Technology Review: POS Tagging for Chinese

Tieh Chu (tiehchu2@illinois.edu)

CS410 Text Information Systems

November 7, 2021

1 Introduction

Part-of-speech (POS) tagging is the process of annotating sentences with part-of-speech tags. It includes multiple sub-tasks since each sentence will have more than one word which is being tagged, so this task can be categorized into 3 sub-tasks named as tagging for single words, tagging for phrases, and part of larger units such as clauses or simple prepositional phrases.

In this review, we want to narrow down our topic to Chinese POS tagging. Chinese POS tagging needs to segment each word apart beforehand. There are approaches that perform word segmentation and POS tagging sequentially, and there are also other approaches that perform these two steps simultaneously. Previous approaches indicated that combining them into a joint task is proved to show better accuracy since POS tags provide information for word segmentation and the outcome of word segmentation also affects POS tagging.

As a result, we will discuss two joint segmentation and POS tagging (perform steps simultaneously) approaches for Chinese. The rest of the report is organized as follows: section 2 gives an overview of the two methods. Section 3 compares the two approaches and analyzes the experimental results. Lastly, section 4 summarizes the review with concluding remarks.

2 Overview of the Approaches

In this section, we will briefly look at the two different approaches and summarize their core values.

1. Bidirectional RNN-CRF [3]
   This approach uses a bi-directional RNN-CRF model with character-level feature vectors to achieve POS tagging. Besides using the pre-trained character embeddings, they represent a sequence of characters as n-grams vectors and incrementally concatenated these vectors in a given context (Figure 1), which shows great improvements over baseline models. Lastly, they further enhance their model by incorporating sub-character level information such as radicals and orthographical features extracted by CNN (Figure 2).

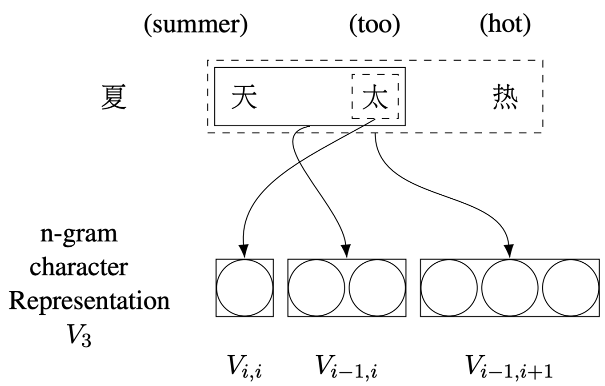

Figure 1: Chinese characters are represented as n-gram vectors.
These vectors are incrementally concatenated in a given context.

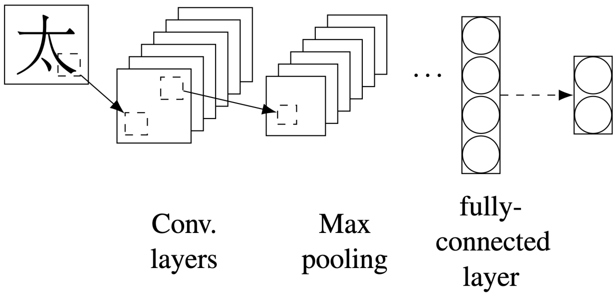

Figure 2: Convolutional Neural Networks are utilized to extract orthographical features. Chinese characters are logograms which contains rich information in the graphical components.

In summary, they believe that (1) instead of relying solely on contextual features extracted from context-free character representations, local information (which will be represented as n-grams vectors) should also be considered. (2) There contains a lot of information (radicals and character figures) in the graphical components of Chinese characters.

2. Two-way Attentions of Auto-analyzed Knowledge [4]
This approach adopts a two-way attention mechanism into their neural network. For each input Chinese character, the first-way attention module gets knowledge instances from corresponding auto-analyzed results, and the second-way attention module extracts the content features of the input Chinese character (Figure 3). The results are then concatenated and fed into their neural network to further guide the tagging process.

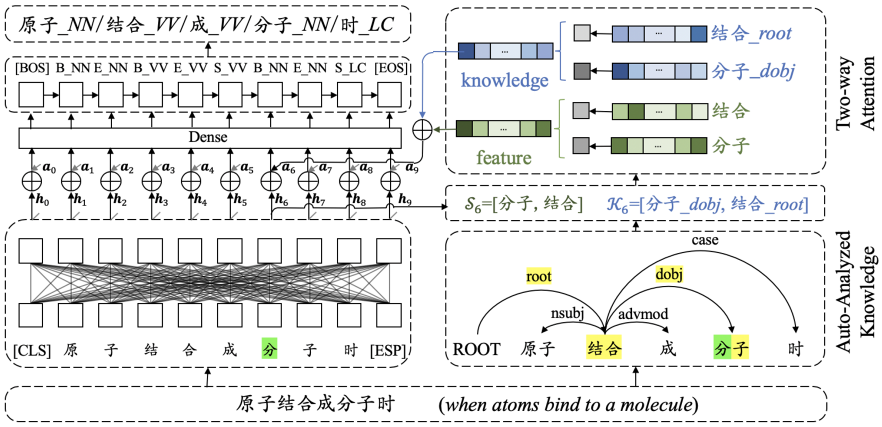

Figure 3: The architecture incorporates the two-way attention mechanism, which is presented with example context features and their dependency knowledge (highlighted in yellow) from auto-analyzed results for a character (highlighted in green) in the given sentence.

In conclusion, they believe that the resulted knowledge instances from auto-processed corpora are not accurate in most cases thus should be wisely used instead of fully relying on them. Hence the two-way attention guides the neural network to distinguish the effectiveness of auto-analyzed knowledge.

3 Comparison of the Approaches

This section provides comparisons of the above two approaches by pointing out the similarity and differences followed by analyzing the pros and cons of their methods.

Firstly, although the above two methods both value additional information. The first method is more dedicated to exploring the additional useful information. They investigate on (1) radicals and (2) orthographical features. Radicals are shared parts of Chinese characters which serve as semantic indicators, that is to say, characters including the same radicals are related. For example, 針(needle), 銀(silver), 鐵(iron) share same radical 金(gold). Furthermore, since Chinese characters are logograms, the graphical features are investigated too by using convolution neural networks (CNNs) to extract additional information. In comparison, the second method does not further discuss sub-character level information. It only utilizes auto-analyzed knowledge of the existing models.

Secondly, the two methods both acknowledge that besides context-free Chinese characters, there is more useful information that should be considered. However, the previous method incorporates the additional information by simply concatenating them together. The second method points out that direct concatenation would be problematic given the fact that the knowledge gained from off-the-shelf toolkits is not always accurate. As a result, the two-way attention module gives guidance to the neural network so that the model is capable of judging inferior information.

To summarize, the pro of the first method is to gain a deeper knowledge of the feasibility of sub-character level information since it investigates and explores more. The con is that the information might not be useful or even harm the results since the method simply concatenated all the extracted information regardless of the quality. In contrast, the advantage of the second method is to avoid error-prone issues by inaccurate knowledge since the method can tell apart the important auto-analyzed knowledge based on their contribution. The disadvantage of the second method is that it lacks further understanding of sub-character level information since it simply utilizes auto-analyzed methods from existing models.

	CTB5		**CTB6**		**CTB7**		**CTB9**		**UD1**		**UD2**	
	Seg	Joint	Seg	Joint	Seg	Joint	Seg	Joint	Seg	Joint	Seg	Joint
Jiang et al. (2008)	97.85	93.41	-	-	-	-	-	-	-	-	-	-
Kruengkrai et al. (2009)	97.87	93.67	-	-	-	-	-	-	-	-	-	-
Sun (2011)	98.17	94.02	-	-	-	-	-	-	-	-	-	-
Wang et al. (2011)	98.11	94.18	95.79	91.12	95.65	90.46	-	-	-	-	-	-
Qian and Liu (2012)	97.85	93.53	-	-	-	-	-	-	-	-	-	-
Shen et al. (2014)	98.03	93.80	-	-	-	-	-	-	-	-	-	-
Kurita et al. (2017)	98.41	94.84	-	-	96.23	91.25	-	-	-	-	-	-
Shao et al. (2017)	98.02	94.38	-	-	-	-	96.67	92.34	95.16	89.75	95.09	89.42
Zhang et al. (2018)	98.50	94.95	96.36	92.51	96.25	91.87	-	-	-	-	-	-
BERT + POS (SCT)	98.77	96.77	97.43	94.82	**97.31**	94.12	97.75	94.87	98.32	95.60	**98.33**	95.46
ZEN + POS (SCT)	**98.81**	**96.92**	**97.45**	**94.87**	97.27	**94.20**	**97.77**	**94.88**	**98.33**	**95.69**	98.18	**95.49**

Figure 4: Comparison (in F-scores of word segmentation and joint tagging) of the second method with the first method on same datasets.

Figure 4 compares the first method (Shao et al. (2017)) and the second method (last two rows) on the same datasets - the Penn Chinese TreeBank [1] (CTB5 and CTB9) and the Chinese part of Universal Dependencies (UD) [2] (UD1 and UD2). We can see that the ability to distinguish auto-analyzed results performs better compared to treating the knowledge as gold references and blindly concatenating the information. The effectiveness of selectively modeling useful information outweighs the disadvantage that they do not have deeper sub-character level information, hence the second method is much better on all datasets.

4 Conclusion

This technology review targets part-of-speech (POS) tagging for Chinese. To be more specific, we narrow down to discuss two joint segmentation and POS tagging approaches. The first method [3] makes further investigation on sub-character level information such as radicals and graphical representations, whereas the second method [4] focuses more on distinguishing the information for the given auto-analyzed knowledge. Experimental results shown from the paper indicate that information from auto-processed corpora is often unreliable, so the ability to distinguish useful information is what should be taken into consideration for all future works.

5 Reference

[1] Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. Natural language engineering, 11(2):207–238.

[2] Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 1659–1666.

[3] Yan Shao, Christian Hardmeier, Jorg Tiedemann, and Joakim Nivre. 2017. Character-based Joint Segmentation and POS Tagging for Chinese using Bidirectional RNN-CRF. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 173–183, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[4] Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. Joint Chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge.In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8286–8296, Online, July 2020. Association for Computational Linguistics.

[5] Hwee Tou Ng and Jin Kiat Low. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 277–284, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[6] Xiong, Y., Wang, Z., Jiang, D. et al. A fine-grained Chinese word segmentation and part-of-speech tagging corpus for clinical text. BMC Med Inform Decis Mak 19, 66 (2019). https://doi.org/10.1186/s12911-019-0770-7