Predicate-Argument Based Bi-Encoder for Paraphrase Identification

Qiwei Peng[1], David Weir[1], Julie Weeds[1], Yekun Chai[2]

[1]University of Sussex , [2]Baidu, Inc

University of Sussex

1 Paraphrase Identification

"Marriage equality law passed in Rhode Island"

In both directions

"Rhode Island becomes the 10th state to enact marriage equality"

Paraphrase Identification:

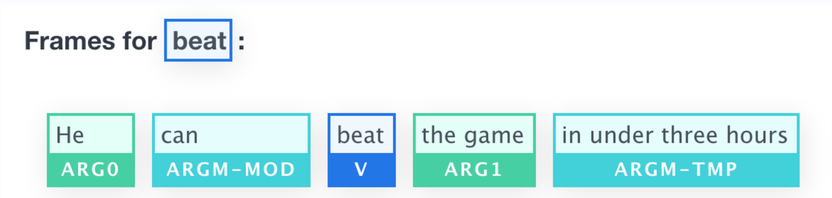
- Generally considered as a symmetric task where the relation holds in both directions (Bhagat and Hovy, 2013)
- Paraphrase pairs are either fully or largely semantically equivalent
- Effective paraphrase models are expected to be structure-aware and word order sensitive

2 Motivation

Background:

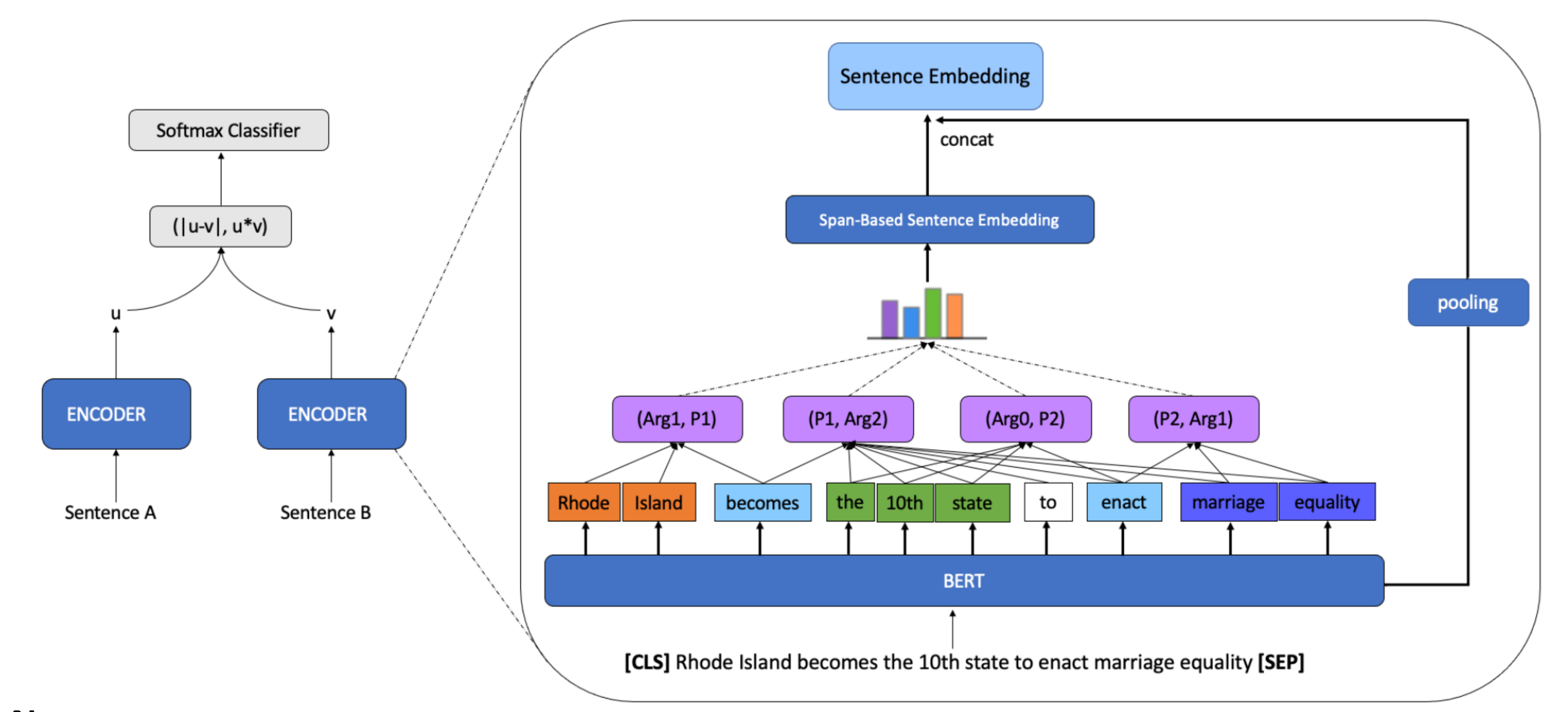
- Various efforts have been put into introducing structural information to pre-trained models (Zhang et al., 2020, Yin et al., 2020 , Wu et al., 2021, Peng et al., 2021), however, many of them introduce **a huge number of** additional parameters
- Cross-encoders face challenges from both extreme computation overhead for many use cases (Thakur et al., 2021) and inconsistent predictions when dealing with symmetric tasks (Chen et al., 2020)

The Question: How can we introduce structural information into pre-trained bi-encoders in a simple but effective way?



Inspired by Sun et al. (2020), we propose a method that effectively introduces sentence structures into bi-encoders via the weighted aggregation of predicate-argument spans with limited additional parameters

3 The Model



Notes:

(The proposed model in a twin structure)

- Span representations are obtained via mean-pooling over all tokens in the span. Learnable weights are applied to the aggregation of predicate-argument spans
- We concatenate the span-based sentence embedding with the original last-avg BERT sentence representation
- The interaction between two sentences is changed from (u, v, |u-v|) to (|u-v|, u ∗ v) to ensure symmetry (compared to the original strategy in SBERT)
- Predicate-Argument spans are obtained via AllenNLP with its semantic role labelling tagger

4 Main Evaluation

	QQP	TwitterURL	MSRP	PAWS_Wiki	PAWS_QQP	PIT2015
SBERT	**90.78±0.09**	70.85±0.28	81.67±0.46	81.57±0.53	66.01±0.45	52.03±1.44
SBERT-RGCN	90.41±0.09	70.40±0.22	81.70±0.17	81.14±0.81	66.22±0.75	59.11±0.93
PAS+SBERT	90.74±0.06	**72.12±0.26**	**83.42±0.23**	**82.60±0.18**	**68.85±0.73**	**59.19±1.85**
SRoBERTa	**90.79±0.09**	70.69±0.23	81.69±0.53	81.42±0.93	67.35±0.97	52.67±2.75
PAS+SRoBERTa	90.76±0.03	**72.04±0.23**	**83.22±0.46**	**82.87±0.35**	**69.68±0.72**	**59.50±2.74**

The proposed model is effective on 5 out of 6 PI tasks and also show competitive performance on QQP

The parameters introduced by PAS is **very limited**

	Params
SBERT-base	109M
PAS only	+768
PAS+SBERT	+3840
SBERT-RGCN	+ 32M

5 Analysis

Compared to Simple Average

	QQP	TwitterURL	MSRP	PAWS_Wiki	PAWS_QQP	PIT2015
PAS+SBERT	90.74±0.06	72.12±0.26	83.42±0.23	82.60±0.18	68.85±0.73	59.19±1.85
- SBERT-only	**90.78±0.09**	70.85±0.28	81.67±0.46	81.57±0.53	66.01±0.45	52.03±1.44
- PAS only	90.70±0.08	**71.64±0.14**	**82.91±0.12**	**82.26±0.34**	**67.38±0.22**	**54.95±1.45**
- PAS only (simple average)	90.11±0.13	71.09±0.30	82.13±0.14	81.85±0.26	66.55±0.41	51.82±1.31

The PAS component plays an important role in performance gain

The learnable weights for aggregation is effective

We report the F1 score of the positive class over 5 random runs with standard error

Compared to Random Span

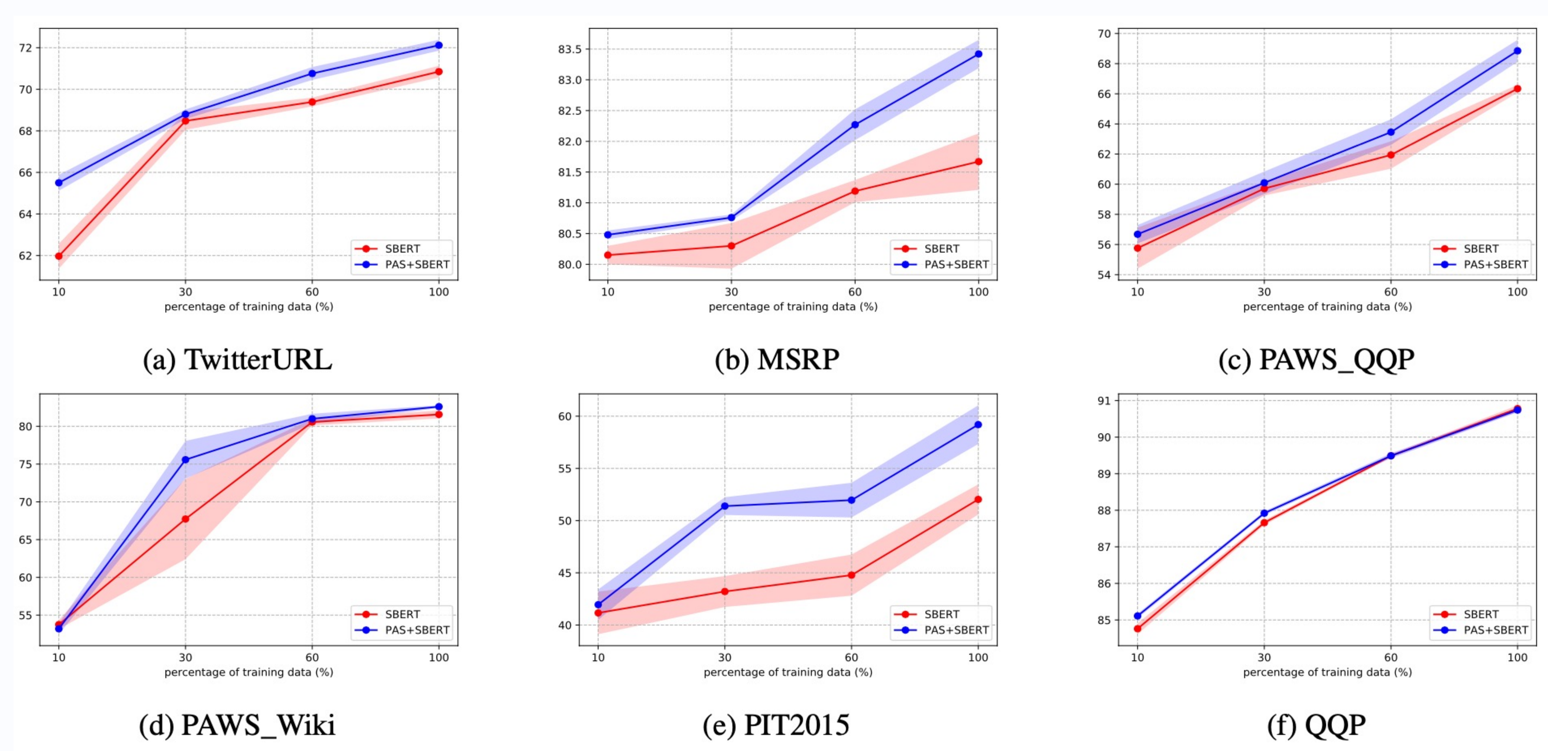
Task	Span Type	Span only	Self-Explain*	SBERT
MSRP	PAS	**82.91±0.12**	81.23±0.27	81.67±0.46
	Continuous Random Span	81.40±0.43		
	Random Span	81.86±0.47		
PAWS_QQP	PAS	**67.38±0.22**	66.88±0.46	66.01±0.45
	Continuous Random Span	65.45±0.44		
	Random Span	65.75±0.74		
PIT2015	PAS	**54.95±1.45**	47.60±1.01	52.03±1.44
	Continuous Random Span	51.62±1.92		
	Random Span	50.85±2.11		

Continuous Random Span -> We randomly sample continuous word sequences from the sentence to build a span

Random Span -> We do not necessarily sample continuous words, but allow word leaps from one to another

It is the **predicate-argument span** that makes the big difference!

Compared to Different Training Size



In spite of limited increased parameters, the proposed model appears to yield **consistent improvements** across different training scales

6 Conclusion

- We propose a method which effectively introduces sentence structure to a sentence embedding via the aggregation of predicate-argument spans (PAS)
- Our model brings improvements on six paraphrase identification tasks
- Upon closer investigation, we show that the PAS component and its learnable weights play a substantial impact in the performance gain
- This PAS component, as demonstrated with SRoBERTa, can be easily extended to other models that require the generation of sentence embeddings
- Compared to RGCN, the PAS component brings in very limited parameters

References and Acknowledgements

[1] Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? Computational Linguistics, 39(3):463–472.

[2] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 9628–9635.

[3] Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sen- tibert: A transferable transformer-based architecture for compositional sentiment semantics. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3695–3706.

[4] Zhaofeng Wu, Hao Peng, and Noah A Smith. 2021. Infusing finetuning with semantic dependencies. Transactions of the Association for Computational Linguistics, 9:226–242.

[5] Qiwei Peng, David Weir, and Julie Weeds. 2021. Structure-aware sentence encoder in bert-based Siamese network. In Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021), pages 57–63, Online. Association for Computational Linguistics.

[6] Zijun Sun, Chun Fan, Qinghong Han, Xiaofei Sun, Yuxian Meng, Fei Wu, and Jiwei Li. 2020. Self- explaining structures improve nlp models. arXiv preprint arXiv:2012.01786.