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

Nicole Peinelt[1,2] and **Dong Nguyen**[1,3] and **Maria Liakata**[1,2]
[1]The Alan Turing Institute, London, UK
[2]University of Warwick, Coventry, UK
[3]Utrecht University, Utrecht, The Netherlands
{n.peinelt, m.liakata}@warwick.ac.uk, dnguyen@turing.ac.uk

Abstract

Semantic similarity detection is a fundamental task in natural language understanding. Adding topic information has been useful for previous feature-engineered semantic similarity models as well as neural models for other tasks. There is currently no standard way of combining topics with pretrained contextual representations such as BERT. We propose a novel topic-informed BERT-based architecture for pairwise semantic similarity detection and show that our model improves performance over strong neural baselines across a variety of English language datasets. We find that the addition of topics to BERT helps particularly with resolving domain-specific cases.

1 Introduction

Modelling the semantic similarity between a pair of texts is a crucial NLP task with applications ranging from question answering to plagiarism detection. A variety of models have been proposed for this problem, including traditional feature-engineered techniques (Filice et al., 2017), hybrid approaches (Wu et al., 2017; Feng et al., 2017; Koreeda et al., 2017) and purely neural architectures (Wang et al., 2017; Tan et al., 2018; Deriu and Cieliebak, 2017). Recent pretrained contextualised representations such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) have led to impressive performance gains across a variety of NLP tasks, including semantic similarity detection. These models leverage large amounts of data to pretrain text encoders (in contrast to just individual word embeddings as in previous work) and have established a new pretrain-finetune paradigm.

While large improvements have been achieved on paraphrase detection (Tomar et al., 2017; Gong et al., 2018), semantic similarity detection in Community Question Answering (CQA) remains a challenging problem. CQA leverages user-generated content from question answering websites (e.g. StackExchange) to answer complex real-world questions (Nakov et al., 2017). The task requires modelling the relatedness between question-answer pairs which can be challenging due to the highly domain-specific language of certain online forums and low levels of direct text overlap between questions and answers.

Topic models may provide additional signals for semantic similarity, as earlier feature-engineered models for semantic similarity detection successfully incorporated topics (Qin et al., 2009; Tran et al., 2015; Mihaylov and Nakov, 2016; Wu et al., 2017). They could be especially useful for dealing with domain-specific language since topic models have been exploited for domain adaptation (Hu et al., 2014; Guo et al., 2009). Moreover, recent work on neural architectures has shown that the integration of topics can yield improvements in other tasks such as language modelling (Ghosh et al., 2016), machine translation (Chen et al., 2016), and summarisation (Narayan et al., 2018; Wang et al., 2018). We therefore introduce a novel architecture for semantic similarity detection which incorporates topic models and BERT. More specifically, we make the following contributions:

1. We propose tBERT — a simple architecture combining topics with BERT for semantic similarity prediction (section 3).[1]

2. We demonstrate that tBERT achieves improvements across multiple semantic similarity prediction datasets against a finetuned vanilla BERT and other neural models in both F1 and stricter evaluation metrics (section 5).

3. We show in our error analysis that tBERT's gains are prominent on domain-specific cases, such as those encountered in CQA (section 5).

[1]Code is available at https://github.com/wuningxi/tBERT.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7047–7055
July 5 - 10, 2020.

2 Datasets and Tasks

We select popular benchmark datasets featuring different sizes (small vs. large), tasks (QA vs. paraphrase detection) and sentence lengths (short vs. long) as summarised in Table 1. Examples for each dataset are provided in Appendix A.

MSRP The Microsoft Research Paraphrase dataset (MSRP) contains pairs of sentences from news websites with binary labels for paraphrase detection (Dolan and Brockett, 2005).

SemEval The SemEval CQA dataset (Nakov et al., 2015, 2016, 2017) comprises three subtasks based on threads and posts from the online expat forum *Qatar Living*.[2] Each subtask contains an initial post as well as 10 possibly relevant posts with binary labels and requires to rank relevant posts above non-relevant ones. In subtask A, the posts are questions and comments from the same thread, in an answer ranking scenario. Subtask B is question paraphrase ranking. Subtask C is similar to A but comments were retrieved from an external thread, which increases the difficulty of the task.

Quora The Quora duplicate questions dataset contains more than 400k question pairs with binary labels and is by far the largest of the datasets.[3] The task is to predict whether two questions are paraphrases. The setup is similar to SemEval subtask B, but framed as a classification rather than a ranking problem. We use Wang et al. (2017)'s train/dev/test set partition.

All of the above datasets provide two short texts (usually a sentence long but in some cases consisting of multiple sentences). From here onward we will use the term 'sentence' to refer to each short text. We frame the task as predicting the semantic similarity between two sentences in a binary classification task. We use a binary classification setup as this is more generic and applies to all above datasets.

Dataset	Task	Len	Size
Quora	paraphrase detection	13	404K
MSRP	paraphrase detection	22	5K
SemEval	(A) internal answer ranking	48	26K
	(B) paraphrase ranking	52	4K
	(C) external answer ranking	45	47K

Table 1: Text pair similarity data sets. Size = number of text pairs. Len = mean sentence length in tokens.

[2]Following convention, we use the 2016 test set as development set and 2017 test set as test set.

[3]https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning

3 tBERT

3.1 Architecture

In this paper, we investigate if topic models can further improve BERT's performance for semantic similarity detection. Our proposed **topic-informed BERT-based model (tBERT)** is shown in Figure 1. We encode two sentences S_1 (with length N) and S_2 (with length M) with the uncased version of $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019), using the C vector from BERT's final layer corresponding to the *CLS* token in the input as sentence pair representation:

$$C = \text{BERT}(S_1, S_2) \in R^d \quad (1)$$

where d denotes the internal hidden size of BERT (768 for $\text{BERT}_{\text{BASE}}$). While other topic models can be used, we experiment with two popular topic models: LDA (Blei et al., 2003) and GSDMM (Yin and Wang, 2014), see section 3.2 for details. Based on previous research which successfully combined word and document level topics with neural architectures (Narayan et al., 2018), we further experiment with incorporating different topic types. For document topics D_1 and D_2, all tokens in a sentence are passed to the topic model to infer one topic distribution per sentence:

$$D_1 = \text{TopicModel}([T_1, ..., T_N]) \in R^t \quad (2)$$

$$D_2 = \text{TopicModel}([T'_1, ..., T'_M]) \in R^t \quad (3)$$

where t indicates the number of topics. Alternatively, for word topics W_1 and W_2, one topic distri-

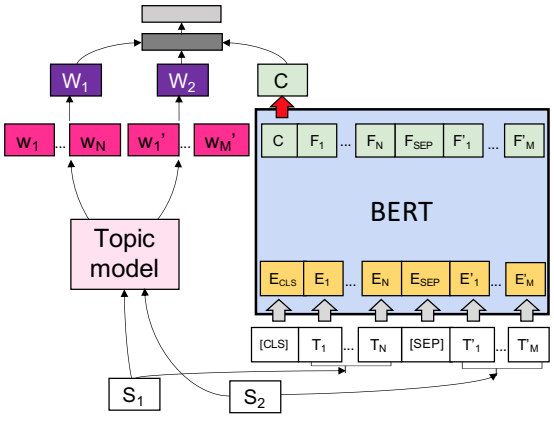

Figure 1: Architecture of tBERT with word topics.

bution w_i is inferred per token T_i

$$w_i = \text{TopicModel}(T_i) \in R^t \quad (4)$$

before averaging them to obtain a fixed-length topic representation on the sentence level:

$$W_1 = \frac{\sum_{i=1}^{N} w_i}{N} \in R^t \quad (5)$$

$$W_2 = \frac{\sum_{i=1}^{M} w'_i}{M} \in R^t \quad (6)$$

We combine the sentence pair vector with the sentence-level topic representations similar to Ostendorff et al. (2019) as

$$\text{F} = [C; D_1; D_2] \in R^{d+2t} \quad (7)$$

for document topics and as

$$\text{F} = [C; W_1; W_2] \in R^{d+2t} \quad (8)$$

for word topics (where ; denotes concatenation). This is followed by a hidden and a softmax classification layer. We train the model for 3 epochs with early stopping and cross-entropy loss. Learning rates are tuned per dataset and random seed.[4]

3.2 Choice of Topic Model

Topic number and alpha value The number of topics and alpha values are important topic model hyper-parameters and dataset dependent. We use the simple topic baseline (section 4) as a fast proxy (on average 12 seconds per experiment on CPU) to identify useful topic models for each dataset without expensive hyper-parameter tuning on the full tBERT model. In our experiments, 70 to 90 topics with alpha values of 1 or 10 worked well.[5]

	MSRP	Quora	SemEval		
			A	B	C
BERT	.906	.906	.714	.754	.414
tBERT with LDA					
+ word topics	.905	**.911**	.744	**.766**	**.439**
+ doc topics	**.907**	.909	**.748**	.761	.419
tBERT with GSDMM					
+ word topics	**.918**	.908	**.752**	**.760**	**.447**
+ doc topics	.915	**.909**	.751	.760	.424

Table 2: F1 scores of BERT-based models with different topic settings on development set. We report average performance for two different random seeds. Bold indicates the selected setting for our final model.

[4] We report tuned hyper-parameters in Appendix E.

[5] See Appendix D for detailed topic model settings.

Topic model and topic type LDA (Blei et al., 2003) is the most popular and widely used topic model, but it has been reported to be less suitable for short text (Hong and Davison, 2010). Therefore, we also experiment with the popular short text topic model GSDMM (Yin and Wang, 2014). To select the best setting for our final model (in Table 3), we evaluated different combinations of tBERT with LDA vs. GSDMM and word (W_1 and W_2) vs. document topics (D_1 and D_2) on the development partition of the datasets (Table 2). The tBERT settings generally scored higher than BERT, with word topics (W_1 and W_2) usually outperforming document topics.

4 Baselines

Topic baselines As a simple baseline, we train a topic model (LDA or GSDMM) on the training portion of each dataset (combining training sets for SemEval subtasks) and calculate the Jensen-Shannon divergence (Lin, 1991) (JSD) between the topic distributions of the two sentences. The model predicts a negative label if JSD is larger than a threshold and a positive label otherwise. We tune threshold, number of topics and alpha value based on development set F1.[5]

Previous systems For SemEval, we compare against the highest performing system of earlier work based on F1 score. As these models rely on hand-crafted dataset-specific features (providing an advantage on the small datasets), we also include the only neural system without manual features (Deriu and Cieliebak, 2017). For MSRP, we show a neural matching architecture (Pang et al., 2016). For Quora, we compare against the Interactive Inference Network (Gong et al., 2018) using accuracy, as no F1 has been reported.

Siamese BiLSTM Siamese networks are a common neural baseline for sentence pair classification tasks (Yih et al., 2011; Wang et al., 2017). We embed both sentences with pretrained GloVe embeddings (concatenated with ELMo for BiLSTM + ELMo) and encode them with two weight-sharing BiLSTMs, followed by max pooling and hidden layers.

BERT We encode the sentence pair with BERT's C vector (as in tBERT) followed by a softmax layer and finetune all layers for 3 epochs with early stopping. Following Devlin et al. (2019), we tune learning rates on the development set of each dataset.[4]

5 Results

Evaluation We evaluate systems based on F1 scores (Table 3) as this is more reliable for datasets with imbalanced labels (e.g. SemEval C) than accuracy. We further report performance on difficult cases with non-obvious F1 score (Peinelt et al., 2019) which identifies challenging instances in the dataset based on lexical overlap and gold labels. Dodge et al. (2020) recently showed that early stopping and random seeds can have considerable impact on the performance of finetuned BERT models. We therefore use early stopping during finetuning and report average model performance across two seeds for BERT and tBERT models.

Overall trends The BERT-based models outperform the other neural systems, while closely competing with the feature-engineered system on the relatively small SemEval A dataset. The simple topic baselines perform surprisingly well in comparison to much more sophisticated models, indicating the usefulness of topics for the tasks.

Do topics improve BERT's performance? Adding LDA topics to BERT consistently improves F1 performance across all datasets. Moreover, it improves performance on non-obvious cases over BERT on all datasets (except for Quora which contains many generic examples and few domain-specific cases, see Table 4). The addition of GSDMM topics to BERT is slightly less stable: improving performance on MSRP, Semeval A and B, while dropping on Semeval C. The largest performance gains regardless of the chosen topic model are observed in the internal question-answering task (SemEval A).

	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 (100 cases per dataset) by manually annotated properties. Number of cases in parenthesis.

Where can topics help? We randomly sampled 100 examples (half only correct by BERT, half only correct by LDA-tBERT) from the development set of each dataset and manually annotated them (500 in total) with binary labels regarding three properties that may be associated with topic-related gains or losses (Table 4). Named entities (e.g. *iPhone*) and domain-specific words (e.g. *murabaha*) occurred frequently in the datasets, while there were too few examples with non-standard spelling (e.g. *thanx*) for meaningful comparisons. tBERT generally performed better than BERT on examples with domain-specific cases. Overall patterns were

	F1					non-obvious F1				
	MSRP	Quora	SemEval A	SemEval B	SemEval C	MSRP	Quora	SemEval A	SemEval B	SemEval C
Previous systems										
Filice et al. (2017) - feature-based	-	-	-	.506	-	-	-	-	.199	-
Wu et al. (2017) - feature-based	-	-	**.777**	-	-	-	-	.707	-	-
Koreeda et al. (2017) - feature-based	-	-	-	-	.197	-	-	-	-	.028
Deriu and Cieliebak (2017) - neural	-	-	.433	-	-	-	-	.352	-	-
Pang et al. (2016) - neural	.829	-	-	-	-	-	-	-	-	-
Gong et al. (2018) (accuracy) - neural	-	(.891)	-	-	-	-	-	-	-	-
Our implementation										
LDA topic baseline	.799	.736	.684	.436	.096	.780	.606	.684	.172	.019
GSDMM topic baseline	.796	.679	.663	.403	.102	.769	.448	.488	.130	.015
Siamese BiLSTM	.763	.813	.671	.349	.126	.781	.740	.597	.168	.049
Siamese BiLSTM + ELMo	.765	.832	.661	.345	.149	.775	.754	.599	.180	.073
BERT	.876	.902	.704	.473	.268	.827	**.860**	.656	.243	.085
tBERT with LDA topics	**.884**	**.905**	**.768**	**.524**	**.273**	**.866**	.859	.708	.258	**.100**
tBERT with GSDMM topics	.883	**.905**	.766	.518	.233	.844	.856	**.714**	**.266**	.081

Table 3: Model performance on test set. The first 6 rows are taken from the cited papers. Bold font highlights the best system overall and our best implementation is underlined. Italics indicate that F1 and accuracy were identical.

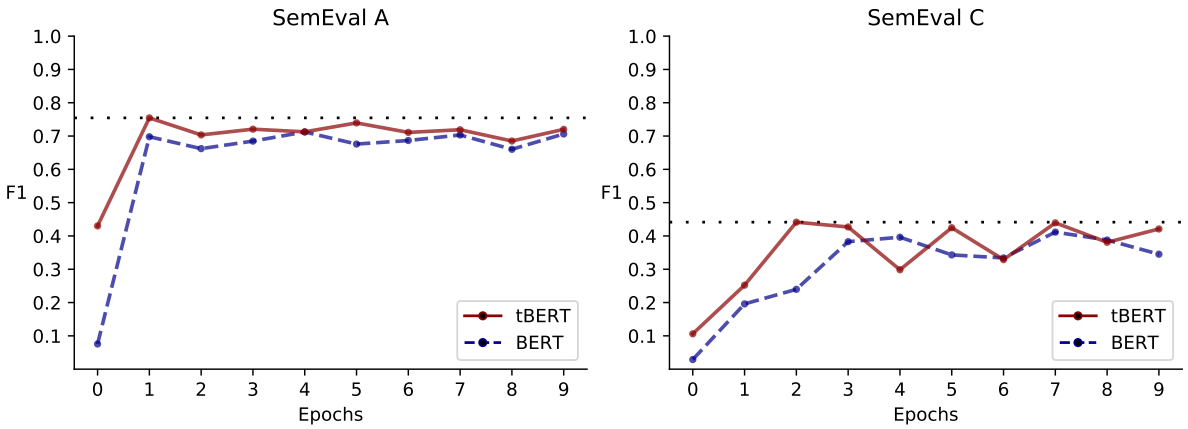

Figure 2: Performance of BERT and tBERT on dev set when trained for up to 9 epochs. The dotted line indicates tBERT's best performance within the first 3 epochs. Plots for the other datasets are provided as Appendix G.

less clear for named entities; based on manual inspection BERT dealt better with common named entities likely to have occurred in pretraining (such as well-known brands), while tBERT improved on dataset-specific named entities. We reason that for domain-specific words which are unlikely to have occurred in pretraining (e.g. *Fuwairit* in Table 5), BERT may not have learned a good representation (even after finetuning) and hence can't make a correct prediction. Here, topic models could serve as an additional source for dataset-specific information. The usefulness of topics for such cases is also supported by previous work, which successfully leveraged topics for domain adaptation in machine translation (Hu et al., 2014) and named entity recognition (Guo et al., 2009).

Could we just finetune BERT longer? Based on our observation that tBERT performs better on dataset-specific cases, one could assume that BERT may simply need to be finetuned longer than the usual 3 epochs to pick up more domain-specific information. In an additional experiment, we finetuned BERT and tBERT (with LDA topics) for 9 epochs (see Figure 2 and Appendix G). On most datasets, BERT reached peak performance within the first 3 epochs. Although training for 4 or 7 epochs achieved marginal gains on Semeval A and C, longer finetuning of BERT could not exceed tBERT's best performance from the first 3 epochs (dotted line) on any dataset. We conclude that longer finetuning does not considerably boost BERT's performance. Adding topics instead is more effective, while avoiding the burden of greatly increased training time (compare Appendix F).

s1	Are there good beaches in the Northern part of Qatar?
s2	Fuwairit is very clean !
gold label	True
predictions	BERT:False, BERT+topics:True
manual annotation	domain-specific word:True, named entity:True, non-standard spelling:False

Table 5: Predictions and annotation for an example from SemEval.

6 Conclusion

In this work, we proposed a flexible framework for combining topic models with BERT. We demonstrated that adding LDA topics to BERT consistently improved performance across a range of semantic similarity prediction datasets. In our qualitative analysis, we showed that these improvements were mainly achieved on examples involving domain-specific words. Future work may focus on how to directly induce topic information into BERT without corrupting pretrained information and whether combining topics with other pretrained contextual models can lead to similar gains. Another research direction is to investigate if introducing more sophisticated topic models, such as named entity promoting topic models (Krasnashchok and Jouili, 2018) into the proposed framework can further improve results.

Acknowledgments

This work was supported by Microsoft Azure and The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning research*, 3:993–1022.

Wenhu Chen, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter. 2016. Guided Alignment Training for Topic-Aware Neural Machine Translation. In *Proceedings of AMTA*, pages 121–134, Austin, USA.

Jan Milan Deriu and Mark Cieliebak. 2017. SwissAlps at SemEval-2017 task 3: Attention-based Convolutional Neural Network for Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation.*, volume 17, pages 334–338, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 4171—4186, Minneapolis, USA. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. *arXiv:2002.06305 [cs]*.

William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP@IJCNLP)*, pages 9–16, Jeju Island, Korea. Asian Federation of Natural Language Processing.

Wenzheng Feng, Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. 2017. Beihang-MSRA at SemEval-2017 Task 3- A Ranking System with Neural Matching Features for Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017*, pages 280–286, Vancouver, Canada. Association for Computational Linguistics.

Simone Filice, Giovanni Da San Martino, and Alessandro Moschitti. 2017. KeLP at SemEval-2017 Task 3- Learning Pairwise Patterns in Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017*, pages 326–333, Vancouver, Canada. Association for Computational Linguistics.

Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. 2016. Contextual LSTM (CLSTM) Models for Large Scale NLP Tasks. *arXiv:1602.06291 [cs]*.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural Language Inference over Interaction Space. In *6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada.

Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang, Xian Wu, and Zhong Su. 2009. Domain Adaptation with Latent Semantic Association for Named Entity Recognition. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09*, pages 281–289, Boulder, USA. Association for Computational Linguistics.

Liangjie Hong and Brian D. Davison. 2010. Empirical Study of Topic Modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, pages 80–88, Washington D.C., USA. ACM Press.

Yuening Hu, Ke Zhai, Vladimir Eidelman, and Jordan Boyd-Graber. 2014. Polylingual Tree-Based Topic Models for Translation Domain Adaptation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1166–1176, Baltimore, Maryland. Association for Computational Linguistics.

Yuta Koreeda, Takuya Hashito, Yoshiki Niwa, Misa Sato, Toshihiko Yanase, Kenzo Kurotsuchi, and Kohsuke Yanai. 2017. Bunji at SemEval-2017 Task 3: Combination of Neural Similarity Features and Comment Plausibility Features. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval@ACL 2017)*, pages 353–359, Vancouver, Canada. Association for Computational Linguistics.

Katsiaryna Krasnashchok and Salim Jouili. 2018. Improving Topic Quality by Promoting Named Entities in Topic Modeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 247–253, Melbourne, Australia. Association for Computational Linguistics.

Jianhua Lin. 1991. Divergence Measures based on the Shannon Entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Todor Mihaylov and Preslav Nakov. 2016. SemanticZ at SemEval-2016 Task 3: Ranking Relevant Answers in Community Question Answering Using Semantic Similarity Based on Fine-tuned Word Embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT 2016)*, pages 879–886, San Diego, USA. Association for Computational Linguistics.

Preslav Nakov, Doris Hoogeveen, Llúis Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 Task 3: Community Question Answering. In *Proceedings of the 11th International Workshop on*

Semantic Evaluation (SemEval@ACL 2017), pages 27–48, Vancouver, Canada. Association for Computational Linguistics.

Preslav Nakov, Lluis Marquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, James Glass, and Bilal Randeree. 2016. SemEval-2016 Task 3: Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT 2016)*, pages 525–545, San Diego, USA. Association for Computational Linguistics.

Preslav Nakov, Lluis Marquez, Magdy Walid, Alessandro Moschitti, James Glass, and Bilal Randeree. 2015. SemEval-2015 task 3: Answer Selection in Community Question Answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT 2015)*, pages 269–281, Denver, USA. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Malte Ostendorff, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, Georg Rehm, and Bela Gipp. 2019. Enriching BERT with Knowledge Graph Embeddings for Document Classification. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*, Erlangen, Germany.

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text Matching as Image Recognition. In *Proceedings of the Thirtieth Conference on Artificial Intelligence (AAAI)*, pages 2793–2799, Phoenix, USA. AAAI Press.

Nicole Peinelt, Maria Liakata, and Dong Nguyen. 2019. Aiming beyond the Obvious: Identifying Non-Obvious Cases in Semantic Similarity Datasets. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2792–2798, Florence, Italy. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2227–2237, New Orleans, USA. Association for Computational Linguistics.

Zengchang Qin, Marcus Thint, and Zhiheng Huang. 2009. Ranking Answers by Hierarchical Topic Models. In *Next-Generation Applied Intelligence, 22nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2009)*, volume 5579, pages 103–112, Tainan, Taiwan. Springer.

Chuanqi Tan, Furu Wei, Wenhui Wang, Weifeng Lv, and Ming Zhou. 2018. Multiway Attention Networks for Modeling Sentence Pairs. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4411–4417, Stockholm, Sweden.

Gaurav Singh Tomar, Thyago Duque, Oscar Täckström, Jakob Uszkoreit, and Dipanjan Das. 2017. Neural Paraphrase Identification of Questions with Noisy Pretraining. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 142–147, Copenhagen, Denmark. Association for Computational Linguistics.

Quan Hung Tran, Vu Tran, Tu Vu, Minh Nguyen, and Son Bao Pham. 2015. JAIST: Combining Multiple Features for Answer Selection in Community Question Answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 215–219, Denver, USA. Association for Computational Linguistics.

Li Wang, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. 2018. A Reinforced Topic-Aware Convolutional Sequence-to-Sequence Model for Abstractive Text Summarization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4453–4460, Stockholm, Sweden.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral Multi-Perspective Matching for Natural Language Sentences. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4144–4150, Melbourne, Australia.

Guoshun Wu, Yixuan Sheng, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 Task 3- Using Traditional and Deep Learning Methods to Address Community Question Answering Task. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval@ACL 2017)*, pages 356–360, Vancouver, Canada. Association for Computational Linguistics.

Wen-tau Yih, Kristina Toutanova, John C Platt, and Christopher Meek. 2011. Learning Discriminative Projections for Text Similarity Measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*, pages 247–256, Portland, USA. Association for Computational Linguistics.

Jianhua Yin and Jianyong Wang. 2014. A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '14*, pages 233–242, New York, USA. ACM Press.

Appendix

A Dataset Examples

Dataset	Sentence pair	L
MSRP	There are only 2,000 Roman Catholics living in Banja Luka now. There are just a handful of Catholics left in Banja Luka.	1
Quora	Which is the best way to learn coding? How do you learn to program?	1
SemEval A	Anybody recommend a good dentist in Doha? Dr Sarah Dental Clinic	1
SemEval B	Where I can buy good oil for massage? Blackheads - Any suggestions on how to get rid of them??	0
SemEval C	Can anybody tell me where is Doha clinic? Dr. Rizwi - Al Ahli Hospital	0

Table 6: Examples from different datasets. Labels (L) indicate if the second sentence is a paraphrase (for paraphrasing tasks) or relevant (for QA tasks).

B LDA Topic Examples

T1:	life purpose important thing real biggest
T2:	drink water coffee tea drinking good
T3:	pokémon flight car ticket train fly
T4:	school university college high students student
T5:	chemical determine formula acid determined san

Table 7: Top key words for example topics learned by an LDA model with 90 topics on the Quora training set.

T1:	regiment cavalry north 3rd passenger fort
T2:	court judge federal district supreme file
T3:	windows server software microsoft 2003 system
T4:	president bush time presidential report george
T5:	hospital condition study risk cancer women

Table 8: Top key words for example topics learned by an LDA model with 80 topics on the MSRP training set.

T1:	gym club pool fitness gyms swimming
T2:	drink good club music night alcohol
T3:	husband sponsorship wife company sponsor work
T4:	day eid holidays days ramadan hours
T5:	time doha bus area morning early

Table 9: Top key words for example topics learned by an LDA model with 70 topics on the training set of all three SemEval tasks combined.

C GSDMM Topic Examples

T1:	difference examples law social science
T2:	effects earthquake major compare cambodia
T3:	arbitration court cards australia world
T4:	panel solar provider installation california
T5:	get best rid skin remove

Table 10: Top key words for example topics learned by a GSDMM model with 90 topics on the Quora training set.

T1:	cases said number year reported sales meeting
T2:	states united wrong sense deal
T3:	two killed united states people government
T4:	condition hospital center taken medical county
T5:	charges commission arrested exchange

Table 11: Top key words for example topics learned by a GSDMM model with 80 topics on the MSRP training set.

T1:	know qatar years many indian qatari
T2:	good qatar live doha know dog
T3:	arabic doha best people time english
T4:	month like 000 car compound villa
T5:	time find visa working company study

Table 12: Top key words for example topics learned by a GSDMM model with 70 topics on the training set of all three SemEval tasks combined.

D Hyper-Parameters for Topic-Aware Models

Topic model hyper-parameters were chosen based on development set F1 scores of the topic baseline. We tried number of topics: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 and alpha values: 0.1, 1, 10, 50. The topic baselines and tBERT models use topic models with the same hyper-parameters as listed in Table 13.

	MSRP	Quora	SemEval		
			A	B	C
# of topics	80	90	70	80	70
LDA alpha values	1	1	50	10	10
GSDMM alpha values	0.1	0.1	0.1	0.1	0.1

Table 13: Tuned topic model hyper-parameters.

E Hyper-Parameters for BERT-Based Models

Table 14 reports additional hyper-parameters for BERT and tBERT. The learning rate was tuned based on development set F1 score per seed and model using grid search (2e-5, 3e-5 or 5e-5).

	MSRP	Quora	SemEval		
			A	B	C
batch size	32	32	16	32	16
BERT					
lrate (1st seed)	5e-5	2e-5	3e-5	2e-5	2e-5
lrate (2nd seed)	5e-5	2e-5	2e-5	2e-5	3e-5
tBERT					
lrate (1st seed)	3e-5	3e-5	2e-5	2e-5	3e-5
lrate (2nd seed)	5e-5	2e-5	2e-5	3e-5	2e-5

Table 14: Tuned hyper-parameters for BERT-based models. lrate = learning rate.

F Training time

	MSRP	Quora	SemEval		
			A	B	C
BERT					
3 epochs	13	839	223	26	340
9 epochs	44	2710	638	75	1047
tBERT					
3 epochs	13	885	211	24	348
9 epochs	42	2916	658	75	1082

Table 15: Average training time on one NVIDIA Tesla K80 GPU in minutes.

G Longer Finetuning Experiment

Longer BERT finetuning does not surpass tBERT's best performance from the first 3 epochs (dotted line) while considerably increasing training time (compare Appendix F).

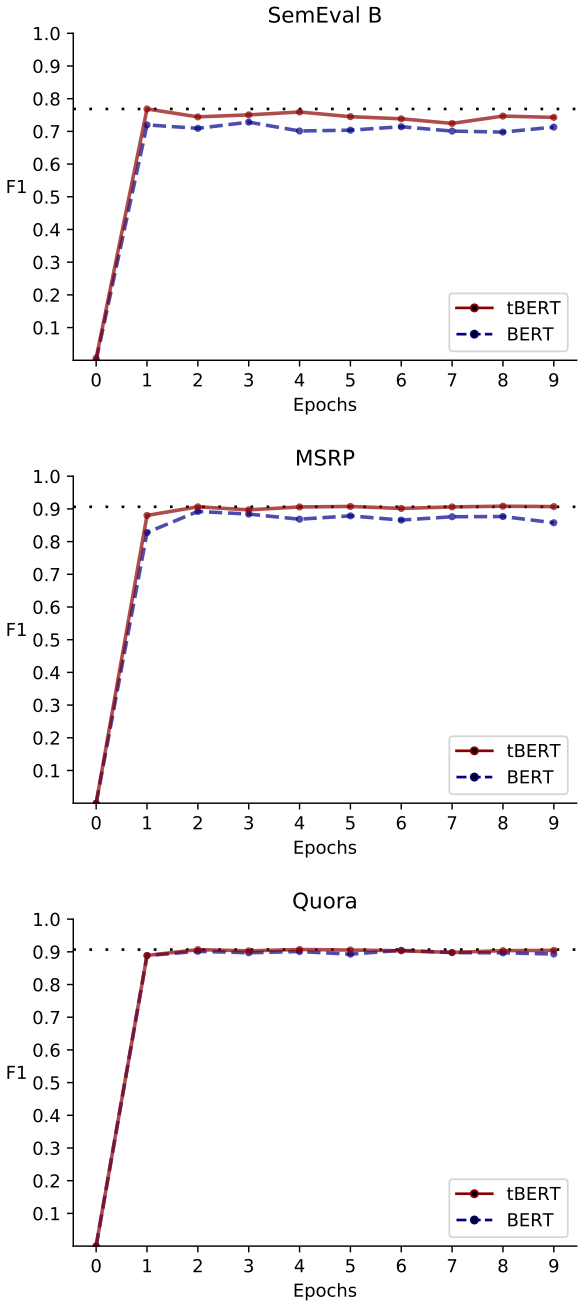

Figure 3: Performance of BERT and tBERT on development set when trained for up to 9 epochs. The dotted line indicates tBERT's best performance within the first 3 epochs.