PKU Paraphrase Bank: A Sentence-Level Paraphrase Corpus for Chinese

Bowei Zhang[1,2,4], Weiwei Sun[1,2,3], Xiaojun Wan[1,2], and Zongming Guo[1]

[1] Institute of Computer Science and Technology, Peking University
[2] The MOE Key Laboratory of Computational Linguistics, Peking University
[3] Center for Chinese Linguistics, Peking University
[4] Center for Data Science, Peking University
{bw_zhang,ws,wanxiaojun,guozongming}@pku.edu.cn

Abstract. One of the main challenges of conducting research on paraphrase is the lack of large-scale, high-quality corpus, which is particularly serious for non-English investigations. In this paper, we present a simple and effective unsupervised learning model that is able to automatically extract high-quality sentence-level paraphrases from multiple Chinese translations of the same source texts. By applying this new model, we obtain a large-scale paraphrase corpus, which contains 509,832 pairs of paraphrased sentences. The quality of this new corpus is manually examined. Our new model is language-independent, meaning that such paraphrase corpora for other languages can be built in the same way.

Keywords: Paraphrase · Paraphrase Extraction · Sentence Embedding · Sentence Similarity

1 Introduction

Paraphrases are linguistic expressions that restate the meaning using different expressions, sentences or phrases, which convey the same meaning using different wording [5]. Paraphrases have proven useful for a wide variety of Natural Language Processing applications, e.g., semantic parsing [4], knowledge based question answering [10], information extraction [30], paraphrase generation [1,17], machine translation [24] and many others.

In this paper, we are concerned with the data bottleneck problem in current paraphrase research—the lack of large-scale, high-quality sentence-level corpora. We present a simple and effective unsupervised method that combines the semantic representation of sentences in high-dimensional sparse spaces with the semantic representations in low-dimensional dense spaces to construct scoring functions that are employed to detect paraphrase candidates. We use this method to automatically extract high-quality paraphrases from multiple Chinese translations of the same source texts. In particular, we use the different Chinese translations of the same foreign novels. The diversity of linguistic expressions exhibited by parallel translations makes them a good source for collecting sentence-level paraphrases. By exploring the semantic and structural correspondence between two parallel translations, we are able to harvest a large set of sentence-level paraphrases.

We introduce a sentence-level paraphrase corpus for Chinese that contains 509,832 sentence pairs, the quality of which is strictly controlled and analyzed. All paraphrase pairs are ranked according to a semantic metric, so we can strike a balance between quantity and quality for different application purposes. Manual evaluation highlights the reliability of this resource: The overall accuracy of the whole set is 92%. When we select the top-60% sentences, this number goes up to 97%. In addition, we compare the three existing English sentence-level paraphrase corpora from the language styles and the diversity of expressions of sentence pairs, which proves that our corpus is not only enough, but the quality of paraphrase is good enough. To the best of our knowledge, this is the first large-scale sentence-level paraphrase corpus for Mandarin Chinese.

We have released[5] the newly created Chinese paraphrase data.

2 Related Work

There have been some studies on constructing high-quality paraphrase data sets. Barzilay and Mckeown extracted sentence pairs from multiple translations of the same material [3]. Lin and Pantel extracted paraphrase from a similar context using an unsupervised algorithm [16]. There are also many people who use multilingual news resources to get paraphrase data sets [8,9]. Both statistical and neural machine translation methods [21,26,27] have been applied to obtain a paraphrase corpus. In addition, some paraphrase data sets are constructed via crowdsourcing platforms [13] or matching the URLs of tweets [14].

Datasets consisting of paraphrases of different granularities have been introduced:

- The sentence-level paraphrase corpora available include: MSR Paraphrase Corpus [8,9] that contains 5,801 pairs of sentences extracted from parallel news corpus. And Twitter Paraphrase Corpus [28,29] that contains about 14,000 sentence pairs that are derived from Twitter's trending topic data. The latest Twitter News URL Corpus [14] contains 51,524 pairs of sentences.
- The phrase-level paraphrase corpora available includes: DIRT [15], PATTY [18], POLY [12] and Paraphrase Database (PPDB [2,11]).

The existing corpora are almost all based on English texts, and the sentence-level corpora are all of a modest scale.

3 Our Method

3.1 Paraphrase Extraction by Alignment

We explore multiple Chinese translations of the same source texts. There are multiple Chinese translations for a number of well-known books that are written in English or other western languages, e.g., *Oliver Twist* and *Gone with the wind*. The diversity of linguistic expressions exhibited by parallel translations makes them a good source for collecting sentence-level paraphrases. Because translations are usually conducted in a

[5] PKU Paraphrase Bank: https://github.com/pkucoli/PKU-Paraphrase-Bank

Fig. 1. A pair of two-sentence sequences as well as their literal translations. The red words are the key elements to distinguish sentences; the blue words have the same meaning but are different expressions, which interfere with the detection of paraphrase sentence pairs.

sentence-by-sentence way, the paraphrase extraction problem turns to be a text alignment problem: Given two sequences of sentences, say T^1 and T^2, the goal is to find the best sentence alignment.

Assume that T^1 and T^2 consist of N^1 and N^2 sentences respectively. We define an alignment matrix C as follows,

$$C_{ij} = \begin{cases} 1 \text{ if } T_i^1 \text{ and } T_j^2 \text{ match} \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

where T_i^1 and T_j^2 denote the i-th sentence in T^1 and the j-th sentence in T^2 respectively.

A score function, viz. SCORE is employed to evaluate the *goodness* of each candidate pair of aligned sentences. By coupling C and SCORE, we can transform the original alignment problem into a constrained optimization problem, defined as follows:

$$\begin{aligned} \max. \ & \sum_{i=1}^{N^1} \sum_{j=1}^{N^2} C_{ij} \times \text{SCORE}(T_i^1, T_j^2) \\ \text{s.t.} \ & \sum_{j=1}^{N^2} C_{ij} = 1 \text{ for all } 1 \le i \le N^1 \\ & \sum_{i=1}^{N^1} C_{ij} = 1 \text{ for all } 1 \le j \le N^2 \\ & C_{i_1 j_2} + C_{i_2 j_1} \le 1 \text{ for all } 1 \le i_1 < i_2 \le N^1 \\ & \qquad \text{and } 1 \le j_1 < j_2 \le N^2 \end{aligned} \tag{2}$$

It is reasonable to assume that paraphrased sentences between parallel texts uniquely exist. The constraints in (2) ensures that there exists one and only one sentence in T^2 that is aligned to a particular sentence in T^1. So is the case for sentences in T^2.

3.2 The Score Functions

We empirically study different strategies to design a good SCORE function and find that an effective SCORE function needs to consider the number of co-occurrences of words in high-dimensional space and the vector similarity of sentences embedding in low-dimensional space. Our SCORE function is defined by Eq. 6, the components of which is introduced as follows.

The Sparse Approach The key idea to measure the consistency of two candidate sentences is calculate how many words between them are the same. See Figure 1 for an example. Although two paraphrased sentences should use different wording, a small portion of words are inevitably shared and thus become an essential evidence to determine the paraphrase relation. Furthermore, different words should be treated differently.

It is very important to assign different weights to different words. Therefore, we emphasize the importance of low-frequency words and introduce a frequency-based weight for each word type, as shown in Eq. 3. In general, the common low-frequency nouns and verbs in T_i^1 and T_j^2 are strongly discriminative, while high-frequency auxiliary words or pronouns are not very effective. In some cases, different adjectives that express the same meaning bring some noises. Take the first sentence pair in Figure 1 for example: "公路/Road" and "田野/field" (marked in red) are two low-frequency nouns that are highly indicative.

$$W_k = \log(\frac{(N^1 + N^2)}{f_{w_k}}) \quad k \in (1, ..., M) \tag{3}$$

M denotes the total number of word types; f_{w_k} denotes the number of the occurrences of word type w_k in T^1 and T^2; W_k is the corresponding weight of w_k.

Denote sets of words for two candidate sentences T_i^1 and T_j^2 as Q_i^1 and Q_j^2 respectively. A intersection-based $S_{sp}(T_i^1, T_j^2)$ is defined as follows:

$$S_{sp}(T_i^1, T_j^2) = \frac{\sum_{t=1}^{|Q_i^1 \cap Q_j^2|} W_t}{2\sum_{t=1}^{|Q_i^1|} W_t} + \frac{\sum_{t=1}^{|Q_i^1 \cap Q_j^2|} W_t}{2\sum_{t=1}^{|Q_j^2|} W_t} \quad i \in (1, ..., N^1), j \in (1, ..., N^2) \tag{4}$$

The Dense Approach The $S_{sp}(T_i^1, T_j^2)$ function is in a rather sparse space. It by itself is able to efficiently identify the majority of pairs of paraphrases. However, this method can only identify semantics that are related to exactly the same words. In the first sentence pair in Figure 1, the two modifiers "直/straight" and "笔直地/straightly" (marked in blue) have the same meaning to some extent. However, due to the difference of their surface forms, they cannot contribute positively to the score function. In the second sentence pair, the coreference relation matters: "它/it" and "羊/the sheep" (marked in blue) refer to the same thing but such semantics is, again, ignored.

To deal with the above problems, we adopt a sentence embedding method to derive low-dimensional but dense representations of semantics at the sentence level. We use the Bert[6] [7] pre-training network. Specifically, we fixed the parameters of the Bert pre-trained network, and use the vectors $Bert(T_i^1)$ and $Bert(T_j^2)$ as the semantic representations of the sentences T_i^1 and T_j^2, which are automatically derived using Bert.

$$S_{de}(T_i^1, T_j^2) = \frac{Bert(T_i^1) \cdot Bert(T_j^2)}{\|Bert(T_i^1)\| \|Bert(T_j^2)\|} \quad i \in (1, ..., N^1), j \in (1, ..., N^2) \tag{5}$$

[6] https://github.com/google-research/bert/
https://github.com/huggingface/pytorch-pretrained-BERT/

Using the parameter λ to linearly blend the scoring function $S_{sp}(T_i^1, T_j^2)$ in the sparse space and the scoring function $S_{de}(T_i^1, T_j^2)$ in the dense space, we arrive at the following scoring function $S(T_i^1, T_j^2)$:

$$S(T_i^1, T_j^2) = \lambda S_{sp}(T_i^1, T_j^2) + (1-\lambda) S_{de}(T_i^1, T_j^2) \quad i \in (1, ..., N^1), j \in (1, ..., N^2) \tag{6}$$

λ is a hyperparameter that is tuned using a small size development data set. In our experiment, it is set to 0.8.

If we only consider a sentence pair, the effectiveness of the above metric is limited. However, it is worth noting that this *local* score function is only one component in our alignment-based solution and the structural information in Eq. 2 will significantly enhance such local alignments.

3.3 Solving the Optimization Problem

According to the underlying idea of our method which is illustrated in §3.1, it is obvious that an ideal extraction algorithm should take all candidate sentences into account. However with the increase of N^1 and N^2, the search space will expand rapidly, and it will be impractical to run such an algorithm. Furthermore, more noises are introduced and then harm the final extraction results. If we excessively restrict the search space for efficiency, we may miss the gold candidate.

In this paper, we employ a greedy search strategy to solve the optimization problem. We improve the extraction efficiency greatly from the following two aspects while ensuring the quality and quantity of extracted paraphrases.

Positional Relationship Since the sentences in a book-style text imply a sequential relationship, we dynamically determine the initial alignment range of each sentence pair (T_i^1, T_j^2). The positional relationship of all candidate sentence pairs in T^1 and T^2 is as shown in (7), where I_i^1 and I_j^2 are the position representations of T_i^1 and T_j^2, respectively. I_i^1 means that T_i^1 is the I_i^1-th sentence in T^1, and so is I_j^2 defined. L is a large constant that is a hyperparameter to ensure that the range for search is large enough.

$$\begin{aligned} -L < I_i^1 - I_j^2 < |N^1 - N^2| + L \quad N^1 \geq N^2 \\ -L < I_j^2 - I_i^1 < |N^1 - N^2| + L \quad N^1 < N^2 \end{aligned} \tag{7}$$

Fast Pruning In order to avoid missing potential paraphrase sentence pairs and to obtain as much data as possible, our initial alignment range is quite large. Therefore it takes a lot of time to evaluate the semantic similarity of possible candidate sentence pairs. Carefully observing the data, we find that most of the negative candidate pairs have a very low score. If sentence pairs with such low semantic relevance can be eliminated before calculating semantic similarity, the efficiency of the paraphrase extraction procedure will be greatly improved. We implement this idea using the inverted index. We remove the high-frequency words of a candidate sentence pair. If the intersection of remaining parts of any two sentences is empty, this candidate pair will be removed.

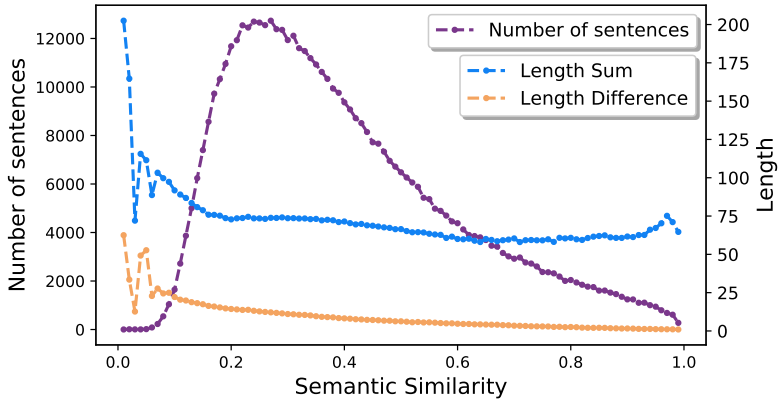

Fig. 2. Corresponding to the left ordinate axis, the purple line represents the number of sentence pairs relative to different semantic similarity. Corresponding to the right ordinate axis, the blue and orange lines represent the sum and the difference of the number of word tokens in the candidate sentence pairs.

4 Our Paraphrase Corpus

4.1 Preprocessing

We applied our paraphrase extraction method to a collection of Chinese translations of books written in English as well as other European languages. This collection contains a total of 95 translations of 40 novels from the Internet [7]. And before searching the best alignment, we conduct 4-step preprocessing:

PDF to Text Conversion We convert images from the original scanned PDF into recognizable texts. Though current state-of-the-art OCR technology is not perfect, the performance of recognizing printed texts is relatively satisfactory.

Data Cleaning Different versions of translations may not have the same headers, footers, page numbers, and annotations. We write some heuristic rules to remove them, reducing the noise for the next step.

Sentence Segmentation and Combination The translation habits of different translators are not completely consistent, which may lead to different sentence segmentation. We determine the sentence boundaries according to three Chinese punctuation markers, viz. '。', '？' and '！'. Very short sentences, which contain less than 6 Chinese characters, are then unified with the previous sentence.

Word Segmentation Since there are no explicit word boundaries in the writing system for Chinese, we need an automatic word segmentation system. In this work, we employ a supervised segmenter introduced in [25] to process raw texts.

[7] The supplementary note gives the detailed information about these books.

4.2 Quality and Quantity

Initially, we collect 707,274 candidate sentence pairs in total. We present an analysis in Figure 2. *Semantic Similarity* denotes the the local score calculated according to Eq. 6. From the blue and orange lines, we can see that the relation between the length of the sentence pairs tends to be stable, as the semantic similarity of the sentence pairs increases. In order to ensure the high quality of the final data set, we remove some sentence pairs based on the relationship between the length and the similarity distribution. After this pruning step, we obtain 509,832 sentence pairs. We divided the sentence

Table 1. Analyzing the length characteristics and quantity of sentence pairs by hierarchical statistics.

Top Rankings	Quantity	#Character (avg.)	#Word (avg.)	Precision
20%	101,966	28.29	18.64	100%
40%	203,933	29.25	19.23	99%
60%	305,899	30.27	19.86	97%
80%	407,866	31.31	20.51	95%
100% (all)	509,832	35.43	23.05	92%

pairs, which are selected, into five groups based on their similarity scores. We randomly select 100 sentence pairs for each set and manually check their correctness. The results are shown in Table 1. It can be clearly seen that the similarity score calculated by Eq. 6 is a good indicator of the quality of the extracted paraphrases. The overall accuracy of the whole set is 92%. When we select the top-60% sentences, this number goes up to 97%. This *goodness* metric allows us to strike a balance between quantity and quality for different application purposes.

5 Comparison and Analysis

5.1 Corpus Comparison

The corpus of this paper is the first large-scale sentence-level paraphrase corpus in Chinese. And there is no reference to our corpus of the same language. Therefore, we compare the corpus of this paper with the only three English sentence-level paraphrase corpora from the size of the data set and the differences in pairs of paraphrase characteristics caused by different language sources.

MSR Paraphrase Corpus [MSRP] [8,9] It was extracted pairs of paraphrase from news articles by an SVM classifier, which contains a total of 5,801 pairs of sentences.

Twitter Paraphrase Corpus [PIT-2015] [28,29] The data for this corpus was derived from popular topics on tweets, and it contains 14,000 pairs of sentences. If the machine automatically filter the theme, the effect will be poor. Therefore, it is important to manually specify a reasonable tweet theme.

Twitter News URL Corpus [14] This corpus construct sentence pairs of paraphrase by comparing similar URL links in Twitter to find similar user comments, which contains 51,524 pairs of sentences. Table 2 shows the size of our corpus is an order of magnitude

Table 2. Compare our corpus with three existing large-scale English sentence-level paraphrase corpus from three dimensions: the size of the corpus, the average length of the sentence pairs, and the language style of the corpus.

Corpus Name	Size	Sentence Length	Genre	Formal
MSR Paraphrase Corpus (MSRP)	5801 pairs	18.9 words	News	Yes
Twitter Paraphrase Corpus (PIT-2015)	14,000 pairs	11.9 words	Twitter	No
Twitter News URL Corpus	51,524 pairs	14.8 words	Twitter	No
Chinese Paraphrase Bank(Our Corpus)	509,832 pairs	23.05 words	Literature	Yes

larger than the three English corpora. Obviously, the amount of data is very important. This opinion can be found in the recent influential papers that introduced ELMo [20], Bert [7] as well as GPT [22,23]. The average sentence length of our corpus is also much larger than the average sentence length of other sentence-level corpora. The average length of our paraphrase is longer, proving that there is more information between each pair of sentences. We can also use these sentences to further construct phrase-level paraphrase or synonym pairs.

5.2 Language Style and The Diversity of Paraphrases

Table 2 shows that the two corpora of PIT-2015 and Twitter URL extract the pairs of paraphrase from the spontaneously generated comments of users in Twitter; the MSRP corpus extracts pairs of paraphrases from news; and our corpus extracts sentence pairs from literary works. In order to compare the influence of different language styles on the data style in the corpus, we score the sentence pairs in the four corpora using the PINC (Paraphrase In N-gram Changes) metric that is defined in [6].

PINC is the opposite of the famous BLEU [19]. The fewer the co-occurrences of the n-grams in the pair of sentences being evaluated, the higher the PINC score, and indicating the greater the difference between the pair of sentences. The paired sentences are denoted as n_i, n_j respectively, then Eq. 9 means PINC [8]. Here we set $N = 4$.

$$W = |\text{n-gram}_{n_i} \cap \text{n-gram}_{n_j}| \tag{8}$$

$$PINC(n_i, n_j) = \frac{1}{N}\sum_{n=1}^{N} 1 - (\frac{2W}{|\text{n-gram}_{n_i}|} + \frac{2W}{|\text{n-gram}_{n_j}|}) \tag{9}$$

[8] Since there is no fixed reference relationship for the sentence pairs in our corpus, the formula for the original PINC formula has been slightly modified. After the two sentences are exchanged, the PINC is calculated again, and the calculation results of the two calculations are averaged.

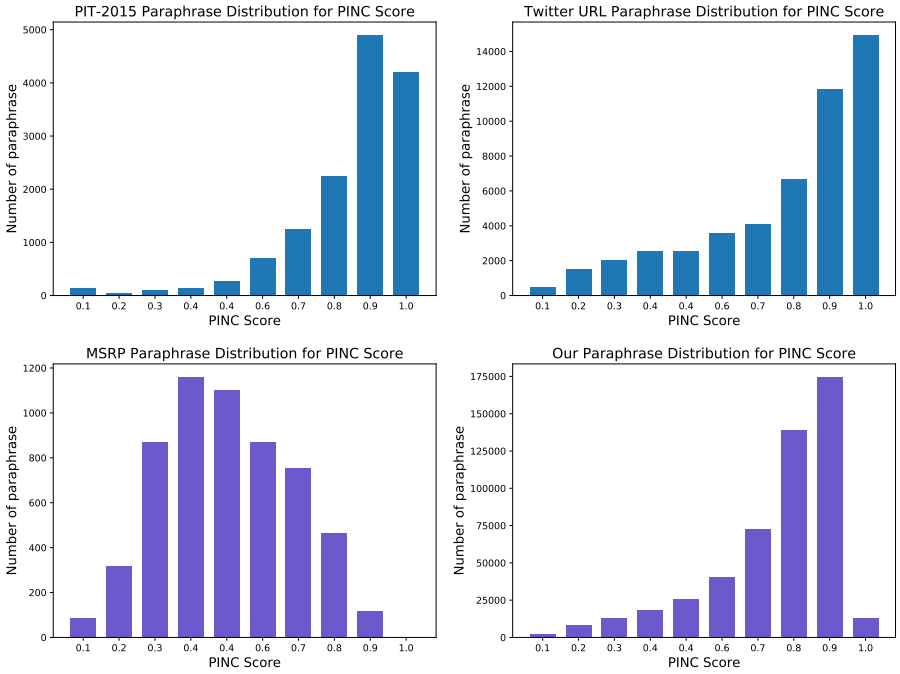

Fig. 3. The figure shows the distribution of PINC scores in three English paraphrase corpora and our paraphrase corpus. The above two blue subgraphs show the distribution of PINC scores for the sentences in informal style obtained from Twitter; the following two purple subgraphs show the distribution of PINC scores for sentence pairs with formal expressions obtained from news or literatures.

The distribution of the PINC scores for the sentence pairs in the four corpora is shown in Figure 3. The two blue subgraphs above indicate that the PINC distribution of the paraphrase sentence pairs extracted from twitter is generally higher. This is because Twitter users are very casual when they express on social networks, preferring to use shorter sentences, which leads to greater differences in paraphrase sentence pairs. The following two purple subgraphs are from the official text which prefer to use a few standard expressions. As the sentences are longer, it is more likely that repeated n-grams appear. These two factors theoretically lead to a decline in the diversity of sentence pairs. But in fact, our corpus not only ensures high precision, but also maintains strong text diversity, showing strong practicability.

6 Conclusion

We introduce a large-scale sentence-level paraphrase corpus for Chinese Language Processing. The manual evaluation and analysis of the corpus highlights the quality of this corpus. With the use of this corpus, we can enhance many NLP tasks, such as paraphrase detection, semantic parsing and natural language generation. The information

source is very general and the method is language-independent, therefore our method can be adapted to extract paraphrase corpora for other languages.

Acknowledgments

This work was supported by National Key R&D Program of China (2018YFC0831900), National Natural Science Foundation of China (61772036) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We thank the anonymous reviewers for their helpful comments. Weiwei Sun is the corresponding author.

A Translation Corpus List

Table 3 shows the details of all the translation resources that are used.

References

1. Androutsopoulos, I., Malakasiotis, P.: A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research **38**, 135–187 (2010)
2. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: Meeting of the Association for Computational Linguistics. pp. 597–604 (2005)
3. Barzilay, R., Mckeown, K.R.: Extracting paraphrases from a parallel corpus. In: Meeting of the Association for Computational Linguistics. pp. 50–57 (2001)
4. Berant, J., Liang, P.: Semantic parsing via paraphrasing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1415–1425. Association for Computational Linguistics, Baltimore, Maryland (June 2014)
5. Bhagat, R., Hovy, E.: What is a paraphrase? Computational Linguistics **39**(3), 463–472 (2013)
6. Chen, D.L., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. pp. 190–200. Association for Computational Linguistics (2011)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
8. Dolan, B., Quirk, C., Brockett, C.: Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: Proceedings of the 20th international conference on Computational Linguistics. p. 350. Association for Computational Linguistics (2004)
9. Dolan, W.B., Brockett, C.: Automatically constructing a corpus of sentential paraphrases. In: Proceedings of the Third International Workshop on Paraphrasing (IWP2005) (2005)
10. Dong, L., Mallinson, J., Reddy, S., Lapata, M.: Learning to paraphrase for question answering. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 875–886. Association for Computational Linguistics, Copenhagen, Denmark
11. Ganitkevitch, J., Van Durme, B., Callison-Burch, C.: Ppdb: The paraphrase database. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 758–764 (2013)

Table 3. This collection contains a total of 95 translations of 40 novels from the Internet.

Chinese Title	Original Title	Quantity
基督山伯爵	Le Comte de Monte-Cristo	3
飘	Gone with the Wind	2
大卫·科波菲尔	David Copperfield	2
堂吉诃德	Don Quijote de la Mancha	3
白鲸	Moby Dick	3
笑面人	L'homme qui rit	3
第二十二条军规	Catch-22	2
三个火枪手	Les Trois Mousquetaires	2
尤利西斯	Ulysses	2
罪与罚	Преступле́ние и наказа́ние	4
福尔摩斯探案集	Adventure of Sherlock Holmes	2
红与黑	Le Rouge et le Noir	2
复活	Воскресение	3
简爱	Jane Eyre	3
苔丝	Tess of the D'Urbervilles	2
童年	Децтво	3
在人间	В людях	3
我的大学	мои университеты	3
小王子	Le Petit Prince	2
麦田里的守望者	The Catcher in the Rye	2
少年维特的烦恼	Die Leiden des jungen Werther	2
雾都孤儿	Oliver Twist	2
包法利夫人	Madame Bovary	2
钢铁是怎样炼成的	Как закалялась сталсб	3
城堡	Das Schlo	2
漂亮朋友	Bel Ami	2
呼啸山庄	Wuthering Heights	2
前夜	Накануне	2
父与子	Отцы и дети	2
海上劳工	Les Travailleurs de la mer	3
神曲	Divine Comedy	3
猎人笔记	Записки охотника	3
双城记	A Tale of Two Cities	2
罗亭	Рудин	2
贵族之家	Дворянское гнездо	2
巴黎圣母院	Notre-Dame de Paris	3
安娜·卡列尼娜	Анна Каренина	2
你往何处去	Quō vādis?	2
了不起的盖茨比	The Great Gatsby	2
战争与和平	Война и мир	2

12. Grycner, A., Weikum, G.: Poly: Mining relational paraphrases from multilingual sentences. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2183–2192 (2016)
13. Jiang, Y., Kummerfeld, J.K., Lasecki, W.S.: Understanding task design trade-offs in crowdsourced paraphrase collection. arXiv preprint arXiv:1704.05753 (2017)
14. Lan, W., Qiu, S., He, H., Xu, W.: A continuously growing dataset of sentential paraphrases. arXiv preprint arXiv:1708.00391 (2017)
15. Lin, D., Pantel, P.: Dirt@ sbt@ discovery of inference rules from text. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 323–328. ACM (2001)
16. Lin, D., Pantel, P.: Discovery of inference rules for question-answering. Natural Language Engineering **7**(4), 343–360 (2001)
17. Madnani, N., Dorr, B.J.: Generating phrasal and sentential paraphrases: A survey of data-driven methods. Computational Linguistics **36**(3), 341–387 (2010)
18. Nakashole, N., Weikum, G., Suchanek, F.: Patty: a taxonomy of relational patterns with semantic types. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 1135–1145. Association for Computational Linguistics (2012)
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
20. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations (2018)
21. Quirk, C., Brockett, C., Dolan, B.: Monolingual machine translation for paraphrase generation (2004)
22. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. URL https://s3-us-west-2. amazonaws. com/openai-assets/research-covers/languageunsupervised/language understanding paper. pdf (2018)
23. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**, 8 (2019)
24. Seraj, R.M., Siahbani, M., Sarkar, A.: Improving statistical machine translation with a multilingual paraphrase database. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1379–1390 (2015)
25. Sun, W., Xu, J.: Enhancing Chinese word segmentation using unlabeled data. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 970–979. Association for Computational Linguistics, Edinburgh, Scotland, UK. (July 2011)
26. Suzuki, Y., Kajiwara, T., Komachi, M.: Building a non-trivial paraphrase corpus using multiple machine translation systems. In: Proceedings of ACL 2017, Student Research Workshop. pp. 36–42 (2017)
27. Wieting, J., Gimpel, K.: Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1 (2018)
28. Xu, W., Callison-Burch, C., Dolan, B.: Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015). pp. 1–11 (2015)
29. Xu, W., Ritter, A., Callison-Burch, C., Dolan, W.B., Ji, Y.: Extracting lexically divergent paraphrases from twitter. Transactions of the Association for Computational Linguistics **2**, 435–448 (2014)
30. Zhang, C., Soderland, S., Weld, D.S.: Exploiting parallel news streams for unsupervised event extraction. Transactions of the Association of Computational Linguistics **3**(1), 117–129 (2015)