A large-scale Chinese Nature language inference and Semantic similarity calculation Dataset

Abstract. Natural language inference (NLI) and semantic similarity calculation are basic research tasks in the field of natural language processing (NLP). In recent years, NLP technology based on deep learning has achieved great success. On account of the complex model structure of deep neural network, a large amount of training data is needed to avoid overfitting. Aiming at the limited scale of Chinese datasets related to this two tasks, a Chinese Natural language inference and Sentence similarity calculation Dataset (CNSD) is constructed in this paper. CNSD comes from four datasets with different characteristics and contains 2,195,000 sentence pairs. CNSD is the first Chinese dataset of millions in the field of NLI and semantic similarity calculation. In this paper, the deep neural network model BERT is applied to the NLP task based on this dataset, and the obtained results are taken as the baseline of it. This baseline result will provide reference for future NLP research based on CNSD. CNSD will be available for download by researchers to contribute to Chinese NLP researches.

Keywords: Natural language inference, Sentence similarity, Chinese dataset, BERT for classification.

1 Introduction

Natural language inference (NLI) is one of the four tasks of natural language processing (NLP) and belongs to sentence relation judgment [1]. NLI uses models to determine whether the sentence pair (premise, hypothesis) has entailment, contradiction or neutral relation. Semantic similarity calculation is a measure of finding resemblance between texts [2]. There are increasing deep learning models provided with certain semantic learning ability. NLI and semantic similarity calculation both need to obtain low-level syntactic information and high-level information of text, then judge the relation between sentence pairs. The difference lies in that NLI is a one-way implication relation, while semantic similarity calculation is a two-way similar relation. But implication relation can be converted into similar relation. Entailment in NLI is equivalent to label 1 in semantic similarity calculation, i.e., the meaning of two sentences is similar. This alleviates the problem of lack of annotated data in semantic similarity calculation to some extent. It has been applied to semantic similarity calculation with NLI datasets[1]. It should be noted that similarity relation cannot be converted into implication relation. As we know, data is the basis of research. The community of NLP has been dedicating to preserve and standardize existing information in the world. Obviously, it contributes to enhance the robustness and accuracy of model with more real data. However, we find that existing large-scale datasets are mainly written in English. Chinese datasets

[1] https://github.com/dhwajraj/deep-siamese-text-similarity

are rare relatively. There exits Chinese natural language reasoning sentence pairs published by Yu Dong in 2018 CCL assessment task[2], Chinese semantic similarity calculation annotation data provided by Ant financial[3], and Chinese problem matching corpus released by Liu, Xin et al. [3].
In this paper, aiming at the problem of small scale of open NLI dataset for Chinese, we construct CNSD, a **C**hinese **N**atural language inference and **S**emantic similarity calculation **D**ataset. It contains 2,191,500 sentence pairs and its corresponding relation. It is the largest open NLI dataset in Chinese among academic community in present. CNSD is derived from four English datasets including SNLI, MNLI, etc. We will introduce more details in section 3. To ensure accuracy, this paper adopts machine translation and manual proofreading. CNSD can not only be used for natural language reasoning and semantic calculation, but also can be used as an extended application in related fields of NLP, such as paragraph extraction and paragraph ordering in Machine Reading Comprehension, and redundancy relation elimination in Knowledge Map fusion.

2 Related work

As important branches of NLP, many Chinese and English related datasets have produced in NLI and semantic similarity calculation. The following section will introduce some of them in chronological order of publication and compare them.

SICK. Sentences Involving Compositional Knowledge is extracted from video and image description [4]. There are about 10k sentence pairs in total. Sentence description in SICK contains text entailment relation, which is similar to SNLI and MNLI, and semantic relation degree mark scores, same to STS-B. In addition, SICK marks pair A-B, as well as pair B-A, respectively. It works uniformly as CNSD exchange premise and hypothesis manually for the data of missing label.

SNLI. The Stanford Natural Language Inference Corpus is the first large manual annotated dataset, which contains 570k pairs of manually written English sentences [5]. The data format is (Text, Hypothesis, Judgments). SNLI is a milestone of NLI datasets.

MNLI. The Multi-Genre Natural Language Inference corpus has the same format as SNLI and contains 433k sentence pairs annotated with textual entailment information. Data in MNLI comes from fiction, telephone, travel, government, slate and so on [6]. Thus it is closer to real life. Compared with SNLI, MNLI is more diverse, more complex and richer in sentence patterns.

[2] https://github.com/blcunlp/CNLI
[3] https://dc.cloud.alipay.com/index#/topic/intro?id=8

QQP. Data in Quora Question Pairs is derived from Quora's real questions. Among them, marked training set is 404k, and unmarked test set is 2.3m [7]. QQP gathers different question expressions to determine whether the semantics expressed by two questions are equal, using 0 or 1 to label them.

STS-B. Semantic Textual Similarity Benchmark covers image and news title as well as text from BBS with 8.5k sentence pairs [8]. STS-B rates sentence pairs on a scale of 1 to 5 based on similarity. The higher the score, the more similar the sentence is.

CNLI. As one of the evaluation tasks of China National Conference on Computational Linguistics 2018, Yu et al. published 8-100k sentence pairs for the task evaluation of Chinese Text Entailment. Similar to our work, data in CNLI come from SNLI and MNLI, but the scale is only one-tenth of ours.

LCQMC. Liu, Xin et al. published manually annotated Chinese problem matching dataset in 2018, with a scale of 260k. They use 0 or 1 to label whether questions are matched, the same way to QQP.

There are 100k pairs of data provided by Ant Financial Problem Similarity Calculation contest 2018. It is from the practical application scenarios of Financial Brain in Ant Financial. The rematch phase of this competition provides massive scale data, but can not be available for public download.

Table 1. Compares between Datasets

Dataset	Time	Language	Source	Task
SICK	2014	English	Description of video and image	NLI & Semantic similarity calculation (SSA)
SNLI	2015	English	Manually generated	NLI
MNLI	2017	English	Spoken and written words	NLI
QQP	2017	English	Questions from users	SSA
STS-B	2017	English	Title and BBS text	SSA
CNLI	2018	Chinese	Machine translation & manual proofreading	NLI
LCQMC	2018	Chinese	Manually annotated	SSA
NLP Contest	2018	Chinese	The Financial Brain	SSA

3 CNSD

3.1 Data Resource

In this paper, SNLI, MNLI, QQP and STS-B are selected as data sources from many English NLI datasets. QQP is divided into training set with annotation and test set

without annotation. We select all data with annotation and some of it with none-annotation. SNLI and MNLI are NLI datasets, while QQP and STS-B are semantic similarity calculation datasets. CNSD integrates this two types. The four original datasets involve the most basic dataset in English, datasets close to real-life scenario and freely expression datasets, etc. It covers various aspects which might involve in.

3.2 Data Process

We tested multiple translation interfaces and compared their accuracy, fluency and vocabulary richness of the translation results. Finally, we decide to use Tencent cloud API[4] as the main machine translation interface.

There are some problems in machine translation, such as literal translation and cannot exactly distinct fuzzy boundary. To improve the smoothness of the translation results, we manually modified wrong words and corrected some inaccurate sentences. The rectified results are more consistent with the Chinese expression habits. And we list some of them in Table 2.

Table 2. Examples of manually rectified data.

Original English sentence	Translation interface results	Rectified results
Dog in pool.	池中狗。	狗在游泳。
A woman is doing a cartwheel.	一个女人在做手推车。	一个女人在做侧手翻。
Island native fishermen reeling in their nets after a long day's work.	在漫长的一天工作之后，岛上的本地渔民们在渔网里打滚。	岛上的土著渔民在一天的工作后，把网收了起来。

Table 3. Examples of new sentence pairs

Premise	Hypothesis	Original label	New label
一对夫妇在外面的一张桌子上吃饭，他指着什么东西。	一对夫妇在一张桌子上吃东西，他指着什么东西。	contradiction	entailment
身穿粉红色头巾的中东妇女正走在一个戴着紫色头巾的妇女旁边。	两个女人在一起走路。	-	entailment
两位医生为病人做手术。	两名医生正在对一名男子进行手术。	neural	
两名医生正在对一名男子进行手术。	两位医生为病人做手术。		entailment

We delete hardly used attributes in the original dataset and only retain label that can be used for semantic similarity judgment and contains text entailment, simplifying the data

[4] https://cloud.tencent.com

format. We correct and supplement some false and missing label in original dataset. We also exchange premise and hypothesis to generate new correct inference relation, which is similar to entailment for A-B order and entailment for B-A order in SICK. The new sentence pairs are shown in Table 3.

3.3 Data Presentation

CNSD comes from four different English natural language processing datasets, including a total of 2,191,500 sentence pairs. Each data consists of a triple: premise, hypothesis, and sentence label. We list some of them in Table 4. And the dataset size is shown in Table 5.

Table 4. Examples of Chinese Natural Language Inference.

Premise	Hypothesis	Label
一只长着长毛的白色狗跳跃着去抓一个红绿相间的玩具。	一只动物跳跃着去捕捉一个物体。	entailment
身穿黑色西服、白色衬衫和黑色保龄球的男子在演奏乐器，周围环绕着他的其他交响乐。	没有人有西装。	contradiction
一名身穿白色t恤的男子在街道中央拍了一张照片，背景是两辆公共汽车。	一名男子在纽约旅游时拍照。	neutral

Table 5. Scale of Chinese Natural Language Inference.

	Train	Dev	Test	Sum
Chinese-SNLI	550k	10k	10k	570k
Chinese-MNLI	390k	12k	13k	415k
Chinese-QQP	390k	8k	800k (without label)	1.1m
Chinese-STS-B	5.7k	1.5k	1.3k	8.5k
Total	1.3m	31.5k	824.3k	2.1m

3.4 Potential application in CNSD

Multi-document Machine Reading Comprehension needs to sort candidate documents and find paragraphs most likely to own correct answers, then extract answers [9]. Using the similarity of questions and sentences in paragraphs can reduce the interference in answer extraction process.

When QA task use traditional information retrieval methods to match existing questions with user questions, it misses answers whose words own the same meaning but with different vocabularies [10]. Obtaining semantic similarity of two problems can solve this problems softly.

In the process of Knowledge Graph, redundant relationships need to be deleted [11]. Besides the relationships with consistent expression, that with consistent expression need to be deleted as well. QA based on Knowledge Graph needs to identify the relation between entities according to semantics so as to answer questions.
Scoring tasks for students' homework require comparing students' answers with standard answers to give text entailments [12]. It can also judge the correct degree of students' answer based on similarity.

4 Benchmark Result

Jacob Devlin, etc. in 2018 proposed **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (we will use BERT when mention it later), which has got state-of-the-art in 11 NLP tasks [13]. We compare the results of BERT with those of other models. BERT achieved the best performance and trained faster. Finally, BERT-base is selected as the benchmark system of this dataset.

4.1 Model

BERT can be applied to four subtasks, namely Sentence Pair Classification Tasks, Single Sentence Classification Tasks, Question Answering Tasks and Single Sentence Tagging Tasks. Considering about the feature of our dataset, we decide to use BERT for Sentence Pair Classification Tasks. The following section describes specific use of the benchmark model.

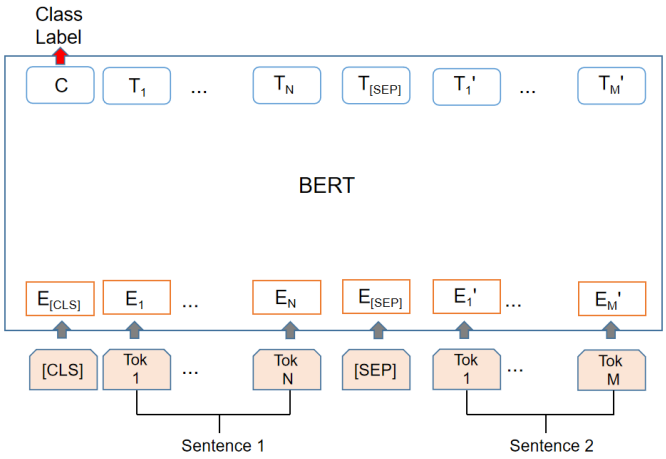

Fig. 1. BERT for Sentence Pair Classification.

As is shown in Figure 1. We use [CLS,$token_i^1$,...$token_N^1$,SEP, $token_i^2$,...$token_M^2$] as input to the model. CLS is a representation of classification output. SEP is a

segmentation representation of non-coherent word sequences. $token_i^1$ and $token_i^2$ mark the i-th word in sentence 1 and 2 respectively.

$$Input = [CLS,S_1,SEP,S_2] \quad (1)$$

$$V^{CLS} = BERT(input) \quad (2)$$

$$Y^{FC} = W \cdot V^{CLS} + b \quad (3)$$

$$P_i = softmax(Y^{FC}) \quad (4)$$

V^{CLS} is the vectorization representation of classification label CLS through BERT. It calculates the possibility to each class by fully-connected layer and softmax layer, then take the maximum probability as the classification result of the final prediction.

4.2 Setup

In our experiment, we use Adam as optimizer. We use the same hyper parameter with BERT-base and other parameters set specially describe in Table 6. And we choose cross entropy loss function as loss function following Equation 5. The accuracy of prediction is used as indicator of evaluation.

Table 6. Hyper parameters in Benchmark model.

Parameter	Value
Batch size	40
Max sequence length	80
Train_epochs	7
Num_attention_heads	12
Num_hidden_layers	12

$$L = -[y \cdot log\tilde{y} + (1-y) \cdot log(1-\tilde{y})] \quad (5)$$

4.3 Result

Table 7. Experiment results.

Model	Chinese-SNLI		Chinese-MNLI		Chinese-QQP	Chinese-STS-B	
	Dev	Test	Mis-matched	Matched	Dev	Dev	Test
Embed+add-attention	74.46	75.05	63.28	62.25	72.56	-	-
BiLSTM+self-attention	81.19	80.96	69.47	67.79	81.45	43.87	41.24
DiSAN	81.32	81.45	69.54	68.13	82.32	44.21	42.09
BERT	**87.39**	**86.95**	**79.76**	**79.39**	**89.08***	**53.84**	**50.26**

We experiment DiSAN [14], Bi-LSTM [15] and BERT models in this dataset respectively. The experiment results are recorded and compared in Table 7 and Figure 2. When models require word vectors, we apply the 300d Chinese word vectors published by Shen Li, a total of 600k words [16].

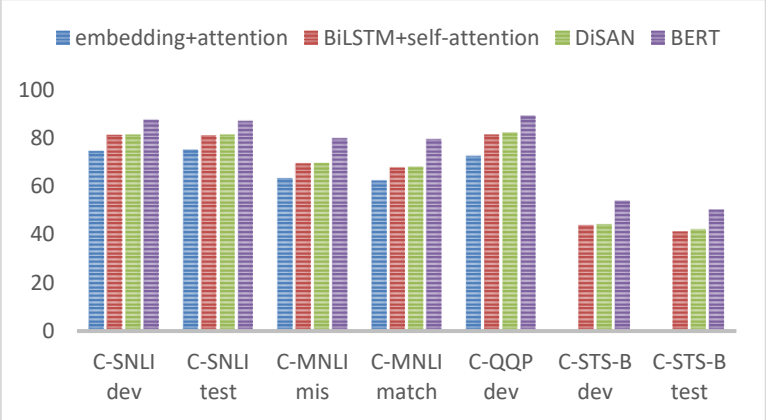

Fig. 2 This figure describes four models used in CNSD, and their separate accuracy.

The experiment shows that, compared with the previous model, BERT has achieved excellent results in all datasets. BRET is used as the baseline model for this dataset because it can be fine-tuned for tasks quickly and well. It also proves the validity of BERT for Chinese pretraining model in Chinese NLI tasks. All experiments are conducted on TITAN Xp.

5 Conclusion

From selection of dataset sources, to translation between Chinese and English, to application of specific models, we reserve the original data format and modify some corpus as well. Although the process is relatively tedious, the experimental results show that the system has great perform and can be effectively applied to the tasks of NLI and semantic similarity calculation. The highest accuracy of BERT-base in Chinese reaches 89.08%, which proves that the quality of corpus improved under the dual action of machine translation and manual modification. It also proves that the size and accuracy of data play a vital role in promoting NLP researches. We hope this dataset will be helpful for Chinese NLP tasks and other related artificial intelligence tasks. We also know what we have done is not far enough. We will do our best to improve CNSD and collect more suitable resources in the future.

6 Acknowledgements

Thanks to lab students for participating in the translation errata, making Chinese sentences more accurate and fluent. Thanks to Tencent cloud for providing computing services.

References

1. Dagan, Ido and Oren Glickman. “PROBABILISTIC TEXTUAL ENTAILMENT: GENERIC APPLIED MODELING OF LANGUAGE VARIABILITY.” (2004).
2. Wu, Hao et al. “BIT at SemEval-2017 Task 1: Using Semantic Information Space to Evaluate Semantic Textual Similarity.” SemEval@ACL (2017).
3. Liu, Xin et al. “LCQMC: A Large-scale Chinese Question Matching Corpus.” COLING (2018).
4. Socher, Richard et al. “Grounded Compositional Semantics for Finding and Describing Images with Sentences.” Transactions of the Association for Computational Linguistics 2 (2014): 207-218.
5. Bowman, Samuel R. et al. “A large annotated corpus for learning natural language inference.” EMNLP (2015).
6. Williams, Adina et al. “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference.” NAACL-HLT (2018).
7. Shankar, Shashi Kant. “Identifying Quora question pairs having the same intent.” (2017).
8. Holmes, David Richard et al. “2012 ACCF/AATS/SCAI/STS expert consensus document on transcatheter aortic valve replacement.” Journal of the American College of Cardiology 59 13 (2012): 1200-54 .
9. Chen, Danqi et al. “Reading Wikipedia to Answer Open-Domain Questions.” ACL (2017).
10. Cui, Wanyun et al. “KBQA: An Online Template Based Question Answering System over Freebase.” IJCAI (2016).
11 Lin, Yankai et al. “Learning Entity and Relation Embeddings for Knowledge Graph Completion.” AAAI (2015).
12. Dzikovska, Myroslava O. et al. “SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge.” SemEval@NAACL-HLT (2013).
13. Devlin, Jacob et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” CoRR abs/1810.04805 (2018): n. pag.
14. Shen, Tao et al. “DiSAN: Directional Self-Attention Network for RNN/CNN-Free Language Understanding.” AAAI (2018).
15. Wang, Shuohang and Jing Jiang. “Learning Natural Language Inference with LSTM.” HLT-NAACL (2016).
16. Li, Shen et al. “Analogical Reasoning on Chinese Morphological and Semantic Relations.” ACL (2018).