The 35th Pacific Asia Conference
on Language, Information and Computation

Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-stage Span Labeling

Authors: Duc-Vu Nguyen, Linh-Bao Vo, Ngoc-Linh Tran
Kiet Van Nguyen, Ngan Luu-Thuy Nguyen
Affiliations: University of Information Technology
Viet Nam National University Ho Chi Minh City

November 7, 2021

Outline

Outline

Introduction

1. Chinese word segmentation and part-of-speech tagging are necessary tasks in terms of computational linguistics and application of natural language processing.

Introduction

1. Chinese word segmentation and part-of-speech tagging are necessary tasks in terms of computational linguistics and application of natural language processing.
2. Many researchers still debate the demand for Chinese word segmentation and part-of-speech tagging in the deep learning era.

Introduction

1. Chinese word segmentation and part-of-speech tagging are necessary tasks in terms of computational linguistics and application of natural language processing.
2. Many researchers still debate the demand for Chinese word segmentation and part-of-speech tagging in the deep learning era.
3. Nevertheless, resolving ambiguities and detecting unknown words are challenging problems in this field.

Introduction

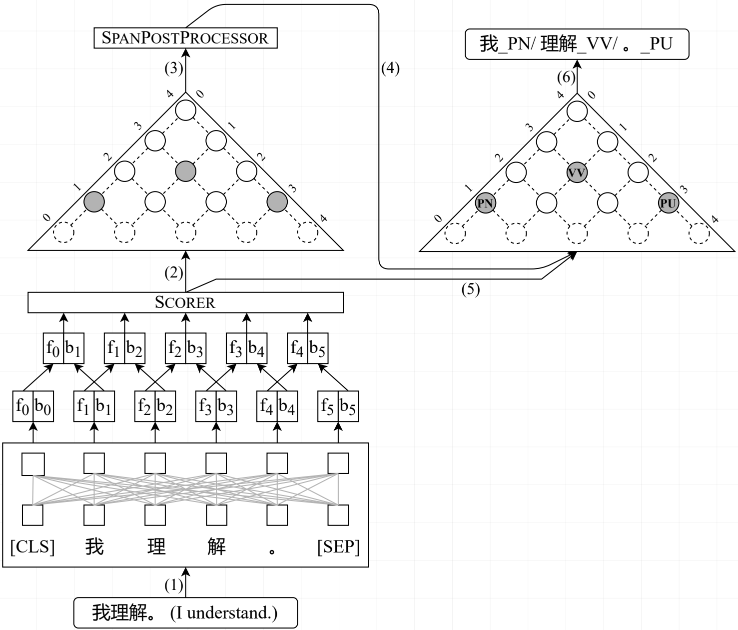
1. Chinese word segmentation and part-of-speech tagging are necessary tasks in terms of computational linguistics and application of natural language processing.
2. Many researchers still debate the demand for Chinese word segmentation and part-of-speech tagging in the deep learning era.
3. Nevertheless, resolving ambiguities and detecting unknown words are challenging problems in this field.
4. Previous studies on joint Chinese word segmentation and part-of-speech tagging mainly follow the character-based tagging model focusing on modeling n-gram features.

Introduction

1. Chinese word segmentation and part-of-speech tagging are necessary tasks in terms of computational linguistics and application of natural language processing.
2. Many researchers still debate the demand for Chinese word segmentation and part-of-speech tagging in the deep learning era.
3. Nevertheless, resolving ambiguities and detecting unknown words are challenging problems in this field.
4. Previous studies on joint Chinese word segmentation and part-of-speech tagging mainly follow the character-based tagging model focusing on modeling n-gram features.
5. Unlike previous works, we propose a neural model named SpanSegTag for joint Chinese word segmentation and part-of-speech tagging following the span labeling in which the probability of each n-gram being the word and the part-of-speech tag is the main problem.

Outline

The proposed architecture for joint CWS & POS tagging



Joint Chinese Word Segmentation and Part-of-speech Tagging as Two Stages Span Labeling

- Input: a sequence of characters $\mathcal{X} = x_1 x_2 \dots x_n$ with the length of n.
- Output:
  1. Word segmentation: a sequence of words $\mathcal{W} = w_1 w_2 \dots w_m$ with the length of m.
  2. Part-of-speech Tagging: a sequence of POS tags $\mathcal{T} = t_1 t_2 \dots t_m$ with the length of m, where $1 \leq m \leq n$.
- Notations:
  1. We use $x_i x_{i+1} \dots x_{i+k-1}$ to denote that the word w_j is constituted by k consecutive characters beginning at character x_i, where $1 \leq k \leq n$.
  2. We get the inspiration of span representation in constituency parsing[1] to use the span $(i-1, i-1+k)$ representing the word constituted by k consecutive characters $x_i x_{i+1} \dots x_{i+k-1}$.

[1] [1] Stern et al. "A Minimal Span-Based Neural Constituency Parser". 2017.

Joint Chinese Word Segmentation and Part-of-speech Tagging as Two Stages Span Labeling (continue)

- Formally, the first stage of our SpanSegTag for CWS can be formalized as:

$$\hat{\mathcal{S}}_{\mathsf{novlp}} = \textsc{SpanPostProcessor}(\hat{\mathcal{S}}) \quad (1)$$

where $\textsc{SpanPostProcessor}(\hat{\mathcal{S}})$ is introduced in the research[2]. $\textsc{SpanPostProcessor}(\hat{\mathcal{S}})$ solely is an algorithm for producing the word segmentation boundary guaranteeing non-overlapping between every two spans.

[2] [2] Nguyen et al. "Span Labeling Approach for Vietnamese and Chinese Word Segmentation". 2021.

Joint Chinese Word Segmentation and Part-of-speech Tagging as Two Stages Span Labeling (continue)

- The $\hat{\mathcal{S}}$ is the set of predicted spans as follows:

$$\hat{\mathcal{S}} = \Big\{ (l, r) \text{ for } 0 \leq l \leq n-1 \text{ and } l < r \leq n \\ \text{and } \textsc{Scorer}(\mathcal{X}, l, r).\textsc{Seg} > 0.5 \Big\} \quad (2)$$

where n is the length of the input sentence. The l and r denote left and right boundary indexes of the specific span. The $\textsc{Scorer}(\mathcal{X}, l, r).\textsc{Seg}$ is the scoring module for the span (l, r) of sentence $\mathcal{X}$. The output of $\textsc{Scorer}(\mathcal{X}, l, r).\textsc{Seg}$ has a value in the range of 0 to 1. We choose the sigmoid function as the activation function at the last layer of $\textsc{Scorer}(\mathcal{X}, l, r).\textsc{Seg}$ module.

Joint Chinese Word Segmentation and Part-of-speech Tagging as Two Stages Span Labeling (continue)

- Next, given the set of predicted spans $\hat{\mathcal{S}}_{\text{novlp}}$ satisfying non-overlapping between every two spans for the input sentence $\mathcal{X}$, the second stage of our SpanSegTag to perform Chinese POS tagging can be formalized as:

$$\hat{\mathcal{Y}} = \left\{ \left((l, r), \underset{\hat{t} \in \mathcal{T}}{\text{argmax}}\, \textsc{Scorer}(\mathcal{X}, l, r).\textsc{Tag}[\hat{t}] \right) \text{ for } (l, r) \in \hat{\mathcal{S}}_{\text{novlp}} \right\} \tag{3}$$

where $\mathcal{T}$ is the union of Chinese POS tag set and the non-word tag since the $\hat{\mathcal{S}}_{\text{novlp}}$ can include the incorrectly predicted span. The $\textsc{Scorer}(\mathcal{X}, l, r).\textsc{Tag}[\hat{t}]$ is the scoring module for the span (l, r) of sentence $\mathcal{X}$ assigned tag $\hat{t}$. To sum up, given the input sentence $\mathcal{X}$, the set $\hat{\mathcal{Y}}$ includes predicted spans with the POS tag. Therefore, the set $\hat{\mathcal{Y}}$ is the result of the second stage of our SpanSegTag and of the joint CWS and POS tagging task.

Decoding Algorithm for Predicted Span

- We inherited the heuristic-based $\textsc{SpanPostProcessor}(\hat{\mathcal{S}})$ algorithm[3].

1. Keeping the spans with the highest score and eliminate the remainder among overlapping spans.
2. Adding the missing word boundary based on all predicted spans $(i-1, i-1+k)$ with $k=1$ to single words to deal with the missing word boundary problem.

[3] [2] Nguyen et al. "Span Labeling Approach for Vietnamese and Chinese Word Segmentation". 2021.

Span Scoring for Word Segmenation

- Inspired by[4], the span scoring module $\textsc{Scorer}(\mathcal{X}, l, r).\textsc{Seg}$ for finding probability of word is computed by using a biaffine operation over the left boundary representation of character x_l and the right boundary representation of character x_r:

$$\textsc{Scorer}(\mathcal{X}, l, r).\textsc{Seg} = \text{sigmoid}\Bigg(\begin{bmatrix} \text{MLP}^{\text{left}}_{\text{seg}}(\mathbf{f}_l \oplus \mathbf{b}_{l+1}) \\ 1 \end{bmatrix}^{\mathsf{T}} \mathbf{W}\big(\text{MLP}^{\text{right}}_{\text{seg}}(\mathbf{f}_r \oplus \mathbf{b}_{r+1})\big)\Bigg) \quad (4)$$

where $\mathbf{W} \in \mathbb{R}^{(d+1)\times d}$ and the symbol $\oplus$ denote the concatenation operation.

[4] [3] Zhang et al. "Fast and Accurate Neural CRF Constituency Parsing". 2020.

Span Scoring for Part-of-speech Tagging

- Similarly, the span scoring module $\textsc{Scorer}(\mathcal{X}, l, r).\textsc{Tag}[\hat{t}]$ for finding score of a POS tag $\hat{t} \in \mathcal{T}$ is computed by:

$$\textsc{Scorer}(\mathcal{X}, l, r).\textsc{Tag}[\hat{t}] = \begin{bmatrix} \mathrm{MLP}^{\mathrm{left}}_{\mathrm{tag}}(\mathbf{f}_l \oplus \mathbf{b}_{l+1}) \\ 1 \end{bmatrix}^{\mathsf{T}} \mathbf{W}_{\hat{t}} \begin{bmatrix} \mathrm{MLP}^{\mathrm{right}}_{\mathrm{tag}}(\mathbf{f}_r \oplus \mathbf{b}_{r+1}) \\ 1 \end{bmatrix} \tag{5}$$

where $\mathbf{W}_{\hat{t}} \in \mathbb{R}^{(d+1)\times(d+1)}$.

Outline

3. Experiments

Statistics of five Chinese benchmark datasets[5]

Datasets		# Sent	# Char	# Word	OOV
CTB5	Train	18,104	804,587	493,930	-
	Dev	352	11,543	6,821	8.1
	Test	348	13,738	8,008	3.5
CTB6	Train	23,420	1,055,583	641,368	-
	Dev	2,079	100,316	59,955	5.4
	Test	2,796	134,149	81,578	5.6
CTB7	Train	31,112	1,160,209	717,874	-
	Dev	10,043	387,209	236,590	5.5
	Test	10,292	398,626	245,011	5.2
CTB9	Train	105,971	2,642,998	1,696,340	-
	Dev	9,850	209,739	136,468	2.9
	Test	15,929	378,502	242,317	3.1
UD	Train	3,997	156,309	98,608	-
	Dev	500	20,000	12,663	12.1
	Test	500	19,206	12,012	12.4

[5] [4] Xue et al. "The Penn Chinese TreeBank: Phrase structure annotation of a large corpus". 2005; [5] Nivre et al. "Universal Dependencies v1: A Multilingual Treebank Collection". 2016.

Experimental results on development sets of six Chinese benchmark datasets

SpanSegTag		CTB5		CTB6		CTB7		CTB9		UD1		UD2	
Encoder	MLP Size	Seg	Tag	Seg	Tag	Seg	Tag	Seg	Tag	Seg	Tag	Seg	Tag
BiLSTM	100	96.71	92.80	94.33	89.43	94.46	89.17	95.64	91.27	91.84	85.21	91.48	84.80
	200	96.90	93.08	94.90	90.06	94.70	89.36	95.96	91.57	92.36	85.92	92.27	85.78
	300	97.03	93.21	95.00	90.06	94.86	89.39	96.05	91.61	92.43	86.14	92.72	85.93
	400	96.82	93.27	95.18	90.16	95.04	89.53	96.15	91.54	93.02	86.45	92.84	86.03
	500	97.30	**93.39**	95.29	**90.19**	95.10	**89.53**	96.27	**91.61**	93.08	**86.74**	93.12	**86.29**
BERT	100	98.76	97.78	97.71	95.25	97.06	94.16	97.75	94.92	98.21	95.51	98.22	95.38
	200	98.78	97.71	97.66	95.25	97.11	94.24	97.78	95.07	98.23	95.64	98.21	**95.50**
	300	98.56	97.54	97.70	95.24	97.12	**94.27**	97.74	95.02	98.35	**95.72**	98.22	95.49
	400	98.57	97.64	97.69	**95.26**	97.05	94.18	97.80	**95.10**	98.28	95.70	98.17	95.44
	500	98.81	**97.78**	97.69	95.23	97.10	94.22	97.80	95.01	98.30	95.66	98.30	95.44

Experimental results on test sets of six Chinese benchmark datasets

	CTB5		CTB6		CTB7		CTB9		UD1		UD2	
	Seg	Tag	Seg	Tag	Seg	Tag	Seg	Tag	Seg	Tag	Seg	Tag
Jiang et al. [6]	97.85	93.41	-	-	-	-	-	-	-	-	-	-
Kruengkrai et al. [7]	97.87	93.67	-	-	-	-	-	-	-	-	-	-
Sun [8]	98.17	94.02	-	-	-	-	-	-	-	-	-	-
Wang et al. [9]	98.11	94.18	95.79	91.12	95.65	90.46	-	-	-	-	-	-
Shen et al. [10]	98.03	93.80	-	-	-	-	-	-	-	-	-	-
Kurita et al. [11]	98.41	94.84	-	-	96.23	91.25	-	-	-	-	-	-
Shao et al. [12]	98.02	94.38	-	-	-	-	96.67	92.34	95.16	89.75	95.09	89.42
Zhang et al. [13]	98.50	94.95	96.36	92.51	96.25	91.87	-	-	-	-	-	-
Tian et al. [14] (BERT)	98.77	96.77	97.39	94.99	**97.32**	94.28	97.75	94.87	98.32	95.60	98.33	95.46
Tian et al. [14] (ZEN)	**98.81**	**96.92**	97.47	95.02	97.31	94.32	97.77	94.88	**98.33**	**95.69**	**98.18**	95.49
SpanSegTag (BERT)	98.67	96.77	**97.53**	**95.04**	97.30	**94.50**‡	**97.86**	**95.22**‡	98.06	95.59	98.12	**95.54**

Table 1: The symbol ‡ denotes that the improvement is statistically significant at $p < 0.01$ compared with TwASP (ZEN)[6] using paired t-test.

[6] [14] Tian et al. "Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge". 2020.

Outline

4 Analysis

Recall of Out-of-vocabulary and in-vocabulary Words

	$R_{POS\text{-}OOV}$			$R_{POS\text{-}iV}$		
	TwASP (BERT)	TwASP (ZEN)	Our (BERT)	TwASP (BERT)	TwASP (ZEN)	Our (BERT)
CTB5	**83.81**	**83.81**	82.73	97.54	**97.55**	97.54
CTB6	83.10	**84.22**	82.69	95.48	95.66	**95.68**
CTB7	79.94	79.39	**80.19**	95.20	95.25	**95.33**
CTB9	**79.93**	78.80	78.52	95.49	95.44	**95.80**
UD1	**88.67**	87.40	86.13	96.64	**96.92**	96.85

Table 2: Recall of out-of-vocabulary words and their POS tags ($R_{POS\text{-}OOV}$) and recall of in-vocabulary words and their POS tags ($R_{POS\text{-}iV}$). Notably, we do not provide scores on UD2 dataset since we can not reproduce result from the pre-trained model of[7].

[7] [14] Tian et al. "Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge". 2020.

Combination Ambiguity String Error

	CTB5	CTB6	CTB7	CTB9	UD1
TwASP (BERT)	**96.43**	93.72	94.26	94.61	96.40
TwASP (ZEN)	**96.43**	94.88	94.23	95.47	**97.30**
Our (BERT)	95.71	**95.30**	**94.72**	**95.56**	**97.30**

Table 3: CWS accuracies of TwASP[8] using BERT and ZEN versus our SpanSegTag on 70 high-frequency two-character CASs.

[8][14] Tian et al. "Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge". 2020.

Model Size and Inference Speed

	CTB5	CTB6	CTB7	CTB9	UD1
TwASP (BERT)	514	699	716	650	435
TwASP (ZEN)	989	1,010	1,170	1,100	909
Our (BERT)	**433**	**434**	**435**	**441**	**413**

Table 4: Model sizes (MB) of TwASP[9] using BERT and ZEN versus our SpanSegTag.

- In theory, our SpanSegTag is a $O(n^2)$ algorithm due to computing of all possible span representations. In practice, when use GPU Tesla V100 via Google Colaboratory, the inference speed of our SpanSegTag (BERT) and TwASP (BERT) are 264 and 239 (sentence/second), respectively. We notice that we did not count the time TwASP [14] consuming by running off-the-shelf toolkits.

[9] [14] Tian et al. "Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge". 2020.

Outline

Conclusion

1. Our experiments show that our BERT-based model SpanSegTag achieved competitive performances on the CTB5, CTB6, and UD, and significant improvements on the CTB7 and CTB9 benchmark datasets compared with the current state-of-the-art method TwASP using BERT and ZEN encoders.

Conclusion

1. Our experiments show that our BERT-based model SpanSegTag achieved competitive performances on the CTB5, CTB6, and UD, and significant improvements on the CTB7 and CTB9 benchmark datasets compared with the current state-of-the-art method TwASP using BERT and ZEN encoders.
2. Our SpanSegTag has the disadvantage of the complexity and time running. For future work, we will explore the architecture of the BERT model [15] for joint CWS and POS tagging because the primitive of BERT also has the complexity of $O(n^2)$ and the self-attention mechanism over the input sentence may be related to span representation.

References I

[1] Mitchell Stern, Jacob Andreas, and Dan Klein. "A Minimal Span-Based Neural Constituency Parser". In: *Proceedings of ACL*. Association for Computational Linguistics, 2017, pp. 818–827.

[2] Duc-Vu Nguyen et al. "Span Labeling Approach for Vietnamese and Chinese Word Segmentation". In: *Proceedings of PRICAI*. 2021.

[3] Yu Zhang, Houquan Zhou, and Zhenghua Li. "Fast and Accurate Neural CRF Constituency Parsing". In: *Proceedings of IJCAI*. International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 4046–4053.

[4] Naiwen Xue et al. "The Penn Chinese TreeBank: Phrase structure annotation of a large corpus". In: *Natural Language Engineering* 11.2 (2005), pp. 207–238.

[5] Joakim Nivre et al. "Universal Dependencies v1: A Multilingual Treebank Collection". In: *Proceedings of LREC*. European Language Resources Association (ELRA), 2016, pp. 1659–1666.

References II

[6] Wenbin Jiang et al. "A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging". In: *Proceedings of ACL*. 2008, pp. 897–904.

[7] Canasai Kruengkrai et al. "An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging". In: *Proceedings of ACL-IJCNLP*. 2009, pp. 513–521.

[8] Weiwei Sun. "A Stacked Sub-Word Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging". In: *Proceedings of ACL*. 2011, pp. 1385–1394.

[9] Yiou Wang et al. "Improving Chinese Word Segmentation and POS Tagging with Semi-supervised Methods Using Large Auto-Analyzed Data". In: *Proceedings of IJCNLP*. Asian Federation of Natural Language Processing, 2011, pp. 309–317.

[10] Mo Shen et al. "Chinese Morphological Analysis with Character-level POS Tagging". In: *Proceedings of ACL*. 2014, pp. 253–258.

References III

[11] Shuhei Kurita, Daisuke Kawahara, and Sadao Kurohashi. "Neural Joint Model for Transition-based Chinese Syntactic Analysis". In: *Proceedings of ACL*. 2017, pp. 1204–1214.

[12] Yan Shao et al. "Character-based Joint Segmentation and POS Tagging for Chinese using Bidirectional RNN-CRF". In: *Proceedings of IJCNLP*. Asian Federation of Natural Language Processing, 2017, pp. 173–183.

[13] Meishan Zhang, Nan Yu, and Guohong Fu. "A Simple and Effective Neural Model for Joint Word Segmentation and POS Tagging". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 26 (2018), pp. 1528–1538.

[14] Yuanhe Tian et al. "Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge". In: *Proceedings of ACL*. 2020, pp. 8286–8296.

References IV

[15] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL*. 2019, pp. 4171–4186.