ExaPPC: a Large-Scale Persian Paraphrase Detection Corpus

Reyhaneh Sadeghi*, Hamed Karbasi, Ahmad Akbari
Exa Company, Tehran, Iran
r.sadeghi@exalab.co, karbasi@exalab.co, karbasi_hamed@ee.sharif.edu, a.akbari@exalab.co, ahmadakbari@ut.ac.ir

2022 8th International Conference on Web Research (ICWR) | 978-1-6654-6626-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICWR54782.2022.9786243

Abstract—This paper describes the creation of Exa Persian Paraphrase Corpus (ExaPPC), a large paraphrase corpus consisting of monolingual sentence-level paraphrases using different sources. ExaPPC is the first large-scale paraphrase dataset used in Persian paraphrase detection to the best of our knowledge. There are 2.3M labeled sentence pairs in the corpus consisting of a 1M paraphrase label and 1.3M non-paraphrase label. Efforts were made manually and semi-automatically to construct this corpus using techniques such as subtitle alignment, translating existing parallel English-Persian corpus and similarity corpus on English tweets. In addition to enriching the corpus, candidate sentence pairs among tweets have been extracted via NLP tools and labeled by two Persian native speakers. The advantages of this corpus compared to the existing ones are the number of pair sentences, sentence Length variation and textual diversity, including formal and dialogue sentences. The result on the provided test corpus shows that ExaPPC achieves 94% accuracy on paraphrase detection task. The corpus is publicly available[1].

Keywords—Paraphrase Identification; Semantic Similarity; Deep Learning; Paraphrasing Corpora

I. INTRODUCTION

Paraphrase detection is a task where a model should identify whether a pair of sentences is a paraphrase. In the past decades, there was a growing trend in paraphrase detection, and the paraphrase dataset is one of the critical resources in natural language processing. Paraphrase detection is helpful for a question-answer forum, like Quora, to prevent users from asking redundant questions by detecting similar ones. Furthermore, it is also used in natural language processing applications such as information retrieval, summarization, answer selection and more [1].

Another use case is exploring twitter to find similar and equivalent tweets to understand how popular a topic is among Twitter users. At the moment, paraphrase detection or paraphrase generation are primarily limited to the unavailability of the rich datasets, so having a reliable and divergent one is necessary to fine-tune required language models such as BERT. For the Persian language, in particular, the currently available corpora is too small to cover a wide and varied range of sentences that can be used by deep learning methods to have an automatic paraphrase detector or paraphrase generator [2].

Depending on the application and its different requirements, it is challenging to define paraphrase terms accurately. However, there is no precise definition of paraphrases without ambiguity, but different categories exist [3].

According to linguistics, the almost identical expressions are defined paraphrases.

The others believe that paraphrases have the equivalent meaning and can be used in the same position. In some applications like summarization, the exact meaning is ultimately unnecessary, and related context is enough [4]. However, in Q&A systems, *paraphrases* are defined as paired sentences with precisely the same meaning [5].

Here, unlike many previous methods of solution between disciplines taken as a result of different applications, two categories of paraphrase and non-paraphrase are considered. To have diverse non-paraphrase pairs, we consider related and non-related as two subcategories.

In our definition, paraphrases are paired sentences with identical meanings that can be used interchangeably; A paraphrase is a restatement of meaning with different expressions [4]. Paraphrase pairs can vary in synonym words, the length of the sentence, or the grammatical aspects like active and passive. The pair sentences which are not exactly with the same meaning but with related subject and context and almost overlapping in some words are called Related. Conversely, non-related are paired sentences that do not have equivalent meaning, are not common in context and words, and are entirely unrelated. As we mentioned before, the corpus includes just two labels: paraphrase and non-paraphrase. So, utilizing related and non-related labels only arises for labeling procedures by annotators, and we consider these labels non-paraphrases.

To clarify the definition, we have listed some examples in Table I.

The main contributions of this article are:

1) Proposing the first large-scale paraphrase dataset in Persian, including 2.3M pairs.

2) Proposing a general method to gather paraphrase pairs from different sources.

3) Fine-tuning ParsBert using presented corpus and developing a paraphrase detection in Persian.

As mentioned before, the lack of a large dataset in Persian, led us to create a reasonably large corpus of naturally occurring, non-handcrafted sentence pairs that can be used as a rich resource for training or testing purposes.

We hope this corpus can provide a rich source of semantic knowledge to improve downstream natural language understanding

[1] https://github.com/exaco/exappc

978-1-6654-6626-4/22/$31.00 ©2022 IEEE

TABLE I. REPRESENTATIVE EXAMPLES OF PARAPHRASE, RELATED AND NON-RELATED SENTENCE PAIRS CONSIDERED ON EXAPPC

Labels	Type	Examples
Paraphrase		روز گذشته وین میزبان کشورها بود تا توافقاتی برای صلح جهانی تنظیم کنند. Yesterday, Vienna hosted countries to have agreements for world peace. توافق های بین المللی صلح جهانی، روز گذشته در وین منعقد شد. International world peace agreements were signed in Vienna yesterday.
Non-paraphrase	Related	روز گذشته وین میزبان کشورها بود تا توافقاتی برای صلح جهانی تنظیم کنند. Yesterday, Vienna hosted countries to have agreements for world peace. توافق برای صلح جهانی یکی از مهمترین جنبه های دیپلماسی برای همه کشورها است. Agreement for world peace is one of the most important aspects of diplomacy for all countries.
Non-paraphrase	Non-related	روز گذشته وین میزبان کشورها بود تا توافقاتی برای صلح جهانی تنظیم کنند. Yesterday, Vienna hosted countries to have agreements for world peace. نمایشگاه غذا با استقبال خوب مردم روبرو شد. The food exhibition was well received by the people.

tasks and be a valuable resource for paraphrase detection and paraphrase generation tasks. Furthermore, having a valuable paraphrase generator could augment the paraphrase corpus in the future.

The current article is organized as follows: Section II presents the Related Work that inspires this study, section III describes the methods to construct dataset from different sources, section IV shows the corpus structure, section V describes the labeling procedure and the experiment that had been conducted to evaluate the prepared corpus, and finally, section VI is the conclusion.

II. RELATED WORK

First, The English paraphrase datasets are surveyed, then the available Persian ones are reviewed to clarify the primary importance of this research.

MSR Paraphrase Corpus [MSRP] is a paraphrase dataset, which consists of 5,801 sentence pairs generated by clustering news articles with an SVM classifier and human annotations [6].

Twitter is a platform which is beneficial for extracting many paraphrase pairs. Twitter Paraphrase Corpus [PIT-2015] contains 14,035 paraphrase pairs on more than 400 different topics [7]. TwitterPPDB corpus [8] was proposed to develop PIT-2015. TwitterPPDB with 56,787 sentence pairs collected from Twitter by linking tweets through shared URLs. MSRP or PIT-2015 encourage a series of work in paraphrase detection task [9].

Quora released a new dataset called QQP in 2017, containing over 400K question duplicate pairs [10] and is mainly used for paraphrase generation [11], [12].

ParaNMT-50M [13] is a dataset of more than 50 million paraphrase pairs constructed by Wieting and Gimpel. The pairs prepared automatically by translating the non-English side of a large parallel corpus.

ParaSCI is the first large-scale paraphrase dataset in the scientific field, consists of 33,981 paraphrase pairs from ACL (ParaSCIACL) and 316,063 pairs from arXiv (ParaSCIarXiv) [14]. This dataset is constructed through intra-paper and inter-paper methods, such as collecting citations to the same paper or aggregating definitions by scientific terms. The advantages of paraphrase pairs in ParaSCI are due to the prominent length and textual diversity. ParaSCI obtained satisfactory results on human evaluation and downstream tasks such as long paraphrase generation [14].

In another method performed in the Russian language, news agencies and crowdsourcing solutions are utilized to label the extracted candidates. Also, examples are labeled in 3 categories of paraphrase, related and non-related [5]. This method uses a semantic similarity criterion based on a similarity matrix. Other corpus benefits from individuals' non-automatic rewriting of existing expressions [15] or automatic extraction produced from bilingual parallel corpus [16] that all are in non-Persian languages.

A related corpus is developed to evaluate plagiarism detection systems in Persian [17]. This corpus is based on copies available on Persian Wikipedia pages which contains 1500 documents that 411 cases are frauds by swapping, deleting, adding and replacing identical words. Another corpus in the plagiarism detection task consists of 11,089 documents and more than 11,603 plagiarism cases constructed manually, semi-automatically, and automatically [18]. Mahak Samim, a plagiarism detection corpus, containing Persian academic texts including more than five thousand artificial plagiarism cases with various lengths and different degrees of obfuscation [19]. This corpus's minimum plagiarism case length is 50, with less than 1000 samples. A bilingual Persian-English plagiarism detection corpus with 11,200 plagiarism cases is provided from a Persian-English sentence-aligned parallel corpus combined with Wikipedia articles [20]. Paired sentences in parallel corpus with different lengths and degrees of obfuscation have a similarity score between 0 and 1, and the length of fragments is distributed between 3 and 15 sentences.

Degarbayan is a Persian paraphrase corpus consisting of 1523 sentence pairs with different degrees of semantic similarity. This corpus is developed via automatic extraction with crowdsourcing methods and is judged and labeled by annotators. Extracted sentences are from different news agencies with a wide range covering various topics. The data of this corpus are presented in three categories: *paraphrase*, *almost paraphrase*, and *irrelevant*. [2]. PARSINLU is prepared as a dataset in 6 distinct NLU tasks, including a paraphrasing question [21]. There are 4644 manually annotated Persian question pairs with paraphrase and non-paraphrase labels based on available natural sentences and translated from the QQP dataset [10], which is a dataset of English question-pairs.

According to the review of related Persian datasets, it is evident that corpora in plagiarism detection tasks are usually not in sentence-level so can not be used in the paraphrase detection task. Those that consist of sentence pairs, are usually long in length and low in amount. Available Persian paraphrase datasets also have formal literature and are low in an amount too. So, to the best of our knowledge, ExaPPC is the first large-scale paraphrase dataset in Persian, including formal and dialogue pairs like tweets, consisting of question and non-question pairs.

A comparison on available paraphrase datasets in English and Persian in the concept of Genre and size of the corpus is shown in Table II.

TABLE II. REVIEW ON EXISTING PARAPHRASE CORPORA IN ENGLISH AND PERSIAN IN THE CONCEPT OF GENRE, SIZE, AND TYPE OF LABELS

Name	Genre	Size (pairs)	Types of labels
MSRP	News	5,801	Paraphrase
PIT	twitter	14,035	Paraphrase Non-Paraphrase Debatable
TwitterPPDB	Twitter	56,787	Paraphrase Non-Paraphrase
paraNMT-50M	Novels, laws	51,409,585	Paraphrase
Quora	Question	404,289	Paraphrase
paraSCI	Scientific paper	350,044	Paraphrase
Degarbayan (Persian)	News	1,523	Paraphrase Non-Paraphrase irrelevant
PARSINLU Question paraphrasing (Persian)	Question	4,644	Paraphrase Non-Paraphrase

III. DEVELOPMENT OF THE CORPUS

A. Resources

Many methods have been developed for generating or finding paraphrase sentence pairs, such as multiple translations of the same source material [22], comparable articles from multiple news sources [23], crowdsourcing [24], using tweets with matching URLs [25], and statistical machine translation to find lexical and phrasal paraphrases in the parallel text [26].

We use different resources and methods to gather paraphrase pairs. One of our methods is following the approach of [13], which uses a large parallel corpus and translates the non-English side of the parallel text to get English-English paraphrase pairs. Similarly, we translate the non-Persian side of the existing English-Persian parallel corpus and existing English similarity corpus to make paraphrase pairs.

Our dataset is constructed based on the following source materials and approaches:

- Subtitle Alignment
- Using existing parallel English-Persian corpus
- Using existing labeled similarity corpus on English tweets
- Using NLP tools to extract candidate sentences among tweets

B. Subtitle Alignment

One of the methods to extract paraphrase corpora is using subtitles. Some advantages of using movie subtitles are listed below [27]:

- They increase daily in an amount: due to high demand, the online databases of movie subtitles are one of the fastest-growing resources.
- They are publicly available and can be downloaded freely from different subtitle websites.

To implement this method, around 100 subtitle files were captured from Open-subtitles[2], a free online collection of movie subtitles. It included subtitles of various versions of the same movie or several copies translated by different subtitle makers. It resulted in about 50 subtitle pairs. Each pair consisted of two textual files (in *SRT* format) containing subtitles of the same movie version in the Persian language by different translators.

Due to the time frame mismatches of the movie subtitles, some preprocessing needed to synchronize them. Using online subtitle tools is beneficial for shifting and synchronizing mismatched subtitle file pairs and merging them. Afterward, we extract pair subtitles simultaneously as paraphrase pair sentences. Most of the subtitle files begin with comments about the subtitle creator/translation team or related content that usually does not match the file pair. These non-matching contents of the subtitle file pair introduce difficulty while aligning their sentences. An automatic process to chop off these contents from subtitle files is utilized. Duplicates were then removed to make the resource unique and avoid redundancy.

Subtitle files do not have a standard for encoding. All files are converted to UTF-8 encoding to address that issue. This intuitive approach could extract 35,000 paraphrase pairs from Persian subtitle resources with a low alignment error rate.

C. Using existing parallel English-Persian corpus

Most current applications use manually collected paraphrases tailored to a specific application to identify

[2] https://www.opensubtitles.org/

paraphrases. However, manually collecting paraphrases on a large scale is time-consuming. This paper presents a corpus-based method for the semi-automatic extraction of paraphrases. Our method for paraphrase extraction builds upon the methodology developed in Machine Translation (MT) [22] due to translations preserving the meaning of the source but may use different words to convey the meaning.

This part describes the extraction procedure of large paraphrase pairs from English-Persian parallel datasets. In the past decades, reliable parallel English-Persian corpus has been constructed like TEP [28], Mizan [29], PEPC [30]. We used these datasets to construct paraphrase pairs by translating the English part of pairs to Persian via Google Translate. TEP is a large-scale (in order of million words) English-Persian parallel corpus extracted from movie subtitles in English and Persian. It consists of around 612K English-Persian parallel sentences from 1600 movies [28].

Mizan is the most considerable Persian-English parallel corpus with more than 1 million sentence pairs collected from masterpieces of literature and manually aligned. To align corresponding sentences of refined books, an alignment aiding software operated by alignment specialists, mainly translators and linguists, eased the process by providing basic operations such as break, merge, delete and edit tools [29].

PEPC corpus consists of about 200,000 sentences that have been sorted by their degree of similarity calculated by the IR system. This corpus is constructed by a bidirectional method to extract parallel sentences from English and Persian documents aligned with Wikipedia. 1200 English-Persian pairs were released for the test that supervisors translated, and the others were extracted from some websites which offered free parallel sentences from paper abstracts [30]. We used a test dataset consisting of 1200 pairs to rely on English-Persian sentences because the other pairs are constructed automatically and parallel by some scores but not precisely. Here is the Formation procedure of the reliable corpus:

1) *Since some sentences may not be translated completely, and some remain in English, we pass the translator's output to a language detector to avoid non-translated or partly-translated sentences.*

2) Furthermore, we removed duplicate sentences.

With this method, we could present around 1.4 M paraphrase pairs consisting of 1,166 from PEPC, 900k from Mizan, and 500k from TEP.

D. Using existing labeled similarity corpus on English tweets

To have a diverse corpus consisting of different types of data, using a labeled similarity corpus of tweets is the other way to make our corpus richer. The similarity corpora presented in English like SemEval-2015 Task 1 [7] and TwitterPPDB [8] have been used to extract paraphrase and non-paraphrase sentences. These resources are all on tweets labeled by annotators.

The SemEval-2015 shared task on paraphrasing and semantic similarity in twitter (PIT) uses a training and development set of 17,790 sentence pairs and a test set of 972 sentence pairs with labels [7]. The TwitterPPDB dataset is a whole new dataset; it is also the most extensive human-labeled paraphrase corpus of 51,524 sentence pairs. This dataset presented a new method to collect large-scale sentential paraphrases from Twitter by linking tweets through shared URLs. Another development of TwitterPPDB consists of 114K paraphrase pairs extracted in 2017. The main advantage of this method is its simplicity, as it gets rid of the classifier or human in the loop needed to select data before annotation [8].

Every pair of sentences has a score that indicates how many annotators believe in paraphrasing. We used a source table of every corpus to extract paraphrase, non-paraphrase, or debatable labels from scores; afterward, chop off debatable pairs and use high probability paraphrase and non-paraphrase pairs to pass through the pipeline of translators. We used Google Translate to translate similarity pairs to Persian, then passed the translated pairs to a language detector to remove uncompleted translated sentences and duplicate sentences like part III-C. Removing duplicates may chop off some sentences but help to have a reliable corpus.

Furthermore, some hashtags in tweets may not translate when they are along with the sentence. We overcome this by extracting the exact word of the hashtags and passing it separately to the translator. Then we added translated hashtags in their position in the sentence. The output of the presented pipeline consists of around 8440 pairs from SemEval 2015(PIT) and 101,640 from TwitterPPDB, totaling 110k.

E. Using NLP tools to extract candidate sentences among tweets

This section aimed to extract paraphrase and non-paraphrase pairs among actual tweets. To achieve this goal semi-automatically, we used Named Entity Recognition (NER) to make candidate sentences.

Exa[3] has developed a manually-annotated Named Entity Recognition (NER) corpus with sufficient size for machine learning methods. This corpus, called Exa_NER, has more than 600,000 tokens. Thirteen types of entities have been annotated. A deep learning model is trained on the Exa_NER corpus that achieved a precision of 97.18% and F1 of 87.92%.

For gathering paraphrase corpora, we passed 2 million tweets crawled from Twitter to the NER model to extract named entities. Afterward, a pipeline is constructed to extract pair sentences as candidates. We tried three semi-automatic approaches:

Approach 1: First, extracting the most occurrence value of each entity type and then preparing some groups of tweets that are common in two entity values. For example, "واکسن"(vaccine) and "کرونا"(coronavirus) are the most occurrence values of "PRODUCT" entities among 2M tweets. We extracted tweets with these two entity values simultaneously as a group. In order to consider sentences of this group as pairwise, we need to filter some non-beneficial pairs. So, we check the appearance similarity metric of candidate pairs from this group and pass the pairs with a threshold of > 0.6. To calculate this metric, we use the Simhash library implemented in python that shows the apparent

[3] www.exalab.co

similarity of each pair. Then, we delivered candidate pairs to the annotators to label these semi-automatic extracted candidate pairs. It causes to gather paraphrase pairs and non-paraphrase pairs with related similarities because candidate pairs have the same topic and common entity values. It could help future models that want to be trained with our corpus being divergent because our non-paraphrase pairs are not just ones with entirely different words and contexts.

We constructed 3000 manually labeled pairs consisting of paraphrase, related and non-related pairs with this method. Finally, related and non-related pairs are merged as non-paraphrases.

Approach 2: We tried additional pre-training of ParsBERT(Persian BERTmodel) [31]. ParsBERT is a BERT model existing in Persian, a monolingual language model based on Google's BERT architecture. This model is pre-trained on large Persian corpora with various writing styles from numerous subjects (e.g., scientific, novels, news) with more than 3.9M documents, 73M sentences, and 1.3B words [31].

We tried to have an Additional pre-training of ParsBERT by MLM(Masked Language Modeling) method on an extra 5M Persian tweets crawled from Twitter. Additional pre-training of ParsBert is used as a pre-trained model for sentence embedding via sentence-transformers library [32]. We prepare the query and corpus datasets to make sentence embeddings, then use cosine similarity on a pair of embeddings of the query and every corpus sentence. Afterward, we extract the top 5 most similar pairs for presenting to the annotators to label them. This method can guarantee that we can have paraphrase or non-paraphrase-related pairs. Query and corpus are prepared as a group of tweets common in 2 named entities extracted like approach1. Finally, we could have 5000 manually labeled pairs consisting of paraphrase, related and non-related pairs.

Approach 3: In this part, after Extracting the most occurrence value of each entity from 2M tweets, we extracted pair tweets as candidates that are different in concept with different entity types and values. For example, tweets gathered with the entity of "EVENT" and value of "غدیر"(Ghadir) are being paired randomly with tweets gathered with the entity of "ORG" and value of "بارسلونا"(Barcelona) We are sure that the tweets of these groups are entirely different from each other. These pair candidates are then filtered to pass pairs with similarity metric (Simhash library) of <0.5. So, by the semi-automatic manner, we could extract 1.3 M pairs as non-paraphrase pairs.

F. *Selecting High-quality Paraphrases*

Table III shows all resources and the number of pairs extracted from each resource.

Since most sentence pairs are prepared automatically, a filtering method has been utilized to obtain the high-quality paraphrase corpus. Making sentences automatically with methods mentioned in III-B to III-D needs to be evaluated. The high number of pair sentences in this corpus makes human judgment impossible, so we tried a semi-automatic method to filter unusable pair sentences. In order to achieve this purpose, We used an unsupervised method to train the Persian BERT model on prepared 2.8M pair sentences obtained so far from the resources listed in III-B to III-E.

We tried to pre-train ParsBERT on prepared sentences using the MLM (Masked Language Modeling) method. Eventually, we pass pair sentences to the ParsBERT model to get their embeddings via sentence-transformers library [32]; then, we use cosine similarity to measure the similarity metric. Intuitive examination of the similarity metric of paraphrase pairs resulted in filtering with >= 0.69. Also, paraphrase pairs with >= 0.98 are considered duplicate pairs and eliminated.

The final corpus size before and after filtering is shown in Table IV to compare how many unreliable paraphrase pairs are eliminated from the corpus.

IV. CORPUS STRUCTURE

The most advantages of the presented corpus are the number of pairs and the text variety, including wiki, subtitle, linguistic, dialogue, and tweets. Also, diversity in the subject (e.g., novels, news, politics, movie) is significant. Our corpus consists of 2.3 M pair sentences with 1 M paraphrase label and 1.3 M non-paraphrase label. Comprehensive information about the size of the corpus and total tokens are listed in Table V.

V. EXPERIMENTS

A. *Annotation instruction*

This section describes the labeling system. First, we prepare an instruction on how annotators may annotate and describe the type of our labels (paraphrase, related, non-related) discussed in section I.

We had five annotators who were completely aware of the labeling procedure and evaluated 200 test data for quality check. Subsequently, 2 of them with a higher accuracy rate were hired to label our candidate data. We had several meetings to ensure that our annotators were professional in context.

UTAG[4] annotation website supplied the annotators with the mentioned labeling system through a reliable, easy-access web application. The sample of labeling on the UTAG is shown in Fig. 1.

Annotators were told to use their best judgment in labeling sentence pairs. Note that many sentence pairs judged to be "non-paraphrase" still overlap in information content and even wording. These pairs reflect a range of relationships, from unrelated semantically to partially overlapping ones.

B. *Evaluation Results*

In this section, we use a method to evaluate the percentage of how our corpus is accurate and reliable. Due to the large dataset prepared in this paper, manual evaluation by humans will not accomplish, so we tried to fine-tune a BERT model on ExaPPC. We chose BERT because it has more accuracy in paraphrase identification task. As mentioned in [1], the result of the BERT tested on the Quora question pairs dataset was better than using LSTM. So we used ParsBERT [31] that had

[4] https://utag.ir

TABLE III. DISTRIBUTION OF EXTRACTED SENTENCE PAIRS BY CONSTRUCTING METHOD

Name	Subtitle (fa-fa)	Translate-parallel data (en-fa)	Translate- Pair Tweet with label (en-en)	Sentence Similarity from ParsBERT	NER Model	Total
Automatically constructed	35K	1.4M	110K	NaN	1.3M	2.8M
Human annotation	NaN	NaN	NaN	5000	3000	8000

TABLE IV. COMPARISON OF CORPUS SIZE BEFORE AND AFTER FILTERING

Filtering Situation	Size(sentence pairs)	Distribution
Before Filtering	2.8M	paraphrase: 1.5 M non-paraphrase: 1.3 M
After Filtering	2.3M	paraphrase: 1 M non-paraphrase: 1.3 M

TABLE V. STATISTICS OF EXAPPC BY NUMBER OF SENTENCE PAIRS AND TOKENS

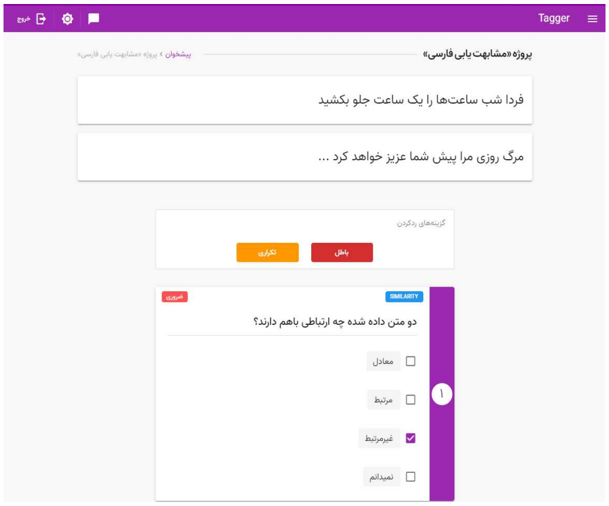
Dataset	ExaPPC
Size(sentence pairs)	2,342,145
Size(tokens)	102,149,576
Average sentence length	22
Distribution	Total paraphrase: 986k Total non-paraphrase: 1.3 M
Number of labels	2(Paraphrase, Non-paraphrase)

TABLE VI. DISTRIBUTION OF TRAIN, VALIDATION AND TEST DATA FOR CONSTRUCTING PARAPHRASE IDENTIFICATION MODEL BY FINE-TUNING PARSBERT

	Sentence Pairs	Paraphrase	Non-paraphrase
train	2.1M	900K	1.2M
validation	220K	86K	135K
test	1500	500	1000 (related:500 non-related:500)

TABLE VII. EVALUATION RESULTS OF FINE-TUNED PARSBERT ON VALIDATION DATA AFTER 5 EPOCHS

Model	Acc	F1	Precision	Recall
Fine-tuned ParsBERT on ExaPPC	0.9451	0.9414	0.9669	0.9172



پروژه «مشابهت یابی فارسی»

فردا شب ساعت‌ها را یک ساعت جلو بکشید

مرگ روزی مرا پیش شما عزیز خواهد کرد ...

دو متن داده شده چه ارتباطی باهم دارند؟

□ معادل

□ مرتبط

☑ غیرمرتبط

□ نمیدانم

Fig. 1. Labeling procedure on exclusive annotation website of Exa

additional pre-train on 5M tweets crawled from Twitter for five epochs. Then we fine-tuned it with ExaPPC by a supervised method to construct a paraphrase identifier to predict and label the pairs, paraphrase or non-paraphrase. We used a Cross-Encoder architecture based on the BERT model [32] with a sigmoid layer at the end for binary classification. As detailed in [32], Cross-Encoder performs better than Bi-Encoders on classification tasks.

ExaPPC has been split to train and validation data. The distribution of train and validation data is mentioned in Table VI.

We did fine-tune on ExaPPC for five epochs and calculated accuracy, precision, f1-measure, and recall on validation data shown in Table VII.

For evaluating the proposed model for the paraphrase identification task, we prepared a test dataset consisting of 1500 pair sentences. The distribution of test data is: We used 500 paraphrase pairs from the Degarbayan corpus [2] and prepared some candidates using the approach2 in III-E and delivered them to the annotators to label them. So, our test dataset is fully labeled by humans and is reliable.

As mentioned before, our classifier is based on two classes: paraphrase and non-paraphrase. In order to have the best evaluation on different kinds of sentences, we consider 500 paraphrases, 500 related pairs, and 500 non-related pairs.

The accuracy of our paraphrase identifier model prediction on 1500 test data is demonstrated in Table VIII.

It is clear that our proposed model can predict 96% of paraphrases correctly and average of ((96% + 91%)/2) = 93.5% correctness of non-paraphrase prediction. The paraphrase detection model on our corpus has an average of 94% of accuracy.

Our training procedures are performed on the two 1080 ti GPUs with 11 GB RAM.

We compared our paraphrase identifier model and the fine-tuned model on the ParsiNLU question paraphrasing dataset [21]. Among ParsiNLU released models, only the mT5 model[5] was available. So we evaluate the mT5-base model on ExaPPC test data. The results show that mT5 has a 70% accuracy on our test dataset.

VI. CONCLUSION

To the best of our knowledge, we present Exa Persian Paraphrase Corpus (ExaPPC), the first large-scale collection of paraphrase and non-paraphrase pairs in Persian consisting of 2.3 M pairs. We described the methods and resources to gather paraphrase and non-paraphrase pairs manually and semi-automatically. We showed how to use ExaPPC to train a paraphrase detection model. Our results suggest that ExaPPC will be helpful in a variety of NLP applications like paraphrase detection tasks with an accuracy of 94% on 1500 test data. The key advantage of our corpus is the large size and variety of data extracted from various sources, including formal and dialogue pairs.

Future releases of ExaPPC will focus on expanding the paraphrase pairs regarding data size usable for paraphrase generation downstream task and increasing the size of related pairs to have a more divergent corpus. Our goal is to provide ExaPPC as a continuous updating and improvement resource.

ACKNOWLEDGMENT

The authors would like to thank Exa Company for providing the infrastructure, human resources, technical and non-technical knowledge that achieves this article and its goals.

TABLE VIII. PARAPHRASE IDENTIFIER EVALUATION RESULTS ON TEST DATA

Model	Paraphrase	Non-Paraphrase	Total
Fine-tuned ParsBERT on ExaPPC	0.96	Related: 0.91 Non-Related: 0.96	0.94

[5] https://huggingface.co/persiannlp/mt5-base-parsinlu-qqp-query-paraphrasing

REFERENCES

[1] A. Chandra and R. Stefanus, “Experiments on paraphrase identification using quora question pairs dataset,” CoRR, vol. abs/2006.02648, 2020. [Online]. Available: https://arxiv.org/abs/2006.02648

[2] R. Maanijou and S. Mirroshandel, “Degarbayan: Developing a persian paraphrase corpus by crowdsourcing,” 10 2017.

[3] C. Boonthum, “iSTART: Paraphrase recognition,” in Proceedings of the ACL Student Research Workshop. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 31–36. [Online]. Available: https://www.aclweb.org/anthology/P04-2006

[4] R. Bhagat and E. Hovy, “What is a paraphrase?” Computational Lin guistics, vol. 39, no. 3, pp. 463–472, 2013.

[5] E. Pronoza, E. Yagunova, and A. Pronoza, “Construction of a russian paraphrase corpus: unsupervised paraphrase extraction,” in Russian Summer School in Information Retrieval. Springer, 2015, pp. 146– 157.

[6] W. B. Dolan and C. Brockett, “Automatically constructing a corpus of sentential paraphrases,” in Proceedings of the Third International Workshop on Paraphrasing (IWP2005), 2005. [Online]. Available: https://www.aclweb.org/anthology/I05-5002

[7] W. Xu, C. Callison-Burch, and B. Dolan, “Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit),” 01 2015, pp. 1–11.

[8] W. Lan, S. Qiu, H. He, and W. Xu, “A continuously growing dataset of sentential paraphrases,” CoRR, vol. abs/1708.00391, 2017. [Online]. Available: http://arxiv.org/abs/1708.00391

[9] J. Mallinson, R. Sennrich, and M. Lapata, “Paraphrasing revisited with neural machine translation,” in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 2017, pp. 881–893.

[10] S. Iyar, N. Dandekar, and K. Csernai, “First quora dataset release: Question pairs,” retrieved at. [Online]. Available: https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

[11] Y. Fu, Y. Feng, and J. P. Cunningham, “Paraphrase generation with latent bag of words,” 2020, vol. abs/2001.01941. [Online]. Available: http://arxiv.org/abs/2001.01941

[12] A. Gupta, A. Agarwal, P. Singh, and P. Rai, “A deep generative framework for paraphrase generation,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[13] J. Wieting and K. Gimpel, “ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine transla tions,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 451–462. [Online]. Available: https://www.aclweb.org/anthology/P18-1042

[14] Q. Dong, X. Wan, and Y. Cao, “Parasci: A large scientific paraphrase dataset for longer paraphrase generation,” CoRR, vol. abs/2101.08382, 2021. [Online]. Available: https://arxiv.org/abs/2101.08382

[15] P. M. McCarthy and D. S. McNamara, “The user-language paraphrase corpus,” in Cross-Disciplinary Advances in Applied Natural Language Processing: Issues and Approaches. IGI Global, 2012, pp. 73–89.

[16] J. Ganitkevitch, B. Van Durme, and C. Callison-Burch, “Ppdb: The para phrase database,” in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 758–764.

[17] K. Khoshnavataher, V. Zarrabi, S. Mohtaj, and H. Asghari, “Developing monolingual persian corpus for extrinsic plagiarism detection using artificial obfuscation,” 2015.

[18] F. Mashhadirajab, M. Shamsfard, R. Adelkhah, F. Shafiee, and C. Saedi, “A Text Alignment Corpus for Persian Plagiarism Detection,” FIRE, 2016.

[19] M.R. Sharifabadi, and S.A. Eftekhari. “Mahak Samim: A Corpus of Persian Academic Texts for Evaluating Plagiarism Detection Systems,” FIRE, 2016.

[20] H. Asghari, K. Khoshnava, O. Fatemi, and H. Faili, “Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus: Notebook for PAN at CLEF 2015,” CLEF, 2015.

[21] D. Khashabi, A. Cohan, S. Shakeri, P. Hosseini, P. Pezeshkpour, M. Alikhani, M. Aminnaseri, M. Bitaab, F. Brahman, S. Ghazarian, M. Gheini, A. Kabiri, R.K. Mahabagdi, O. Memarrast, A. Mosallanezhad, E. Noury, S. Raji, M.S. Rasooli, S. Sadeghi, E.S. Azer, N.S. Samghabadi, M. Shafaei, S. Sheybani, A. Tazarv, and Y. Yaghoobzadeh, "ParsiNLU: A Suite of Language Understanding Challenges for Persian," Transactions of the Association for Computational Linguistics, 2021.

[22] R. Barzilay and K. McKeown, "Extracting paraphrases from a parallel corpus," in Proceedings of the 39th annual meeting of the Association for Computational Linguistics, 2001, pp. 50–57.

[23] B. Dolan, C. Quirk, and C. Brockett, "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources," in Proceedings of the 20th international conference on Computational

[24] Y. Jiang, J. K. Kummerfeld, and W. S. Lasecki, "Understanding task design trade-offs in crowdsourced paraphrase collection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers*, Vancouver, Canada, pp. 103–109,.

[25] W. Lan, S. Qiu, H. He, and W. Xu, "A continuously growing dataset of sentential paraphrases," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 1224–1234,.

[26] C. Bannard and C. Callison-Burch, "Paraphrasing with bilingual parallel corpora," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*.

[27] E. Itamar and A. Itai, "Using movie subtitles for creating a large-scale bilingual corpora." in *LREC*, 2008.

[28] M. T. Pilevar, H. Faili, and A. H. Pilevar, "Tep: Tehran english persian parallel corpus," in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 68–79.

[29] O. Kashefi, "MIZAN: A large persian-english parallel corpus," *CoRR*, vol. abs/1801.02107, 2018. [Online]. Available: http://arxiv.org/abs/1801.02107

[30] A. Karimi, E. Ansari, and B. S. Bigham, "Extracting an english-persian parallel corpus from comparable corpora," *CoRR*, vol. abs/1711.00681, 2017. [Online]. Available: http://arxiv.org/abs/1711.00681

[31] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, "Parsbert: Transformer-based model for persian language understanding," *CoRR*, vol. abs/2005.12515, 2020. [Online]. Available: https://arxiv.org/abs/2005.12515

[32] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *CoRR*, vol. abs/1908.10084, 2019. [Online]. Available: http://arxiv.org/abs/1908.10084