OD-LSTM: An Efficient Network For Named Entity Recognition

Junxiang Ge, Qi Zhang, Haizhou Zhao
AI Research Department, Sogou
Beijing, China

Abstract

With the risen use of AI application, we need to understand some specific type of message from user's query. Named Entity Recognition(NER) is a basic task for that use. Gazzete is very useful when doing NER, howerver, most of the current method use iit in word embedding format, which will bring out the case, we need to retrain model when the word in gazzete is greatly changed. In this paper, we propose a conception, called Outer Dictionary LSTM (OD-LSTM), which means to use the outer type info of gazzete for each word, without using the specific word. Experiments shows it useful when doing NER task, and acheive a great improvement on task with which gazzete is heavily depended on.

Introduction

Named Entity Recognition (NER) is a base task when doing many Natural Language Processing (NLP) tasks. Usually, we need to extract mainly information from users' sentences to help us understand their meanings, and then use the extracted information to do further process. For example, in weather domain, we may need to extract **date and position** to help us understand when and where the user is asked towards the weather; in music type service, we may need to track **singer, song, style** entities to suit user's command; for many commom use, we need **person, location, organization** message to understand a sentence.

To accomplish this task, many method has been explored. Convolution Neural Network (CNN) is used in Image Classification (Krizhevsky, Sutskever and Geoffrey 2012), and then proved also powerful in Text Classification (Kim 2014). CNN is also useful in NER tasks, whether to capture character-level features (Chiu and Nichols 2016; Ma and Hovy 2016; Peters et al. 2017), or to do word-level sequence labelling (Santos, Xiang, and Zhou 2015; Strubell et al. 2017). To learn the correlation between words in sentence, Recurrent Neural Network (RNN) (Sutskever, Vinyals and Le 2014) and Long Short-Term Memory (LSTM) (Sak, Senior and Beaufays 2014) are proposed, and they achieve great success on machine translation area and so on. They are then used in NER tasks and achieve significant improvement (Chiu and Nichols 2016; Ma and Hovy 2016; Peters et al. 2017). These structures can be use in different situations (Yang, Liang, and Zhang 2018). As the depth of Nerual Network (NN) usually improve the final performance, more and more complicated method has been proposed in recent years. Transformer structure (Vaswani et al. 2017) and its derivative methods, like OpenAI GPT (Radford et al. 2018), BERT (Devlin et al. 2018), XLNet (Yang et al. 2019), make innovative attempt to NN structures.

Among these years, many method focus on base layer of NN structure, usually the embedding input layer, and then directly use it to do NLP tasks (Peters et al. 2018; Radford et al. 2018; Devlin et al. 2018; Yang et al. 2019). Though will it make improvement, it usually pays a long time to train the model. But in these way, it tell us that embedding input can directly infect the result of NLP tasks.

There exists a little difference between English NER tasks and Chinese NER tasks. English word is composed by acharacters, for which character level CNN can help improve NER tasks (Chiu and Nichols 2016; Ma and Hovy 2016; Peters et al. 2017; Yang, Liang, and Zhang 2018); but that can not be use in Chinese NER tasks, which is Hieroglyphics. For this reason, Chinese NER tasks is usually done by word baseline or character baseline structure. Word baseline structure refer to using word segmentor to split sentence into word tokens and then use their word vector (Peng and Dredez 2015; Peng and Dredez 2016); while character structure simply treat each sentence into combination of single Chinese character, and then using character embedding to do NLP tasks (Zhang and Yang 2018; Gui et al. 2019). In these methods, it seems useful when using word embedding within character baseline structure, or the oppisite.

Most of state-of-the-art NER methods uses Conditional Random Fields (CRF) (Lafferty, McCallum, and Pereira 2011) to decode the sequence labelling result. CRF can work even without any NN cells. Hand-writing trait can sometimes already achieve high performance (Finkel, Grenager, and Manning 2005; Okazaki 2007).

In some situations, word backgound knowledge, namely dictionary (the same meaning as lexicon and gazette), is neccessary to do NER jobs. For example, in music service case, You Raise Me Up" would be treated more likely as a chat information rather than a song, unless the model is given that knowledge. Most of the state-of-the-art methods using word embedding to complete these task. But it then gives two problems:

- Out Of Vacabulary (OOV) problem. As for **song** entity in music service, word embedding size will be too large; and sometimes will the length of song be too long to make word vector.
- Retrain problem. When the content of dictionary changes abundantly, the trained model would have poor performance doing the same jobs. Still use the **song** entity as an example, the dictionary would be changed monthly or even daily, thus we need to retrain our model synchronously.

Faced with this problem, we proposed a new conception, called Outer Dictionary (OD) models, which means using the dictionary's outer trait, without using its specific content. For example, we use the word type " song ", rather than a specific song name, to help model understand the outer information. In these way, word embedding is no longer needed for Chinese NER task. Exprement result show that in gazzete depended-on task, OD model gives the best result among state-of-the-art method. And when the content of dictionary changed, the model can synchronously update its result without retraining.

Contributions

In summary, we make these contributions:

- We proposed a new conception: Outer Dictionary (OD) model, which is used to update the content of dictionary without retraining the model.
- In order to realize OD model, we explored Tag Embedding LSTM (TE-LSTM), using the word type, rather the word itself, to improve the performance of NER tasks.
- Experiment result shows our model gives significant improvements on gazzete depended-on tasks among state-of-the-art models.
- We release the music service dataset, which we use in our paper, for further scientific research.

Related Work

Success of BERT (Devlin et al. 2018) and XLNet (Yang et al. 2019) show that word embedding is important and useful for many upper NLP tasks. More useful information is there in the vector, more efficient the model would work. To a great degree, embedding layer decides the performance. Same conclusion can be made from a lot of NER models, as for character level message, word level message, position message, handwriting message and so on. Especially in some specific NER task, like music service, gazzete is heavilly depended on to make better performance.

Most of the recent models use gazette as word embedding input, or its string format, to give the model related message. Lattice LSTM (Zhang and Yang 2018) use the embedding as additional input to character baseline LSTM, which can give the model more sentence information. Lexivon Rethinking CNN (Gui et al. 2019), corparate the word embedding vector to correlated CNN layer, and the coflict word segment could be learned by rethinking. Stanford NER (Finkel, Grenager, and Manning 2005) can use the gazette information, as string format, in CRF model.

However, in some pratical use, the content of the gazette will change greatly and constantly. In this way, most of the recent model need to be retrained after every abundantly change of gazette. To solve this problem, we need to construct a model which can still be used after such change. We then came out with the OD models, and use TE-LSTM to realize these model. As a result, with no more word embedding is needed, and model needed only to be trained once. What we need is the word tag information to help the model better understand user's intention.

Outer Dictionary LSTM

Character-Base LSTM

pass

Tag Embedding Model

pass

Experiments

pass

Conclusion

pass

Acknowledgments

Especially grateful to Can Cui for denoting the music ner data, to Jindou Wu for advice on data processing of CCKS 2018 music dataset.

References

Krizhevsky, A.; Sutskever, I.; and Geoffrey E. H. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of Advances in Neural Information Processing Systems 25 (NIPS 2012).

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. arXiv:1408.5882.

Sutskever, I.; Vinyals, O.; and Le, Q.V. 2014. Sequence to Sequence Learning with Neural Networks. In Proceedings of Advances in Neural Information Processing Systems 27 (NIPS 2014).

Sak, H.; Senior, A.; and Beaufays, F. 2014. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. arXiv:1402.1128.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research 12 (2011) 2493-2537.

Santos. C. N.; Xiang, B.; and Zhou, B. 2015. Classifying Relations by Ranking with Convolutional Neural Networks. arXiv:1504.06580.

Strubell, E.; Verga, P.; Belanger, D.; and McCallum, A. 2017. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. arXiv:1702.02098.

Peters, M. E.; Ammar, W.; Bhagavatula, C.; and Power. R. 2017. Semi-supervised sequence tagging with bidirectional language models. arXiv:1705.00108.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. arXiv:1802.05365.

Ma, X.; and Hovy, E. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. arXiv:1603.01354.

Yang, J.; Liang, S.; and Zhang, Y. 2018. Design Challenges and Misconceptions in Neural Sequence Labeling. arXiv:1806.04470.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, Open AI.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv:1906.08237.

Sang, E. F.; and Meulder, F. D. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. arXiv:cs/0306050.

Chiu, J. P.C.; and Nichols, E. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, Volume 4, 2016, p.357-370.

Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv:1508.01991.

Peng, N.; and Dredze, M. 2016. Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning. arXiv:1603.00786.

Chen, Q.; Zhuo, Z.; and Wang, W. 2019. BERT for Joint Intent Classification and Slot Filling. arXiv:1902.10909.

Zhang, Y.; and Yang, J. 2018. Chinese NER Using Lattice LSTM. arXiv:1805.02023.

Gui, T.; Ma, R.; Zhang, Q.; Zhao, L.; Jiang, Y.; and Huang, Y. 2019. CNN-Based Chinese NER with Lexicon Rethinking. In Proceedings of the International Joint Conference on Artificial Intelligence 2019 (IJCAI 2019).

Lafferty, J.; McCallum, A.; and Pereira, F. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001), pages 282-289.

Socher, R.; Manning, C. D.; and Andrew Y. Ng. 2010. Learning Continuous Phrase Representations and Syntactic Parsing with Recursive Neural Networks. In Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop.

Finkel, J. R.; Grenager, T.; and Manning, C. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370.

Okazaki, N. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). URL http://www. chokkan. org/software/crfsuite.

Peng, N.; and Dredze, M. 2015. Named Entity Recognition for Chinese Social Media with Jointly Trained Embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS.

Mikolov, T.; Yih, W.; and Zweig, G. 2013. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL HLT.

Pennington, J.; Socher, R.; and Manning, C.D. 2014. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 15321543.

Mikolov, T.; Grave, E.; Bojanowski, P.; Puhrsch, C.; and Joulin, A. 2017. Advances in Pre-Training Distributed Word Representations. arXiv:1712.09405.