The BQ Corpus: A Large-scale Domain-specific Chinese Corpus For Sentence Semantic Equivalence Identification

Jing Chen[†], **Qingcai Chen**[#*], **Xin Liu**[†], **Haijun Yang**[‡], **Daohe Lu**[‡], **Buzhou Tang**[†]
[†#]Shenzhen Calligraphy Digital Simulation Technology Lab,
Harbin Institute of Technology, Shenzhen, China
[‡]WeBank Inc.
[†]{mcdh.chenjing,hit.liuxin,tangbuzhou}@gmail.com
[#]qingcai.chen@hit.edu.cn
[‡]{navyyang,leslielu}@webank.com

Abstract

This paper introduces the Bank Question (BQ) corpus, a Chinese corpus for sentence semantic equivalence identification (SSEI). The BQ corpus contains 120,000 question pairs from 1-year online bank custom service logs. To efficiently process and annotate questions from such a large scale of logs, this paper proposes a clustering based annotation method to achieve questions with the same intent. First, the deduplicated questions with the same answer are clustered into stacks by the Word Mover's Distance (WMD) based Affinity Propagation (AP) algorithm. Then, the annotators are asked to assign the clustered questions into different intent categories. Finally, the positive and negative question pairs for SSEI are selected in the same intent category and between different intent categories respectively. We also present six SSEI benchmark performance on our corpus, including state-of-the-art algorithms. As the largest manually annotated public Chinese SSEI corpus in the bank domain, the BQ corpus is not only useful for Chinese question semantic matching research, but also a significant resource for cross-lingual and cross-domain SSEI research. The corpus is available in public[1].

1 Introduction

As the semantic matching task, sentence semantic equivalence identification (SSEI) is a fundamental task of natural language processing (NLP) in question answering (QA), automatic customer service and chat-bots. In customer service systems, two questions are defined as semantically equivalent if they convey the same intent or they could be answered by the same answer. Because of rich expressions in natural languages, SSEI is really a challenging NLP task.

Compared with other NLP tasks, the lack of large-scale SSEI corpora is one of the biggest obstacles for SSEI algorithm development. To address this issue, several corpora have been provided in recent years, including the Microsoft Research Paraphrase (MSRP) Corpus (Dolan et al., 2004; Dolan and Brockett, 2005), the Twitter Paraphrase Corpus (PIT-2015 corpus) (Xu et al., 2014, 2015), the Twitter URL corpus (Lan et al., 2017) and the Quora dataset [2].

In the early stage, the MSRP corpus was used to validate paraphrase identification algorithms based on a set of linguistic features (Kozareva and Montoyo, 2006; Mihalcea et al., 2006; Rus et al., 2008). Then, MSRP was also used to validate the deep models within a long duration. The deep convolutional neural networks (DCNNs), recurrent neural networks (RNNs), and their variants, such as Arc-I, Arc-II and BiMPM etc., have been developed and verified on it, even though it contains only thousands of sentence pairs (Hu et al., 2014; Yin and Schütze, 2015; Wang et al., 2016, 2017). Until 2015, the SemEval 2015 released a larger corpus, the PIT-2015 corpus for paraphrase and semantic similarity identification tasks. On this corpus, participants adopted SVM classifiers, logistic regression models, referential translation machines (RTM) and neural networks (Xu et al., 2015). In 2017, a large-scale SSEI corpus named Quora was released, which greatly boost the development of deep matching algorithms. Tomar et al. (2017) proposed a variant of the decomposable attention model. Gong et al.(2018) proposed a Densely Interactive Inference Network (DIIN) by hierarchically extracting semantic features from interaction space. However, the Quora corpus comes from social network sites. Consider-

*Corresponding author
[1]http://icrc.hitsz.edu.cn/Article/show/175.html
[2]https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs
[3]57,037 pairs out of them are manually labeled.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4946–4951
Brussels, Belgium, October 31 - November 4, 2018.

Corpus	Language	Source	Scale	Sentence Length	pos:neg
MSRP	English	news	5801 sentence pairs	18.9 words	2.05:1
PIT-2015 corpus	English	tweets	18,762 sentence pairs	11.9 words	-
The Twitter URL corpus	English	tweets	676,050 sentence pairs[3]	15 words	1:4.93
The Quora dataset	English	Quora	404,290 question pairs	11.1 words	1:1.71
The Bank Question corpus	Chinese	bank	120,000 question pairs	11.9 words	1:1

Table 1: The comparison of public corpus related to paraphrase or semantic similarity.

ing the request for specific domains and real context in SSEI, there still lacks corpora from different domains and corpora with features of non-English languages.

In this paper, we present a large-scale Chinese SSEI corpus constructed from real bank customer service logs. The main contributions of this paper include: 1) we present a large-scale domain-specific Chinese SSEI corpus, which contains 120,000 manually annotated sentence pairs; 2)we propose the Affinity Propagation (AP) (Frey and Dueck, 2007) clustering based method for SSEI corpus construction from a large number of sentences; 3)we provide the benchmark performance of 5 representative algorithms on our corpus. Hopefully, these contributions are useful in promoting the research on Chinese SSEI methods and the transferring methods for cross languages or cross domains.

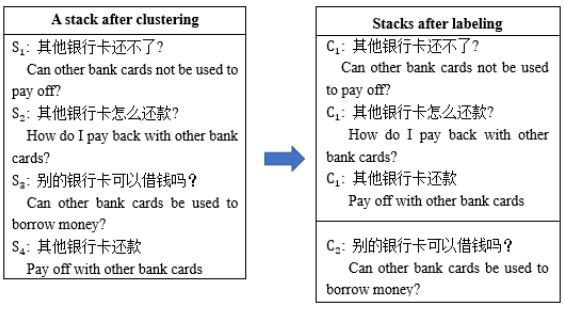


Figure 1: The examples from clustering to labeling

2 The Bank Question Corpus

2.1 Features of the Bank Corpus

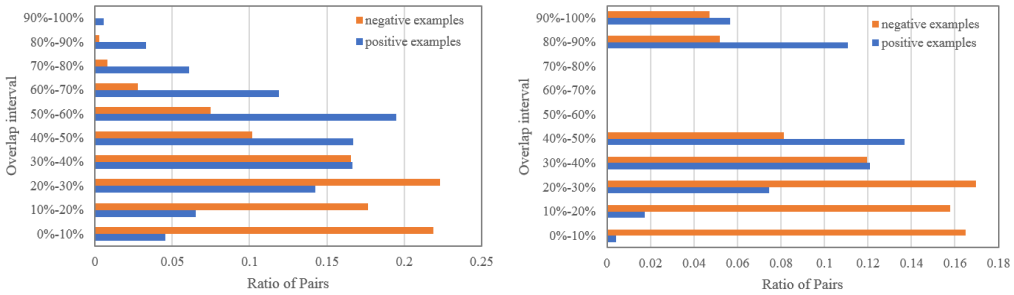
As the first domain-specific large-scale Chinese SSEI corpus, the Bank Corpus contains 120,000 question pairs. It is split into three parts: 100,000 pairs for training, 10,000 pairs for validation, and 10,000 pairs for test. There is no sentence overlap among training, validation and test sets. The last line in Table 1 shows the main features of our corpus. We also highlight features of the most popular SSEI corpus from line 1 to 4 in Table 1. The further analysis of the Bank Corpus will be shown in following sections.

2.2 Construction of the Bank Corpus

The original data came from the 1-year customer service logs with more than 20 millions of questions provided by a Chinese bank. To manually annotate so many questions is unimaginable, so we conducted three steps to get the SSEI corpus, including the clustering of questions, the intent-based annotation and the combination of semantic equivalent question pairs.

Grouping and Clustering At first, two definitions are given: a set of questions replied by the same answer is called a *group*; The clusters generated by an automatic clustering algorithm in a group is called a *stack*. Here, a stack is a subset of a group with questions have the same intent. First, the users' questions were divided into *groups* by their respective answers. The de-duplication is then executed on each group. Next, we used the Word Mover's Distance (WMD) (Kusner et al., 2015) based Affinity Propagation (AP) clustering algorithm to split the questions within each group into multiple question *stacks*. After filtering some emoijs and sentences which are standard answers from the questions, we finally got 799 distinct groups and selected total of 55724 questions from all groups for annotation.

Annotation We adopted two steps to annotate the question stacks. First, we recruited 12 annotators to categorize questions in each clustered stack into different intent classes. Here, if the questions express the same intent, we think that they belong to an intent class. For each labeled stack, the classes of intent are the same. If a question is chit-chat or it can not be combined with other questions into an intent class, it will be put into a specific class called "other". Second, the experts related to this specific domain were requested to check and correct the annotated intent classes. After annotation, we got 953 groups and 18002 questions. Among the questions, 16680



(a) The overlap distribution of our corpus (b) The overlap distribution of the Quora corpus

Figure 2: The overlap distribution of positive and negative data

	Examples		
most common:	一次性还完可以继续借吗 Can I continue to borrow money when I repay the loan at one time	一笔借款提前还清还可以借款吗 Can I borrow money again when a loan be pay off ahead of time	1
low overlap:	这是什么公司的产品 What's the company of the product	企业名称 the enterprise name	1
most common:	我几号还款 When should I repay my loan	今天的还款，可以推迟么? Can I postpone today's repayment?	0
high overlap:	能不能取消申请 Can I cancel the application	能不能取消，重新申请 Can I cancel the application and reapply	0

Table 2: The most common examples, the positive example with low overlap and the negative example with high overlap. Chinese sentences are original-form examples and English sentences below them are their corresponding translations.

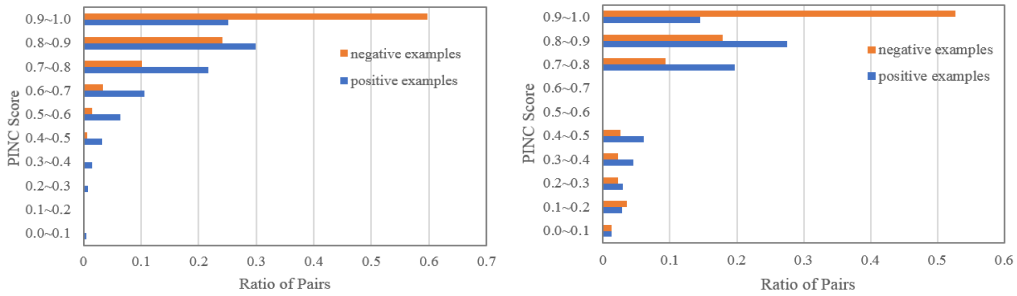
questions are with meaningful intents and can be used to create semantic pairs. There are average 9 stacks in each group. The annotation process from clustering examples to labeled examples is shown as Figure 1. From the clustering results, we find the clustering algorithm cluster the word "借(borrow)" and the word "还(pay back)" together. Actually, they convey different intentions and we need to distinguish them.

Generation Based on the labeled stacks, we combine the questions in each stack which have the same intention to create the positive question pairs, and select questions from different stacks in each group which have different intentions to create the negative question pairs.

2.3 Quality of the Corpus

To verify the quality of the corpus, we analyze the word overlap (Dolan et al., 2004) distribution and the PINC (Paraphrase In N-gram Changes) (Kim, 2014) distribution in the positive pairs and the negative pairs respectively.

The overlap is defined as the number of common words between two sentences divided by the average length of them. As shown in Figure 2, the overlap ratio of positive samples on the intervals appears a normal distribution on our new corpus, while the Quora corpus has no examples on overlaps between 50% and 80%. The positive question pairs with overlap ratio below 50% account for 58.67% and the negative question pairs with overlap ratio above 50% account for 11.36%. Here, we just give some examples with the largest ratio among the overlap intervals, some positive examples with low overlap and some negative examples with high overlap as shown in Table 2. For example, the positive question pair ("这是什么公司的产品(What's the company of the product)" and "企业名称(the enterprise name)") expresses the same intention while they have low overlap, where "什么公司(what company)" has the same meaning as "企业名称(the enterprise name)". The negative pair("能不能取消申请(Can I cancel the application)" and "能不能取消，重新申请(Can I cancel the application and reapply)") is different only on the word "重新(again)" while they nearly convey the contrary meaning. The statistics and examples indicate that except common examples we also have some difficult examples especially the positive pairs with low overlap and our new corpus is meaningful for research on learning methods.



(a) The PINC distribution of our corpus

(b) The PINC distribution of the Quora corpus

Figure 3: The PINC score distribution of positive and negative data

PINC (Kim, 2014) is a score of n-gram differences to measure lexical dissimilarity of sentence pairs. In essence, it is the inverse of BLEU (Papineni et al., 2002). As shown in Figure 3, most of the PINC scores on our corpus are between 0.7 and 1.0 which reveals that our corpus contains more lexical dissimilar question pairs. It contains rich expressions for the same user intention and it is challenging for machine learning methods to identify the semantic equivalence of the question pairs automatically.

3 Semantic Equivalence Identification

For this new public corpus, we provide a benchmark on the question semantic equivalence identification task to better understand its characteristic and provide further evidence for its value.

3.1 Models

Text-CNN (Kim, 2014) is a typical Convolutional Neural Network (CNN) model for sentence classification. We respectively feed each sentence of the question pair into the model with 300-dimensional word vectors and concatenate the sentence representation for SSEI. Here, we trained the word vectors on our new corpus by gensim[4].

BiLSTM (Graves and Schmidhuber, 2005) is an variant of RNN which considers both long and short dependency in context from forward and backward. We use the same structure but constitute the CNN with BiLSTM to model the sentence representation.

BiMPM (Wang et al., 2017) is a bilateral multi-perspective matching model of well performance for natural language sentence matching. The model uses the BiLSTM to learn the sentence representation, matches two sentences from two directions and multi-perspectives, aggregates the matching results with BiLSTM and finally predicts through a fully connected layer.

DIIN (Gong et al., 2018) is a Densely Interactive Inference Network (DIIN) for Natural Language Inference (NLI). It hierarchically extracts semantic features from interaction space to achieve the high-level understanding of sentence pairs. It achieves the state-of-the-art performance on large-scale NLI copora and Quora corpus.

3.2 Results and Discussion

The benchmark performance on our new corpus is shown in Table 3. The performace on the Quora corpus is shown in Table 4. The random method achieves 50.43% which indicates that our new corpus is balanced and meets the basic requirements for SSEI model research.

The TF-IDF method just models the surface features of sentences according to the vocabulary frequency. It can not learn the dependency features in the word sequences and the synonym or near-synonym according to the word meanings. Therefore, it performs not so well, which indicates that the new corpus can not be learned by simple surface features and the deep semantic relationsn need to be mined by deep models.

Here, we use four deep neural network models to verify the new constructed corpus, including two basic and representative models (Text-CNN and BiLSTM) and two latest and well-used models (BiMPM and DIIN) which perform well on the natural language sentence matching task. The results show that the BiLSTM model can learn the dependency features between words in the sentences better than the Text-CNN model. The Accuracy of BiMPM is 81.85% and that of DIIN is 81.41%. Compared with the performance on the Quora corpus, the performance on the BQ corpus

[4]https://radimrehurek.com/gensim

Models	Precision	Recall	F1	Accuracy
Random	50.43	50.56	50.49	50.43
TF-IDF	64.68	60.94	62.75	63.83
Text-CNN	67.77	70.64	69.17	68.52
BiLSTM	75.04	70.46	72.68	73.51
BiMPM	**82.28**	**81.18**	**81.73**	**81.85**
DIIN	81.58	81.14	81.36	81.41

Table 3: The comparison of BQ corpus related to paraphrase or semantic similarity.

Models	Precision	Recall	F1	Accuracy
Random	50.73	51.84	51.29	50.72
TF-IDF	63.66	85.16	72.85	68.24
Text-CNN	82.89	71.58	76.82	78.38
BiLSTM	83.44	73.96	78.41	79.62
BiMPM (Wang et al., 2017)	—	—	—	88.17
DIIN (Gong et al., 2018)	—	—	—	**89.06**

Table 4: The comparison of Quora corpus related to paraphrase or semantic similarity.

is lower, which reveals that our new corpus is challenging for semantic matching model research.

4 Conclusion and Future Work

In this paper, we present a large-scale Chinese corpus for question semantic equivalence identification in the bank domain. The construction procedure and benchmark performance are given. To the best of our knowledge, this corpus is the largest manually annotated public Chinese SSEI corpus in the bank domain. Compared with existing corpora, it is of high quality and challenging, and is hopefully useful for research on SSEI, cross-lingual and cross-domain learning.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. This work is supported by Natural Science Foundation of China (Grant No. 61473101, 61573118), Strategic Emerging Industry Development Special Funds of Shenzhen(Grant No. JCYJ20170307150528934, JCYJ20160531192358466) and the joint project foundation of WeBank.

References

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, page 350. Association for Computational Linguistics.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315(5814):972–976.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. *ICLR*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems*, pages 2042–2050.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

Zornitsa Kozareva and Andrés Montoyo. 2006. Paraphrase identification on the basis of supervised machine learning techniques. In *Advances in Natural Language Processing, International Conference on Nlp, Fintal 2006, Turku, Finland, August 23-25, 2006, Proceedings*, pages 524–533.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *International Conference on International Conference on Machine Learning*, pages 957–966.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of The 2017 Conference on*

Empirical Methods on Natural Language Processing (EMNLP), pages 1235–1245. Association for Computational Linguistics.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, Usa*, pages 775–780.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Vasile Rus, Philip M. Mccarthy, Mihai C. Lintean, Danielle S. Mcnamara, and Arthur C. Graesser. 2008. Paraphrase identification with lexico-syntactic graph subsumption. In *International Florida Artificial Intelligence Research Society Conference, May 15-17, 2008, Coconut Grove, Florida, Usa*, pages 201–206.

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, Copenhagen, Denmark, September 7, 2017*, pages 142–147.

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 2017, Melbourne, Australia, August 19-25, 2017*, pages 4144–4150.

Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. 2016. Sentence similarity learning by lexical decomposition and composition. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 1340–1349.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *International Workshop on Semantic Evaluation*, pages 1–11.

Wei Xu, Alan Ritter, Chris Callisonburch, William B Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from twitter. pages 435–448.

Wenpeng Yin and Hinrich Schütze. 2015. Convolutional neural network for paraphrase identification. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 901–911.