SegBot: A Generic Neural Text Segmentation Model with Pointer Network

Jing Li, Aixin Sun and Shafiq Joty

School of Computer Science and Engineering
Nanyang Technological University (NTU), Singapore

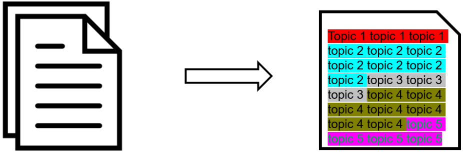
IJCAI-ECAI 2018

Motivation

- **Text segmentation** has been a fundamental task in NLP that has been addressed at different levels of granularity.

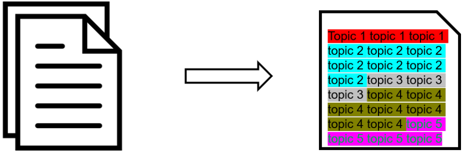
Motivation

- **Text segmentation** has been a fundamental task in NLP that has been addressed at different levels of granularity.
  - **Topic Segmentation**



Motivation

- **Text segmentation** has been a fundamental task in NLP that has been addressed at different levels of granularity.
  - **Topic Segmentation**



  - **Elementary Discourse Unit (EDU) Segmentation**

[A person]$_{\text{EDU}}$ [who never made a mistake]$_{\text{EDU}}$ [never tried anything new]$_{\text{EDU}}$

Motivation

- **Unsupervised** topic segmentation models exploit the strong correlation between topic and lexical usage. However, they fail to utilize training data and unable to learn an accurate model.

  ✓ *Hearst et al., Texttiling: Segmenting text into multi-paragraph subtopic passages. Computational linguistics, 1997.*
  ✓ *Choi et al., Advances in domain independent linear text segmentation. NAACL, 2000.*

Motivation

- **Unsupervised** topic segmentation models exploit the strong correlation between topic and lexical usage. However, they fail to utilize training data and unable to learn an accurate model.

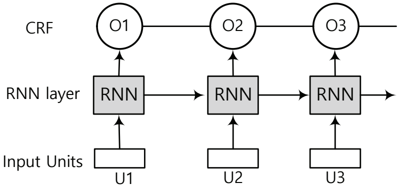
  ✓ *Hearst et al., Texttiling: Segmenting text into multi-paragraph subtopic passages. Computational linguistics, 1997.*
  ✓ *Choi et al., Advances in domain independent linear text segmentation. NAACL, 2000.*

- **Supervised** models are more flexible in using more features (*e.g.*, cue phrases, length and similarity scores) and generally perform better than unsupervised models, however, come with the price of efforts to manually design informative features.

  ✓ *Hsueh et al., Automatic segmentation of multiparty dialogue. EACL, 2006.*
  ✓ *Joty et al., Codra: A novel discriminative framework for rhetorical analysis. Computational Linguistics, 2015.*

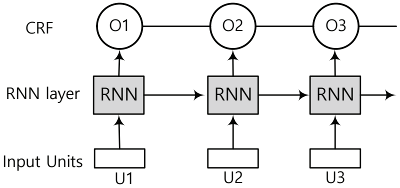
Motivation

- Due to the sparsity of 'yes' boundary tags in EDU and topic segmentation tasks, CRFs did not provide any additional gain over simple classifiers like MaxEnt.

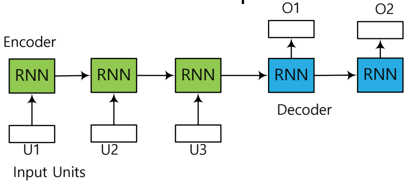


Motivation

- Due to the sparsity of 'yes' boundary tags in EDU and topic segmentation tasks, CRFs did not provide any additional gain over simple classifiers like MaxEnt.



- The limitation of Sequence-to-sequence model is that the output vocabulary (*i.e.*,, from which O1 and O2 are drawn) is fixed so that it needs to train different models with respect to different vocabularies.



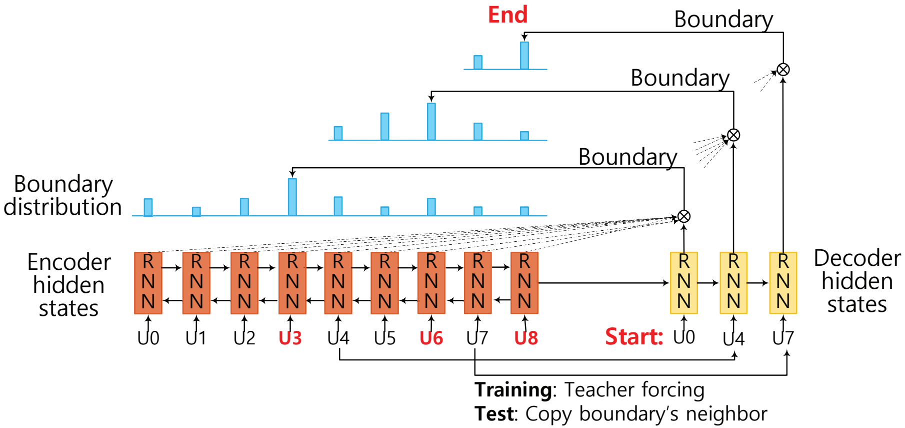
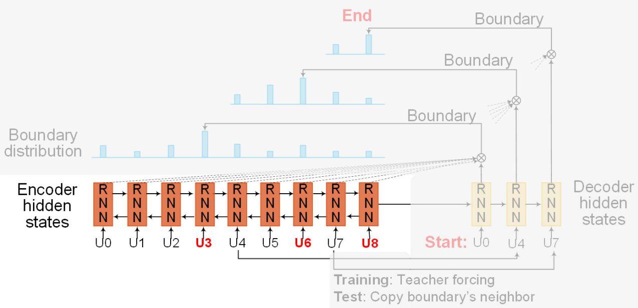


Figure 1: The model architecture of SegBot. Input sequence: $U0, U1, ..., U8$. Identified boundaries: $U3, U6, U8$.

SegBot: Neural Text Segmentation - Encoding Phase



$$\boldsymbol{z}_n = \sigma(\boldsymbol{W}_z \boldsymbol{x}_n + \boldsymbol{R}_z \boldsymbol{h}_{n-1} + \boldsymbol{b}_z) \quad (1)$$

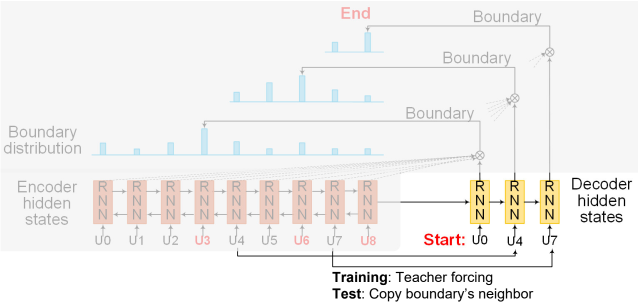
$$\boldsymbol{r}_n = \sigma(\boldsymbol{W}_r \boldsymbol{x}_n + \boldsymbol{R}_r \boldsymbol{h}_{n-1} + \boldsymbol{b}_r) \quad (2)$$

$$\boldsymbol{n}_n = \tanh(\boldsymbol{W}_h \boldsymbol{x}_n + \boldsymbol{R}_h (\boldsymbol{r}_n \odot \boldsymbol{h}_{n-1}) + \boldsymbol{b}_h) \quad (3)$$

$$\boldsymbol{h}_n = \boldsymbol{z}_n \odot \boldsymbol{h}_{n-1} + (1 - \boldsymbol{z}_n) \odot \boldsymbol{y}_n \quad (4)$$

$$\boldsymbol{h}_n = \overrightarrow{\boldsymbol{h}}_n \oplus \overleftarrow{\boldsymbol{h}}_n \quad (5)$$

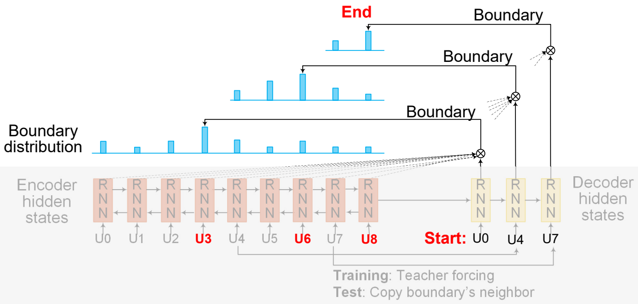
SegBot: Neural Text Segmentation - Decoding



$$d_m = GRU(x_m, \theta) \tag{6}$$

In decoding phase, at each step, the decoder takes a start of a segment in the input sequence as input.
The decoder hidden state at a time step is computed by a unidirectional RNN.

SegBot: Neural Text Segmentation - Pointing



$$u_j^m = \boldsymbol{v}^T \tanh(\boldsymbol{W}_1 \boldsymbol{h}_j + \boldsymbol{W}_2 \boldsymbol{d}_m), \text{ for } j \in (m, \ldots, M) \quad (7)$$

$$p(y_m|\boldsymbol{x}_m) = \text{softmax}(\boldsymbol{u}^m) \quad (8)$$

where $j \in [m, M]$ indicates a possible position in the input sequence, and *softmax* normalizes u_j^m indicating the probability that the unit U_j is a boundary given the start unit U_m.

SegBot: Neural Text Segmentation - Model Training

The loss function $\mathcal{L}$ is the negative log likelihood of boundary distribution over the whole training set $\mathcal{D}$, and can be written as:

$$\mathcal{L}(\boldsymbol{\omega}) = \sum_{\mathcal{D}} \sum_{m=1}^{M} -\log p(y_m|\boldsymbol{x}_m;\boldsymbol{\omega}) + \frac{\lambda}{2}||\boldsymbol{\omega}||_2^2 \quad (9)$$

Experiments

1. Topic Segmentation

2. Elementary Discourse Unit (EDU) Segmentation

Experiment: Topic Segmentation

Data: Choi Dataset

- 700 documents, each being a concatenation of 10 segments.
- A segment of a document is the first n (*s.t.* $3 \leq n \leq 11$, totally 4 subsets) sentences of a randomly selected document from the Brown corpus.

Evaluation Metric: P_k

$$P_k = \sum_{1 \leq s \leq t \leq T} \mathbf{1}(\delta_{tru}(s,t) \neq \delta_{hyp}(s,t)) \tag{10}$$

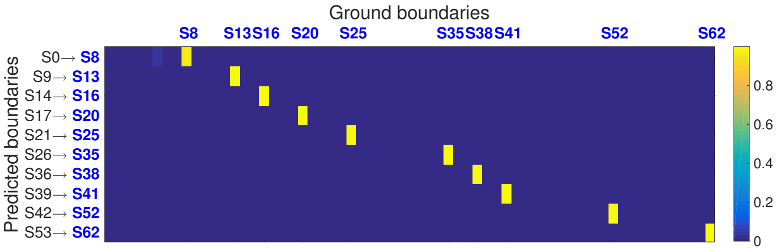
The error metric P_k, which is the commonest metric to evaluate topic segmentation models.
P_k compares the inferred segmentation with gold-standard.
Note that lower P_k means higher accuracy.

Experiment: Topic Segmentation

Group	Method	P_k (%)
A	TextTiling [Hearst, 1997]	45.25
	C99 [Choi, 2000]	10.50
	U00 [Utiyama and Isahara, 2001]	7.75
	ADDP [Ji and Zha, 2003]	5.68
	TSM [Du et al., 2013]	0.92
	GraphSeg [Glavaš et al., 2016]	6.64
B	TopSeg [Brants et al., 2002]	8.22
	F04 [Fragkou et al., 2004]	4.20
	M09 [Misra et al., 2009]	2.72
	SegBot (our model)	**0.33**
C	TopicTiling [Riedl and Biemann, 2012]	0.88
	BiLSTM-CRF [Lample et al., 2016]	<u>0.67</u>
	SegBot (our model)	**0.11***

Table 1: Segmentation results on Choi dataset. Significant improvement over BiLSTM-CRF is marked with * (p-value < 0.01).

Experiment: Topic Segmentation



⇓ For the sake of space, we show three segments from S14∼ S25 for illustration: S14 ⟶ S16; S17 ⟶ S20; S21 ⟶ S25

S14 Some who have written on Utopia have treated it as "a learned diversion of a learned world", "a phantasy with which More amused himself", "a holiday work , a spontaneous overflow of intellectual high spirits , a revel of debate , paradox , comedy and invention". S15 With respect to this view, two points are worth making. S16 First , it appears to be based on the fact that on its title page Utopia is described as "festivus", " gay". S17 The Midwest , oxidation ponds are used extensively for the treatment of domestic sewage from suburban areas. S18 The high cost of land and a few operational problems resulting from excessive loadings have created the need for a wastewater treatment system with the operational characteristics of the oxidation pond but with the ability to treat more organic matter per unit volume. S19 Research at Fayette, Missouri on oxidation ponds has shown that the BOD in the treated effluent varied from 30 to 53 mg with loadings from 8 to 120 lb. S20 Since experience indicates that effluents from oxidation ponds do not create major problems at these BOD concentrations, the goal for the effluent quality of the accelerated treatment system was the same as from conventional oxidation ponds. S21 A proton magnetic resonance study of polycrystalline Afj as a function of magnetic field and temperature is presented. S22 Afj is paramagnetic, and electron paramagnetic dipole as well as nuclear dipole effects lead to line broadening. S23 The lines are asymmetric and over the range of field Afj gauss and temperature Afj the asymmetry increases with increasing Afj and decreasing T. S24 An isotropic resonance shift of Afj to lower applied fields indicates a weak isotropic hyperfine contact interaction. S25 The general theory of resonance shifts is used to derive a general expression for the second moment Afj of a polycrystalline paramagnetic sample and is specialized to Afj.

Experiment: EDU Segmentation

Data: Rhetorical Structure Theory Discourse Treebank (RST-DT)

- A training set of 347 articles (6,132 sentences)
- A test set of 38 articles (991 sentences)

Evaluation Metric: P, R, F1

Following previous work [Hernault et al., 2010, Joty et al., 2015], we measure segmentation accuracy with respect to the sentence-internal segmentation boundaries.

$$\text{Precision} = \frac{c}{h}, \text{Recall} = \frac{c}{g}, \text{and F–score} = \frac{2c}{g+h} \tag{11}$$

Experiment: EDU Segmentation

Method	Precision	Recall	F-score
HILDA [Hernault et al., 2010]	77.9	70.6	74.1
SPADE [Soricut and Marcu, 2003]	83.8	86.8	85.2
F&R [Fisher and Roark, 2007]	91.3	89.7	90.5
DS [Joty et al., 2015]	88.0	92.3	90.1
BiLSTM-CRF [Lample et al., 2016]	89.1	87.8	88.5
SegBot (our model)	**91.6**[*]	**92.8**[*]	**92.2**[*]

Table 2: Segmentation results on RST-DT Dataset. Significant improvements over BiLSTM-CRF is marked with * (p-value < 0.01).

Experiment: EDU Segmentation

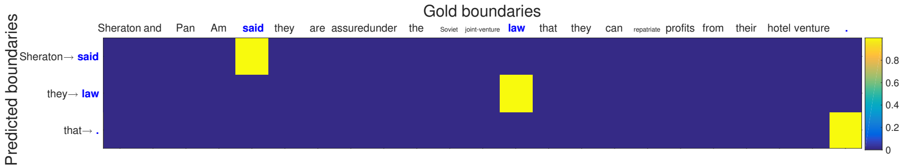


Figure 2: Visualization of EDU segmentation.

Sheraton and Pan Am **said**
they are assured under the Soviet joint-venture **law**
that they can repatriate profits from their hotel venture**.**

Web Application

- **Interface**: http://138.197.118.157:8000/segbot/

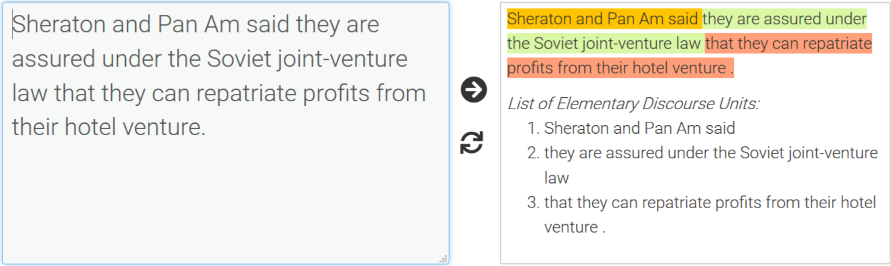


Figure 3: Interface of SegBot for EDU Segmentation.

- **User API**: http://138.197.118.157:8000/segbot/api/?q=
  import urllib
  response = urllib.request.urlopen(url).read()
  print(response)

Conclusion

- SegBot, an end-to-end neural model for text segmentation at different levels of granularity.
- SegBot does not need hand-crafted features or any prior knowledge of the given texts.
- We have demonstrated the effectiveness of SEGBOT against state-of-the-art solutions.
- A web application.

Q&A

Thanks!

jli030@e.ntu.edu.sg

Brants, T., Chen, F., and Tsochantaridis, I. (2002).
Topic-based document segmentation with probabilistic latent semantic analysis.
In *CIKM*, pages 211–218.

Choi, F. Y. (2000).
Advances in domain independent linear text segmentation.
In *NAACL*, pages 26–33.

Du, L., Buntine, W. L., and Johnson, M. (2013).
Topic segmentation with a structured topic model.
In *HLT-NAACL*, pages 190–200.

Fisher, S. and Roark, B. (2007).
The utility of parse-derived features for automatic discourse segmentation.
In *ACL*, volume 45, page 488.

Fragkou, P., Petridis, V., and Kehagias, A. (2004).
A dynamic programming algorithm for linear text segmentation.
JIIS, 23(2):179–197.

Glavaš, G., Nanni, F., and Ponzetto, S. P. (2016).
Unsupervised text segmentation using semantic relatedness graphs.
In *SEM@ACL*.

Hearst, M. A. (1997).
Texttiling: Segmenting text into multi-paragraph subtopic passages.
Computational linguistics, 23(1):33–64.

Hernault, H., Bollegala, D., and Ishizuka, M. (2010).
A sequential model for discourse segmentation.
In *CICLing*, pages 315–326.

Ji, X. and Zha, H. (2003).
Domain-independent text segmentation using anisotropic diffusion and dynamic programming.
In *SIGIR*, pages 322–329.

Joty, S., Carenini, G., and Ng, R. T. (2015).
Codra: A novel discriminative framework for rhetorical analysis.
Computational Linguistics, 41:385–435.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016).
Neural architectures for named entity recognition.
arXiv preprint arXiv:1603.01360.

Misra, H., Yvon, F., Jose, J. M., and Cappe, O. (2009).
Text segmentation via topic modeling: an analytical study.
In *CIKM*, pages 1553–1556.

Riedl, M. and Biemann, C. (2012).
Topictiling: a text segmentation algorithm based on lda.
In *ACL*, pages 37–42.

Soricut, R. and Marcu, D. (2003).
Sentence level discourse parsing using syntactic and lexical information.
In *NAACL*, pages 149–156.

Utiyama, M. and Isahara, H. (2001).
A statistical model for domain-independent text segmentation.
In *ACL*, pages 499–506.