Predicate-Argument Based Bi-Encoder for Paraphrase Identification

Qiwei Peng, David Weir, Julie Weeds, Yekun Chai

UNIVERSITY OF SUSSEX

Paraphrase Identification

What is a paraphrase pair?

Paraphrases are sentences that express the same or similar meanings with different wording (Bhagat and Hovy, 2013):

- They are either **fully or largely** semantically equivalent.
- It is generally considered to be a **symmetric** task where the paraphrase relation holds in both directions.

For example:

a) Marriage equality law passed in Rhode Island

↕ in both direction

b) Rhode Island becomes the 10th state to enact marriage equality

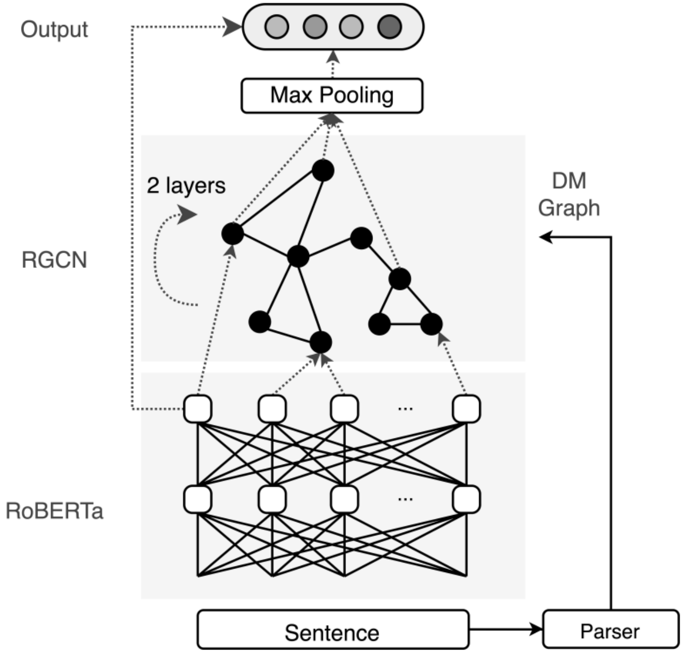
Structural Awareness

- Word order and sentence structure are crucial in determining sentence meaning [the meaning of a sentence can get changed completely by a simple swapping of two words in a sentence]
- Effective paraphrase models are expected to be structure-aware and word order sensitive

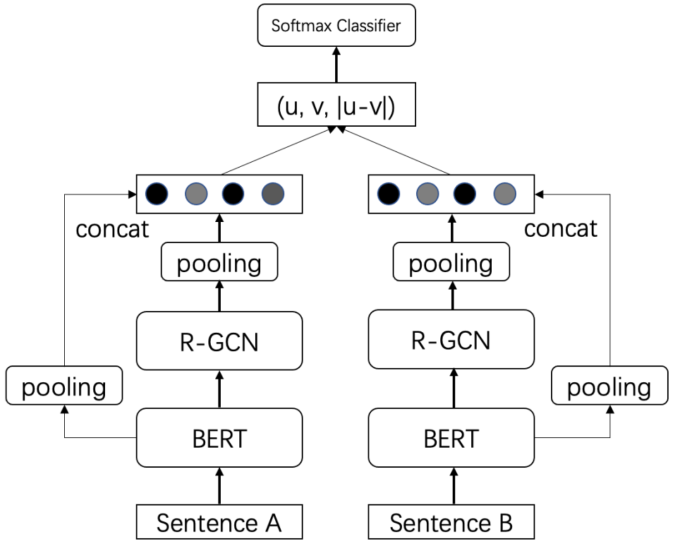
Various efforts put into introducing structural information to pre-trained models.

Datasets concerns structural differences (e.g., PIT2015, PAWS)

non-paraphrase pairs that are lexically similar but semantically dissimilar.



(Wu et al. 2021)



(Peng et al., 2021)

They bring in a large number of additional parameters!

Two main approaches: Cross-Encoder VS Bi-Encoder

Cross-Encoder is widely used for various sentence-pair tasks, but it faces challenges:

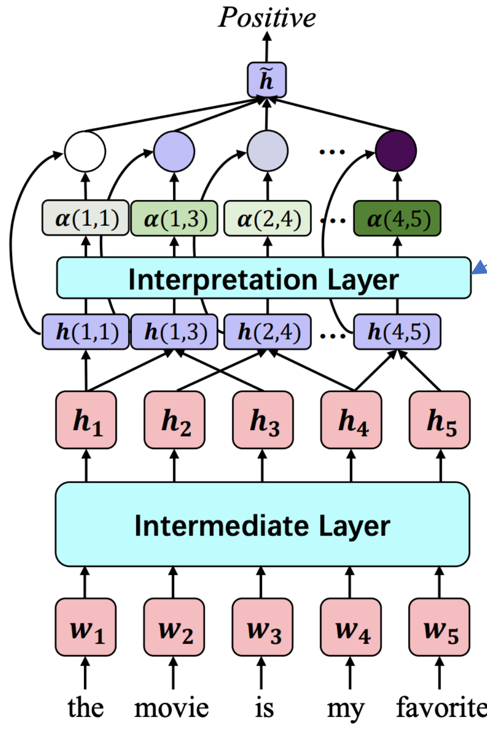
- Extreme computation overhead for many use cases (Reimers and Gurevych, 2019, Thakur et al., 2021)
- Inconsistent predictions when dealing with symmetric tasks (Chen et al., 2020)

In this work, we stick with pre-trained bi-encoders (e.g., SBERT (Reimers and Gurevych, 2019))!

Motivation

- Structural information is important in determining sentence meaning and meaning comparison
- RGCNs are useful in incorporating structural information but bring in a large number of parameters

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



(Model proposed in Sun et al., 2020)

Sun et al. (2020) generate sentence embedding by aggregating all possible ***continuous text spans***

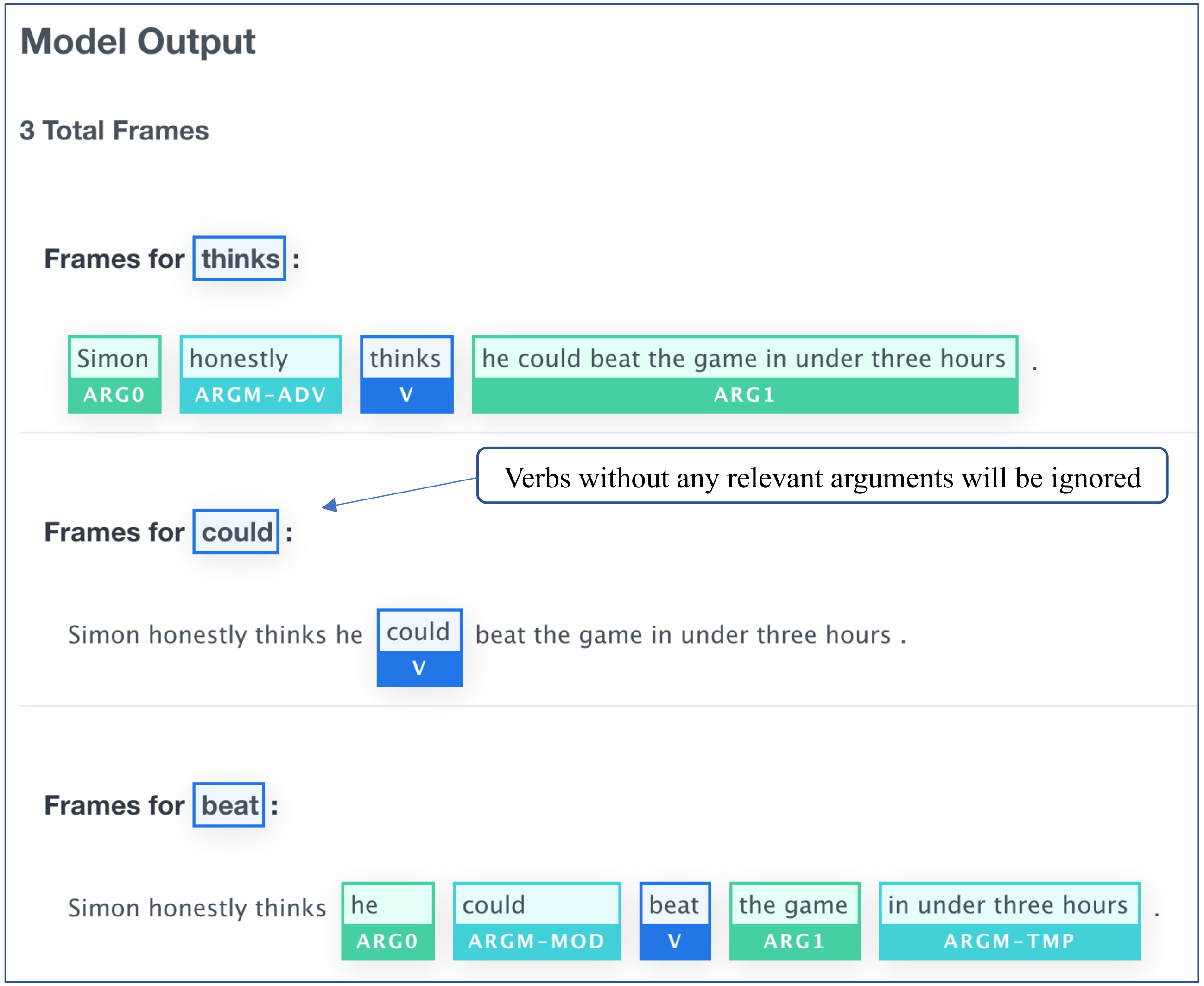
Inspired by Sun et al., we propose a method that effectively introduces sentence structures into bi-encoders via the weighted aggregation of **predicate-argument spans**.

continuous text spans → **Predicate-Argument Spans**

Predicate-Argument Spans

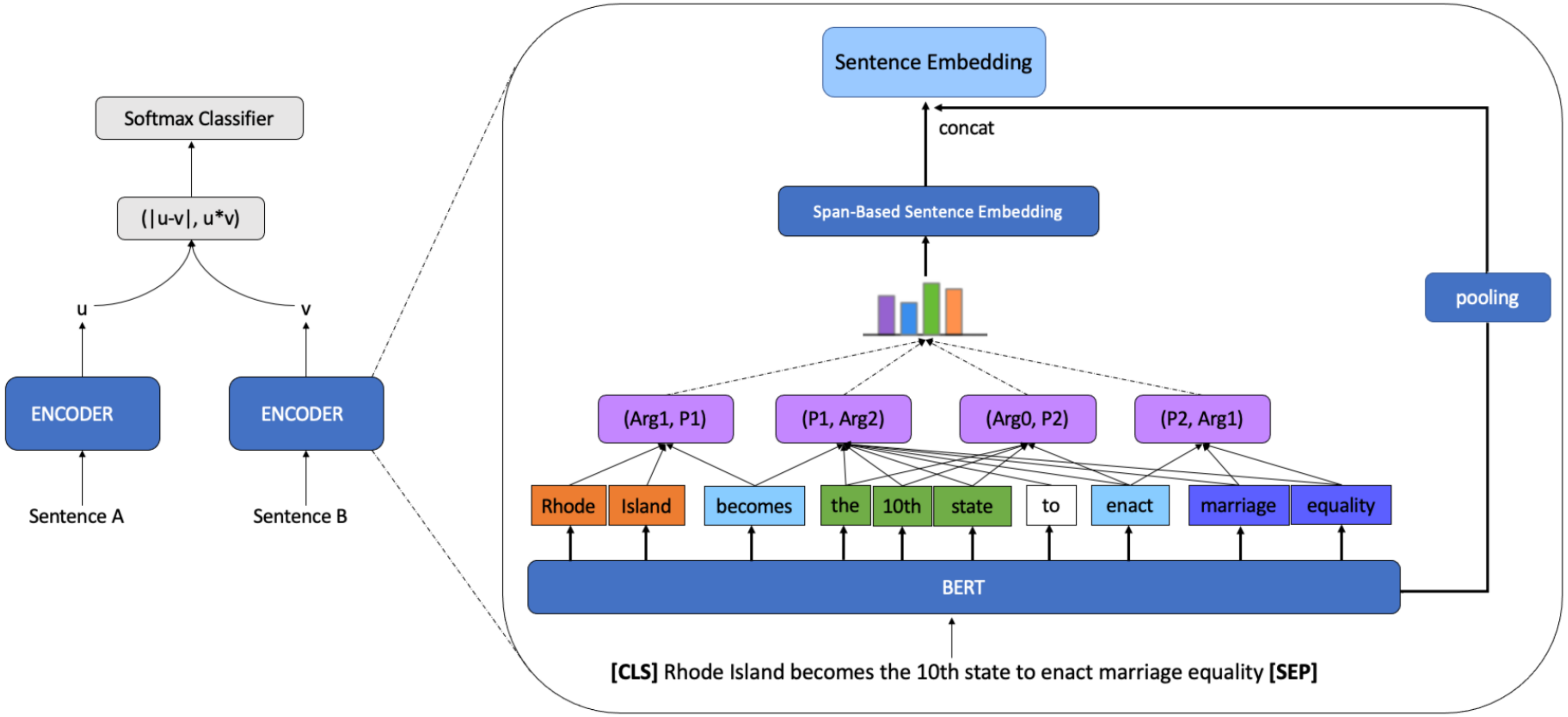
How do we obtain such spans?

Simon honestly thinks he could beat the game in under three hours



We use SRL tagger from AllenNLP (Gardner et al., 2018) to obtain predicate-argument spans

Our Approach



- Group related predicate arguments
- Mean-pooling based span representation
- Weighted aggregation (learnable)
- Concatenation with original BERT representation
- Change from (u, v, |u-v|) to (|u-v|, $u * v$) to ensure symmetry

Experiments

Datasets	Train	Dev	Test
MSRP	3,668	408	1,725
TwitterURL	37,976	4,224	9,334
PIT2015	11,530	4,142	838
QQP	384,348	10,000	10,000
PAWS_QQP	11,986	8,000	677
PAWS_Wiki	49,401	8,000	8,000

- Evaluate on 6 PI datasets
- Report F1 score of the positive class
- 5 runs

Evaluations

Main experiments

	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 achieves best performance on 5 out of 6 PI tasks and also shows competitive performance on QQP

The parameters introduced by PAS is very little

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

Parameter comparison between different models

Compared to Simple Average

When changing weighted sum 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

Random Spans

When changing PAS to random spans

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		

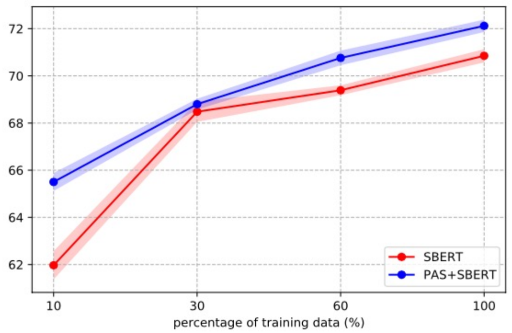
Continuous Random Span -> We randomly sample continuous word sequences from the sentence to build a span

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

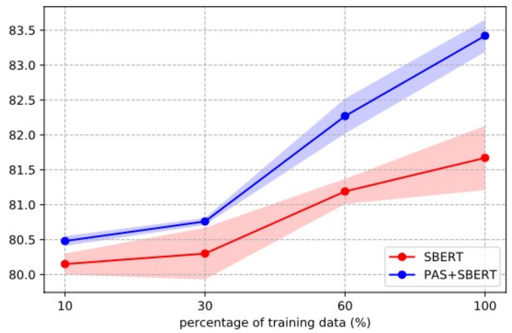
It is the predicate-argument span that makes the big difference!!!

Training Size

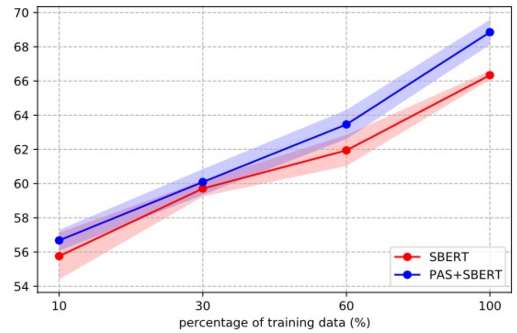
When only limited training data is available



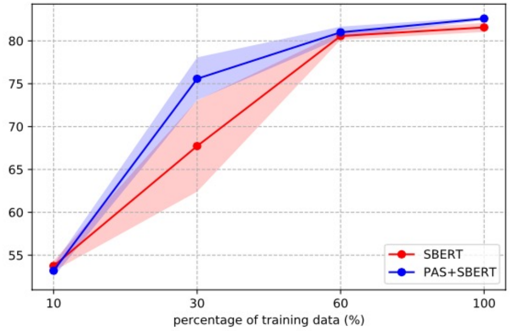
(a) TwitterURL



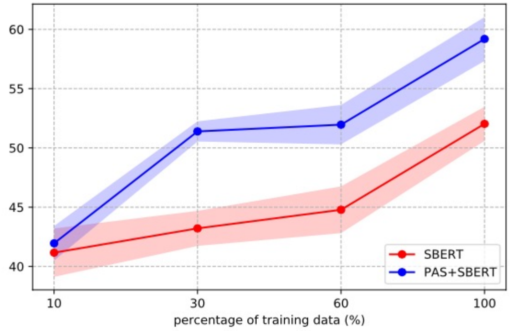
(b) MSRP



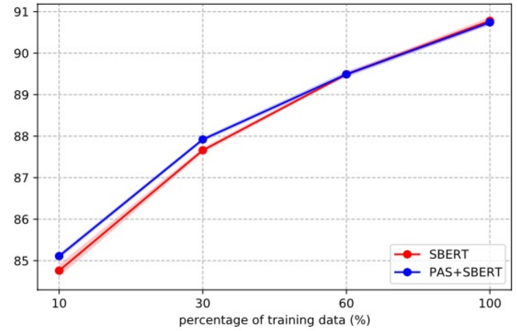
(c) PAWS_QQP



(d) PAWS_Wiki



(e) PIT2015



(f) QQP

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

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 substantial impacts 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

Rahul Bhagat and Eduard Hovy. 2013. What is a para- phrase? Computational Linguistics, 39(3):463–472.

Hannah Chen, Yangfeng Ji, and David K Evans. 2020. Pointwise paraphrase appraisal is potentially problematic. In Proceedings of the 58th Annual Meet ing of the Association for Computational Linguistics: Student Research Workshop, pages 150–155.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992.

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

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.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sentibert: 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.

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.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Pe- ters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language pro- cessing platform. In Proceedings of Workshop for NLP Open Source Software (NLP-OSS), pages 1– 6, Melbourne, Australia. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. In Proceedings of the 2021 Conference of the North American Chap- ter of the Association for Computational Linguistics: Human Language Technologies, pages 296–310.