminative undirected probabilistic graphical model. This method is used to encode the relationships between each of the observations. And CRF is used for constructing the stable interpretations and these method is used for labeling the data and also for the parsing of the sequential data. Especially the CRF’S is used in the POS tagging and also in the name entity recognition (NER) as well as parsing and also for the shallow parsing. CRF consists of two layer input layer and the output layer.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 8, Issue 4, 2019

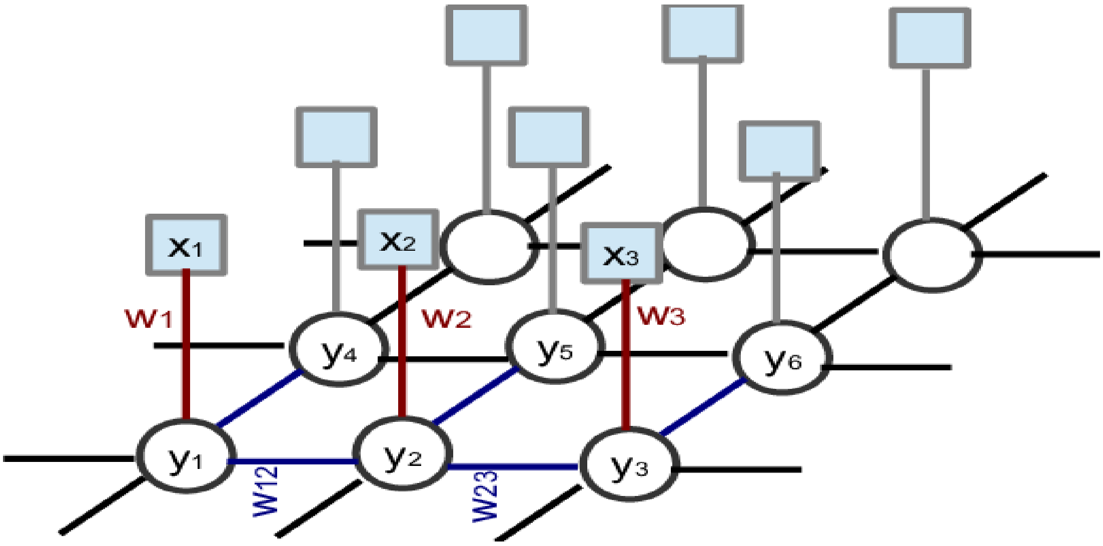


Fig 2.2: CRF

2.3 SVM

The accumulation of experience and estimated memory of the job forecast model based on different weather types are still used in weather forecast work. The accumulation of forecasting expertise is a long-term process, and climate evolution's complexity and non-linearity makes it difficult to define forecasting information. The evolution of any climate or meteorological variable is basically the result of a combination of certain meteorological elements ' conditional combinations, and the combination of these factors is varied and complex. Smart machine recognition skills have been well developed to communicate complex nonlinear relationships between meteorological elements in real time and space with advances in computer technology and smart machines. Support Vector Machine (SVM) is commonly used for many problems of machine learning. Support Vector Machine is one of the key multi-layer feed forward network classification. Like multi-layer perceptrons and radial function networks, support vector machines can be used for identification of patterns and non-linear regression. The SVM is used for performing the following functionalities

a) Better potential for generalization than other NN models.
b) The SVM solution is the same, efficient and absent from local minima.
c) Used for non-vectorial data.

The main purpose of the SVM algorithm is to scan in an N-Dimensional space for a hyperplane so that a collection of data points can be classified. We are given a set of data points in an SVM to be classified, but it is quite difficult to find the right hyperplane for the test. There are plenty of hyperplanes to choose from. If the wrong hyperplane is picked, however, the results may be disastrous due to the incorrect classification of the training data. The hyperplane now operates by considering the total hyperplane range, which is basically the largest distance from the data points. A linear curve that classifies data points based on where they lie on the hyperplane can be basically visualized to the SVM algorithm. The downside with SVM is that outlier datapoints can seriously impede the complete algorithm's accuracy rate.

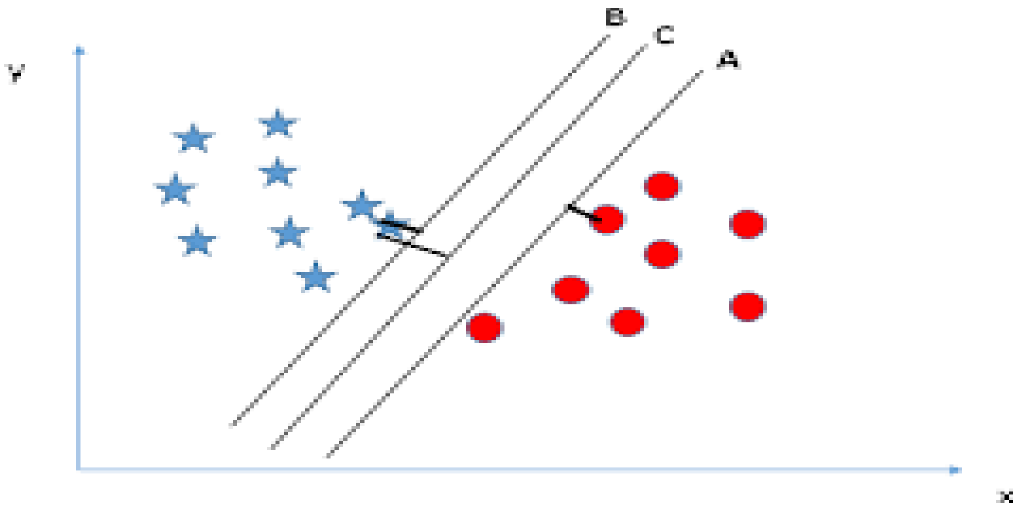


Fig 2.3: SVM

3. Theoretical Analysis

In this Literature survey eight papers are read and compared several techniques present in them published by different authors. Traditional techniques and also the deep learning techniques are discussed in those papers there are advantages and also disadvantages present in those techniques. The techniques used are SVM, CRF, BI-LSTM, CNN etc. Some of the papers state that future scope. Many papers are discussing that deep learning techniques are more efficient when compared to the traditional methods. Accuracy is also more for the deep learning techniques only. Deep learning techniques are CNN, RNN, BI-LSTM etc.

4. Experimental Investigation

4.1 Algorithm

Step 1: Reading the dataset

Step 2: Pre-processing techniques

Step 3: If there are Null Values, fill those values By preceding values

Step 4: Segmentation

Step 5: Tokenization

Step 6: Splitting the dataset into train and test sets

Step 7: Applying the algorithms to the trained and testing modules

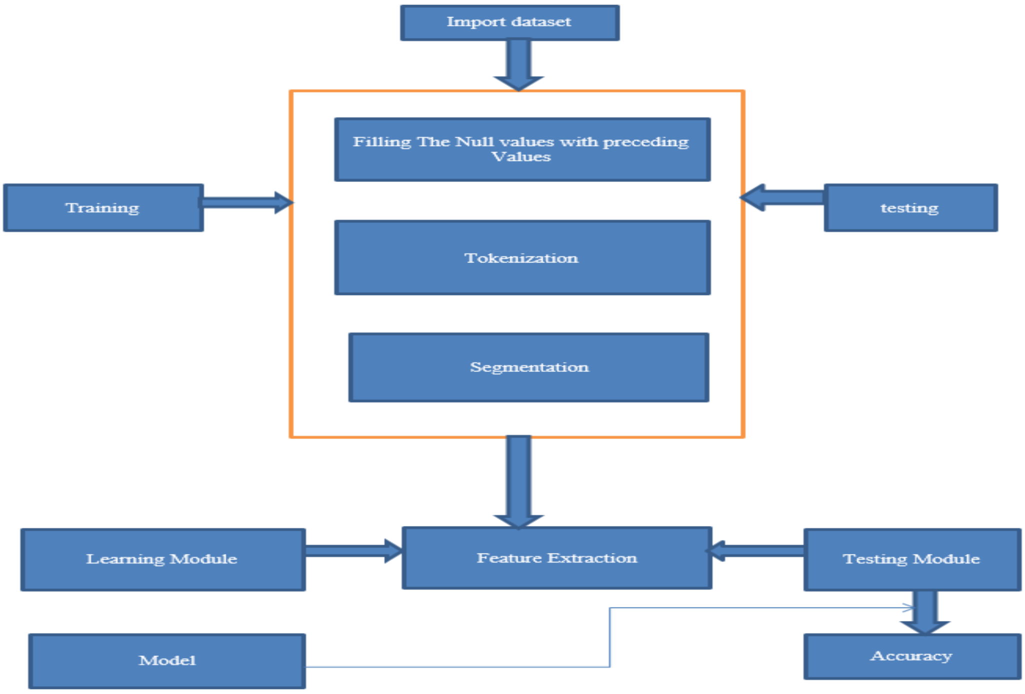
Step 8: Apply the CRF, BI-LSTM CNN on the dataset

Step 9: Train a CRF and also the BI-LSTM CNN model using Sklearn Crfusuite on dataset

Step 10: Feature extraction most of the features are like word parts, simplified POS tags, lower, title, upper flags, features of nearby words

Step 11: Evaluated the model by observing the scores of Precession, Recall, F1-Score.

4.2 Flow chart



4.3 Proposed Methods

4. 3.1 Bi-LSTM

This method is a combination of 2 techniques. They are LSTM and Bi-RNN (Bi-directional recurrent neural networks). Bi-LSTM is an advancement or special development of the Artificial Neural networks (ANN). This method came into existence because in the previous methods for larger sequence of the data the traditional methods are not suitable to solve the problems so to overcome this disadvantage this method is used[12]. And also RNN is not supportable for this longer sequence of data. Bi-LSTM is supportable for the longer sequence of data and it is advancement to the RNN. However, these Bi-LSTM consists of three layers Input layer, hidden layer, Output layer. Bi-LSTM can do both forward and backward operations. So this part becomes the main advantage in this model. Long short term model is capable of storing the information for longer period of time. And also it stores the previous state information and sends to the next layer in a sequence format. If the current layer wants the previous information this BI-LSTM is used. In this BI-LSTM there will be forward hidden layer and also the backward hidden layer. Which helps us in storing the past information of the input layer? Nodes will be present in each layer of the Bi-LSTM. The nodes in the hidden layer are connected this is how the information will be stored in the layers.

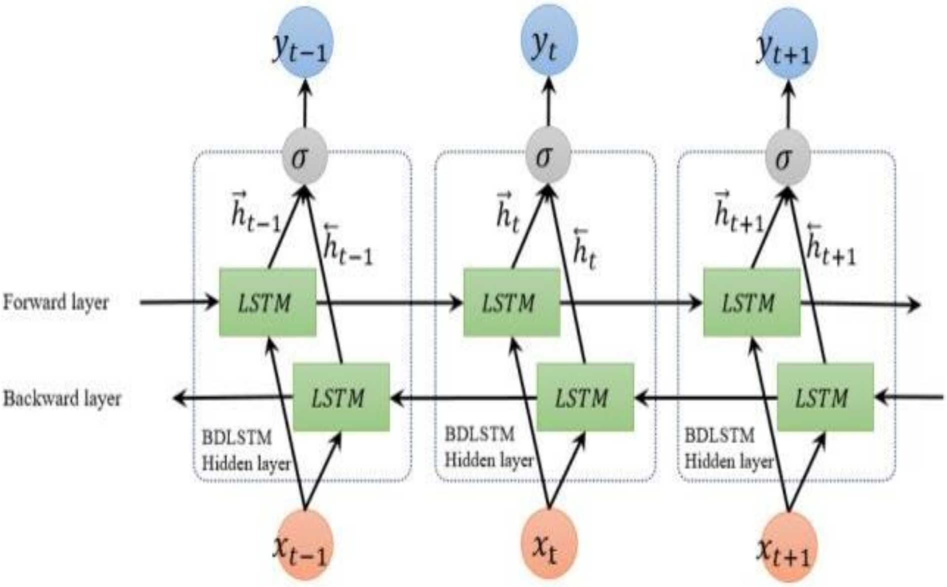


Fig 4.1: Bi-Lstm

4. 3.2 CNN

This one type of artificial neural network(ANN). This method is used in pattern recognition and also used in the sequence labeling part. As this method consists of 3 layers they are Convolution layer and the fully connected layer and also the output layer [14]. This method is mainly used in the name entity recognition(NER). The main task is that it detects the words or the characters present in the text. So by using this BI-LSTM method we can correctly characterize the words and also the characters into the levels format.In this CNN back propagation technique is also available. This becomes the main advantage to the CNN. After getting the result to the output layer. We will check the result and compare with the actual result if the result does not match with the actual result then by using back propagation method we will update the values and we will try to achieve the better and the accurate results [15].

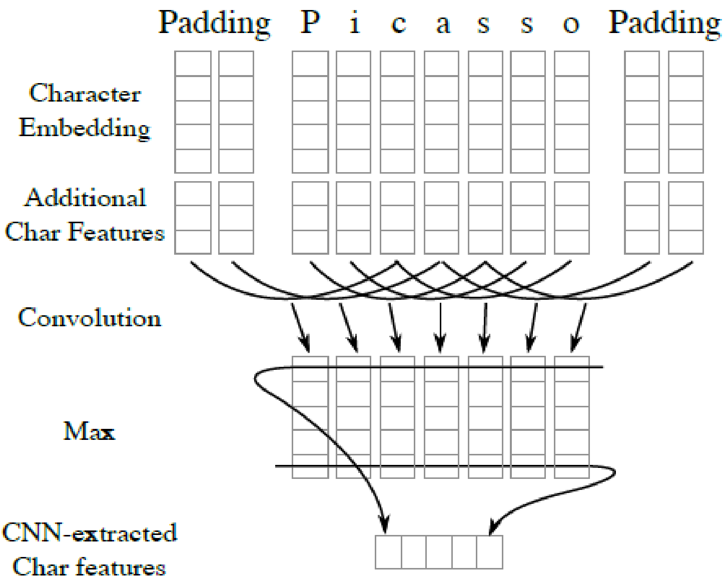


fig 4.2: CNN

4.3.3 Bi-LSTM-CNN

It is the combination of both the neural network models can use for getting better accuracy in NER. It is the hybrid model because it is the combination of both bi-directional LSTMs and CNNs. These models learns about both word level and character level features. This model is inspired by the work of Colobert et al.(2011). Here the lookup tables will transform the discrete features like characters and words into contionous vector representations and then these are concatenated and fed into a BiLSTM network. After that for inducing character level features use a CNN and then at the output layers it decodes the output for each category into a score.

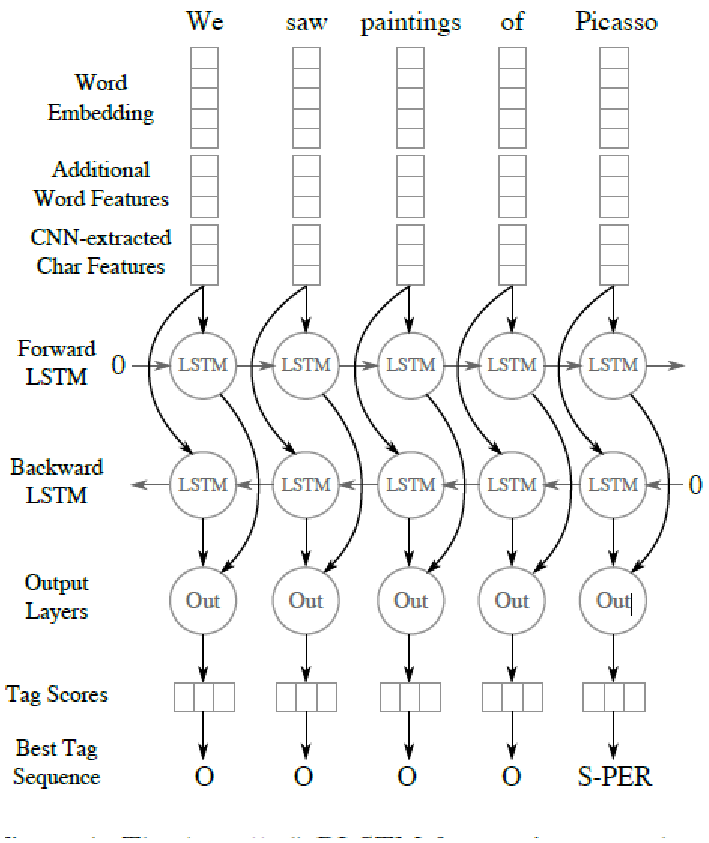


Fig 4.3: BI-LSTM-CNN

The above figure explains how the Bi-LSTM used for named entity recognition. CNN character features are extracted and by using BiLSTM the output sequence and the performance scores are obtained.

5. Conclusion

In this paper we have compared the two models. This comparison is done to find out the accuracy among the methods. So in this paper the techniques what we compared are Bi-LSTM CNN and CRF models. BI-LSTM CNN got more accuracy when compared to CRF. After getting the results this paper finally concludes that deep learning methods are more suitable for sequence labelling because the accuracy is more to the deep learning methods when compared to the traditional methods.

REFERENCES

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In COLING, pages 1638–1649, 2018.

[2] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. arXiv preprint arXiv:1603.06042, 2016. 14

[3] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. The annals of mathematical statistics, 37(6):1554–1563, 1966.

[4] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. Sub-event detection from twitter streams as a sequence labeling problem. NAACL, 2019.

[5] Oliver Bender, Franz Josef Och, and Hermann Ney. Maximum entropy models for named entity recognition. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, pages 148–151. Association for Computational Linguistics, 2003.

[6] Gaabor Berend. Sparse coding of neural word embeddings for ´ multilingual sequence labeling. Transactions of the Association for Computational Linguistics, 5:247–261, 2017.

[7] Daniel M Bikel, Richard Schwartz, and Ralph M Weischedel. An algorithm that learns what’s in a name. Machine learning, 34(1-3):211– 231, 1999.

[8] Bernd Bohnet, Ryan McDonald, Goncalo Simoes, Daniel Andor, Emily Pitler, and Joshua Maynez. Morphosyntactic tagging with a meta-bilstm model over context sensitive token encodings. ACL, 2018.

[9] Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. Adversarial transfer learning for chinese named entity recognition with self-attention mechanism. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 182–192, 2018.

[10] Hui Chen, Zijia Lin, Guiguang Ding, Jianguang Lou, Yusen Zhang, and Borje Karlsson. Grn: Gated relation network to enhance convolutional neural network for named entity recognition. AAAI, 2019.

[11] Luoxin Chen, Weitong Ruan, Xinyue Liu, and Jianhua Lu. Seqvat: Virtual adversarial training for semi-supervised sequence labeling. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8801–8811, 2012.

[12] Hai Leong Chieu and Hwee Tou Ng. Named entity recognition: a maximum entropy approach using global information. In Proceedings of the 19th international conference on Computational linguisticsVolume 1, pages 1–7. Association for Computational Linguistics, 2002.

[13] Jason PC Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. Transactions of the Association for Computational Linguistics, 4:357–370, 2016.

[14] Kyunghyun Cho, Bart Van Merrienboer, Dzmitry Bahdanau, and ¨ Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259, 2014.

[15] Kevin Clark, Minh-Thang Luong, Christopher D Manning, and Quoc V Le. Semi-supervised sequence modeling with cross-view training. EMNLP, 2018.

[16] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 1–8. Association for Computational Linguistics, 2002.

[17] Ronan Collobert, Koray Kavukcuoglu, Jason Weston, Leon Bottou, Pavel Kuksa, and Michael Karlen. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(1):2493–2537, 2011.

[18] Leyang Cui and Yue Zhang. Hierarchically-refined label attention network for sequence labeling. EMNLP, 2019.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[20] Timothy Dozat, Peng Qi, and Christopher D Manning. Stanford's graph-based neural dependency parser at the conll 2017 shared task. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages 20–30, 2017.

[21] Sean R Eddy. Hidden markov models. Current opinion in structural biology, 6(3):361–365, 1996.

[22] Xiaocheng Feng, Xiachong Feng, Bing Qin, Zhangyin Feng, and Ting Liu. Improving low resource named entity recognition using crosslingual knowledge transfer. In IJCAI, pages 4071–4077, 2018.