TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

Đồ án tốt nghiệp

ỨNG DỤNG HỌC SÂU TRONG BÀI TOÁN TÁCH TỪ TIẾNG VIỆT

Sinh viên thực hiện : Lương Tuấn Dũng
Lớp CNTT-TT 2.03 - K59
Giáo viên hướng dẫn: **TS. Nguyễn Kiêm Hiếu**

Nội dung trình bày

1. Giới thiệu đề tài
2. Các phương pháp giải quyết bài toán
3. Thử nghiệm và đánh giá
4. Kết luận

1. Giới thiệu đề tài

Bài toán tách từ

- Tách từ là bài toán chia nhỏ văn bản đầu vào thành các từ thành phần
- Từ là đơn vị nhỏ nhất, có cấu tạo ổn định, mang nghĩa hoàn chỉnh, được dùng để cấu tạo nên câu
- Ví dụ:

input: *học sinh học sinh học*

output: *học_sinh học sinh_học*

1. Giới thiệu đề tài

Mục tiêu đề tài

- Thử nghiệm, áp dụng các mô hình học sâu để giải quyết bài toán tách từ trong tiếng Việt
- Đề xuất mô hình mới dựa trên học sâu áp dụng cho bài toán
- Đưa ra mô hình học sâu có kết quả cạnh tranh được với các mô hình tách từ hiện tại

2. Các phương pháp giải quyết bài toán

Tổng quan

Mô hình tách từ tốt nhất cho tiếng Việt:

- **UETSegmenter** (*f1-score = 97.87*) kết hợp longest matching và linear regression [1]
- **RDRSegmenter** (*f1-score = 97.90*) dựa trên cây SCRDR (*Single Classification Ripple Down Rules*)[2]

Mô hình sử dụng deeplearning kết hợp thông tin từ điển:

- Long Short-Term Memory for Japanese Word Segmentation [3]
- Neural Networks Incorporating Dictionaries for Chinese Word Segmentation [4]

[1] T. *P. Nguyen and A. C. Le, A hybrid approach to Vietnamese word segmen- tation, in IEEE RIVF International Conference on Computing Communica- tion Technologies, Research, Innovation, and Vision for the Future (RIVF), pages 114-119, 2016*

[2] Dat Quoc Nguyen and Dai Quoc Nguyen and Thanh Vu and Mark Dras and Mark Johnson, A Fast and Accurate Vietnamese Word Segmenter, Pro- ceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), pages 2582-2587.

[3] Yoshiaki Kitagawa, Mamoru Komachi, Long Short-Term Memory for Japanese Word Segmentation, PACLIC 2018, 10 pages

[4] Qi Zhang, Xiaoyu Liu, Jinlan Fu. Neural Networks Incorporating Dictionaries for Chinese Word Segmentation. AAAI 2018

2. Các phương pháp giải quyết bài toán

Hướng tiếp cận

Xử lý bài toán tách từ như một bài toán gán nhãn chuỗi

Ví dụ: nhãn **B-I**

input: Thành_phố Hồ_Chí_Minh là thành_phố lớn thứ nhất Việt_Nam

output: B I B I I B B I B B B B I

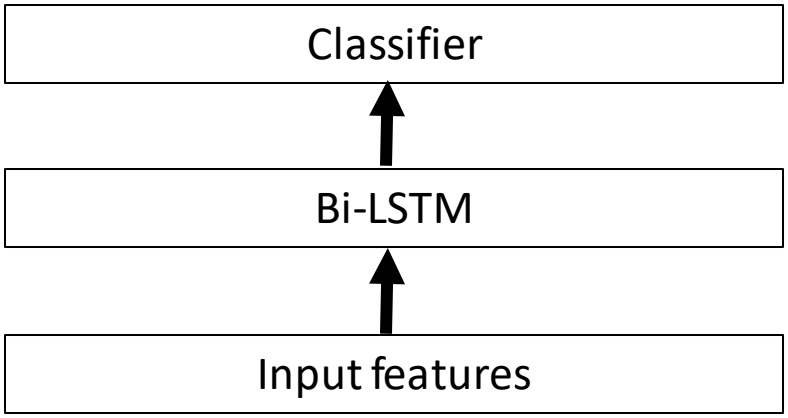
Ví dụ: nhãn **B-I-E**

input: Thành_phố Hồ_Chí_Minh là thành_phố lớn thứ nhất Việt_Nam

output: B E B I E B B E B B B B E

2. Các phương pháp giải quyết bài toán

Mô hình



- Mô hình chính sử dụng *Bi-directional Long Short-Term Memory (BiLSTM)*
- Thay đổi mô hình: thay đổi *biểu diễn đặc trưng đầu vào, bộ phân loại* và *thiết lập tham số mô hình*

2. Các phương pháp giải quyết bài toán

Đặc trưng đầu vào

Biểu diễn mức âm tiết:

- Fasttext vector
- Language model vector

Biểu diễn mức ký tự:

- CharCNN vector

Làm giàu thông tin đầu vào:

- Vector từ điển
- Vector Pointwise Mutual Information (PMI)

Bộ phân loại

- Softmax
- Conditional Random Fields (CRFs)

3. Thử nghiệm và đánh giá

Tập dữ liệu

- **Tập *48k***: gồm 47836 câu được lấy từ bộ dữ liệu VLSP 2013 sau khi đã loại bỏ đi header và chỉ giữ lại nội dung văn bản. Tập dữ liệu được chia thành train, dev, test với kích thước tương ứng 33060, 7367, 7509 câu
- **Tập *75k***: dùng toàn bộ tập VLSP làm tập train và dev. Tập test gồm 2120 câu được lấy từ 10 files 800001.seg tới 800010.seg từ tập test cho bài toán gán nhãn từ loại được công bố trong hội thảo VLSP năm 2013
- **Tập *zing news***: gồm 8,7 triệu câu được thu thập từ nguồn zing news
- **Từ điển**: *vietnamese lexicon* [1] và *vietnamese wordlist* [2] gồm 75827 từ

[1] Le, H. P., Nguyen, T. M. H., Roussanaly, A., and Ho, T. V., A hybrid approach to word segmentation of Vietnamese texts, In Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, pages 240–249.

[2] Van-Duyet Le, vietnamese-wordlist, https://vietnamese-wordlist.duyet.net

3. Thử nghiệm và đánh giá

Phương pháp đánh giá

Sử dụng độ đo precision, recall và f1 score:

- ***Precision*** được tính bằng số lượng từ gán nhãn đúng trên số lượng từ mô hình dự đoán ra
- ***Recall*** được tính bằng số lượng từ gán nhãn đúng trên số lượng từ thực tế có trong câu
- ***F1 score***:

$$\mathrm{F1} = 2 * \frac{\mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

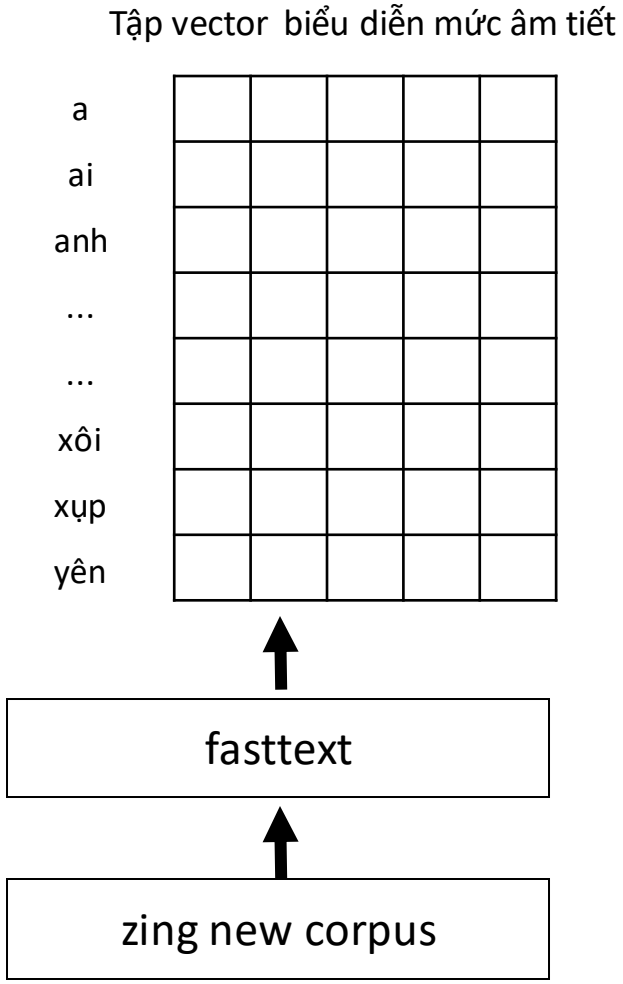
3. Thử nghiệm và đánh giá

Thử nghiệm 1

- Huấn luyện mô hình fasttext thông qua thư viện Gensim trên tập *zing news*
- Sử dụng thư viện tensorflow xây dựng kiến trúc mạng

Kết quả trên tập 48k nhãn B-I:

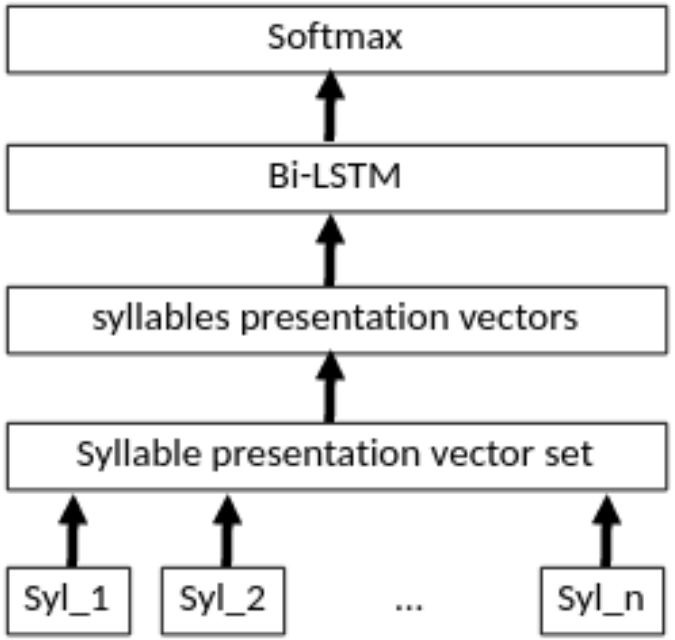
precision	recall	F1 score
96.07	96.88	96.48



3. Thử nghiệm và đánh giá

Thử nghiệm 1

- Huấn luyện mô hình fasttext thông qua thư viện Gensim trên tập *zing news*
- Sử dụng thư viện tensorflow xây dựng kiến trúc mạng

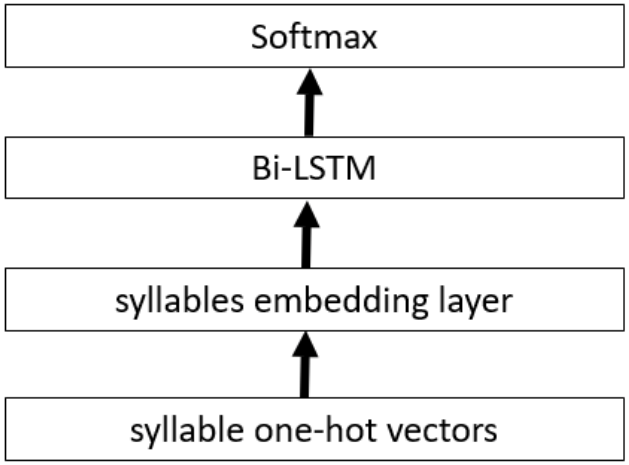


Kết quả trên tập 48k nhãn B-I:

precision	recall	F1 score
96.07	96.88	96.48

3. Thử nghiệm và đánh giá

Thử nghiệm 2



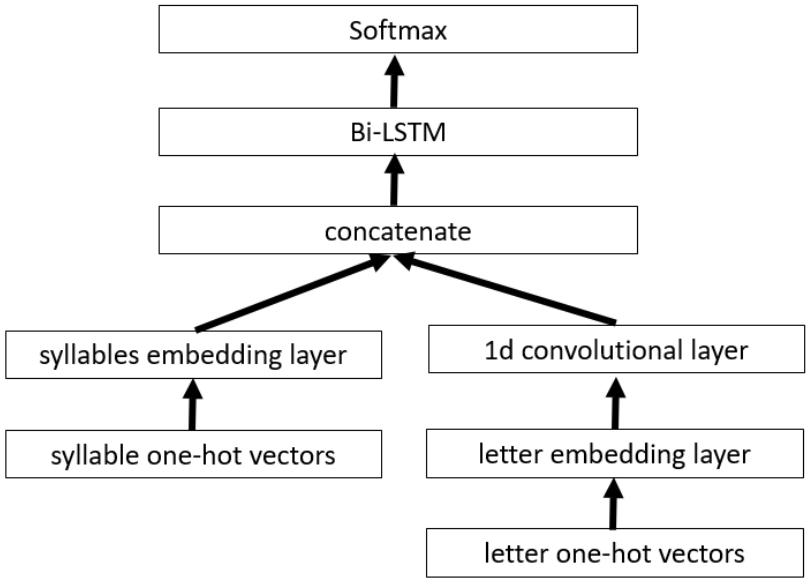
- Vector biểu diễn mức âm tiết được huấn luyện cùng mô hình thông qua tầng nhúng mức âm tiết

Kết quả trên tập 48k nhãn B-I:

precision	recall	F1 score
96.54	96.83	96.68

3. Thử nghiệm và đánh giá

Thử nghiệm 3



- Vector biểu diễn mức ký tự được học thông qua tầng nhúng mức ký tự
- Vector đặc trưng mức ký tự của từng âm tiết được trích rút thông qua mạng 1D CNN

Kết quả trên tập 48k nhãn B-I:

precision	recall	F1 score
96.95	97.32	97.13

3. Thử nghiệm và đánh giá

Thử nghiệm 3

Thống kê trên tập test ta thấy mô hình còn tồn tại 3 dạng lỗi:

- Từ chưa từng xuất hiện trong tập train
- Từ đã xuất hiện trong tập train
- Từ xuất hiện trong tập train nhưng được viết dưới dạng khác (viết hoa hoặc viết thường) chiếm 1401/4514 từ sai

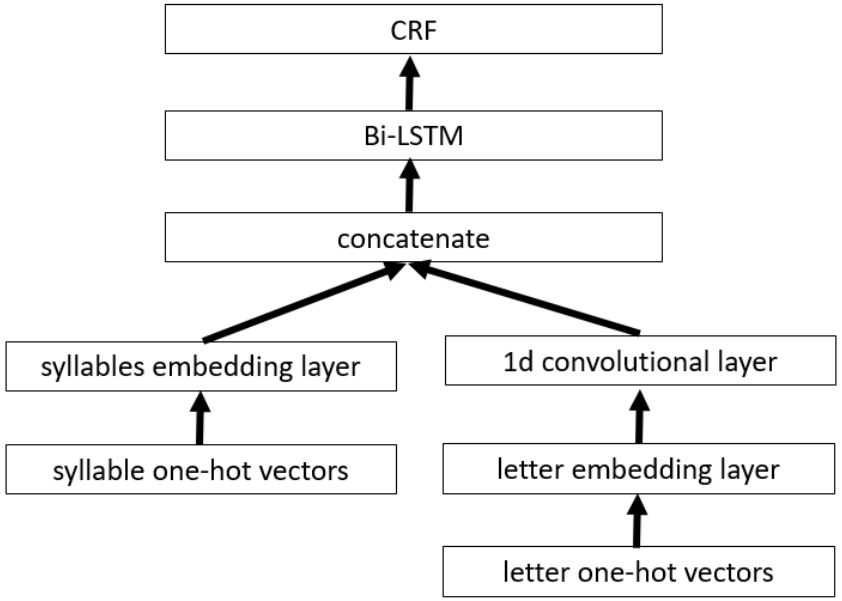
Thử nghiệm đưa toàn bộ dữ liệu về cùng dạng (viết thường) trên nửa đặc trưng đầu vào (đầu vào ký tự hoặc đầu vào âm tiết)

Kết quả trên tập 48k nhãn B-I:

	precision	recall	F1 score
Lower character	96.57	97.20	96.89
Lower syllable	96.74	97.42	97.08

3. Thử nghiệm và đánh giá

Thử nghiệm 4



- Cải tiến mô hình bằng cách thay thế bộ phân loại softmax bằng CRFs

Kết quả trên tập 48k nhãn B-I

precision	recall	F1 score
96.84	97.26	97.05

3. Thử nghiệm và đánh giá

Thử nghiệm 4

Việc sử dụng nhãn B-I không đem lại thông tin quan hệ trên nhãn ngoại trừ nhãn I không thể xuất hiện ở đầu câu

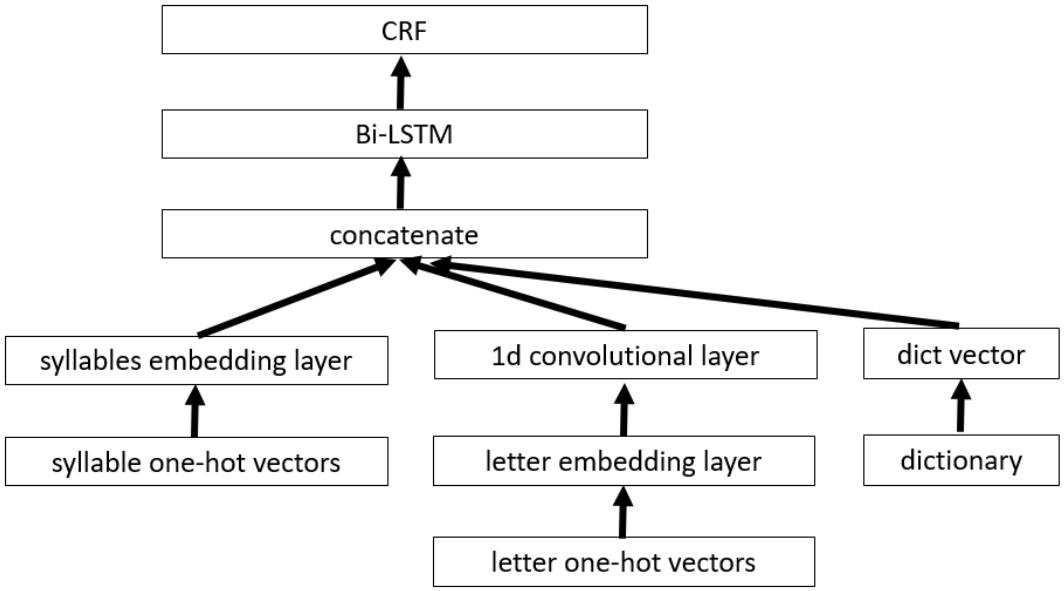
=> Sử dụng nhãn B-I-E cho huấn luyện và đánh giá

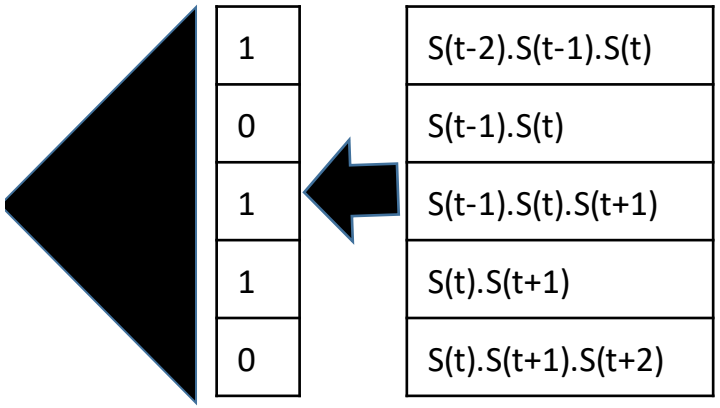
Kết quả trên tập 48k nhãn B-I-E:

	precision	recall	F1 score
Nhãn B-I	96.84	97.26	97.05
Nhãn B-I-E	97.11	97.47	97.29

3. Thử nghiệm và đánh giá

Thử nghiệm 5





- Đưa thêm thông tin từ điển làm đầu vào cho mô hình thông qua vector từ điển

Kết quả trên tập 48k nhãn B-I-E

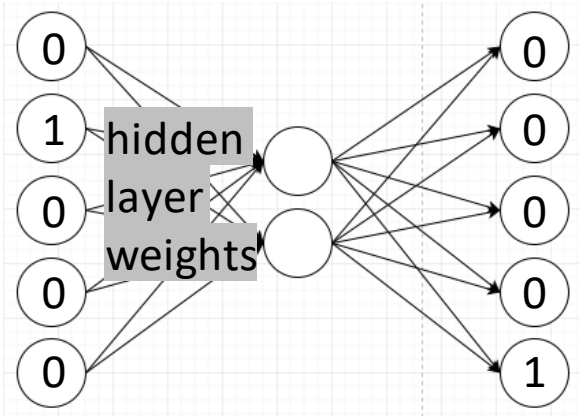
precision	recall	F1 score
98.07	98.40	98.23

3. Thử nghiệm và đánh giá

Thử nghiệm 6

one-hot vector of syllable Si

Tôi	yêu	Việt	Nam
Tôi	Yêu	Việt	Nam
Tôi	Yêu	Việt	Nam
Tôi	Yêu	Việt	Nam



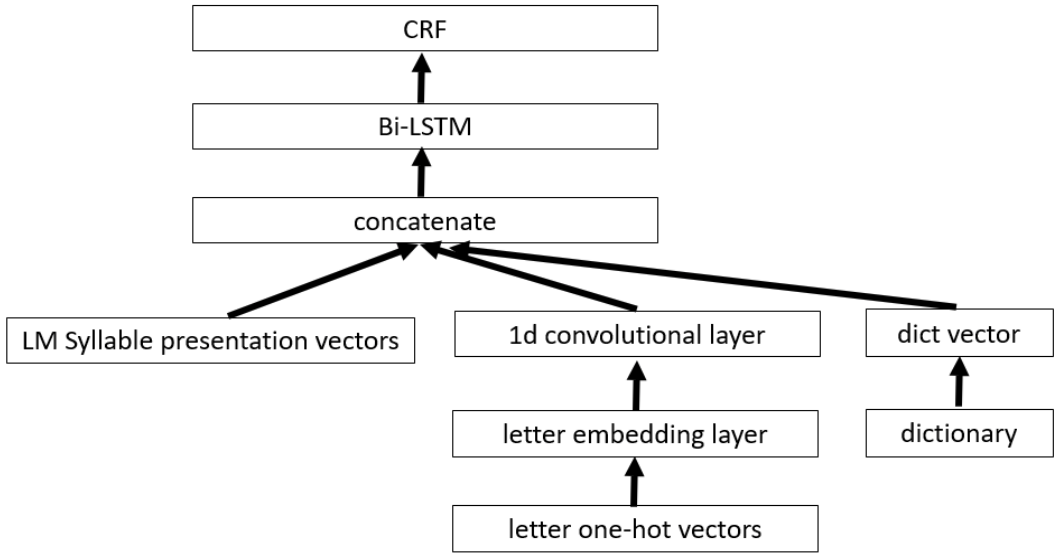
one-hot vector of syllable Si+1

yêu
Việt
Nam
END

- Pretrain vector biểu diễn mức âm tiết trên tập *zingnews* thông qua xây dựng mô hình language model
- Sử dụng hidden layer weights làm tập vector biểu diễn mức âm tiết

3. Thử nghiệm và đánh giá

Thử nghiệm 6

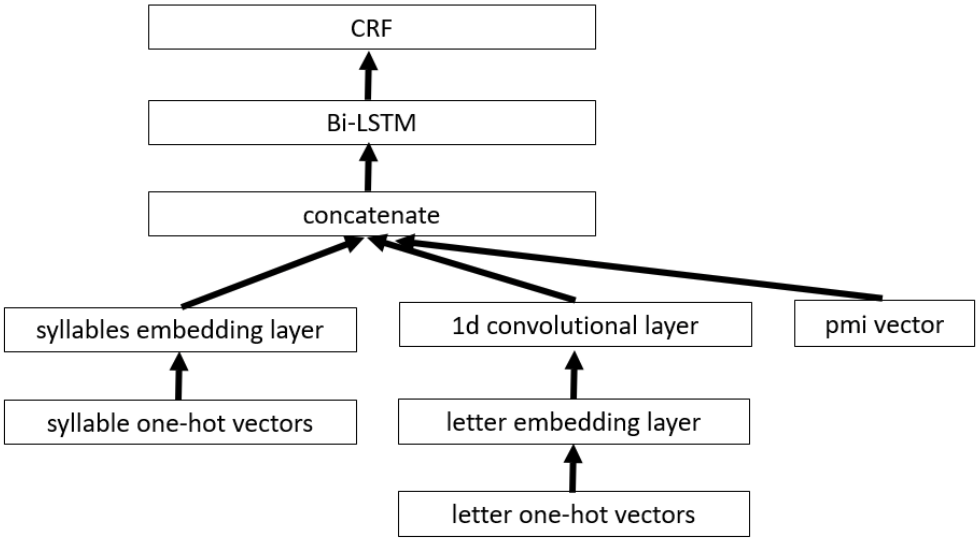


- Pretrain vector biểu diễn mức âm tiết trên tập *zingnews* thông qua xây dựng mô hình language model
- Sử dụng hidden layer weights làm tập vector biểu diễn mức âm tiết
- Kết quả trên tập 48k nhãn B-I-E

precision	recall	F1 score
98.25	98.56	98.41

3. Thử nghiệm và đánh giá

Thử nghiệm 7



- Mô phỏng lại vector từ điển bằng vector pmi

$$pmi(x, y) = log \frac{p(x,y)}{p(x)p(y)}$$

- Giá trị xác suất để tính pmi được thống kê trên tập *zing news*

3. Thử nghiệm và đánh giá

Thử nghiệm 7

Kết quả trên tập 48k nhãn B-I-E

precision	recall	F1 score
95.66	96.90	96.28

Nguyên nhân do dựa vào giá trị pmi không có khả năng phân biệt được kết hợp âm tiết nào sẽ tạo ra từ

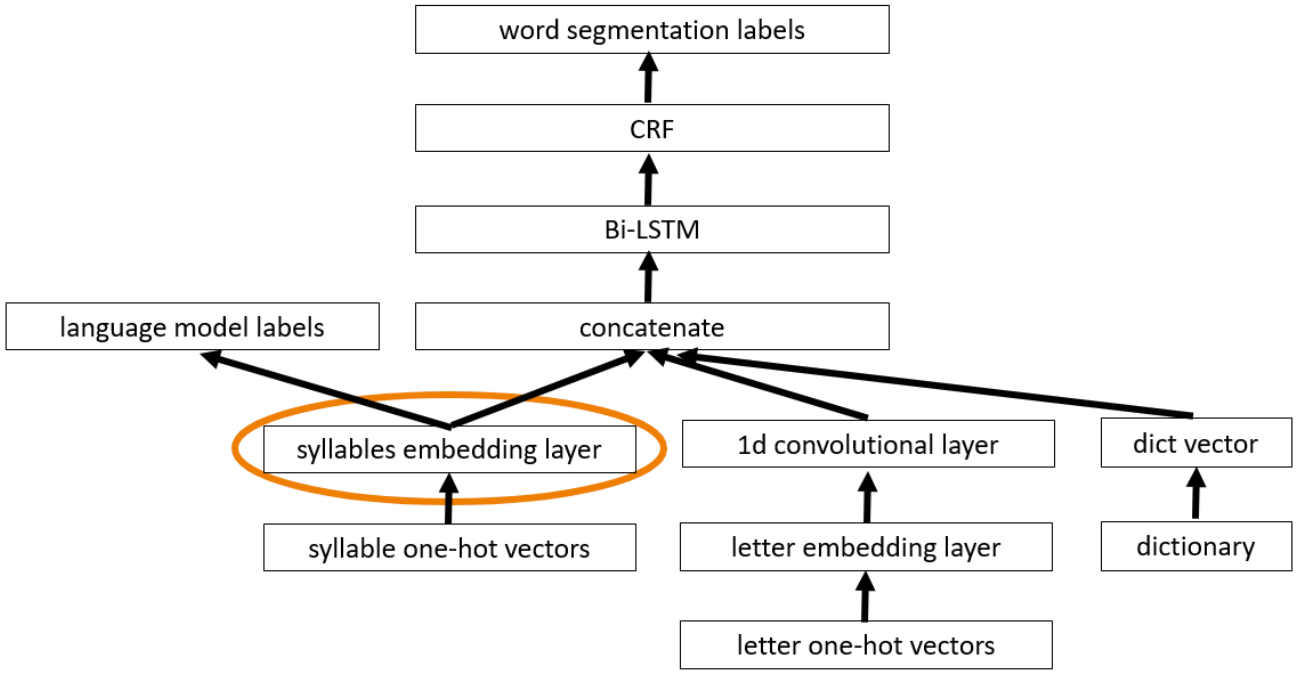
ví dụ:

pmi(cắm cờ) = 6.21, pmi(mang tên) = 4.69

pmi(bộ_đội) = 2.11, pmi(nước_mắt) = 3.57

3. Thử nghiệm và đánh giá

Thử nghiệm 8



- Xây dựng mô hình multi task học đồng thời mô hình ngôn ngữ và tách từ

Kết quả trên tập 48k nhãn B-I-E:

precision	recall	F1 score
98.17	98.53	98.34

3. Thử nghiệm và đánh giá

Tổng hợp kết quả

exp	approach	precision	recall	F1 score
1	BiLSTM + softmax	96.07	96.88	96.48
2	Syllable Embedding Layer + BiLSTM + softmax	96.54	96.83	96.68
3	Syllable Embedding Layer + Character Embedding Layer + CharCNN + BiLSTM + softmax	96.95	97.32	97.13
4	Syllable Embedding Layer + Character Embedding Layer + CharCNN + BiLSTM + CRF	97.11	97.47	97.29
5	Syllable Embedding Layer + Character Embedding Layer + CharCNN + Dict- vector + BiLSTM + softmax	97.93	98.39	98.16
6	Syllable Embedding Layer + Character Embedding Layer + CharCNN + Dict- vector + BiLSTM + CRF	98.07	98.40	98.23
7	Character Embedding Layer + CharCNN + fasttext vector + Dict-vector + BiLSTM + CRF	98.10	98.52	98.31
8	**Character Embedding Layer + CharCNN + LM vector + Dict-vector + BiLSTM + CRF**	**98.25**	**98.56**	**98.41**
9	Character Embedding Layer + CharCNN + LM vector + pmi-vector + BiLSTM + CRF	95.66	96.90	96.28
10	Multi tasks	98.17	98.53	98.34

3. Thử nghiệm và đánh giá

Tổng hợp kết quả

So sánh với bộ tách từ UETSegmenter trên tập 48k

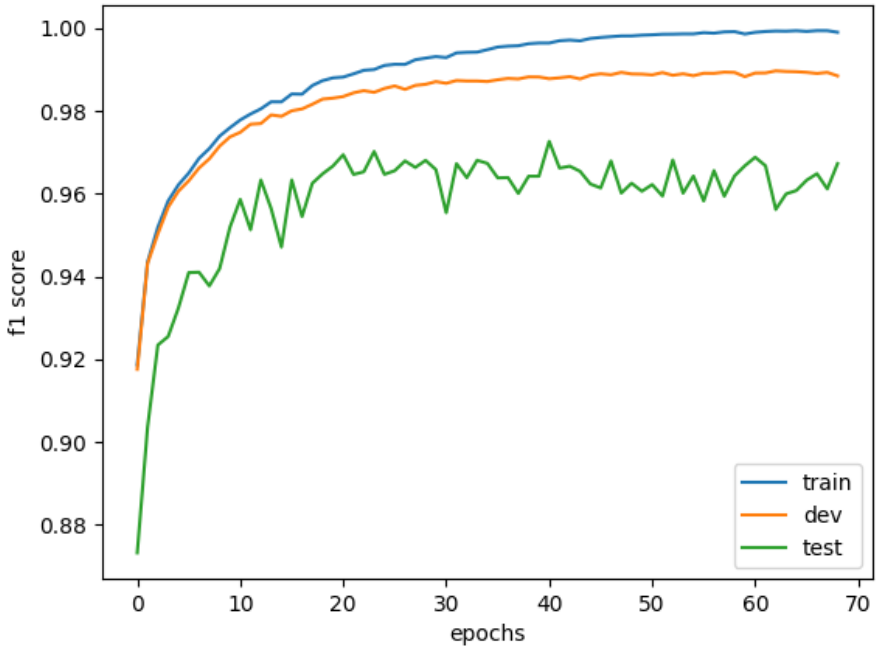
	precision	recall	F1 score
UETSegmenter	97.90	98.33	98.11
My approach	98.25	98.56	98.41

So sánh với bộ tách từ UETSegmenter và RDRSegmenter trên tập 75k

	precision	recall	F1 score
UETSegmenter	**97.51**	98.23	97.87
RDRSegmenter	97.46	**98.35**	**97.90**
My approach	96.69	97.84	97.26

3. Thử nghiệm và đánh giá

Tổng hợp kết quả



Mô hình đề xuất đạt kết quả cao hơn trên tập 48k tuy nhiên lại không đạt kết quả tốt trên tập 75k

Nguyên nhân:

- Tập 48k gồm dữ liệu cùng domain về báo và truyện
- Tập 75k sử dụng train và dev từ VLSP thuộc domain báo và truyện trong khi test thuộc domain văn bản pháp luật

3. Thử nghiệm và đánh giá

Nhận xét

- Mô hình học sâu phù hợp với bài toán tách từ khi đạt kết quả cao
- Vector từ điển đem lại hiệu quả rất tốt cho mô hình (tăng > 1%)
- Bộ phân loại CRFs luôn cho kết quả cao hơn softmax
- Mô hình đề xuất thu được kết quả rất tốt với dữ liệu thuộc cùng domain, kết quả thu được tốt hơn mô hình UETSegmenter
- Mô hình đề xuất cho kết quả chưa đủ cạnh tranh với mô hình tách từ tốt nhất hiện tại là UETSegmenter và RDRSegmenter trên dữ liệu không cùng domain

4. Kết luận

Kết quả đạt được

- Tìm hiểu về bài toán tách từ và lý thuyết về từ trong tiếng Việt
- Tìm hiểu về các mô hình học máy và học sâu, đặc biệt là các mô hình được áp dụng vào bài toán tách từ
- Thử nghiệm những mô hình đã có và đề xuất một vài kiến trúc mô hình mới nhằm giải quyết bài toán tách từ trên bộ dữ liệu tiếng Việt.
- Kết quả thu được cao hơn bộ tách từ UETSegmenter trên tập dữ liệu sử dụng

4. Kết luận

Hướng phát triển trong tương lai

- Tìm cách mô phỏng véc-tơ từ điển thông qua mô hình học không giám sát trên dữ liệu lớn hoặc thay đổi mô hình để có khả năng biểu diễn được thông tin này
- Sử dụng thêm các phương pháp biểu diễn thông tin mức âm tiết hoặc dưới âm tiết
- Sử dụng đặc trưng bằng cách fine-tune các mô hình biểu diễn ngôn ngữ mới và đang rất nổi vào thời điểm hiện tại (BERT, ELMo)
- Sử dụng thêm post processing kết quả đầu ra của mô hình nhằm tăng độ chính xác
- Nghiên cứu và cải tiến mô hình nhằm đạt kết quả cao hơn với dữ liệu khác domain

CẢM ƠN THẦY CÔ VÀ CÁC BẠN ĐÃ LẮNG NGHE!

Tài liệu tham khảo

[1] Nguyễn Thị Minh Huyền, Hoàng Thị Tuyền Linh, Vũ Xuân Lương, Hướng dẫn nhận biết đơn vị từ trong văn bản tiếng Việt, Vietnamese Language and Speech Processing workshop, 2013.

[2] Prajit Ramachandran, Barret Zoph, Quoc V. Le, Searching for Activation Functions, arXiv preprint arXiv:1710.05941, 16 Oct 2017.

[3] Florian Schroff, Dmitry Kalenichenko, James Philbin, FaceNet: A Unified Embedding for Face Recognition and Clustering, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2015.

[4] Baidu Research – Silicon Valley AI Lab, Deep Speech 2: End-to-End Speech Recognition in English and Mandarin, arXiv preprint arXiv:1512.02595, 8 Dec 2015.

[5] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, in Journal of Machine Learning Research volume 15(Jun): 1929- 1958, 2014.

[6] LeCun, Yann; Léon Bottou; Yoshua Bengio; Patrick Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE. 86 (11): 2278–2324.

[7] Sepp Hochreiter, Ju¨rgen Schmidhuber, Long Short-Term Memory, in Neural Computation Journal, Volume 9 Issue 8: 1735-1780 November 15, 1997.

[8] Tobias Glasmachers, Limits of End-to-End Learning, arXiv preprint arXiv:1704.08305v1, 26 Apr 2017.

Tài liệu tham khảo

[9] Sebastian Ruder, An Overview of Multi-Task Learning in Deep Neural Networks, arXiv preprint arXiv:arXiv:1706.05098v1, 15 Jun 2017.

[10] Marco Baroni, Georgiana Dinu, German Kruszewski, Don't count, predict! A systematic comparison of context-counting vs. context-predicting seman- tic vectors, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), June 2014.

[11] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, FastText.zip: Compressing text classification models, arXiv preprint arXiv:1612.03651, 12 Dec 2016.

[12] T. P. Nguyen and A. C. Le, A hybrid approach to Vietnamese word segmen- tation, in IEEE RIVF International Conference on Computing Communica- tion Technologies, Research, Innovation, and Vision for the Future (RIVF), pages 114-119, 2016.

[13] Dat Quoc Nguyen and Dai Quoc Nguyen and Thanh Vu and Mark Dras and Mark Johnson, A Fast and Accurate Vietnamese Word Segmenter, Pro- ceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), pages 2582-2587.

[14] Le, H. P., Nguyen, T. M. H., Roussanaly, A., and Ho, T. V., A hybrid ap- proach to word segmentation of Vietnamese texts, In Proceedings of the 2nd International Conference on Language and Automata Theory and Applica- tions, pages 240–249.

[15] Van-Duyet Le, vietnamese-wordlist, https://vietnamese-wordlist.duyet.net.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understand- ing, arXiv preprint arXiv:1810.04805, 11 Oct 2018.

[17] Peters, Matthew E. and Neumann, Mark and Iyyer, Mohit and Gardner, Matt and Clark, Christopher and Lee, Kenton and Zettlemoyer, Luke, Deep contextualized word representations, NAACL 2018.