Turkish Paraphrase Corpus

Seniz Demir, İlknur Durgar El-Kahlout, Erdem Unal, Hamza Kaya

TUBITAK-BILGEM
Gebze, Kocaeli, TURKEY
{senizd,idurgar,unal,hamzaky}@uekae.tubitak.gov.tr

Abstract

Paraphrases are alternative syntactic forms in the same language expressing the same semantic content. Speakers of all languages are inherently familiar with paraphrases at different levels of granularity (lexical, phrasal, and sentential). For quite some time, the concept of paraphrasing is getting a growing attention by the research community and its potential use in several natural language processing applications (such as text summarization and machine translation) is being investigated. In this paper, we present, what is to our best knowledge, the first Turkish paraphrase corpus. The corpus is gleaned from four different sources and currently contains 1270 paraphrase pairs. All paraphrase pairs are carefully annotated by native Turkish speakers with the identified semantic correspondences between paraphrases. The work for expanding the corpus is still under way.

Keywords: Paraphrasing, Corpus Creation, Turkish

1. Introduction

A semantic meaning can be expressed by different expressions in a language. Variations in syntactic surface forms referring to the differences of texts with the same or almost the same meaning are usually called as paraphrases. Paraphrasing is inherent to speakers of all languages who can subconsciously use or recognize paraphrases. Paraphrases are frequently observed in natural texts at three different levels which correspond to different units of language bearing similar meaning (i.e., *lexical, phrasal,* and *sentential*). Lexical paraphrasing replaces words within a phrase with other words such as synonyms (e.g., "Rich *buys* the tickets from the box office" and "Rich *gets* the tickets from the box office") and phrasal paraphrasing substitutes phrasal fragments by other phrases (e.g., "My mentor *wrote* that book." and "My mentor *was the author of* that book"). On the other hand, sentential paraphrasing rephrases entire sentences. For example, the sentence "*May I borrow your textbook?*" can be rephrased by changing its modality as "*I am wondering if I could borrow your textbook.*".

Since exploring language variability and eliciting semantic equivalences are critical for many natural language applications, paraphrasing has been extensively studied in the field (Androutsopoulos and Malakasiotis, 2010). In multi-document summarization, paraphrasing has been shown to be of great help in avoiding redundancy in relevant sentences selected for inclusion in the summary (Barzilay and McKeown, 2005). A question answering (QA) system should deal with the linguistic variability of questions and answers since the input question may be phrased differently than its candidate answers. Unfortunately, QA systems often return substantially different answers for semantically equivalent input questions (Duboue and Chu-Carroll, 2006). Previous research has demonstrated that taking paraphrases into account significantly improved the performance of QA systems (France et al., 2003; Riezler et al., 2007). Paraphrasing techniques have been leveraged by statistical machine translation systems (SMTs) in order to improve the translation quality. One particular use was to populate the set of reference translations with automatically generated sentential paraphrases of human-authored reference translations (Madnani et al., 2007). In addition, several SMTs have benefited from paraphrasing to use the available translations of paraphrases for unknown source language phrases (Callison-Burch et al., 2006). Other common applications of paraphrasing include query expansion (Jones et al., 2006), information extraction (Sekine, 2006), and language generation (Power and Scott, 2005).

In the last decade, four different types of corpora (Madnani and Dorr, 2010) have been used by data-driven paraphrasing approaches: i) single monolingual corpus consisting of a very large collection of documents, ii) monolingual parallel corpus that consists of semantically equivalent (or almost equivalent) sentence pairs (e.g., multiple translations of the same literary text), iii) monolingual parallel corpus consisting of sentence pairs which overlap in the information or topic they convey (e.g., news articles about the same event published by different agencies), and iv) bilingual parallel corpus that consists of semantically equivalent parallel sentences in two (or more) languages (e.g., parallel materials in in-flight magazines). There are a number of publicly available paraphrase corpora for different languages with varying levels of detail (i.e., paraphrase annotations) and shortcomings such as (Dolan and Brockett, 2005) [5801 paraphrase pairs] and (Cohn et al., 2008) [900 pairs] for English, and (Fujita and Inui, 2005) [2301 pairs] for Japanese. However, to our best knowledge, there is not any available Turkish paraphrase corpus in the literature.

This paper presents our efforts aiming at building the first Turkish paraphrase corpus on a large scale. We have created the corpus by drawing parallel sentences from four different sources. These were multiple translations of a liter-

ary text, two different subtitles of a movie, multiple reference translations of a parallel corpus, and human-written paraphrases of news sentences. Although we have collected a very large amount of paraphrase pairs, the current version of the corpus contains 1270 paraphrastic sentences with human-annotated word and phrase alignments. We argue that our paraphrase corpus, which is continually expanded, will be of great use for the development and evaluation of Turkish paraphrasing systems. The rest of this paper is organized as follows. Section 2. describes the sources that were used for collecting paraphrase pairs. Section 3. presents the methodology that was followed for annotating paraphrase pairs. Section 4. describes the representation of the annotated paraphrases pairs. Finally, Section 5. concludes the paper and discusses future research.

2. Collecting Paraphrases

We compiled our paraphrase corpus from four different sources: i) Turkish translations of a famous novel, ii) Turkish subtitles of a foreign movie, iii) Turkish reference translations from an English-Turkish parallel corpus, and iv) Turkish articles from a news website.

Our first source is a famous English novel "For whom the Bell Tolls" written by Ernest Hemingway in 1940. We gathered two Turkish translations of the novel which differ in the number of sentences (both have approximately 14K sentences) and groupings of these sentences into paragraphs. In order to save time and effort, we first automatically sentence aligned (Moore, 2002) the translations to obtain a set of parallel sentences. Since automatic alignments may be inaccurate in terms of the semantic overlap between corresponding sentences, we asked a native speaker who has expertise in natural language processing to carefully examine all alignments and eliminate sentence pairs which diverge semantically more than some degree (i.e., those that most probably would not be aligned by a native speaker). The remaining more or less semantically overlapped parallel sentences formed our first set of paraphrase pairs.

Subtitlers not only translate but also paraphrase the textual version of a movie in such a way that the viewers will understand the movie. Thus, different subtitles produced for the same movie are a rich source for acquiring paraphrases. We collected two Turkish subtitles of the 1991 thriller movie "The Silence of the Lambs". In this case, the biggest advantage is that sentence pairs are renderings of the same semantic content by different subtitlers. Our second set of paraphrase pairs consisted of parallel sentences from that already sentence-aligned parallel corpora.

There exist a number of multilingual parallel corpora used for developing machine translation systems for different language pairs. Such multilingual corpora with multiple reference translations offer diverse examples of paraphrases. For extracting paraphrase pairs, we exploited the Turkish-English conversational phrases of the BTEC 2004 corpus (Basic Travel Expression Corpus) which consists of a collection of tourism-related sentences. For the Turkish-English language pair, the BTEC 2004 corpus contains 500 English sentences in the test set along with 16 Turkish reference translations for each sentence. We produced all possible Turkish paraphrases (i.e., 120 paraphrase pairs) for each test sentence by pairing the reference translations of that sentence and populated our corpus with those paraphrases.

We finally collected paraphrase pairs from Turkish native speakers by asking them to paraphrase the given sentences. For this, we assembled a corpus of Turkish news articles (approximately 29K) from the Southeast European Times website which publishes articles on daily events, business, politics, and sports from across and about the region in ten languages. Each collected article had 10-30 sentences on average. We presented 12 native speakers with a set of sentences randomly drawn from the collected articles and asked them to paraphrase each sentence so that it remains the same information. The participants were told to use only the information contained in the sentence and not to rely on commonsense knowledge. A different set of 20 sentences was given to each participant where each set contained at most one sentence from the same article. After all, this study produced 240 paraphrase pairs from the news domain.

We had to ensure that all paraphrase pairs that we collected are semantically equivalent or contain almost the same meaning in different wording. As a final step, we eliminated paraphrase pairs from our corpus where one of the sentences implies the other, but not the other way around. Three PhD graduates with mother-tongue Turkish and a background in natural language processing addressed this task. Each identified paraphrase pair was examined by 2 of these native speakers and a judgement was made whether the corresponding sentences should be considered as paraphrases of each other or not. The agreement between these speakers was moderate with a kappa of 0.416. The disagreements were resolved by the third speaker.

3. Annotating Paraphrases

We collected a very large amount of Turkish paraphrase pairs from different domains. Parallel sentences in those pairs convey the same (or almost the same) meaning in different wording, thus should have parts (e.g., words or phrases) in correspondence. In order to identify such correspondences and to annotate each pair accordingly, we developed an easy-to-use annotation tool which is shown in Figure 1. The tool displays paraphrase pairs on the left and allows the user to select one pair at a time in order to mark semantic correspondences within the pair. The user can click on a word or a sequence of words from both sentences and select the strength of correspondence between the highlighted words (via a different color) by pressing either the "certain alignment" or the "possible alignment" button. We offer two strength types in order to enable users to differentiate sequences of words that are strongly in correspondence than those having a loose correspondence. The tool does not allow a word to be a part of two different alignments. The "unalign" button

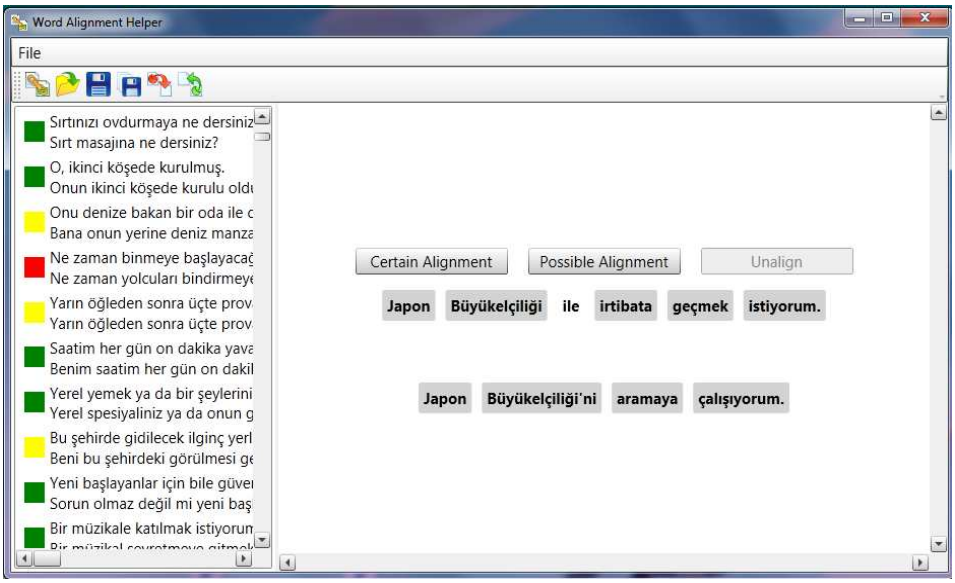

Figure 1: Word alignment tool.

enables the user to unalign previously aligned sequences of words in the selected pair.

We asked 14 native Turkish speakers, who did not participate in earlier studies, to annotate the presented paraphrase pairs with word and phrase alignments. Prior to the study, each participant was trained with an annotation guideline describing what is expected in the study and how the tool can be used for annotation task. The participants were told that three kinds of alignments (i.e., one to one, one to many, and many to many) can be used to mark the semantic correspondences. The participants were also told that certain alignments (if applicable) and smaller alignments (whenever possible) should be preferred. For example, a number of small certain alignments between parallel sentences should be preferred to one big possible alignment where a long sequence of words in one sentence is aligned with another long sequence of words in the other sentence. Moreover, the participants were told to align as many words as possible (ideally all words in parallel sentences). The participants were divided into two groups of seven and each group annotated one half of the paraphrase pairs (i.e., 635 pairs) contained in the current corpus. Table 1 shows the number of annotated paraphrase pairs from each source[1].

Source	Number of Pairs
Literary Text	482
Subtitle	108
Parallel Corpus	440
News Articles	240

Table 1: Annotated paraphrase pairs in the corpus

For each paraphrase pair, we finally examined all alignments produced by the participants and determined a common alignment for that pair. The common alignment of a pair contains alignments of word sequences marked by at least 4 participants (out of 7) that were presented with that paraphrase pair. If the same alignment was both annotated with certain and possible correspondences, the strength type selected by the highest number of participants was used. In cases of equality, the alignment was annotated with a possible correspondence.

4. Representing Paraphrases

The corpus provides two different representations for paraphrase pairs. Each representation presents a paraphrase pair along with its common alignment in GIZA++ format. The first representation (**Txt_R**) is a plain text representation whereas the second representation (**Xml_R**) is an XML-based representation. Consider, for example, the paraphrase pair shown in Figure 1:

- Japon$_1$ Büyükelçiliği$_2$ ile$_3$ irtibata$_4$ geçmek$_5$ istiyorum$_6$.
  I'd like to contact the Japanese Embassy.
- Japon$_1$ Büyükelçiliği'ni$_2$ aramaya$_3$ çalışıyorum$_4$.
  I am trying to call the Japanese Embassy.

The Txt_R and Xml_R representations of that paraphrase pair are given in Figure 2. The common alignment of the pair shows that the first two words of the first sentence (*Japon* and *Büyükelçiliği*) have certain correspondences with the first two words of the second sentence (*Japon* and *Büyükelçiliği'ni*) respectively. On the other hand, the fourth and fifth words of the first sentence (*irtibata geçmek*) are aligned to the third word of the second sentence (*aramaya*) with a possible alignment. Similarly, the sixth word of the first sentence (*istiyorum*) is aligned to the fourth word of the second sentence (*çalışıyorum*) with a possible alignment. The third word of the first sentence (*ile*) is not aligned to any word of the second sentence.

[1]The annotation task is in progress for the remaining paraphrase pairs.

Txt_R Representation:

Paraphrase Pair (#1):
Source length: 6 Target length: 4
Sentence (#1): Japon Büyükelçiliği ile irtibata geçmek istiyorum.
Sentence (#2): Japon Büyükelçiliği'ni aramaya çalışıyorum.
Alignment: NULL ({} CERTAIN) Japon ({1} CERTAIN) Büyükelçiliği ({2} CERTAIN) ile ({}) irtibata ({3} POSSIBLE) geçmek ({3} POSSIBLE) istiyorum. ({4} POSSIBLE)

Xml_R Representation:

```
<Paraphrase Pair="1" Source_length="6" Target_length="4">
 <Sentence Num="1">
  Japon Büyükelçiliği ile irtibata geçmek istiyorum.
 </Sentence>
 <Sentence Num="2">
  Japon Büyükelçiliği'ni aramaya çalışıyorum.
 </Sentence>
 <Alignment>
  NULL ({} CERTAIN) Japon ({1} CERTAIN) Büyükelçiliği ({2} CERTAIN) ile ({})
  irtibata ({3} POSSIBLE) geçmek ({3} POSSIBLE) istiyorum. ({4} POSSIBLE)
 </Alignment>
</Paraphrase>
```

Figure 2: The Txt_R and Xml_R representations.

	0	1	2	3	4	
0	*	*	*	*	*	NULL
1	*	C	*	*	*	Japon
2	*	*	C	*	*	Büyükelçiliği
3	*	*	*	*	*	ile
4	*	*	*	P	*	irtibata
5	*	*	*	P	*	geçmek
6	*	*	*	*	P	istiyorum.
	NULL	Japon	Büyükelçiliği' ni	aramaya	çalışıyorum.	

Figure 3: Alignment matrix.

For each paraphrase pair, the corpus also provides an alignment matrix which visually presents the common alignment of the pair. For example, Figure 3 shows the alignment matrix provided for the paraphrase pair given in Figure 1.

5. Conclusion

In this paper, we present our efforts towards building the first large-scale Turkish paraphrase corpus which we believe will trigger indepth studies on Turkish paraphrasing in the future. The corpus contains 1270 paraphrastic sentences drawn from four different sources. Each paraphrase pair is annotated by native speakers with word and phrase alignments. All paraphrase pairs along with their common alignments are represented via plain text and XML-based representations. For visualization purposes, an alignment matrix is also provided for each paraphrase pair. We consider other possible directions towards further developments of this work. For instance, we currently work on extending the corpus to other domains as well as enhancing the corpus with semantically related but not paraphrastic sentence pairs. Such non-paraphrase pairs would be helpful for the development of machine learning systems on this corpus.

6. Acknowledgement

The authors would like to thank the study participants, especially the members of the MTRD group, for their willingness to participate in the study. The authors would also like to extend their appreciation to Coskun Mermer for the script used for generating alignment matrices. This work was funded inpart by FP7-REGPOT-2008-1 project, MULTISAUND (MULTilingualism Integrated to Speech and Audio UNDerstanding).

7. References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *JAIR*, 38:135–187.

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31:297–328.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne.

2006. Improved statistical machine translation using paraphrases. In *HLT-NAACL '06*.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*, 34:597–614.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *IWP2005*.

Pablo Ariel Duboue and Jennifer Chu-Carroll. 2006. Answering the question you wish they had asked: the impact of paraphrasing for question answering. In *HLT-NAACL*.

Florence Duclaye France, Franois Yvon, and Olivier Collin. 2003. Learning paraphrases to improve a question-answering system. In *EACL Workshop NLP for Question-Answering*.

Atsushi Fujita and Kentaro Inui. 2005. A class-oriented approach to building a paraphrase corpus. In *IWP2005*.

Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. 2006. Generating query substitutions. In *WWW*.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, pages 341–387.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie J. Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *StatMT*.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *AMTA*.

Richard Power and Donia Scott. 2005. Automatic generation of large-scale paraphrases. In *IWP*.

Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu O. Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *ACL*.

Satoshi Sekine. 2006. On-demand information extraction. In *COLING-ACL*.