Neural Chinese Word Segmentation as Sequence-to-sequence Translation

Xuewen Shi, Heyan Huang, Ping Jian, Yuhang Guo, Xiaochi Wei and Yikun Tang
{xwshi,hhy63,pjian,guoyuhang,wxchi,tangyk}@bit.edu.cn

报告人：史学文

北京理工大學
BEIJING INSTITUTEOF TECHNOLOGY

Outline

- Chinese word segmentation
- Chinese spelling correction
- Related works
- Conclusion

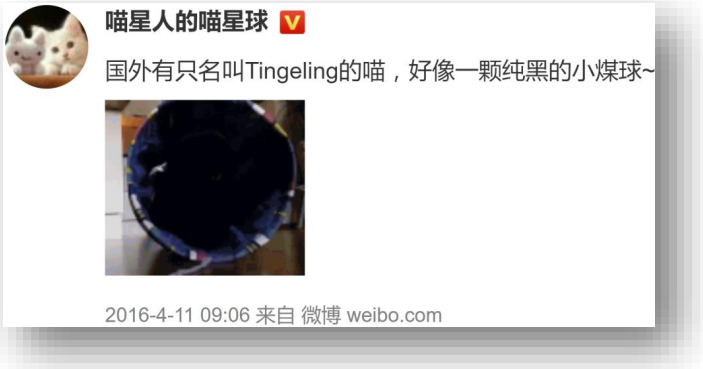
Background

- Chinese Word Segmentation (CWS)
  - Basic task of Chinese NLP
  - Popular algorithm：HMM、CRF etc.
  - Popular open source CWS toolkit：NLPIR，HIT-LTP，jieba etc.
- Chinese Weibo Text
  - New words; foreign language; informal grammar

CWS

国 外 / 有 / 只 / 名 叫 / T I n g e l I n g / 的 / 喵 / , / 好 像 / 一 / 颗 / 纯 黑 / 的 / 小 / 煤 球 / ~

国 外 有 只 名 叫 T I n g e l I n g 的 喵 ， 好 像 一 颗 纯 黑 的 小 煤 球 ~



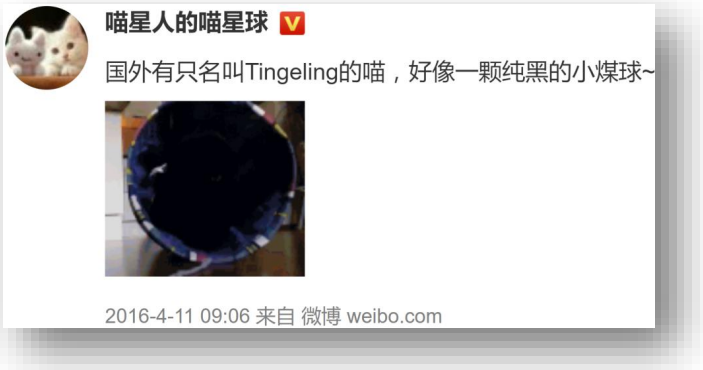
CWS: Popular Approaches

b e s s b e b m ..... m e s s s b e s s b e s s b e s

国外/有/只/名叫/TIngelIng/的/喵/,/好像/一/颗/纯黑/的/小/煤球/~

Sequence labeling

国外有只名叫TIngelIng的喵，好像一颗纯黑的小煤球~



CWS: Our Approach

Target characters sequence with delimiters ‘/’ :

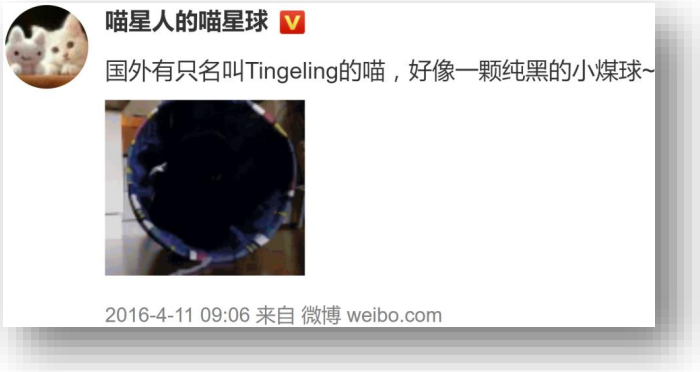
b e s s b e b m ..... m e s s s b e s s b e s s b e s

国外/有/只/名叫/TIngelIng/的/喵/,/好像/一/颗/纯黑/的/小/煤球/~

~~Sequence labeling~~ → Sequence-to-sequence translation

Source characters sequence :

国外有只名叫TIngelIng的喵，好像一颗纯黑的小煤球~



Model Comparison

- Popular approaches

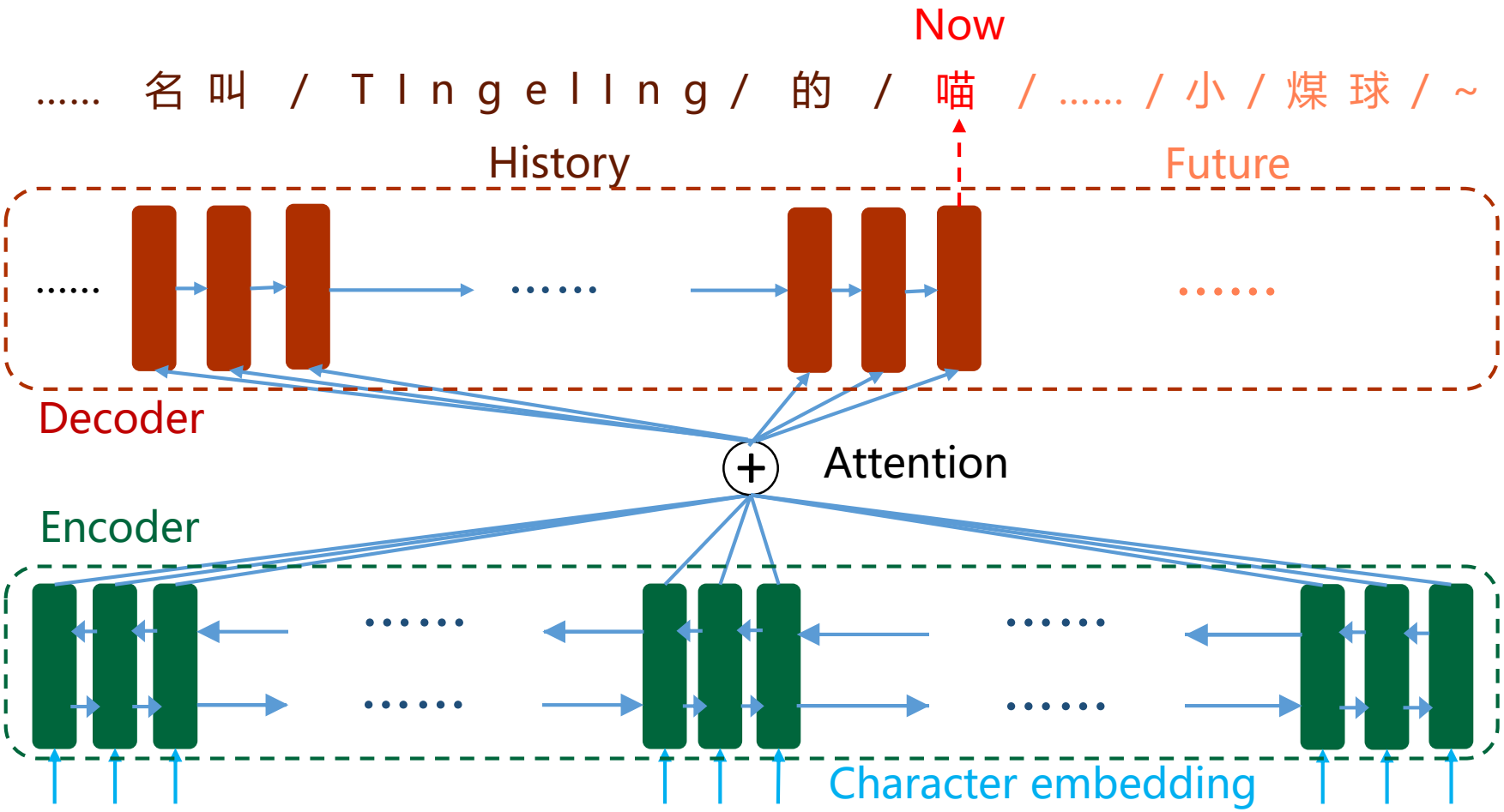
$$P(T|X) = ?$$

- Our approach:

$$P(Y|X) = \prod_{t}^{T} p(y_y|y_1, \dots, y_{t-1}, X)$$

$$p(y_t|y_1, \dots, y_{t-1}, X) = g(y_{t-1}, s_t, X)$$

Model Architecture



Model architecture: attention-based encoder-decoder [Bahdanau et al., 2015]

Formulation

- Decoder: $p(y_t|y_1, \dots, y_{t-1}, X) = g(y_{t-1}, s_t, X)$

$$s_t = f(s_{t-1}, y_{t-1}, c_t)$$

$$c_t = \sum_{i=1}^{N} \alpha_{t,i} h_j$$

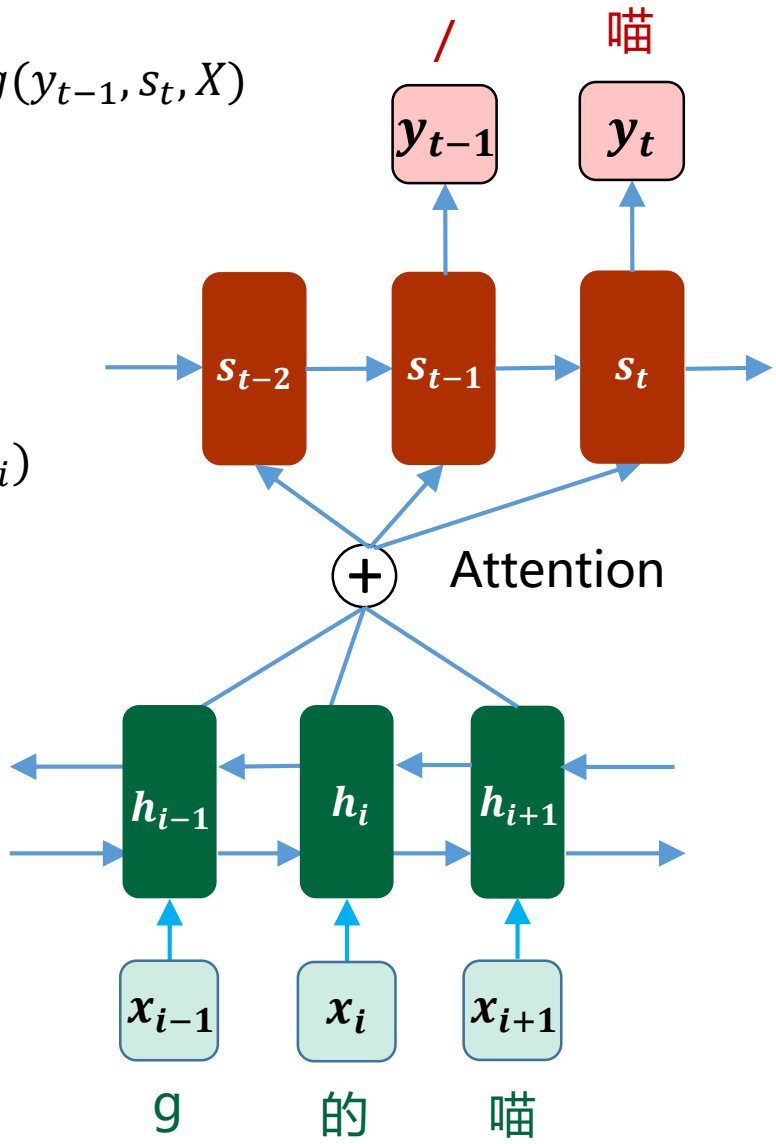
$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{N} \exp(e_{t,k})} \qquad e_{t,i} = a(s_{t-1}, h_i)$$

- Encoder :

Maps $X = [x_1, \dots, x_N]$
into $H = [h_1, \dots, h_N]$

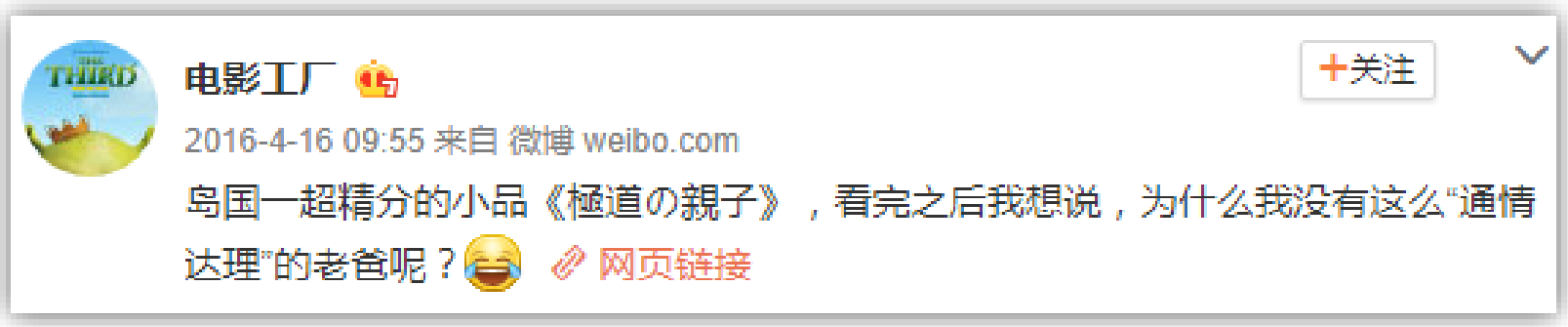
$$\overleftarrow{h}_i = \overleftarrow{f}(\overleftarrow{h}_{i+1}, x_i) \qquad \vec{h}_i = \vec{f}(\vec{h}_{i-1}, x_i)$$

$$h_i = \begin{bmatrix} \vec{h}_i \\ \overleftarrow{h}_i \end{bmatrix}$$



Post-editing

- *WHY*: translating errors
- Example:



Outputs: 岛国/一/超/精分/的/小品/《/UNK道/UNK/UNK子/》/，/看完/之后/我/想/说/，/为什么/我/没有/这么/“/通情达理/” /的/老爸/呢/？

Post-editing

- *Longest common subsequence* (LCS) based post-editing method

Step1	…… S B E S B E S ……	Labels
	……《/UNK道/UNK/UNK子/》……	Outputs
Step2	…… S B E S B E S ……	Labels
	……《 UNK道 UNK UNK子 》……	Outputs
	……《 極 道 の 親 子 》……	Inputs
	……《 道 子 》……	LCS
Step3	…… S B E S B E S ……	Labels
	……《 / 極 道 / の / 親 子/》……	PE outputs

The source code is available at
https://github.com/SourcecodeSharing/CWSpostediting

Setup

- Model Setup
  - Source code: GroundHog
    https://github.com/lisa-groundhog/GroundHog
  - Vocabulary size 7190, hidden size 1000, embedding size 620
  - Seq2seq baseline: Moses phrase-based statistical machine translation model
- Dataset
  - Weibo text[1] [Qiu et al., 2016]
  - PKU and MSRA[2] [Emerson et al., 2005]

Datasets	Training			Testing		
	Sentences	Words	Characters	Sentences	Words	Characters
Weibo	20,135	421,166	688,743	8,592	187,877	315,865
PKU	43,475	1,109,947	1,826,448	4,261	104,372	172,733
MSRA	86,924	2,368,391	4,050,469	3,985	106,873	184,355

1: https://github.com/FudanNLP/NLPCC-WordSeg-Weibo
2: http://sighan.cs.uchicago.edu/bakeoff2005/

Pre-training

- *WHY*:
  - Large scale free parameters
  - Small training data
  - To avoid ***underfitting*** or ***overfitting***

Datasets	P	R	F
Weibo	89.8	89.5	89.6
PKU	87.0	88.6	87.8
MSRA	95.1	93.2	94.1

- Pseudo Data:
  - Dataset : UN1.0 [Ziemski et al., 2016]
  - Chinese word segmentation toolkit: LTP [Che et al., 2010]
  - P, R and F score on pseudo test data: 98.2, 97.1 and 97.7

Weibo Results

Groups	Models	Standard Scores			Weighted Scores		
		P	R	F	P	R	F
A	LTP [3]	83.98	90.46	87.09	69.69	80.43	74.68
B [16]	S1	94.13	94.69	94.41	79.29	81.62	80.44
	S2	94.21	95.31	94.76	78.18	81.81	79.96
	S3	94.36	95.15	94.75	78.34	81.34	79.81
	S4	93.98	94.78	94.38	78.43	81.20	79.79
	S5	93.93	94.80	94.37	76.24	79.32	77.75
	S6	93.90	94.42	94.16	75.95	78.20	77.06
	S7	93.82	94.60	94.21	75.08	77.91	76.47
	S8	93.74	94.31	94.03	74.90	77.14	76.00
	S9	92.89	93.65	93.27	71.25	73.92	72.56
M	Moses PB w/ 3-gram LM	92.42	92.26	92.34	76.74	77.23	76.98
	Moses PB w/ 5-gram LM	92.37	92.26	92.31	76.58	77.25	76.91
	RNNsearch w/o fine-tuning	86.10	88.82	87.44	68.88	75.20	71.90
	RNNsearch	92.09	92.79	92.44	75.00	78.27	76.60
	RNNsearch w/ post-editing	93.48 (>S9)	94.60 (>S6)	94.04 (>S8)	76.30 (>S5)	79.99 (>S5)	78.11 (>S5)

Weibo Examples

Reference	安心/持股/，/别/一惊一乍/。
This work	安心/持股/，/别/一惊一乍/。
LTP	安心/持/股/，/别/一/惊/一/乍/。

Not in the training data

Reference	网传/《/唐/十八陵/石人/石马/洗澡/千年/包浆/被/清洗 》/。
This work	网传/《/唐/十八陵/石人/石马/洗澡/千/年/包浆/被/清洗/》/。
LTP	网传/《 唐/十八/陵石/人/石马/洗澡/千/年/包浆/被/清洗/》/。

Reference	据/报道/，/该/案/疑点/众多/，/且/刘/受过/刑讯逼供 。
This work	据/报道/，/该/案/疑点/众多/，/且/刘/受/过/刑讯/逼供 。
LTP	据/报道/，/该案/疑点/众多/，/且/刘受/过/刑讯/逼供 。

Weibo Examples

Input	几年后才发现。。。原来夏川真凉的初恋男友是486啊。。。[晕][晕][晕] #我女友与青梅竹马的惨烈修罗场#
Output	几/年/后/才/发现/。/。/。/原来/夏川/真凉/的/初恋/男友/是/486/啊/。/。/。/ [/晕/] /[/晕/] /[/晕/] /#/我/女友/与/青梅/竹马/的/惨烈/修罗场/#
LTP	几/年/后/才/发现/。。。/原来/夏川真凉/的/初恋/男友/是/486/啊/。。。/ [/晕/] /[/晕/] /[/晕/] /#/我/女友/与/青梅竹马/的/惨烈/修罗场/#

PKU & MSRA Results

Groups	Models	PKU			MSRA		
		P	R	F	P	R	F
A	LTP [3]	95.9	94.7	95.3	86.8	89.9	88.3
B	Zheng et al., 2013 [25]	93.5	92.2	92.8	94.2	93.7	93.9
	Pei et al., 2014 [13]	94.4	93.6	94.0	95.2	94.6	94.9
	Chen et al., 2015 [4]	96.3	95.9	96.1	96.2	96.3	96.2
	Chen et al., 2015 [5]	96.3	95.6	96.0	96.7	96.5	96.6
	Cai and Zhao, 2016 [2]	95.8	95.2	95.5	96.3	96.8	96.5
C	Zhang et al., 2013 [23]	-	-	96.1	-	-	97.4
M	Moses PB w/ 3-gram LM	92.9	93.0	93.0	96.0	96.2	96.1
	Moses PB w/ 5-gram LM	92.7	92.8	92.7	95.9	96.3	96.1
	Moses PB w/ 3-gram LM w/ CSC	92.9	93.0	92.9	95.3	96.5	95.9
	Moses PB w/ 5-gram LM w/ CSC	92.6	93.2	92.9	95.9	96.3	96.1
	RNNsearch w/o fine-tuning	93.1	92.7	92.9	84.1	87.9	86.0
	RNNsearch	94.7	95.3	95.0	96.2	96.0	96.1
	RNNsearch w/ post-editing	94.9	95.4	95.1	96.3	96.1	96.2
	RNNsearch w/ CSC	95.2	94.6	94.9	96.1	96.1	96.1
	RNNsearch w/ CSC and post-editing	95.3	94.7	95.0	96.2	96.1	96.2

Outline

- Chinese word segmentation
- Chinese spelling correction
- Related works
- Conclusion

Chinese Spelling Correction (CSC)

- Motivation
  - Sequence-to-sequence CWS is unnecessary and bring into other problems
  - End-to-end NLP preprocessing framework
- Manual constructed data
  - Correct-to-wrong dictionary counting from SIGHAN 2014 [Yu et al., 2014]

Original input	在这个基础上，公安机关还从原料采购等方面加以严格控制，统一发放“准购证”。
Modified input	在这个基础上，公安机关还从源料采购等方面加以严格控制，统一发放“准购证”。
Gold standard	在/这个/基础/上/，/公安/机关/还/从/原料/采购/等/方面/加以/严格/控制/，/统一/发放/“/准/购/证/”/。

CSC Results

PKU

Models	P	R	F
Modified testing data	99.0	99.0	99.0 (-1.0)
LTP [3]	94.0	93.2	93.6 (-1.7)
Moses PB w/ 3-gram LM	90.8	91.5	91.2 (-1.8)
Moses PB w/ 3-gram LM w/ CSC	92.7	92.9	92.8 (-0.1)
Moses PB w/ 5-gram LM	90.6	91.3	91.0 (-1.7)
Moses PB w/ 5-gram LM w/ CSC	92.3	93.0	92.6 (-0.3)
RNNsearch	93.2	93.2	93.2 (-1.8)
RNNsearch w/ CSC	**95.0**	**94.5**	**94.8** (-0.1)

MSRA

Models	P	R	F
Modified testing data	98.5	98.5	98.5 (-1.5)
LTP [3]	84.8	88.4	86.6 (-1.7)
Moses PB w/ 3-gram LM	93.7	94.6	94.2 (-1.9)
Moses PB w/ 3-gram LM w/ CSC	95.0	96.3	95.6 (-0.3)
Moses PB w/ 5-gram LM	93.7	94.7	94.2 (-1.9)
Moses PB w/ 5-gram LM w/ CSC	94.6	95.9	95.3 (-0.7)
RNNsearch	93.8	94.7	94.2 (-1.9)
RNNsearch w/ CSC	**96.0**	**96.0**	**96.0** (-0.1)

CSC Examples

Gold standard	(/新华社/北京/12月/31日/电/)
Modified input	(新华社背景12月31日电)
Output	(/新华社/北京/12月/31日/电/)

Gold standard	正是/在/这样/的/国际/和/国内/经济/社会/背景/下/，/加纳/民众/思/变/心切/，/特别/是/向来/支持/全国/民主/大会党/的/广大/农民/倒戈/，/直接/导致/执政党/在/这次/大选/中/败北/。
Modified input	正是在这样的国际和国内经济社会背景下，加纳民众思变心切，特别是向来支持全国民主大会党的广大农民倒戈，直接导致执政党在这次大选中败北。
Output	正是/在/这样/的/国际/和/国内/经济社会/背景/下/，/加纳/民众/思变/心切/，/特别/是/向来/支持/全国/民主/大会党/的/广大/农民/倒戈/，/直接/导致/执政党/在/这次/大选/中/败北/。

CSC Examples

Gold standard	港澳/回归/，/台湾/父老/，/统一/人心/正义/稠/。
Modified input	港澳回归，台弯父老，统一人心正义稠。
Output	港澳/回归/，/台湾/父老/，/统一/人心/正义稠/。

Gold standard	令/人/疑惑/的/是/，/直到/11月/30日/还/坚持/罢免书/无效/的/乡/政府/，/突然/来/了/个/180/度/的/大/转弯/，/不再/提/核实/问题/，/仓促/决定/于/12月/5日/召开/罢免/大会/。
Modified input	令人疑惑的是，至到11月30日还坚持罢免书无效的乡政府，突然来了个180度的大转弯，不再提核实问题，仓促决定于12月5日召开罢免大会。
Output	令/人/疑惑/的/是/，/直到/11月/30日/还/坚持/罢免书/无效/的/乡政府/，/突然/来/了/个/180/度/的/大/转弯/，/不再/提/核实/问题/，/仓促/决定/于/12月/5日/召开/罢免/大会/。

Outline

- Chinese word segmentation
- Chinese spelling correction
- Related works
- Conclusion

Related Works

- Neural CWS
  - Neural sequence labelling [Zheng et al., 2013; Pei et al., 2014; Chen et al., 2015a; Chen et al., 2015b]
  - Direct segmentation learning [Cai and Zhao, 2016]

- Sequence-to-sequence model
  - Neural machine translation [Sutskever et al., 2014; Bahdanau et al., 2015]
  - Sequence-to-dependency [Wu et al., 2017]
  - AMR [Konstas et al., 2017]

Outline

- Chinese word segmentation
- Chinese spelling correction
- Related works
- Conclusion

Conclusion

- First treat CWS as a sequence-to-sequence translation task and introduce an attention-based encoder-decoder framework into CWS
- Propose an LCS based post-editing method to deal with possible translation errors
- Let our sequence-to-sequence CWS model simultaneously tackle CSC in an end-to-end mode, and well validate its applicability in our experiments.

Thanks for Attention

References

- Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
- Cai, D., Zhao, H.: Neural word segmentation learning for chinese. arXiv preprint arXiv:1606.04300 (2016)
- Che, W., Li, Z., Liu, T.: Ltp: A chinese language technology platform. In: Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. pp. 13{16. Association for Computational Linguistics (2010)
- Chen, X., Qiu, X., Zhu, C., Huang, X.: Gated recursive neural network for Chinese word segmentation. In: ACL (1). pp. 1744{1753 (2015a)
- Chen, X., Qiu, X., Zhu, C., Liu, P., Huang, X.: Long short-term memory neural networks for chinese word segmentation. In: EMNLP. pp. 1197{1206 (2015b)
- Emerson, T.: The second international chinese word segmentation bakeoff. In: Proceedings of the fourth SIGHAN workshop on Chinese language Processing. vol. 133 (2005)
- Konstas, I., Iyer, S., Yatskar, M., Choi, Y., & Zettlemoyer, L. (2017). Neural AMR: Sequence-to-Sequence Models for Parsing and Generation. arXiv preprint arXiv:1704.08381.
- Pei, W., Ge, T., Chang, B.: Max-margin tensor neural network for chinese word segmentation. In: ACL (1). pp. 293{303 (2014)
- Qiu, X., Qian, P., Shi, Z.: Overview of the NLPCC-ICCPOL 2016 shared task: Chinese word segmentation for micro-blog texts. In: Proceedings of The Fifth Conference on Natural Language Processing and Chinese Computing & The Twenty Fourth International Conference on Computer Processing of Oriental Languages (2016)
- Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104{3112 (2014)
- Wu, S., Zhang, D., Yang, N., Li, M., & Zhou, M. (2017). Sequence-to-Dependency Neural Machine Translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 698-707).
- Yu, L.C., Lee, L.H., Tseng, Y.H., Chen, H.H., et al.: Overview of sighan 2014 bake-off for chinese spelling check. In: Proceedings of the 3rd CIPSSIGHAN Joint Conference on Chinese Language Processing (CLP' 14). pp. 126{132 (2014)
- Zheng, X., Chen, H., Xu, T.: Deep learning for chinese word segmentation and postagging. In: EMNLP. pp. 647{657 (2013)
- Ziemski, M., Junczys-Dowmunt, M., Pouliquen, B.: The united nations parallel corpus v1. 0. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC. pp. 23{28 (2016)