DeepDup: Duplicate Question Detection in Community Question Answering

MOHOMED SHAZAN MOHOMED JABBAR

Alberta Machine Intelligence Institute

LUKE KUMAR

Alberta Machine Intelligence Institute

HAMMAN SAMUEL*

Department of Computing Science, University of Alberta, Edmonton, Canada

MI-YOUNG KIM

Department of Science, Augustana Faculty, University of Alberta

SANKALP PRABHAKAR

Alberta Machine Intelligence Institute

RANDY GOEBEL

Department of Computing Science, University of Alberta, Edmonton, Canada

OSMAR ZAÏANE

Department of Computing Science, University of Alberta, Edmonton, Canada

Duplicate question detection is an ongoing challenge in community question answering because semantically equivalent questions can have significantly different words and structures. The identification of duplicate questions can reduce the resources required for retrieval and increase findability of the associated community forums. This ongoing study presents DeepDup, a deep learning model for duplicate question detection. Our research also explores the possibility of domain adaptation with transfer learning to improve the under-performing target domains for the text-pair duplicates classification task, using heterogeneous datasets from the Stack Exchange sub-communities for Ubuntu and English. Ultimately, our study investigates the null hypothesis that there is no significant difference between a base model and a transfer-learned model.

CCS CONCEPTS • Machine Learning • Natural Language Processing • Social Media

Additional Keywords and Phrases: Community Question Answering, Duplicate Detection, Deep Learning

ACM Reference Format:

Mohomed Shazan Mohomed Jabbar, Luke Kumar, Hamman Samuel, Mi-Young Kim, Sankalp Prabhakar, Randy Goebel, Osmar Zaiane. 2021. DeepDup: Cross-Domain Duplicate Question Detection in Community Question Answering. In ICDLT '21: International Conference on Deep Learning Technologies, July 23–25, 2021, Qingdao, China.

* Correspondence email: hwsamuel@ualberta.ca

1 INTRODUCTION

To increase findability on Community Question Answering (CQA) forums, users ought to avoid re-posting questions that have been already asked and answered. The term “findability” is taken from information architecture literature, meaning the ease of locating relevant information [1], and identification of duplicate questions in CQA forums provides three main advantages. Firstly, finding duplicate questions saves users' time because they do not have to wait for responses. Secondly, users searching for questions will be presented with better results with duplicates pruned. Thirdly, the overall performance of the CQA forum will be enhanced because of reduced unnecessary content.

However, identifying two questions as duplicates is challenging because the choice of words, structure of sentences, and even context, can vary significantly between questions, even if the intended semantics are near identical. In addition, questions with similar verbiage are not necessarily duplicates. Traditional computational Information Retrieval (IR) and Natural Language Processing (NLP) methods have achieved only limited success in detecting semantically identical text-pairs. When comparing state-of-the-art machine learning methods for this task, an interesting observation is that a classification model trained on a dataset from one domain might not achieve the same performance to predict text-pair duplicates in another domain. For instance, the similarity between two question pairs could be completely different depending on the domain of the dataset which was used to train the classification model. As an example, the question pair “Where can I find a place to eat pizza?” and “What's the closest Italian restaurant?” can be classified as duplicate or not, depending on the domain of the dataset used in model training. When using the SpaCy Python library, on the Quora CQA website, the similarity for these sentences was reported as 6%, while on the Stack Exchange CQA website, it was 46% [2], demonstrating a numerical difference. Whether this observation could be generalized a research question we explore in our study. Another illustration of the challenge of duplicate detection is shown in Figure 1.

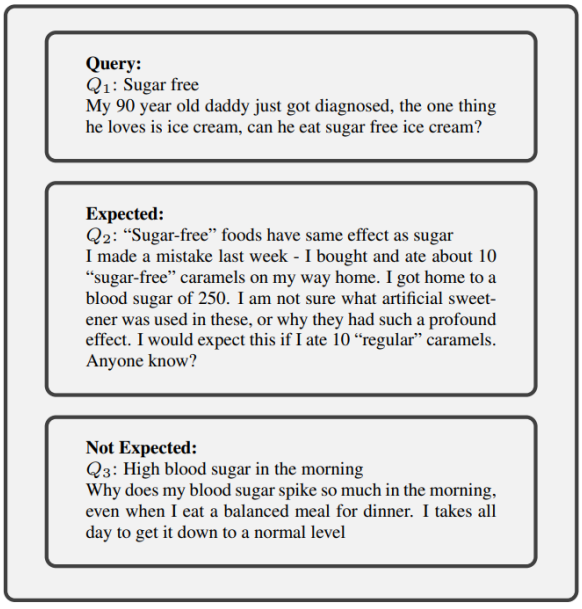

Figure 1: Question retrieval examples demonstrating the challenge of duplicate question detection.

In this research work, we present DeepDup, a Siamese Neural Network (SNN) architecture that performs duplicate question detection. In addition, we explore the possibility of cross-domain adaptation to increase the performance of under-performing domains using Transfer Learning (TL). We address three questions with our research goals. Firstly, we investigate the research challenge of text-pair duplicate detection. Secondly, we explore the possibility of a general-purpose cross-domain duplicate detection approach for heterogeneous datasets. Ultimately, we determine whether the domain of the training dataset affects the outcomes of the trained model, and evaluate the null hypothesis that there is no significant difference between a base model and a transfer-learned model.

2 METHODOLOGY

To evaluate DeepDup and compare with other approaches, we used publicly available datasets from Stack Exchange's sub-communities for Ubuntu and English forums. The language used in these sub-communities is diverse enough for exploring duplicate detection across domains. The datasets have labels for duplicate and non-duplicate question pairs annotated by forum moderators. Dataset properties are summarized in Table 1.

Table 1: Dataset Properties. WPQ = Words Per Question.

	Question Pairs	**Max WPQ**	**Mean WPQ**
Ubuntu	131,271	33	8.7
English	33,661	32	8.9

2.1 Data Preprocessing

For preprocessing, each question was tokenized, and question pairs whose data types do not match were filtered. Non-English questions were removed by checking for non-English vowels. We also performed stop word removal, lemmatization, and stemming. Finally, abbreviated forms such as “what's”, “i'm”, etc. were transformed to their unabbreviated forms, i.e. “what is”, “i am”, etc.

2.2 Deep Learning

Underlying semantic similarity between questions can be learned with a better numerical representation of the texts, such as the ones learned through neural network models. The datasets we used have sufficient attributes to be used with a variety of deep learning models. Siamese neural networks (SNN) have been popularly used to compare two objects and find similarity relationships between them [3]. A salient feature of these networks is that they employ two sub-networks, which share parameters, thus reducing the number of parameters to learn, and give a consistent representation for the two objects being compared.

We adapted a similar architecture for DeepDup to compare question pairs, and to determine whether they are duplicates. In Figure 2 we illustrate a detailed view of this adapted architecture to the duplicate question problem. The Lambda layer given in Figure 2 combines the two representations we get from the two sentences. The researchers in [3] proposed the contrastive loss function for the Siamese network for images, and we propose a new loss function shown in Equation 1, based on an absolute error, which we found to give better empirical results for the Quora CQA dataset. Here, x_1 and x_2 are the representations of the two sentences.

$$e^{-1\times\|x_1-x_2\|} \quad (1)$$

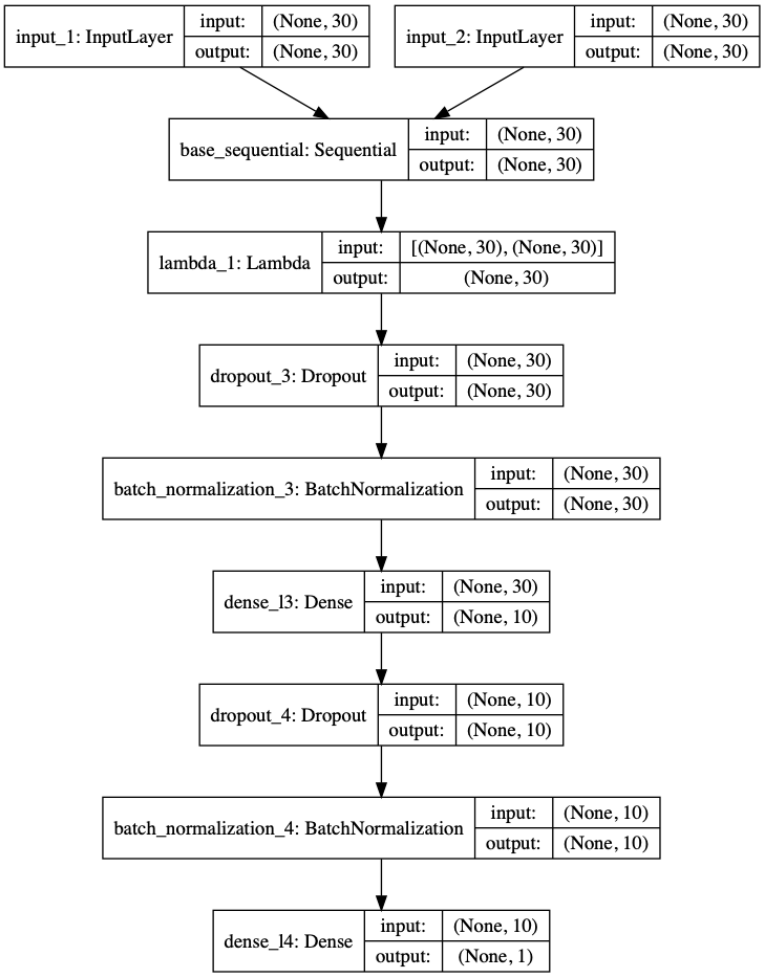

Figure 2: Siamese Neural Network (SNN) for duplicate question detection, showing input and output dimensions of each layer starting from two input sentences with 30 tokens. Architecture of the `base_sequential` layer is given in Figure 3.

2.3 Transfer Learning

Knowledge learned from one domain can be transferred to another domain using TL. With this approach, trained models from a better-performing source domain are used to increase the performance of an under-performing target domain with insufficient or sparsely labeled examples [4]. Prior work in deep learning-based computer vision models indicates that TL can be successfully utilized [5]. NLP applications have used similar ideas to improve performance in certain target domains [4,6]. For our research, we explore the possibility of transferring and utilizing knowledge learned from one dataset to improve the performance in other target domains. Our intention is that this will lead to generally improved duplicate question detection across domains.

Specifically, we adopt the INIT TL approach [6], which uses parameters trained on a source domain to initialize parameters of the target domain's model. Figure 3 provides full details for the common base layers that build representation for both the sequences, i.e. shared parameter labelled as `base_sequential` in Figure 2.

All sentence representations are trained with this shared network starting from the pre-trained GloVe word-embeddings [7] and later we feed it to a Long-Short Term Memory (LSTM) network to learn the sequential representations.

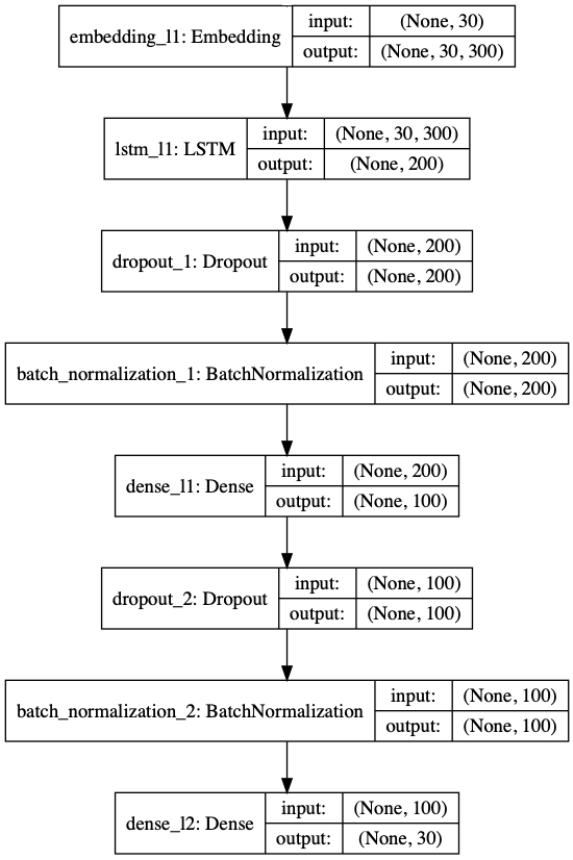

Figure 3: Common base layers for TL, where each sentence will be represented by a 30 dimension vector for later layers.

2.4 Hypothesis Testing

For testing our hypothesis, we use the McNemar test statistic, shown in Equation 2, where χ^2 is the test statistic, based on a chi-squared distribution with 1 degree of freedom, and the values for b and c are retrieved from a contingency table, which is a summarized tabulation of outcomes from two tests [8]. The McNemar test has been widely used for evaluation of models where it would be expensive or impractical to train multiple models [9]. The null hypothesis is that there is no significant difference between the base approach of using DeepDup and the target approach that uses TL. This hypothesis is tested separately for each of the Ubuntu and the English datasets.

$$\chi^2 = \frac{(b-c)^2}{b+c} \qquad (2)$$

3 RESULTS AND DISCUSSION

As a baseline, the base model's performance on a dataset taken from the Quora CQA website was 94.1% AUC. Having gotten good performance for the Quora dataset, our research question is to determine how we could utilize these results on more specialized datasets for the same task. Results for the English and Ubuntu datasets are presented in Figure 4.

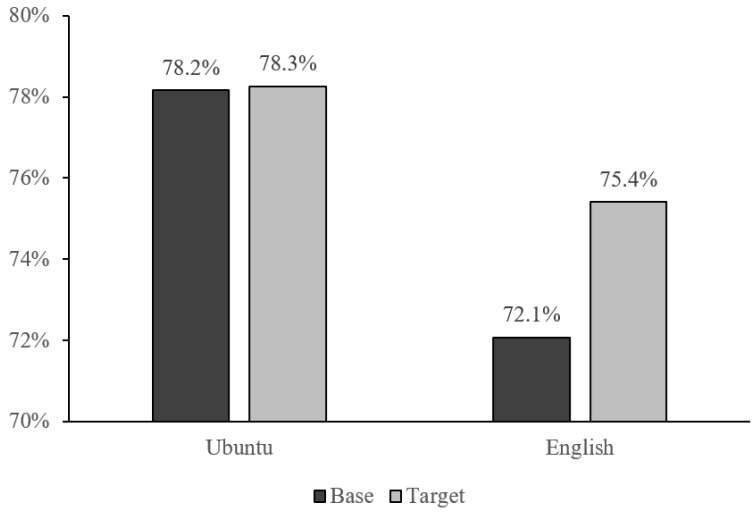

Figure 4: Accuracy Results from Heterogeneous Datasets using SNN (base) and TL (target)

For comparative evaluation of the performance, we employed the accuracy metric. For the Ubuntu dataset, the accuracy was slightly better for the target model. Also for the English dataset, the accuracy was noticeably better for the target model. Overall, the accuracy tends to shows that the target model gives better performance on true positives and true negatives. However, to perform a more rigorous comparison, we used the McNemar statistical test to compare the two results.

For hypothesis testing, α was set at 5%, in order to compare with the p-value from the McNemar test statistic. For the case of the English dataset, the test statistic was 212.00 and p-value ≈ 0.000. Similarly, for the Ubuntu dataset, p-value ≈ 0.013 from the test statistic of 87.00. Hence, in both cases, the null hypothesis is rejected. This implies that there is evidence supporting the alternative hypotheses for a statistically significant difference in error proportions of the two models in each case, and by extension, both the models are different. Together with the higher accuracy values we obtained for the target models in Ubuntu and English, we can say that the amount of knowledge that can be positively transferred across domains is significant.

Comparing results between the two TL models, it can be seen that the performance on the English dataset is better than the Ubuntu dataset. We postulate that the domain's language complexity could affect the performance of the duplicate text-pairs prediction task. The Ask Ubuntu forums tend to have questions with technical jargon and acronyms, while the English forums are more general-purpose or academic.

4 RELATED LITERATURE

We can classify the techniques for similar or duplicate question retrieval into three categories: i) translation models, ii) topic models, and iii) deep learning approaches. For methods using translation models, phrase-based translation models in community-based question retrieval have proven more effective because they seem to capture contextual information in modeling the translation of phrases as a whole, rather than translating single words in isolation [10,11]. While these studies showed some promising results for detecting similar text-pairs, we did not gain any improvements in performance in exploring translation models on our duplicate datasets.

Some research has explored the use of topic models: first latent topics are identified for each question pair; and then four similarity scores are computed using their titles, descriptions, latent topics and tags [12]. Others have proposed a supervised question-answer topic modeling approach, which assumes that questions and answers share some common latent topics [13]. Deep learning approaches have been investigated for identifying semantically equivalent questions as well as duplicate postings [14–16]. Experiments have shown that a Convolutional Neural Network (CNN) can achieve high performance when word embeddings are pre-trained on in-domain data [17].

SNN have also been used for similar question retrieval, where the network learns the similarity metric for question pairs by leveraging question-answer pairs available in CQA archives [18]. The Siamese architecture approach has also been used for duplicate pairs classification specifically on the Quora data [19]. Our study exploits that research, and explores the general-purpose application of various machine learning approaches to heterogeneous datasets, including experimenting with TL to adapt for different domains.

5 CONCLUSION

We presented the results of our ongoing research work on DeepDup, a duplicate text-pair detection deep neural network with applications of community question answering. Our goal was to determine if a robust pipeline or model could be built that can predict duplicate question pairs across heterogeneous datasets. We tested the null hypothesis that there is no significant difference between a base model and a transfer-learned model. Our selected learning methods included deep neural networks and transfer learning, and our empirical analysis and statistical testing showed support for the alternative hypothesis that the amount of knowledge that can be positively transferred across domains is significant. For future work, we intend to investigate more complex configurations for transfer learning approaches using a larger collection of domain-diverse Q&A datasets while also tackling the data imbalance problem and working on increasing DeepDup's accuracy.

ACKNOWLEDGMENTS

The authors wish to thank the Alberta Machine Intelligence Institute, Amii, for supporting this research work. Amii is one of Canada’s preeminent centres of artificial intelligence, with a unique role in bridging world-leading academic research and professional industries.

REFERENCES

[1] Peter Morville. 2005. *Ambient Findability*. O’Reilly Media.

[2] Matthew Honnibal. 2017. Supervised Similarity: Learning Symmetric Relations from Duplicate Question Data.

[3] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1735–1742.

[4] Tushar Semwal, Gaurav Mathur, Promod Yenigalla, and Shivashankar B Nair. 2018. A Practitioners' Guide to Transfer Learning for Text Classification using Convolutional Neural Networks. *arXiv Prepr. arXiv1801.06480* (2018).

[5] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. 2016. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* 35, 5 (2016), 1285–1298.

[6] Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How Transferable are Neural Networks in NLP Applications? *arXiv Prepr. arXiv1603.06111* (2016).

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

[8] Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12, 2 (June 1947), 153–157.

[9] Thomas G. Dietterich. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* 10, 7 (October 1998), 1895–1923.

[10] 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 Annual Meeting of the Association for Computational Linguistics (ACL): Human Language Technologies (HLT)*, 653–662.

[11] Xiaobing Xue, Jiwoon Jeon, and W Bruce Croft. 2008. Retrieval Models for Question and Answer Archives. In *ACM International Conference on Research and Development in Information Retrieval*, 475–482.

[12] Yun Zhang, David Lo, Xin Xia, and Jian Ling Sun. 2015. Multi-Factor Duplicate Question Detection in Stack Overflow. *J. Comput. Sci. Technol.* 30, 5 (September 2015), 981–997.

[13] Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, and Ming Zhou. 2014. Question retrieval with high quality answers in community question answering. In *CIKM 2014 - Proceedings of the 2014 ACM International Conference on Information and Knowledge Management*, Association for Computing Machinery, Inc, 371–380. DOI:https://doi.org/10.1145/2661829.2661908

[14] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, PA, USA, 632–642.

[15] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Sen Wang, and Chengqi Zhang. 2018. Reinforced self-attention network: A hybrid of hard and soft attention for sequence modeling. In *IJCAI International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence, 4345–4352.

[16] Di Liang, Fubao Zhang, Weidong Zhang, Qi Zhang, Jinlan Fu, Minlong Peng, Tao Gui, and Xuanjing Huang. 2019. Adaptive multi-attention network incorporating answer information for duplicate question detection. In *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, Inc, 95–104.

[17] Dasha Bogdanova, Cicero dos Santos, Luciano Barbosa, and Bianca Zadrozny. 2015. Detecting Semantically Equivalent Questions in Online User Forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Stroudsburg, PA, USA, 123–131.

[18] Arpita Das, Harish Yenala, Manoj Chinnakotla, and Manish Shrivastava. 2016. Together we stand: Siamese networks for similar question retrieval. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, Association for Computational Linguistics (ACL), 378–387.

[19] Yushi Homma, Stuart Sy, and Christopher Yeh. *Detecting Duplicate Questions with Deep Learning*. Retrieved March 30, 2021 from https://zhiguowang.github.io.