White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/190604/

Version: Published Version

Proceedings Paper:
Rücklé, A., Moosavi, N.S. and Gurevych, I. (2019) Neural duplicate question detection without labeled training data. In: Inui, K., Jiang, J., Ng, V. and Wan, X., (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019), 03-07 Nov 2019, Hong Kong, China. Association for Computational Linguistics, pp. 1607-1617. ISBN: 9781950737901.

https://doi.org/10.18653/v1/D19-1171

© 2019 Association for Computational Linguistics. Available under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

Neural Duplicate Question Detection without Labeled Training Data

Andreas Rücklé and **Nafise Sadat Moosavi** and **Iryna Gurevych**
Ubiquitous Knowledge Processing Lab (UKP)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Abstract

Supervised training of neural models to duplicate question detection in community Question Answering (cQA) requires large amounts of labeled question pairs, which are costly to obtain. To minimize this cost, recent works thus often used alternative methods, e.g., adversarial domain adaptation. In this work, we propose two novel methods: (1) the automatic generation of duplicate questions, and (2) weak supervision using the title and body of a question. We show that both can achieve improved performances even though they do not require any labeled data. We provide comprehensive comparisons of popular training strategies, which provides important insights on how to 'best' train models in different scenarios. We show that our proposed approaches are more effective in many cases because they can utilize larger amounts of unlabeled data from cQA forums. Finally, we also show that our proposed approach for weak supervision with question title and body information is also an effective method to train cQA answer selection models without direct answer supervision.

1 Introduction

The automatic detection of question duplicates in community Question Answering (cQA) forums is an important task that can help users to more effectively find existing questions and answers (Nakov et al., 2017; Cao et al., 2012; Xue et al., 2008; Jeon et al., 2005), and to avoid posting similar questions multiple times. Neural approaches to duplicate detection typically require large quantities of labeled question pairs for supervised training—i.e., labeled pairs of duplicate questions that can be answered with the same information.[1]

In practice, it is often difficult to obtain such data because of the immense manual effort that is required for annotation. A large number of cQA forums thus do not contain enough labeled data for supervised training of neural models.[2] Therefore, recent works have used alternative training methods. This includes weak supervision with question-answer pairs (Qiu and Huang, 2015), semi-supervised training (Uva et al., 2018), and adversarial domain transfer (Shah et al., 2018). An important limitation of these methods is that they still rely on substantial amounts of labeled data—either thousands of duplicate questions (e.g., from a similar source domain in the case of domain transfer) or large numbers of question-answer pairs. Furthermore, unsupervised methods rely on encoder-decoder architectures that impose limitations on the model architectures and they often fall short of the performances that are achieved with supervised training (Lei et al., 2016), or they need to be combined with complex features to achieve state-of-the-art results (Zhang and Wu, 2018). To train effective duplicate question detection models for the large number of cQA forums without labeled duplicates we thus need other methods that do not require any annotations while performing on-par with supervised in-domain training.

In this work, we propose two novel methods for scenarios where we only have access to unlabeled questions (title-body), including (1) automatic duplicate question generation (*DQG*); and (2) weak supervision with the title-body pairs (*WS-TB*). Because a question body typically provides additional important information that is not included in the title (Wu et al., 2018), we hypothesize that titles and bodies have similar properties as duplicate ques-

[1] For example, "Nautilus shortcut for new blank files?" and "How do you create a new document keyboard shortcut?" are titles of labeled duplicate questions from AskUbuntu.com.

[2] Shah et al. (2018) argue that even larger StackExchange sites do not offer enough duplicates for supervised training. Further, there exist many platforms that do not contain any labeled duplicates (e.g., https://gutefrage.net).

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1607–1617,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

tions. For instance, they are only partially redundant but fundamentally describe the same question (see Figure 1 for an example). As a consequence, we can use the information from titles and bodies together with their relations to train our models.

In DQG, we use question generation models to generate a new question title from the body and then consider the generated title as a duplicate to the question's original title. In WS-TB, we take this one step further and directly train models on title-body pairs—i.e., learning to predict whether both texts belong to the same question. The advantage of our proposed methods, compared to previous work, is that they can make use of the large number of unlabeled questions (titles and bodies) in cQA forums, which is typically an order of magnitude more data than is available for supervised training.[3]

In our experiments, we evaluate common question retrieval and duplicate detection models such as RCNN (Lei et al., 2016) and BiLSTM and compare a wide range of training methods: DQG, WS-TB, supervised training, adversarial domain transfer, weak supervision with question-answer pairs, and unsupervised training. We perform extensive experiments on multiple datasets and compare the different training methods in different scenarios, which provides important insights on how to 'best' train models with varying amounts training data. We show that:

1. Training models with title-body information is very effective. With larger amounts of unlabeled questions, WS-TB and DQG outperform adversarial domain transfer from similar source domains by more than 5.8pp on average. Because the amounts of labeled question duplicates is often limited, WS-TB and DQG can in some cases achieve better performances than supervised training.

2. DQG transfers well across domains, i.e., question generation models can be applied to novel target domains to obtain generated duplicates that are suitable for model training.

3. Our training methods are effective when being used to fine-tune more recent models such as BERT (Devlin et al., 2018a).

[3] Question titles and bodies are common in all StackExchange sites, popular platforms in other languages (e.g., GuteFrage.net), and forums such as Reddit. A counterexample is Quora, which only contains question titles. However, there exists a large annotated corpus of question pairs for this forum.

TITLE

How to customize each Firefox window icon individually?

BODY (1st PARAGRAPH)

I'm a tab hoarder and I admit it. But at least I've sorted them into contextual windows now, and I'd love to have different icons for each window in the Windows task bar (not the tab bar, which is governed by the favicons). How can this be achieved?

ANSWER

This can be done using the free AutoHotkey. Create a .ahk text file and enter these contents: (. . .)

Figure 1: An example question, the first paragraph of its body, and the first answer (from SuperUser[4]).

4. WS-TB can also be used to train cQA answer selection models without direct answer supervision. This shows that our methods can have broader impact on related tasks and beyond duplicate question detection.

2 Related Work

Duplicate question detection is closely related to question-question similarity and question retrieval. Early approaches use translation models (Jeon et al., 2005; Xue et al., 2008; Zhou et al., 2011) that were further enhanced with question category information (Cao et al., 2012) and topic models (Ji et al., 2012; Zhang et al., 2014).

More recent works in the context of the SemEval cQA challenges (Nakov et al., 2017) improve upon this and use tree kernels (TK) (Da San Martino et al., 2016), TK with neural networks (Romeo et al., 2016), neural networks with multi-task learning (Bonadiman et al., 2017), and encoder-decoder architectures together with shallow lexical matching and mismatching (Zhang and Wu, 2018). Common neural models such as CNNs achieved superior performance compared to TK when they were trained on sufficiently large numbers of labeled question pairs (Uva et al., 2018).

Similarly, neural representation learning methods have proved to be most effective in technical cQA domains. Santos et al. (2015), for example, learn representations of questions with CNNs and compare them with cosine similarity for scoring. Lei et al. (2016) propose RCNN, which extends CNN with a recurrent mechanism (adaptive gated

[4] https://superuser.com/q/1393090

Method	Duplicates	Answers	Bodies
Supervised	×	-	(×)
WS-QA	-	×	(×)
Domain Transfer	×*	-	(×)
DQG	-	-	×
WS-TB	-	-	×

Table 1: The different training methods and the data they use. Models typically also use text from the bodies during training and evaluation, which we indicate with (×). ×* = domain transfer requires duplicates from a sufficiently similar source domain.

decay). This approach was further extended with question-type information (Gupta et al., 2018).

If in-domain training data is scarce—i.e., if the cQA platform does not offer enough labeled duplicates—alternative training strategies are required. If there exist some labeled question pairs (thousands), one can first train a less data-hungry non-neural model and use it for supervised training of neural models (Uva et al., 2018). Further, if there exist large numbers of labeled question-answer pairs, we can use them for weakly-supervised training (Wang et al., 2017; Qiu and Huang, 2015).

More related to our work are methods that do not rely on any labeled data in the target domain. Existing methods use unsupervised training with encoder-decoder architectures (Lei et al., 2016; Zhang and Wu, 2018), and adversarial domain transfer where the model is trained on a source domain and adversarially adapted to a target domain (Shah et al., 2018). However, such approaches typically fall short of the performances that are being achieved with in-domain supervised training.

In contrast, we propose two novel methods, DQG and WS-TB, that do not require any annotations for model training and in some cases perform better than in-domain supervised training with duplicate questions. While WS-TB is related to the approaches mentioned before, DQG is is also related to question generation (QG). Most of the previous work in QG is in the context of reading comprehension (e.g., Du et al., 2017; Subramanian et al., 2018; Zhao et al., 2018; Du and Cardie, 2018) or QG for question answering (Duan et al., 2017). They substantially differ from our approach because they generate questions based on specific answer spans, while DQG generates a new title from a question's body that can be used as a question duplicate.

3 Training Methods

Given a pair of questions, our goal is to determine whether they are duplicates or not. In practice, the model predictions are often used to rank a list of potentially similar questions in regard to a new user question, e.g., to retrieve the most likely duplicate for automatic question answering.

To train models, we obtain a set of examples $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ in which each $x_n \in \mathcal{X}$ is an instance, i.e., a tuple containing texts such as two questions, and $y_n \in \{-1, +1\}$ is its corresponding binary label, e.g., duplicate/no-duplicate. Obtaining instances with positive labels $\mathcal{X}^+ = \{x_n^+ \in \mathcal{X} | y_n = 1\}$ is generally more difficult than obtaining $\mathcal{X}^-$ because instances with negative labels can be automatically generated (e.g., by randomly sampling unrelated questions).

In the following, we outline three existing training methods that use different kinds of instances, and in §3.2 we present our two novel methods: duplicate question generation, and weak supervision with title-body pairs. Both do not require any annotations in $\mathcal{X}^+$, and can therefore use larger amounts of data from the cQA forums. Table 1 gives an overview of the different training methods.

3.1 Existing Methods

Supervised (in-domain) training is the most common method, which requires labeled question duplicates, i.e., $x_n^+ = (q_n, \tilde{q}_n)$. Unrelated questions can be randomly sampled. With this data, we can train representation learning models (e.g., Lei et al., 2016) or pairwise classifiers (e.g., Uva et al., 2018). Most models combine the titles and bodies of the questions during training and evaluation (e.g., by concatenation), which can improve performances (Lei et al., 2016; Wu et al., 2018).

Weak supervision with question-answer pairs (WS-QA) is an alternative to supervised training for larger platforms without duplicate annotations (Qiu and Huang, 2015). WS-QA trains models with questions q_n and accepted answers a_n, and therefore $x_n^+ = (q_n, a_n)$. Instances in X^- can be obtained by randomly sampling unrelated answers for a question. An advantage of this method is that there typically exist more labeled answers than duplicate questions. For instance, Yahoo! answers has accepted answers but it does not contain labeled duplicate questions.

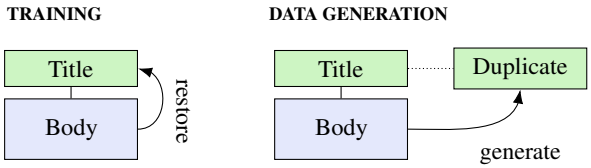

Figure 2: During training we restore the original question title from its body. During data generation we consider the generated title as a new duplicate question.

Domain transfer performs supervised training in a source domain and applies the trained model to a different target domain in which no labeled duplicate questions exist. Shah et al. (2018) use this method with adversarial training to learn domain-invariant question representations prior to transfer. They show that adversarial training can considerably improve upon direct transfer, but their method requires sufficiently similar source and target domains. For instance, they could not successfully transfer models between technical and other non-technical domains.

3.2 Proposed Methods with Unlabeled Data

The disadvantage of the existing methods is that they require labeled question duplicates, accepted answers, or similar source and target domains for transfer. We could alternatively use unsupervised training within an encoder-decoder framework, but this imposes important limitations on the network architecture, e.g., a question can only be encoded independently (no inter-attention).

Our proposed methods do not suffer from these drawbacks, i.e., they do not require labeled data and they do not impose architectural limitations.

Duplicate question generation (DQG) generates new question titles from question bodies, which we then consider as duplicates to the original titles. Our overall approach is depicted in Figure 2. First, we train a question generation model QG to maximize $P(\text{title}(q_n)|\text{body}(q_n))$. This is similar to news headline generation or abstractive summarization (Rush et al., 2015; Chopra et al., 2016) because QG needs to identify the most relevant aspects in the body that best characterize the question. However, restoring the exact title is usually not possible because titles and bodies often contain complementary information (see, e.g., Figure 1). We therefore consider $\text{QG}(\text{body}(q_n))$ as a duplicate of $\text{title}(q_n)$ and obtain positive labeled instances $x_n^+ = (\text{title}(q_n), \text{QG}(\text{body}(q_n)))$.

Dataset	Train / Dev / Test	|Q|	|A|
AskUbuntu-Lei	12,584 / 200 / 200	288k	84k
AskUbuntu	9106 / 1000 / 1000	288k	84k
SuperUser	9106 / 1000 / 1000	377k	142k
Apple	- / 1000 / 1000	89k	29k
Android	- / 1000 / 1000	47k	14k

Table 2: The dataset statistics. Numbers for Train/Dev/Test refer to the number of questions with duplicates. |Q| refers to the number of unlabeled questions, and |A| refers to the number of accepted answers.

Because DQG requires no annotated data, we can use this method to train duplicate detection models for all cQA forums that offer a reasonable number of unlabeled title-body pairs to obtain a suitable QG model (the smallest number of questions we tried for training of question generation models is 23k, see §5). An important advantage is that we can make use of *all questions* (after some basic filtering), which is often an order of magnitude more training data than annotated duplicates.

We can use any sequence to sequence model for QG, and we performed experiments with a Transformer (Vaswani et al., 2017) and MQAN (McCann et al., 2018).

Weak supervision with title-body pairs (WS-TB) takes the assumption of DQG one step further. If question titles and question bodies have similar attributes as duplicates, we could also just train duplicate detection models directly on this data without prior question generation.

In WS-TB, we thus train models to predict whether a given title and body are related, i.e., whether they belong to the same question. Therefore, $x^+ = (\text{title}(q_n), \text{body}(q_n))$.

This method considerably simplifies the sourcing of training data because it requires no separate question generation model. However, it also means that the duplicate detection model must be able to handle texts of considerably different lengths during training (for instance, bodies in SuperUser.com have an average length of 125 words). This might not be suitable for some text matching models, e.g., ones that were designed to compare two sentences.

4 Experiments

4.1 Experimental Setup

We use models and data from previous literature to obtain comparable results for evaluation, and we rely on their official implementations, default

hyperparameters, and evaluation measures. An overview of the datasets is given in Table 2, which also shows that they considerably differ in the amounts of data that is available for the different training methods.

The evaluation setup is the same for all datasets: given a user question q and a list of potentially related questions, the goal is to re-rank this list to retrieve duplicates of q (one or more potential related questions are labeled as duplicates). Even though some training methods do not use bodies during training, e.g., WS-DQG, during evaluation they use the same data (annotated pairs of questions with titles and bodies).[5]

AskUbuntu-Lei. First, we replicate the setup of Lei et al. (2016), which uses RCNN to learn dense vector representations of questions and then compares them with cosine similarity for scoring. Besides supervised training, this also includes unsupervised training with the encoder-decoder architecture. We report precision at 5 (P@5), i.e., how many of the top-5 ranked questions are actual duplicates. The dataset is based on the AskUbuntu data of Santos et al. (2015) with additional manual annotations for dev/test splits (user questions have an average of 5.7 duplicates).

Android, Apple, AskUbuntu, and Superuser. Second, we replicate the setup of Shah et al. (2018), which uses BiLSTM to learn question representations. This setup also includes adversarial domain transfer. The data is from the AskUbuntu, Superuser, Android, and Apple sites of StackExchange, and different to AskUbuntu-Lei, each question has only one duplicate. We measure AUC(0.05), which is the area under curve with a threshold for false positives—Shah et al. (2018) argue that this is more stable when there are many unrelated questions.

Questions and answers. To train the models with WS-TB and WS-QA, we use questions and answers from publicly available data dumps[6] of the StackExchange platforms. We obtain our new training sets as specified in §3.2. For instance, for WS-TB we replace every annotated duplicate $(q_n, \tilde{q}_n)$ from the original training split with $(\text{title}(q_n), \text{body}(q_n))$, and we randomly sample unrelated bodies to obtain training instances with negative labels.

It is important to note that the number of questions and answers is much larger than the number of annotated duplicate questions. Therefore, we can add more instances to the training splits with these methods. However, if not otherwise noted, we use the same number of training instances as in the original training splits with duplicates.

DQG setup. To train question generation models, we use the same StackExchange data. We filter the questions to ensure that the bodies contain multiple sentences. Further, if a body contains multiple paragraphs, we only keep the one with the highest similarity to the title. Details of the filtering approach are included in the Appendix. Less than 10% of the questions are discarded on average.

We train a MQAN (Multi-task Question Answering Network) model, which was proposed as a very general network architecture to solve a wide variety of tasks as part of the Natural Language Decathlon (McCann et al., 2018). The model first encodes the input with LSTMs and applies different attention mechanisms, including multi-headed self-attention. MQAN also includes pointer-generator networks (See et al., 2017), which allow it to copy tokens from the input text depending on the attention distribution of an earlier layer.

We performed the same experiments with a Transformer sequence to sequence model (Vaswani et al., 2017), but on average MQAN performed better because of its ability to copy words and phrases from the body. We include the Transformer results and a comparison with MQAN in the Appendix.

We use all available questions from a cQA forum to train the question generation model. We perform early stopping using BLEU scores to avoid overfitting. To generate duplicate questions, we then apply the trained model on all questions from the same cQA forum. We do not use a separate heldout set because this would considerably limit both the question generation training data and the number of generated duplicates. We did not observe negative effects from using this procedure.

4.2 Experimental Results

The results are given in Table 3. For domain transfer, we report the best scores from Shah et al. (2018), which reflects an optimal transfer setup from a similar source domain.

[5] It has been shown that including bodies in the experimental setup can lead to improved performances (Lei et al., 2016). In initial experiments, we found that the performances are mostly impacted by having access to bodies during evaluation.

[6] https://archive.org/download/stackexchange

Measuring P@5. Results (dev / test) for RCNN	**AskUbuntu-Lei**	**Android**	**Apple**	**AskUbuntu**	**Superuser**	Average
		Measuring AUC(0.05). Results for BiLSTM				
Trained on 1x data (all methods use the same number of training instances as in supervised training)						
Supervised (in-domain)	**48.0** / 45.0	-	-	**0.848**	**0.944**	-
Unsupervised	42.6 / 42.0	-	-	-	-	-
Direct Transfer (best)	-	0.770	0.828	0.730	0.908	0.809
Adversarial Transfer (best)	-	0.790	0.861	0.796	0.911	0.840
WS-QA	47.2 / 45.3	0.780	**0.894**	0.790	0.919	0.846
DQG	46.4 / 44.8	0.793	0.870	0.801	0.921	0.846
WS-TB	46.4 / **45.4**	**0.811**	0.866	0.804	0.913	**0.849**
Trained on all available data						
Unsupervised	43.0 / 41.8	-	-	-	-	-
WS-QA	47.3 / 44.2	0.814	0.901	0.828	0.951	0.874
DQG	**47.4** / 44.3	0.833	0.911	0.855	0.944	0.886
WS-TB	47.3 / **45.3**	0.852	0.910	**0.871**	**0.952**	0.896
DQG + WS-TB (combined)	46.4 / 44.0	**0.863**	**0.916**	0.866	0.946	**0.898**

Table 3: Results of the models with different training strategies. Android and Apple datasets do not contain labeled duplicates for supervised in-domain training.

Supervised training. As we expect, supervised in-domain training with labeled duplicates achieves better scores compared to other training methods when we consider the same number of training instances. An exception is on AskUbuntu-Lei where DQG, WS-TB, and WS-QA can achieve results that are on the same level on test or marginally worse on dev.

One reason for the better performances with labeled duplicates is that they contain more information, i.e., a pair of questions consist of two titles and two bodies compared to just one title and body for each training instance in WS-TB. However, the results show that all weakly supervised techniques as well as DQG are effective training methods.

DQG, WS-TB, and WS-QA. All methods outperform direct transfer from a similar source domain as well as the encoder-decoder approach on AskUbuntu-Lei. On average, WS-TB is the most effective method, and it consistently outperforms adversarial domain transfer (0.9pp on average).

We otherwise do not observe large differences between the three methods DQG, WS-TB, and WS-QA, which shows that (1) the models we use can learn from different text lengths (title-body, question-answer); and (2) the information that we extract in DQG is suitable for training (examples are given in §6). The good results of WS-TB might suggest that question generation as separate step is not required, however we argue that it can be important in a number of scenarios, e.g., when we need to train sentence matching models that would otherwise not be able to handle long texts.

Using all available data. One of the biggest advantages of our proposed methods is that they can use larger amounts of training data. This greatly improves the model performances for BiLSTM, where we observe average improvements of up to 4.7pp (for WS-TB). In many cases our methods now perform better than supervised training. We observe smaller improvements for WS-QA (2.8pp on avg) because it has access to fewer training instances. The performances for RCNN on AskUbuntu-Lei are mostly unchanged with minor improvements on dev. The reason is that the performances were already close to supervised training with the same data sizes.

In Figure 3 we plot the performance scores of BiLSTM averaged over the four StackExchange datasets in relation to the available training data with WS-TB. We see that the model performance consistently improves when we increase the training data (we observe similar trends for DQG and WS-QA). Thus, it is crucial to make use of all available data from the cQA forums.

We also explored a combination of our two proposed approaches where we merge their respective training sets. We find that this helps mostly for smaller cQA platforms with fewer questions (where larger training sets would be most necessary), e.g., the performances on Android and Apple improve by 0.6–1.1pp compared to WS-TB. Even though the combination does not introduce new information because both use the same question data, complementing WS-TB with DQG can provide additional variation with the generative component.

In summary, our results show that even when we

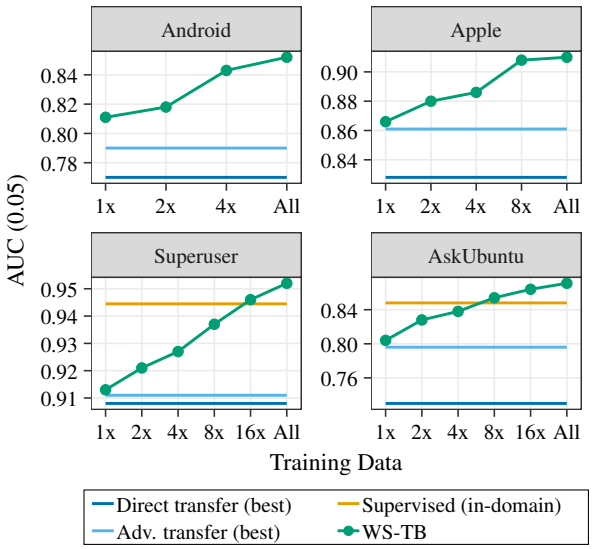

Figure 3: Performances of BiLSTM as a function of the available training data. '2x' means that there are twice as many (unlabeled) questions available to WS-TB than there are annotated duplicate questions in the original dataset (1x = 9106).

have access to sufficient numbers of labeled duplicates, the 'best' method is not always supervised training. When we use larger numbers of title-body pairs, DQG and WS-TB can achieve better performances.

5 Further Application Scenarios

To test if our methods are applicable to other scenarios with high practical relevance, we explore (1) whether DQG can be used in cQA forums with fewer unlabeled title-body pairs, (2) if we can use WS-TB to train answer selection models without labeled question-answer pair, and (3) how well large pre-trained language models perform when being fine-tuned with our methods.

5.1 DQG for Small-Scale cQA Forums

In our previous experiments, we assumed that there exist enough unlabeled questions to train the question generation model (at least 47k questions, see Table 2). To simulate a more challenging scenario with fewer in-domain questions, we explore the effects of cross-domain question generation. This is highly relevant for DQG because in such scenarios the generated duplicates could be combined with WS-TB to obtain more training data.

We replicate the transfer setup of Shah et al. (2018) where they originally transfer the duplicate question detection model from a source to a target domain. For DQG we instead train the question generation model on the source domain and generate duplicates for the target domain, with which we then train the duplicate detection model. To provide a fair comparison against adversarial domain transfer, we always use the same number of 9106 duplicates to train the duplicate detection models.

Source	Target	Adv. DT	DQG	Δ
AskUbuntu	Android	0.790	**0.797**	+0.004
	Apple	0.855	**0.861**	−0.009
	SuperUser	0.911	**0.916**	−0.005
SuperUser	Android	0.790	**0.794**	+0.001
	Apple	**0.861**	**0.861**	−0.009
	AskUbuntu	0.796	**0.809**	+0.008
Academia	Android	-	0.776	−0.017
	Apple	-	0.854	−0.016
	SuperUser	-	0.912	−0.009
	AskUbuntu	-	0.760	−0.039

Table 4: The domain transfer performances. Δ denotes the difference to the setup with in-domain DQG.

Results for the transfer from SuperUser and AskUbuntu to other domains are given in Table 4. They show that the question generation model for DQG can be successfully transferred across similar domains with only minor effects on the performances. Importantly, DQG still performs better than adversarial domain transfer with the same number of training instances.

To test an even more extreme case, we also transfer from StackExchange Academia (only 23k title-body pairs to train question generation) to the technical target domains. This could, e.g., be realistic for other languages with fewer cQA forums. Most notably, the performance of DQG decreases only mildly, which demonstrates its practical applicability in even more challenging scenarios. This is mostly due to the copy mechanism of MQAN, which is stable across domains (see §6).

5.2 Answer Selection

In answer selection we predict whether a candidate answer is relevant in regard to a question (Tay et al., 2017; Nakov et al., 2017; Tan et al., 2016; Rücklé and Gurevych, 2017), which is similar to duplicate question detection.

To test whether our strategy to train models with title-body pairs is also suitable for answer selection, we use the data and code of Rücklé et al. (2019a) and train two different types of models with WS-TB on their five datasets that are based on StackExchange Apple, Aviation, Academia, Cooking, and Travel. We train (1) a siamese BiLSTM, which

Model	Supervised	WS-TB	WS-TB (all)
BiLSTM	35.3	37.5	42.5
COALA	44.7	45.2	44.5

Table 5: Answer selection performances (averaged over five datasets) when trained with question-answer pairs vs. WS-TB.

learns question and answer representations; and (2) their neural relevance matching model COALA. Both are evaluated by how well they re-rank a list of candidate answers in regard to a question.

The results are given in Table 5 where we report the accuracy (P@1), averaged over the five datasets. Interestingly, we do not observe large differences between supervised training and WS-TB for both models when they use the same number of positive training instances (ranging from 2.8k to 5.8k). Thus, using title-body information instead of question-answer pairs to train models without direct answer supervision is feasible and effective. Further, when we use all available title-body pairs, the BiLSTM model substantially improves by 5pp, which is only slightly worse than COALA (which was designed for smaller training sets). We hypothesize that one reason is that BiLSTM can learn improved representations with the additional data. Further, title-body pairs have a higher overlap than question-answer pairs (see §6) which provides a stronger training signal to the siamese network.

These results demonstrate that our work can have broader impact to cQA, e.g., to train models on other tasks beyond duplicate question detection.

5.3 BERT Fine-Tuning

Large pre-trained language models such as BERT (Devlin et al., 2018b) and RoBERTa (Liu et al., 2019) have recently led to considerable improvements across a wide range of NLP tasks. To test whether our training strategies can also be used to fine-tune such models, we integrate BERT in the setups of our previous experiments.[7] We fine-tune a pre-trained BERT-base (uncased) model with supervised training, WS-TB (1x), and WS-TB (8x).

The results are given in Table 6. We observe similar trends as before but with overall better results. When increasing the number of training examples, the model performances consistently improve. We note that we have also conducted preliminary experiments with larger BERT models where we observed further improvements.

[7]We add the AskUbuntu-Lei dataset to the framework of Rücklé et al. (2019a) for our BERT experiments. Details are given in the Appendix.

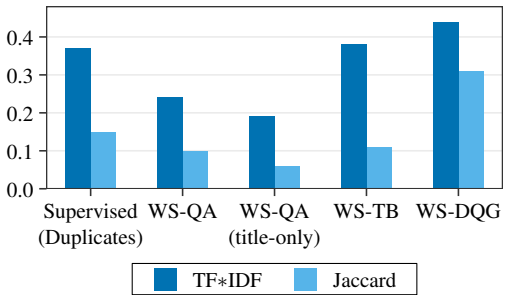

Figure 4: Average overlap of texts from positive training instances (words were stemmed and lowercased).

6 Analysis

6.1 Overlap

To analyze the differences in the training methods we calculate the overlap between the texts of positive training instances (e.g., question-question, title-body, question-answer etc.). For questions, we concatenate titles and bodies.

Figure 4 shows the Jaccard coefficient and the TF∗IDF score averaged over all instances in the four StackExchange datasets of §4.2. We observe that the overlap in WS-TB is similar to the overlap of actual duplicate questions in supervised training. The WS-DQG overlap is higher, because generated titles only contain relevant content (e.g., no conversational phrases). We also found that the BLEU scores of the MQAN model for QG are not very high (between 13.3–18.9 BLEU depending on the dataset), which shows that the texts are still sufficiently different. The overlap shows that both our methods use suitable training data with sufficiently similar, but not fully redundant texts.

Interestingly, the overlap scores of question-answer pairs are lower, especially when considering title-answer pairs as it is the case in the answer selection experiments (§5.2). This could explain one factor that may contribute to the better scores that we achieve with WS-TB for BiLSTM in this scenario. Because the overlap of title-body pairs is higher, the siamese network can receive a stronger training signal for positive instances, which could lead to better representations for similarity scoring.

6.2 Qualitative Analysis

To better understand the results for DQG and WS-QA, we manually checked a random sample of 200 generated questions and title-body pairs from multi-

	AskUbuntu-Lei	Android	Apple	AskUbuntu	Superuser	Answer Selection
	Measuring P@5. Results for dev / test	Measuring AUC(0.05)				Measuring Accuracy
Supervised (in-domain)	**54.0 / 52.3**	-	-	0.862	0.954	56.8
WS-TB (1x)	47.8 / 47.2	0.857	0.908	0.841	0.932	55.5
WS-TB (8x)	50.4 / 49.6	**0.896**	**0.933**	**0.897**	**0.971**	**59.7**

Table 6: Results of fine-tuned BERT models with different training strategies.

Title: 14.10, 15.04 - HDMI audio not working on Dell Vostro 3750 - nVidia card not detected by aplay -l

DQG: ALSA not detected in nVidia

Title: Installing ubuntu 12.04.02 in uefi mode

DQG: Ubuntu 16.04 LTS boot loader not working

Title: Grub2 not updating

DQG: How to fix Grub2 error

Figure 5: Random samples of titles and DQG output.

ple platforms. Three titles and generated duplicates from AskUbuntu are shown in Figure 5.

For DQG we found that most of the generated duplicates are sensible, and most of the error cases fall into one of the following two categories:

(1) Some generated questions are somewhat off-topic because they contain information that was generated from a body that has minimal overlap with the title (see example 4 in the Appendix).

(2) A number of questions include wrong version numbers or wrong names (see example 5 in the Appendix, or the second example in Figure 5).

Generally, however, we find that many of the generated titles introduce novel information, as can be seen in Figure 5 (e.g., 'ALSA', 'boot loader' etc). The same drawbacks and benefits also apply to title-body information in WS-TB, with the exception that they are less noisy (i.e., not generated) but contain conversational phrases and many details.

We also checked the training data of the difficult DQG domain transfer case to explore reasons for the small performance decreases when transferring the question generation model. Most importantly, we find that the model often falls back to copying important phrases from the body and sometimes generates additional words from the source domain. We note that this is not the case for models without copy mechanisms, e.g., Transformer often generates unrelated text (examples are in the Appendix).

7 Conclusion

In this work, we have trained duplicate question detection models without labeled training data. This can be beneficial for a large number of cQA forums that do not contain enough annotated duplicate questions or question-answer pairs to use existing training methods. Our two novel methods, duplicate question generation and weak supervision with title-body pairs, only use title-body information of unlabeled questions and can thus utilize more data during training. While both are already highly effective when using the same number of training instances as other methods (e.g., outperforming adversarial domain transfer), our experiments have shown that we can outperform even supervised training when using larger amounts of unlabeled questions.

Further, we have demonstrated that weak supervision with title-body pairs is well-suited to train answer selection models without direct answer supervision. This shows that our work can potentially benefit a much wider range of related tasks beyond duplicate question detection. For instance, future work could extend upon this by using our methods to obtain more training data in cross-lingual cQA setups (Joty et al., 2017; Rücklé et al., 2019b), or by combining them with other training strategies, e.g., using our methods for pre-training.

The source code and the data of our experiments are publicly available: `http://github.com/UKPLab/emnlp2019-duplicate_question_detection`.

Acknowledgements

This work has been supported by the German Research Foundation (DFG) as part of the QA-EduInf project (grant GU 798/18-1 and grant RI 803/12-1), by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 03VP02540 (ArgumenText), and by the German Federal Ministry of Education and Research (BMBF) as part of the Software Campus program under the promotional reference 01IS17050.

References

Daniele Bonadiman, Antonio Uva, and Alessandro Moschitti. 2017. Effective shared representations with multitask learning for community question answering. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, pages 726–732.

Xin Cao, Gao Cong, Bin Cui, Christian S. Jensen, and Quan Yuan. 2012. Approaches to exploring category information for question retrieval in community question-answer archives. *ACM Transactions on Information Systems (TOIS)*, 30(2):7:1–7:38.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*, pages 93–98.

Giovanni Da San Martino, Alberto Barrón Cedeño, Salvatore Romeo, Antonio Uva, and Alessandro Moschitti. 2016. Learning to re-rank questions in community question answering using advanced features. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM 2016)*, pages 1997–2000.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 1907–1917.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 1342–1352.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 866–874.

Deepak Gupta, Rajkumar Pujari, Asif Ekbal, Pushpak Bhattacharyya, Anutosh Maitra, Tom Jain, and Shubhashis Sengupta. 2018. Can taxonomy help? improving semantic question matching using question taxonomy. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 499–513.

Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. 2005. Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM 2005)*, pages 84–90.

Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. 2012. Question-answer topic model for question retrieval in community question answering. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM 2012)*, pages 2471–2474.

Shafiq Joty, Preslav Nakov, Lluís Màrquez, and Israa Jaradat. 2017. Cross-language learning with adversarial neural networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 226–237.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Lluís Màrquez. 2016. Semi-supervised question retrieval with gated convolutions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*, pages 1279–1289.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Preslav Nakov, Doris Hoogeveen, Lluís 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-2017)*, pages 27–48.

Xipeng Qiu and Xuanjing Huang. 2015. Convolutional neural tensor network architecture for community-based question answering. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 15)*, pages 1305–1311.

Salvatore Romeo, Giovanni Da San Martino, Alberto Barrón-Cedeño, Alessandro Moschitti, Yonatan Belinkov, Wei-Ning Hsu, Yu Zhang, Mitra Mohtarami, and James Glass. 2016. Neural attention for learning to rank questions in community question answering. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 1734–1745.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated power mean embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*.

Andreas Rücklé and Iryna Gurevych. 2017. Representation learning for answer selection with lstm-based importance weighting. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017)*.

Andreas Rücklé, Nafise Sadat Moosavi, and Iryna Gurevych. 2019a. Coala: A neural coverage-based approach for long answer selection with small data. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI 2019)*, pages 6932–6939.

Andreas Rücklé, Krishnkant Swarnkar, and Iryna Gurevych. 2019b. Improved cross-lingual question retrieval for community question answering. In *Proceedings of the World Wide Web Conference (WWW 2019)*, pages 3179–3186.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 379–389.

Cicero dos Santos, Luciano Barbosa, Dasha Bogdanova, and Bianca Zadrozny. 2015. Learning hybrid representations to retrieve semantically equivalent questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015)*, pages 694–699.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 1073–1083.

Darsh Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. 2018. Adversarial domain adaptation for duplicate question detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 1056–1063.

Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. 2018. Neural models for key phrase extraction and question generation. In *Proceedings of the Workshop on Machine Reading for Question Answering (MRQA 2018)*, pages 78–88.

Ming Tan, Cicero Dos Santos, Bing Xiang, and Bowen Zhou. 2016. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 464–473.

Yi Tay, Minh C. Phan, Luu Anh Tuan, and Siu Cheung Hui. 2017. Learning to rank question answer pairs with holographic dual lstm architecture. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017)*, pages 695–704.

Antonio Uva, Daniele Bonadiman, and Alessandro Moschitti. 2018. Injecting relational structural representation in neural networks for question similarity. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 285–291.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, ukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS 2017)*, pages 5998–6008.

Pengwei Wang, Yong Zhang, Lei Ji, Jun Yan, and Lianwen Jin. 2017. Concept embedded convolutional semantic model for question retrieval. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM 2017)*, pages 395–403.

Wei Wu, Xu Sun, and Houfeng Wang. 2018. Question condensing networks for answer selection in community question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 1746–1755.

Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, pages 475–482.

Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, and Ming Zhou. 2014. Question retrieval with high quality answers in community question answering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM 2014)*, pages 371–380.

Minghua Zhang and Yunfang Wu. 2018. An unsupervised model with attention autoencoders for question retrieval. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI 2018)*, pages 4978–4986.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 3901–3910.

Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. 2011. Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2011)*, pages 653–662.