

Neural Semantic Role Labeling

Haonan Li

Student Number: 955022

haonanl5@student.unimelb.edu.au

Introduction

Semantic Role Labeling (SRL) is a basic Natural Language Processing task that aims at recovering the structure of predicted-argument in a sentence. Specifically, to determine essentially “who did what to whom”, “when” and “where”. This task gradually draws more attention because semantic roles provide the intermediate level of semantic representation of a sentence so that benefits many NLP tasks such as Question Answering (Moschitti et al. 2007), text classification and Machine Translation (Liu and Gildea 2010; Bazrafshan and Gildea 2013).

For a given sentence, SRL is to extract all arguments of a target verb and classify them to their semantic roles. For example, for the sentence “Harry lent a pen to Rone last month” and the target verb is *lent*, SRL yields the following outputs:

[ARG0 Harry] [V lent] [ARG1 a pen] [ARG2 to Rone] [AM-TMP last month].

Here, ARG0 represents the *lender*, ARG1 represents the *thing lent*, ARG2 represents the *entity lent to*, AM-TMP is an adjunct indicating the *timing of the action* and V represents the *verb*.

SRL attracts attention since the CoNLL-2004 Share Task¹, in CoNLL-2005 (Carreras and Màrquez 2005), the training set is enlarged, task definition becomes more clearly and evaluation methods are more reasonable. The CoNLL-2012 enlarged the data again and later works evaluation almost based on the two datasets of CoNLL-2005 and CoNLL-2012. The three papers we are going to introduce naturally reported their experiment results on these two datasets.

The remainder of this paper is structured as follows: Section 2 describes three latest SRL models. Conclusions and future work are in Section 3.

Models and Performances

In this section, we introduce three latest deep neural network models for semantic role labeling. Two of them (He et al. 2017) (Tan et al. 2018) treat SRL as a BIO tagging problem while the last one (He et al. 2018) is a span-based model

which can jointly predict all predicates and arguments spans and relations.

Deep BiLSTM Model

(He et al. 2017) propose an 8 layers bidirectional long short term memory neural network (BiLSTM) model with highway connections, RNN-dropouts and constrained decoding. This is the first deep neural model for SRL that achieves state of the art performance. In their paper, they not only provide detailed descriptions of the model and inspiration of the design, but also give extensive analyses of the common questions of deep neural networks on the particular task. The comprehensive analysis play a guiding role in subsequent researches and analyses.

Their model performs good is mainly due to the following points. First, they apply the latest advances in training deep LSTMs such as highway connections (Srivastava, Greff, and Schmidhuber 2015) and RNN-dropouts (Gal and Ghahramani 2016), which makes their model successfully captures long-distance dependencies and reduce over-fitting. Second, they use constrained A* decoding algorithm to reduce the computational complexity.

Another main contribution of the paper is the analyses of the deep model from 4 aspects:

- (a) **Error Analysis** They indicate that their model detects more arguments compared with two non-neural baseline models although the final labeling errors they made are similar. They think the reason may be that parsing processing makes some arguments hard to recover for syntax-based model.
- (b) **Long-range Dependencies** They further analyzed the different models’ ability to capture long-range dependencies by setting different distances of semantic dependence. The results indicate that deeper networks always perform better than shallow networks, such gaps become more obvious as the dependency distances become larger.
- (c) **Structural Consistency** This paper enforces some structural consistencies to the deep BiLSTM model though the result is not perfect. They summarize two reasons. First, the number of violations of the consistency is relatively small so that the tricks they used is only applied to small number of cases. Second, the gold SRL structures do not

¹The official CoNLL-2004 shared task web page <http://www.lsi.upc.edu/srlconll>

Model	CoNLL-2015				CoNLL-2012			
	P	R	F1	Comp.	P	R	F1	Comp.
(Pradhan et al. 2013)	–	–	–	–	78.5	76.6	77.5	55.8
(Punyakanok et al. 2008)	82.3	76.8	79.4	53.8	–	–	–	–
(He et al. 2017)	85.0	84.3	84.6	66.5	83.5	83.3	83.4	68.5
(Tan et al. 2018)	85.9	86.3	86.1	69.0	83.3	84.5	83.9	69.3
(He et al. 2018)	–	–	87.4	–	–	–	85.5	–

Table 1: Comparison of three models with two baseline models on CoNLL-2005 dataset and CoNLL-2012 dataset. The results are reported in terms of precision (P), recall (R), F1 and percentage of completely correct predicates (Comp.)

completely obey the rules they set so that performance will be hurt due to the enforcement of consistencies.

- (d) **Syntax Assistance** They argue that though deep neural model can learn syntax by itself, joint training with gold syntactic tree benefits more to the models.

Self-Attention Model

(Tan et al. 2018) propose a deep attention neural network model which takes advantages of both the high performance of deep neural networks and the ability of the attention model to draw global dependencies. They argue that recurrent neural network (RNN) based models have two main limitations though such model achieves success in recent years. The first is such model requires large memory capacity due to the storage of the entire history of sentence information. The second intrinsic drawback is that sequential processing of the input sentences is hard to capture tree-like syntactic structures.

Therefore, they propose an attention model that can address these problems. The most novel thing about their model is the use of self-attention layer that uses a multi-head scaled dot product attention mechanism. This kind of design is a combination of standard attention model (Lin et al. 2017) and multiplicative attention model (Luong, Pham, and Manning 2015). Different from both of them, their attention mechanism allows faster computation by using matrix production.

This paper illustrates the advantages of self-attention model compared with traditional RNN and CNN models through experiments and further explanations, which is worth learning to other tasks and further improvements. First, different from RNN whose distance between two input is n , the input elements’ distance is 1. Second, the gradient propagation is much easier than standard attention mechanism because of the uses of weighted sum of product output vectors. Last but not least, the product is parallel computable which makes it compute faster than RNNs.

Besides, authors have done many relevant experiments that give us more guidances and inspirations in our future work. They employ three kinds of non-linear sub-layers to transform the raw input to attention layer’s input, include recurrent LSTMs, convolutional Gated Linear Unit (GLU) (Dauphin et al. 2017) and feed-forward ReLU sub-layers. They also verify increasing model depth can improve the performance for a relatively shallow network. Furthermore, they demonstrate the influence of various model width, us-

ing pre-trained word embedding (Pennington, Socher, and Manning 2014), position encoding and so on.

Span-based Joint Model

(He et al. 2018) proposed another end-to-end model which can jointly predict all predicates, arguments spans and relations between them. Different from BIO-based models (He et al. 2017)(Tan et al. 2018) that incurs error-propagation because of pipelined processing, span-based graph model can predict all needed result in one forward pass. Compared with predict relations with given chunks, jointly predict arguments and relations is based on a more realistic setting. Which makes it more and more popular in recent years in many tasks such as knowledge extraction (Nguyen, Cho, and Grishman 2016; Zheng et al. 2017) and parsing (Swayamdipta et al. 2016; Srivastava, Labutov, and Mitchell 2017).

Similar with two previous works, (He et al. 2018) use deep BiLSTM neural architecture together with attention architecture. They also use latest contextualized word representations ELMo (Peters et al. 2018) to further improve the performance. The result shows that ELMo is helpful in this sequence processing task.

Results

All of above papers measured the performance of their system on two datasets: CoNLL-2005 (Carreras and Màrquez 2005) and CoNLL-2012 (Pradhan et al. 2013) that are extracted from Wall Street Journal (WSJ) corpus and OneNotes v5.0 corpus respectively. The results are shown in Table 1. From the table we can easily find that three reported deep neural models perform much better than other two baseline models. Among them, the self-attention model performs better than pure LSTM layers model while span-based attention model performs best.

Conclusion and Future work

As computational capability improve, deep learning model is now becoming more and more feasible and popular. Through all three models’ performance we conclude the following conclusions. First, deeper networks with reasonable structures usually perform better than shallow networks. Second, adding attention mechanism in recurrent neural network can effectively reduce long dependencies lost though

it is a very simple structure. Third, applying the latest advances such as highway connections and recurrent dropout to models may help.

In the future, We will build deep neural models on our Spatial Role Labeling task, which is an in-domain task with different difficulties with SRL. Many advances such as attention mechanism, highway connections will be test on our models to get best current results. In addition, we will do experiments on jointly tagging name entities and their relations using a graph model.

References

- Bazrafshan, M., and Gildea, D. 2013. Semantic roles for string to tree machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, (Volume 2: Short Papers)*, volume 2, 419–423.
- Carreras, X., and Màrquez, L. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning, CoNLL 2005, Ann Arbor, Michigan, USA, June 29-30, 2005*, 152–164.
- Dauphin, Y. N.; Fan, A.; Auli, M.; and Grangier, D. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 933–941.
- Gal, Y., and Ghahramani, Z. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, NIPS 2016, December 5-10, 2016, Barcelona, Spain*, 1019–1027.
- He, L.; Lee, K.; Lewis, M.; and Zettlemoyer, L. 2017. Deep semantic role labeling: What works and whats next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 473–483.
- He, L.; Lee, K.; Levy, O.; and Zettlemoyer, L. 2018. Jointly predicting predicates and arguments in neural semantic role labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, 364–369.
- Lin, Z.; Feng, M.; dos Santos, C. N.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017. A structured self-attentive sentence embedding. *Computing Research Repository, CoRR 2017 abs/1703.03130*.
- Liu, D., and Gildea, D. 2010. Semantic role features for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 716–724.
- Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, 1412–1421.
- Moschitti, A.; Quarteroni, S.; Basili, R.; and Manandhar, S. 2007. Exploiting syntactic and shallow semantic kernels for question answer classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, ACL 2007, June 23-30, 2007, Prague, Czech Republic*, 776–783.
- Nguyen, T. H.; Cho, K.; and Grishman, R. 2016. Joint event extraction via recurrent neural networks. In *NAACL 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics, San Diego California, USA, June 12-17, 2016*, 300–309.
- Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP 2014*, 1532–1543.
- Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, 2227–2237.
- Pradhan, S.; Moschitti, A.; Xue, N.; Ng, H. T.; Björkelund, A.; Uryupina, O.; Zhang, Y.; and Zhong, Z. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, 143–152.
- Punyakanok, V.; Roth, D.; Yih, W.; and emmm. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics* 34(2):257–287.
- Srivastava, R. K.; Greff, K.; and Schmidhuber, J. 2015. Training very deep networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, NIPS 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 2377–2385.
- Srivastava, S.; Labutov, I.; and Mitchell, T. M. 2017. Joint concept learning and semantic parsing from natural language explanations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, 1527–1536.
- Swayamdipta, S.; Ballesteros, M.; Dyer, C.; and Smith, N. A. 2016. Greedy, joint syntactic-semantic parsing with stack lsm. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, 187–197.
- Tan, Z.; Wang, M.; Xie, J.; Chen, Y.; and Shi, X. 2018. Deep semantic role labeling with self-attention. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*.
- Zheng, S.; Wang, F.; Bao, H.; Hao, Y.; Zhou, P.; and Xu, B. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 1227–1236.