The Alan Turing Institute

WARWICK
THE UNIVERSITY OF WARWICK

Queen Mary
University of London

Utrecht University

tBERT: Topic Models and BERT Joining Forces for Semantic Similarity Detection

Nicole Peinelt, Dong Nguyen, Maria Liakata

ACL, 6-8 July 2020

Authors

Nicole Peinelt

The Alan Turing Institute, UK
University of Warwick, UK

Dr. Maria Liakata

The Alan Turing Institute, UK
University of Warwick, UK
Queen Mary University, UK

Dr. Dong Nguyen

The Alan Turing Institute, UK
Utrecht University, The Neverlands

Semantic Similarity Detection for English

– Given a sentence pair, predict binary relatedness

Dataset	Task
Quora	paraphrase detection
MSRP	paraphrase detection
SemEval	(A) internal answer ranking
	(B) paraphrase ranking
	(C) external answer ranking

Semantic Similarity Detection for English

– Given a sentence pair, predict binary relatedness

Dataset	Task
Quora	paraphrase detection
MSRP	paraphrase detection
SemEval	(A) internal answer ranking (B) paraphrase ranking (C) external answer ranking

Example from Quora:

Sentence Pair:

Which is the best way to learn coding?
How do you learn to program?

Label:
is_paraphrase

Motivation

Feature-engineering

Qin et al. (2009),
Tran et al. (2015),
Mihaylov and Nakov (2016),
Filice et al. (2017)

Motivation

Feature-engineering

Qin et al. (2009),
Tran et al. (2015),
Mihaylov and Nakov (2016),
Filice et al. (2017)

Neural architectures

Wang et al. (2017),
Deriu and Cieliebak (2017),
Tan et al. (2018),
Gong et al. (2018),
Devlin et al. (2019)

Motivation

Feature-engineering

Qin et al. (2009),
Tran et al. (2015),
Mihaylov and Nakov (2016),
Filice et al. (2017)

Hybrid approaches

Feng et al. (2017),
Koreeda et al. (2017),
Wu et al. (2017)

Neural architectures

Wang et al. (2017),
Deriu and Cieliebak (2017),
Tan et al. (2018),
Gong et al. (2018),
Devlin et al. (2019)

Motivation

Feature-engineering

Qin et al. (2009),
Tran et al. (2015),
Mihaylov and Nakov (2016),
Filice et al. (2017)

Hybrid approaches

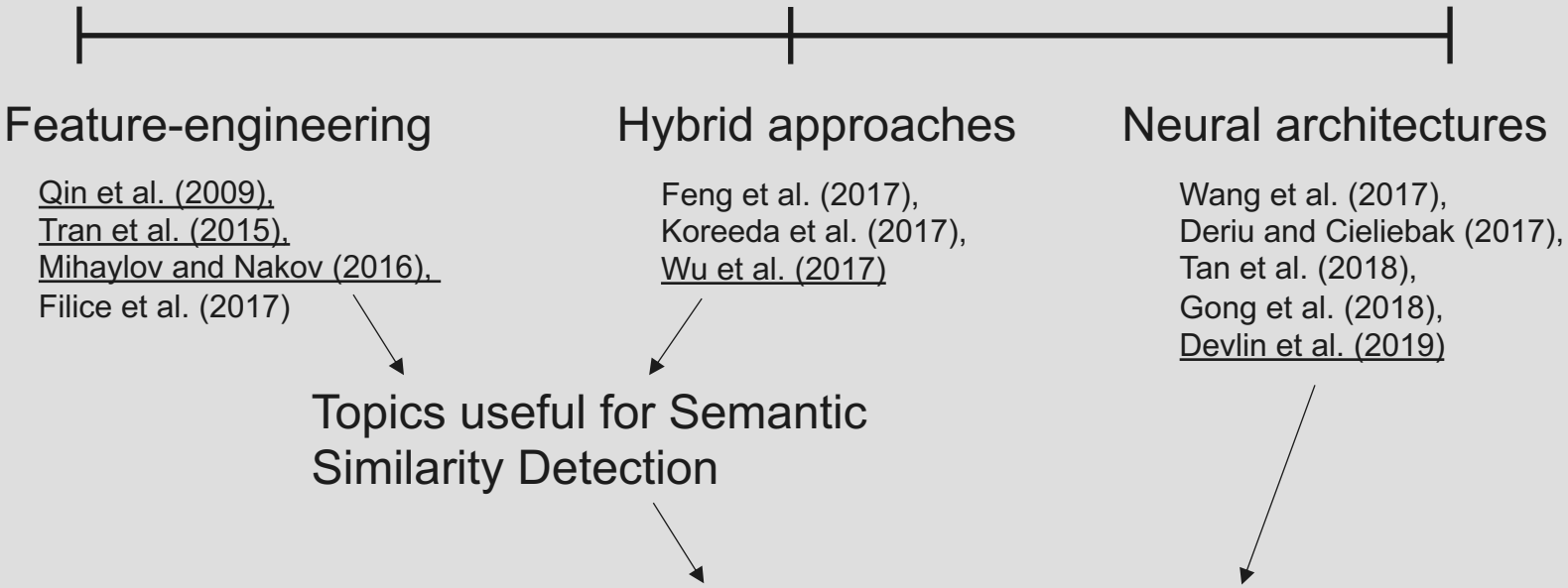
Feng et al. (2017),
Koreeda et al. (2017),
Wu et al. (2017)

Neural architectures

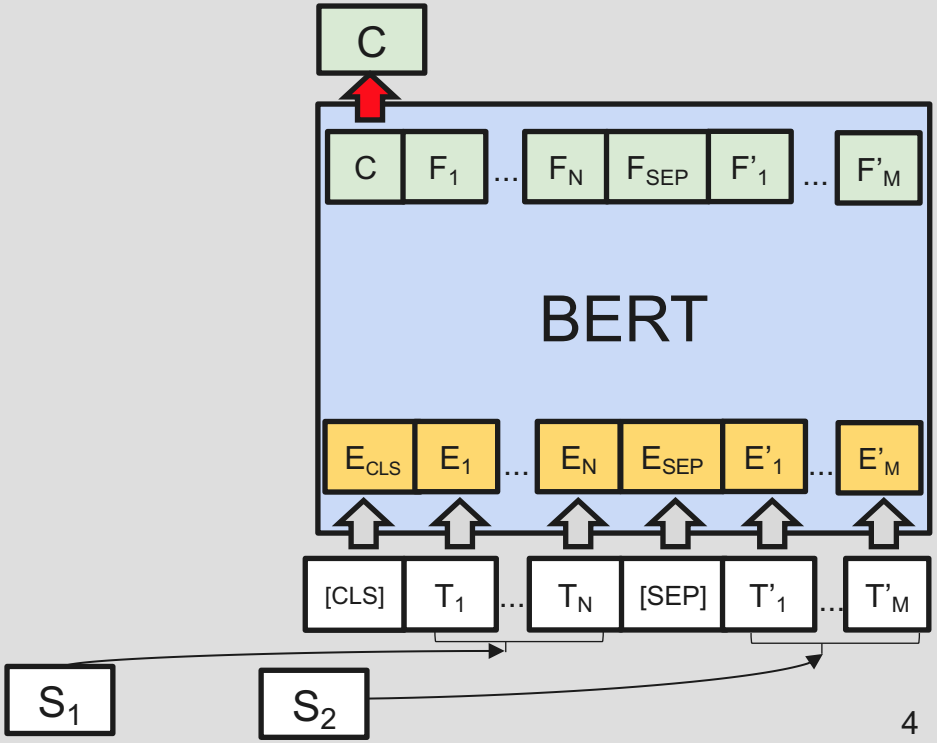
Wang et al. (2017),
Deriu and Cieliebak (2017),
Tan et al. (2018),
Gong et al. (2018),
Devlin et al. (2019)

Topics useful for Semantic Similarity Detection

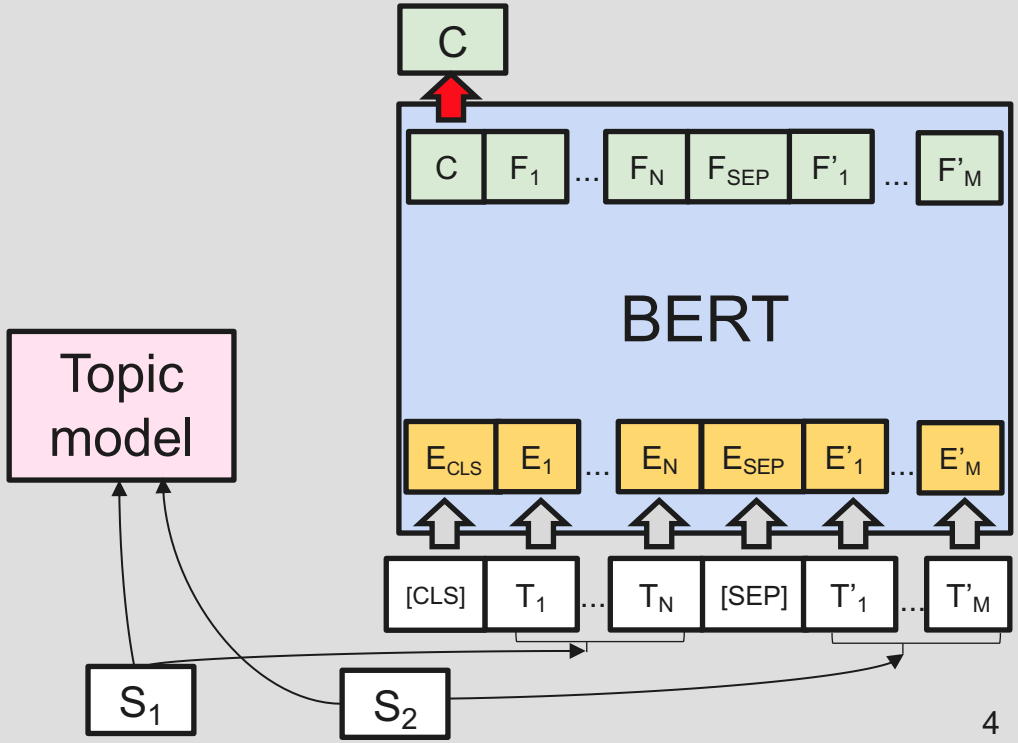
Motivation


Feature-engineering
Qin et al. (2009),
Tran et al. (2015),
Mihaylov and Nakov (2016),
Filice et al. (2017)
Hybrid approaches
Feng et al. (2017),
Koreeda et al. (2017),
Wu et al. (2017)
Neural architectures
Wang et al. (2017),
Deriu and Cieliebak (2017),
Tan et al. (2018),
Gong et al. (2018),
Devlin et al. (2019)
Topics useful for Semantic Similarity Detection
→Idea: Combining topic models and BERT

tBERT

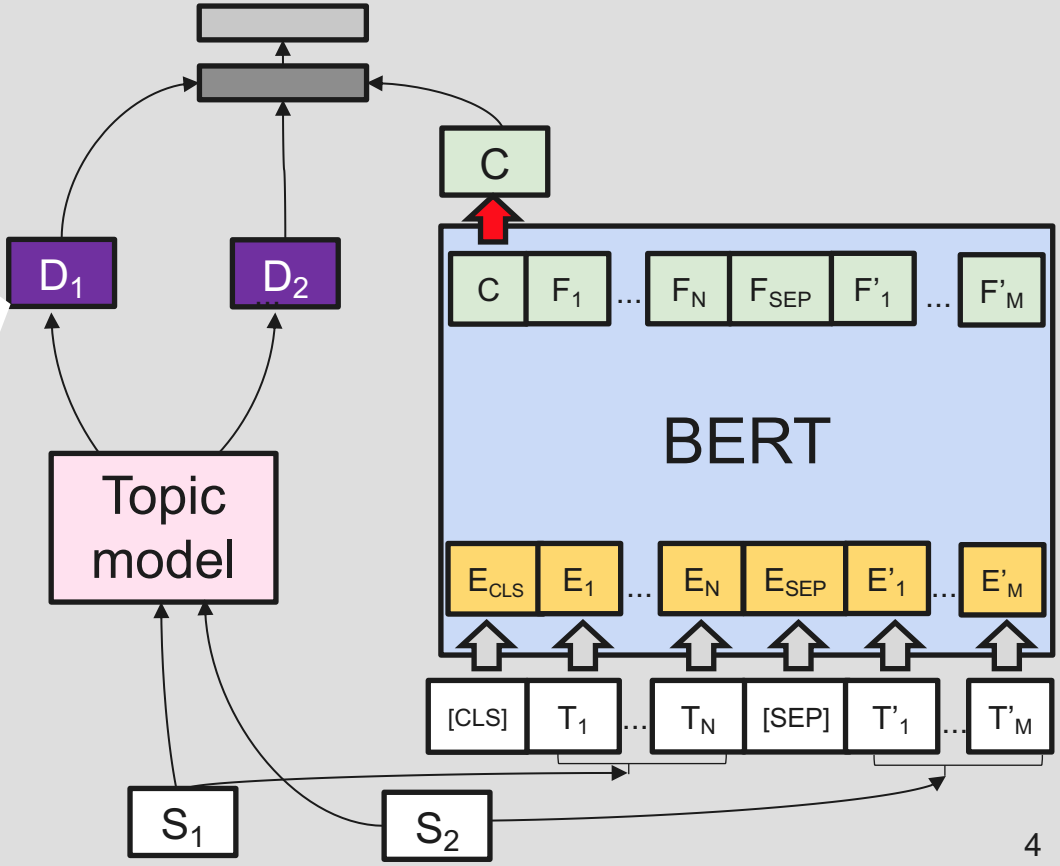

C
C
F1
...
FN
FSEP
F'1
...
F'M
BERT
ECLS
E1
...
EN
ESEP
E'1
...
E'M
[CLS]
T1
...
TN
[SEP]
T'1
...
T'M
S1
S2

tBERT


C
C
F1
FN
FSEP
F'1
F'M
BERT
Topic model
ECLS
E1
EN
ESEP
E'1
E'M
[CLS]
T1
TN
[SEP]
T'1
T'M
S1
S2

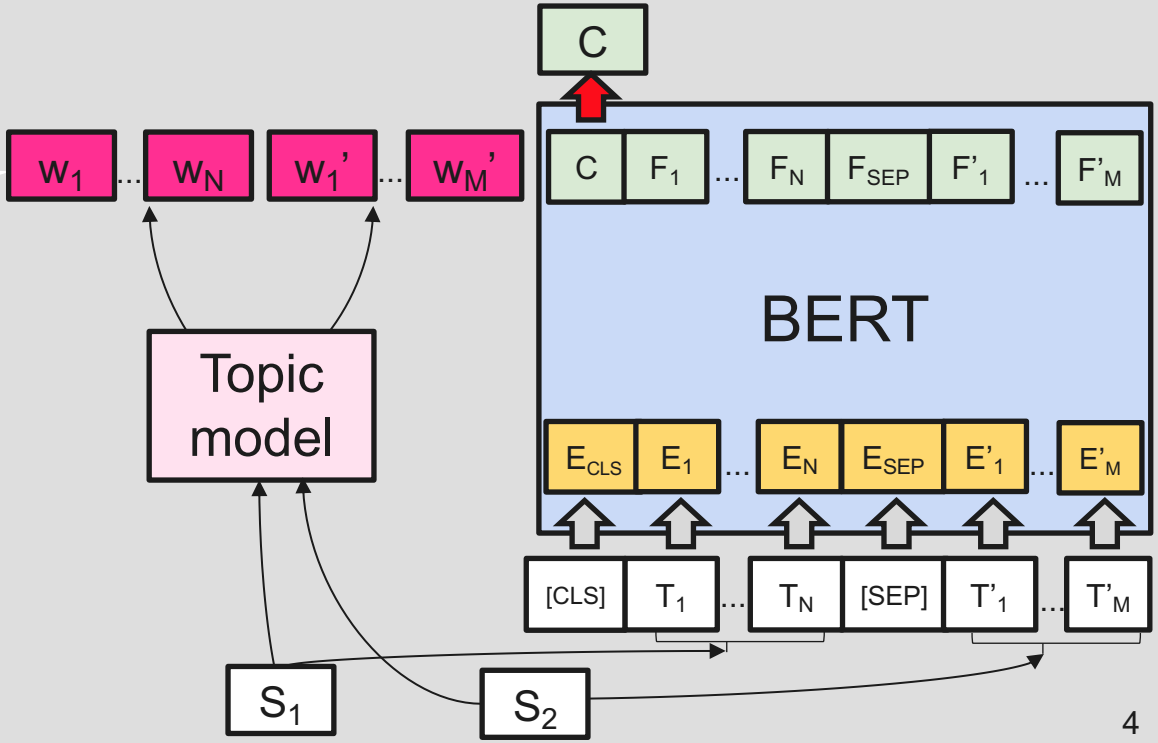
tBERT with document topics

$$D_1 = \text{TopicModel}([T_1, ..., T_N]) \in R^t$$
$$D_2 = \text{TopicModel}([T'_1, ..., T'_M]) \in R^t$$



tBERT with word topics

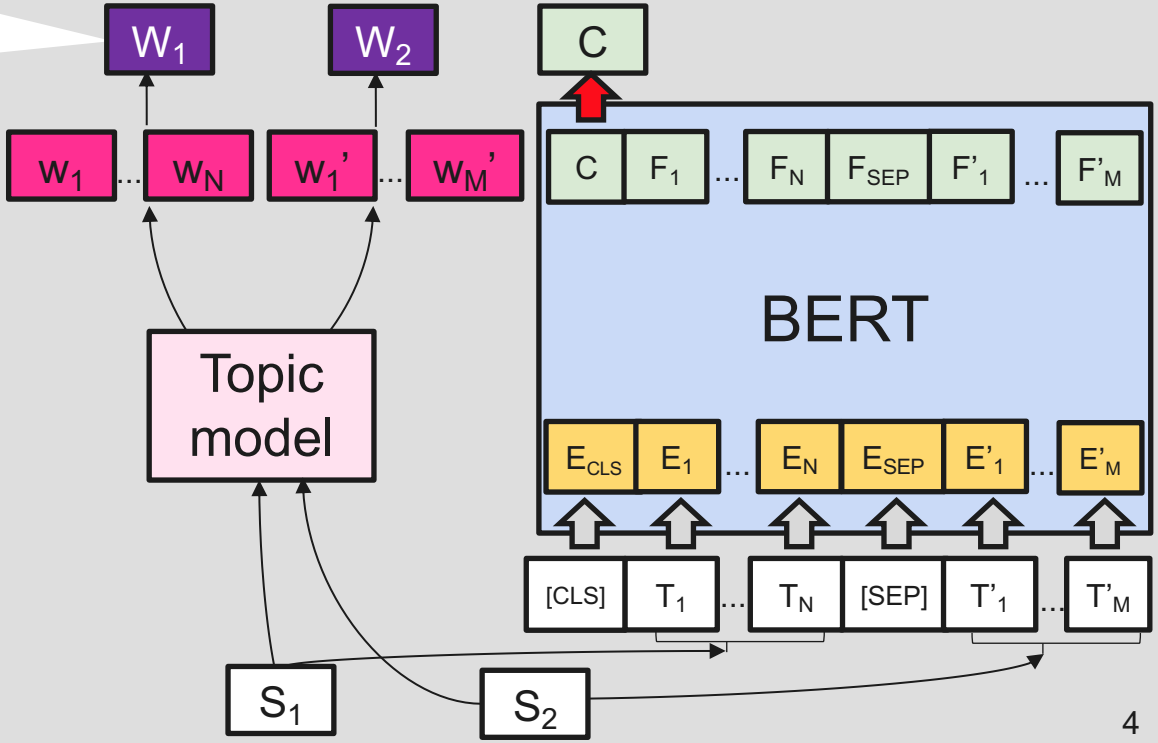
$$w_i = \text{TopicModel}(T_i) \in R^t$$



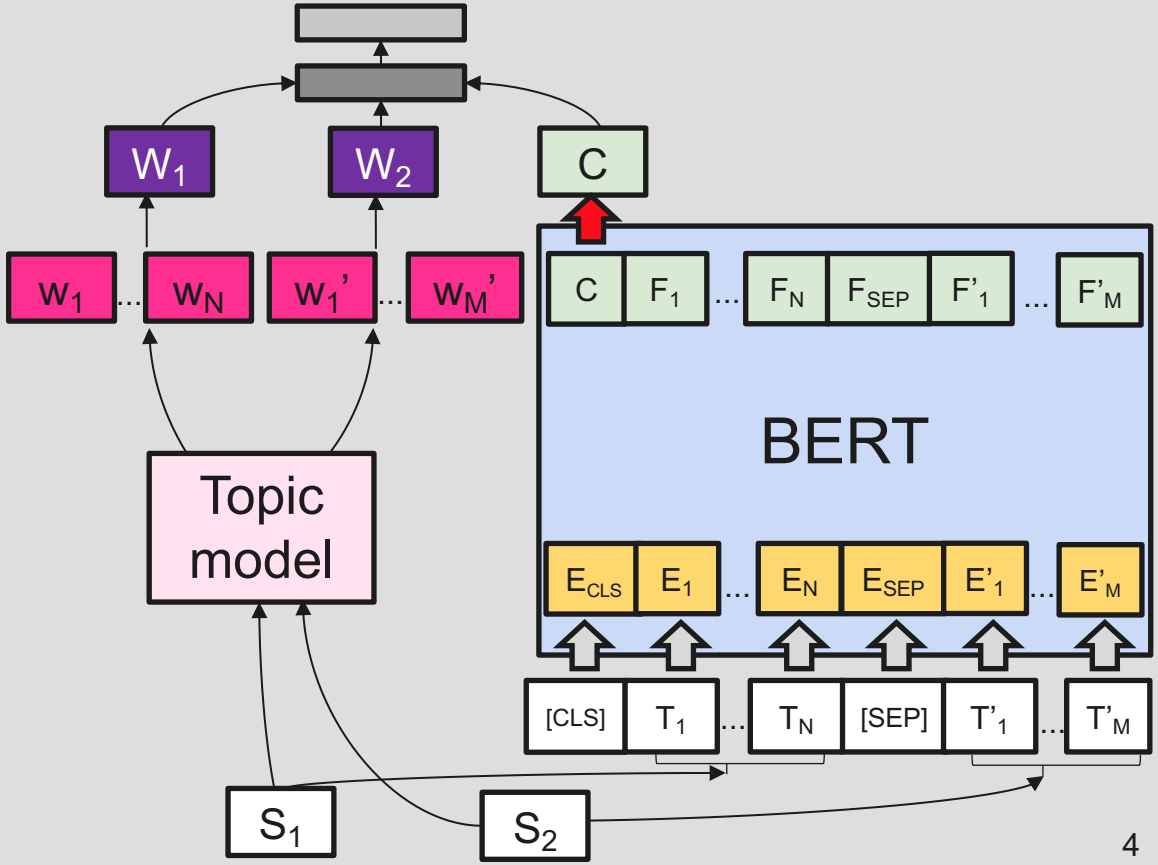
tBERT with word topics

$$W_1 = \frac{\sum_{i=1}^{N} w_i}{N} \in R^t$$

$$W_2 = \frac{\sum_{i=1}^{M} w'_i}{M} \in R^t$$



tBERT with word topics



Evaluation

			F1		
	MSRP	Quora	SemEval		
			A	B	C

Evaluation

	F1				
	MSRP	Quora	SemEval		
			A	B	C
Previous systems					
Filice et al. (2017) - feature-based	-	-	-	.506	-
Wu et al. (2017) - feature-based	-	-	**.777**	-	-
Koreeda et al. (2017) - feature-based	-	-	-	-	.197
Deriu and Cieliebak (2017) - neural	-	-	.433	-	-
Pang et al. (2016) - neural	.829	-	-	-	-
Gong et al. (2018) (accuracy) - neural	-	(.891)	-	-	-

Evaluation

	F1				
	MSRP	Quora	SemEval		
			A	B	C
Previous systems					
Filice et al. (2017) - feature-based	-	-	-	.506	-
Wu et al. (2017) - feature-based	-	-	**.777**	-	-
Koreeda et al. (2017) - feature-based	-	-	-	-	.197
Deriu and Cieliebak (2017) - neural	-	-	.433	-	-
Pang et al. (2016) - neural	.829	-	-	-	-
Gong et al. (2018) (accuracy) - neural	-	(.891)	-	-	-
Our implementation					
LDA topic baseline	.799	.736	.684	.436	.096
GSDMM topic baseline	.796	.679	.663	.403	.102

Evaluation

	F1				
	MSRP	Quora	SemEval		
			A	B	C
Previous systems					
Filice et al. (2017) - feature-based	-	-	-	.506	-
Wu et al. (2017) - feature-based	-	-	**.777**	-	-
Koreeda et al. (2017) - feature-based	-	-	-	-	.197
Deriu and Cieliebak (2017) - neural	-	-	.433	-	-
Pang et al. (2016) - neural	.829	-	-	-	-
Gong et al. (2018) (accuracy) - neural	-	(.891)	-	-	-
Our implementation					
LDA topic baseline	.799	.736	.684	.436	.096
GSDMM topic baseline	.796	.679	.663	.403	.102
BERT	.876	.902	.704	.473	.268
tBERT with LDA topics	**.884**	**.905**	.768	**.524**	**.273**
tBERT with GSDMM topics	.883	**.905**	.766	.518	.233

Evaluation

	F1				
	MSRP	Quora	SemEval		
			A	B	C
Previous systems					
Filice et al. (2017) - feature-based	-	-	-	.506	-
Wu et al. (2017) - feature-based	-	-	**.777**	-	-
Koreeda et al. (2017) - feature-based	-	-	-	-	.197
Deriu and Cieliebak (2017) - neural	-	-	.433	-	-
Pang et al. (2016) - neural	.829	-	-	-	-
Gong et al. (2018) (accuracy) - neural	-	(.891)	-	-	-
Our implementation					
LDA topic baseline	.799	.736	.684	.436	.096
GSDMM topic baseline	.796	.679	.663	.403	.102
BERT	.876	.902	.704	.473	.268
tBERT with LDA topics	**.884**	**.905**	.768	**.524**	**.273**
tBERT with GSDMM topics	.883	**.905**	.766	.518	.233

Evaluation

	F1				
	MSRP	Quora	SemEval		
			A	B	C
Previous systems					
Filice et al. (2017) - feature-based	-	-	-	.506	-
Wu et al. (2017) - feature-based	-	-	**.777**	-	-
Koreeda et al. (2017) - feature-based	-	-	-	-	.197
Deriu and Cieliebak (2017) - neural	-	-	.433	-	-
Pang et al. (2016) - neural	.829	-	-	-	-
Gong et al. (2018) (accuracy) - neural	-	(.891)	-	-	-
Our implementation					
LDA topic baseline	.799	.736	.684	.436	.096
GSDMM topic baseline	.796	.679	.663	.403	.102
BERT	.876	.902	.704	.473	.268
tBERT with LDA topics	**.884**	**.905**	.768	**.524**	**.273**
tBERT with GSDMM topics	.883	**.905**	.766	.518	.233

Evaluation

	MSRP	Quora	SemEval		
			A	B	C
F1 on cases with **named entities** (total: 230/500)					
BERT	.20	**.54**	.50	**.53**	.32
tBERT	**.35**	.49	**.52**	.21	**.56**
(# of cases)	(23)	(31)	(58)	(60)	(58)
F1 on cases with **domain-specific words** (total: 159/500)					
BERT	.18	.00	.36	.36	.26
tBERT	**.67**	**.50**	**.62**	**.40**	**.58**
(# of cases)	(14)	(7)	(36)	(41)	(61)
F1 on cases with **non-standard spelling** (total: 53/500)					
BERT	.00	N/A	.20	**.71**	.43
tBERT	.00	N/A	**.80**	.00	**.62**
(# of cases)	(1)	(0)	(20)	(19)	(13)

Table 4: F1 for BERT and tBERT on annotated development set examples by manually annotated properties.

Evaluation

	MSRP	Quora	SemEval		
			A	B	C
F1 on cases with **named entities** (total: 230/500)					
BERT	.20	**.54**	.50	**.53**	.32
tBERT	**.35**	.49	**.52**	.21	**.56**
(# of cases)	(23)	(31)	(58)	(60)	(58)
F1 on cases with **domain-specific words** (total: 159/500)					
BERT	.18	.00	.36	.36	.26
tBERT	**.67**	**.50**	**.62**	**.40**	**.58**
(# of cases)	(14)	(7)	(36)	(41)	(61)
F1 on cases with **non-standard spelling** (total: 53/500)					
BERT	.00	N/A	.20	**.71**	.43
tBERT	.00	N/A	**.80**	.00	**.62**
(# of cases)	(1)	(0)	(20)	(19)	(13)

Table 4: F1 for BERT and tBERT on annotated development set examples by manually annotated properties.

Evaluation

	MSRP	Quora	SemEval A	SemEval B	SemEval C
F1 on cases with **named entities** (total: 230/500)					
BERT	.20	**.54**	.50	**.53**	.32
tBERT	**.35**	.49	**.52**	.21	**.56**
(# of cases)	(23)	(31)	(58)	(60)	(58)
F1 on cases with **domain-specific words** (total: 159/500)					
BERT	.18	.00	.36	.36	.26
tBERT	**.67**	**.50**	**.62**	**.40**	**.58**
(# of cases)	(14)	(7)	(36)	(41)	(61)
F1 on cases with **non-standard spelling** (total: 53/500)					
BERT	.00	N/A	.20	**.71**	.43
tBERT	.00	N/A	**.80**	.00	**.62**
(# of cases)	(1)	(0)	(20)	(19)	(13)

Table 4: F1 for BERT and tBERT on annotated development set examples by manually annotated properties.

s1	Are there good beaches in the Northern part of Qatar?
s2	Fuwairit is very clean!
gold label	is_relevant

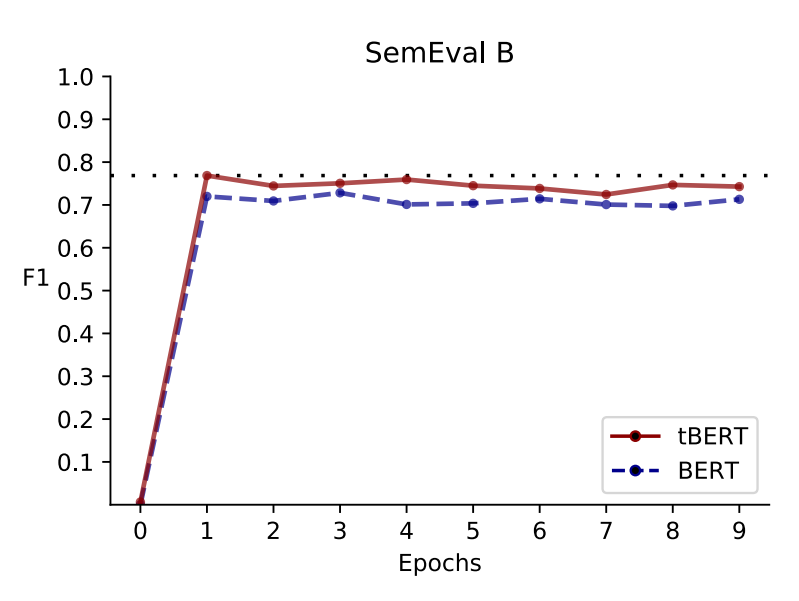
Evaluation

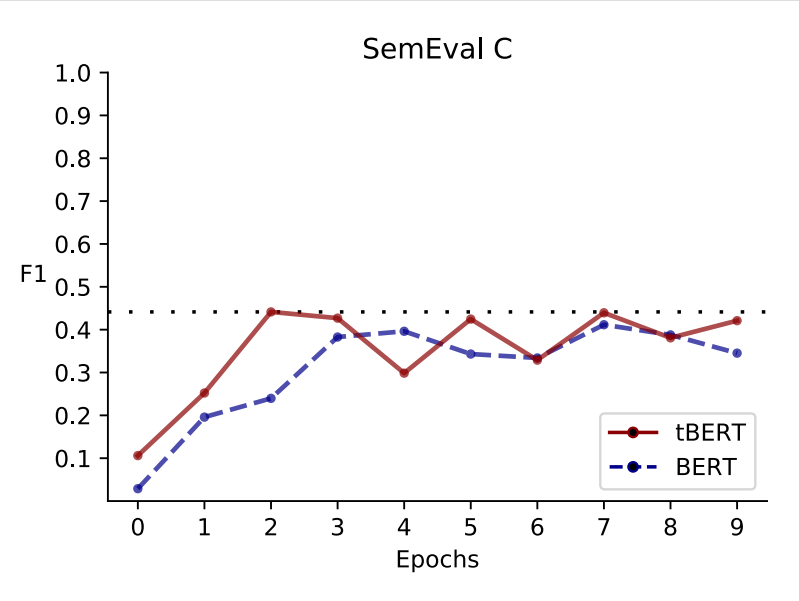
	MSRP	Quora	SemEval A	SemEval B	SemEval C
F1 on cases with **named entities** (total: 230/500)					
BERT	.20	**.54**	.50	**.53**	.32
tBERT	**.35**	.49	**.52**	.21	**.56**
(# of cases)	(23)	(31)	(58)	(60)	(58)
F1 on cases with **domain-specific words** (total: 159/500)					
BERT	.18	.00	.36	.36	.26
tBERT	**.67**	**.50**	**.62**	**.40**	**.58**
(# of cases)	(14)	(7)	(36)	(41)	(61)
F1 on cases with **non-standard spelling** (total: 53/500)					
BERT	.00	N/A	.20	**.71**	.43
tBERT	.00	N/A	**.80**	.00	**.62**
(# of cases)	(1)	(0)	(20)	(19)	(13)

Table 4: F1 for BERT and tBERT on annotated development set examples by manually annotated properties.

s1	Are there good beaches in the Northern part of Qatar?
s2	Fuwairit is very clean!
gold label	is_relevant

Could we just train BERT longer?





Conclusion

1. Flexible framework for combining topic models with BERT
2. Adding LDA topics to BERT improves performance
3. Improvements mainly on domain-specific examples

Thank you

@wuningxi
n.peinelt@warwick.ac.uk
github.com/wuningxi/tBERT