Using Multiple Metrics in Automatically Building Turkish Paraphrase Corpus

Bahar Karaoğlan[1], Tarık Kışla[2], Senem Kumova Metin[3],
Ufuk Hürriyetoğlu[1], Katira Soleymanzadeh[1]

[1] Ege University, International Computer Institute, İzmir, Turkey
bahar.karaoglan@ege.edu.tr, {ufuk.hurriyetoglu,
katira.sole}@gmail.com
[2] Ege University, Department of Computer Education and Instructional Technology
tarik.kisla@ege.edu.tr
[3] İzmir University of Economics, Department of Software Engineering
senem.kumova@ieu.edu.tr

Abstract. Paraphrasing is expressing similar meanings with different words in different order. In this sense it is viewed as translation in the same language. It is an important issue in natural language processing for automatic machine translation, question answering, text summarization and language generation. Studies in paraphrasing can be classified as paraphrase extraction, paraphrase generation, paraphrase recognition. In this paper we present automatic sentential paraphrase extraction from comparable texts downloaded from Turkish newspapers related to similar news. We applied seven text similarity metrics and assumed the two most similar ones as candidates. Through an interface these are shown to 3 human annotators to be labelled as paraphrase, entailing, entailed, opposite in meaning and not paraphrase. In this paper we only present results driven from a single topic. The sentences in the other topics will be processed based on the experience gained in the current work. This will be the first automatically built and golden standard tagged Turkish paraphrase corpus.

Keywords: Paraphrase extraction, Turkish paraphrase corpus

1 Introduction

Identification of paraphrasing is an important issue in natural language understanding and information retrieval. As the first requirement golden standard tagged corpus for the assessment purposes is needed. Paraphrase identification is not much studied in Turkish and there is no corpus developed for Turkish paraphrase identification. In this paper we present an incremental methodology for selecting candidate paraphrase sentences for the human annotators. Basic idea is using different text similarity metrics to measure the similarity of each sentence to all the other sentences within the same topic. Then, take the two most similar sentences obtained from each metric as the candidates to be shown to annotators.

To create a paraphrase corpus different types of sources can be used to extract data. Source of data to be selected depends on the granularity of paraphrasing which can be at the phrase [1] [2] [3] sentence [4] [5] [6] [7] or paragraph level [8] [9]. Most of the studies have been at phrase level for automatic machine translation, information retrieval and information extraction purposes or sentence level for question answering, text summary and such. Comparable texts (Newspaper articles for the same news from different sources), parallel texts (bilingual or monolingual translations of the same text, answers given to the same question (FAQ, exams, customer opinions, rephrasing), text modifications (Wikipedia) are some of the data sources.

Once the data is collected and pre-processed, the next task is to tag the text pairs as paraphrase by the human annotators to achieve golden standard or training purposes. To ease the task of the annotators, text similarity techniques are exploited to choose the candidates automatically. These candidates are then marked as paraphrase or not paraphrase by the annotators.

Paraphrasing is a vague concept by itself and its understanding may vary from person to person. Some definitions [10] are as follows:

Wikipedia:	A restatement of a text or passage using different words
WordNet:	Express the same message in different words, rewording for the purpose of clarification
Purdue's OWL:	Your own rendition of essential information and ideas expressed by someone else, presented in a new form.
Pearson's glossary:	To record someone else's words in the writer's own words

To achieve a standard to some extent, guidelines are set for the annotation [5]. Even then, the boundary is not clear. In some studies paraphrases are labelled with degree of confidence in a graded fashion. For example in STS [11] and ULPC [7] textual similarity is annotated on 6 scale from exact semantic equivalence to complete un-relatedness. The annotation task is either done through field experts [7] or through crowdsourcing [11].

Along with paraphrasing studies, some researchers [12] have made distinction between paraphrasing and entailment in the labelling. Paraphrasing is interpreted as a bidirectional relation where the same meaning is derived from both texts. Textual entailment is interpreted as a directional relation where one text can be inferred from the other, but the reverse is not true. "Precise paraphrase" is addressed if the relation is bidirectional: Text A is a paraphrase of text B if and only if A entails B and B entails A [13].

In this study the sentences are annotated on five scale as 1) Paraphrase, 2) Entails, 3) Entailed, 4) Opposite and 5) Not paraphrase. We considered the first 3 as paraphrase and the last two as not paraphrase. Sentence 5, given as example below, is taken from our news database. Sentence 79 is labelled as paraphrase and sentence 1 is marked as entailed. Both sentences 5 and 79 give approximately the same message. Sentence 1 can be deduced from these sentences.

Sentence 05:
In the implementation, it is foreseen that shareholders use credit approximately 115% of their accretion that they get in four instalments up till now.

(Turkish: Uygulamada, hak sahiplerinin şimdiye kadar dört taksitte aldıkları nemanın yaklaşık ``yüzde 115`i kadar`` kredinin kullandırılması öngörüldü)

Sentence 79 (Paraphrase):
Credit will be given up to 115% of the total accretion that is received in 4 instalments till today.
(Turkish: Bugüne kadar 4 taksitte alınan toplam nemanın yüzde 115`ine kadar kredi verilecek)

Sentence 01 (Entailed):
To those who wish, Ziraat Bank will advance money on their receivable accretion.
(Turkish: Ziraat Bankası, isteyenlere alacakları nemalar karşılığında kredi kullandıracak.)

The assessment of the most linguistic studies heavily depends on the tagged corpus on which it is carried on. To increase the reliability of the golden standard, annotation is done by several experts and inter annotator agreement is calculated. The performance of paraphrase recognition approach is assessed with regard to the extent by which it correlates with human annotators as human annotators correlate with each other. For paraphrase studies, even in the cases where annotation rules are rigid, as in MSRP, high inter-annotator agreements cannot be achieved.

2 Related Work

Androutsopoulos and Malakasiotis [14] classify the studies in paraphrasing field with respect to two dimensions: 1) whether paraphrasing or entailment, 2) processing of paraphrases: generation, extraction or recognition. Since the main goal of this study is to create a Turkish gold-standard paraphrase corpus, we will focus on methodologies for paraphrase extraction. Paraphrase corpora have been developed for different purposes from different sources. We look at these studies from aspects given in Table 1.

Table 1. Classification of paraphrase corpora

Aspects	**Explanation**
Source	Comparable corpora (eg. News about the same event); bilingual corpora (using one language as a pivot to find paraphrases in other documents.), monolingual corpora (parallel translations of the same source), users (question answers, rephrases, twitter, Wikipedia, etc.)
Annotation	Automatic, experts, crowdsourcing
Granularity	Paraphrase, sentence, paragraph
Recognition basis	Syntactic similarity, semantic similarity, text alignment, word overlap
Rating	Binary, scaled
Purpose	Information retrieval, Automatic machine translation, Language generation, Question answering, Summarization

Microsoft Research Paraphrase corpus (MSRP) can be considered as the first major public paraphrase corpus [5] [15] annotated by humans in binary mode as paraphrase or not paraphrase. Two methods: string similarity measure and discourse-based heuristic are used to draw candidate sentences from news after applying support vector machine classifier. It consists of 5801 pairs of sentences, of which 67% are judged to be paraphrases.

User Language Paraphrase Corpus (ULPC) [7] is composed of 1998 sentence pairs taken from students rephrases in response to target sentences. To describe the quality of user response, 10 dimensions (garbage, frozen expression, irrelevant, elaboration, writing quality, semantic similarity, lexical similarity, entailment, syntactic similarity, paraphrase quality) of paraphrasing is considered. The annotators were asked to rate between 1-6 interval (1: minimum, 6: maximum) with equal distance that is, 1 and 6 denote negative or positive with maximum confidence, whereas 3 and 4 denote negative or positive with minimum confidence. The main purpose in posing this challenge was to facilitate intelligent tutoring systems to provide users with feedback comparable to those of experts.

Barzilay and McKeown [16] is an example to monolingual technique for corpus construction for the purpose of paraphrase extraction where, multiple English translations of the same literary text are used. Sub-sentential paraphrases were labelled as true and false by human annotators. They achieved 69% of accuracy in extracting paraphrases.

Two corpora in different languages are used to extract paraphrases in bilingual approach, taking one language as the pivoting language. Translations for phrases in the targeted language are found in the pivoting language using statistical and automatic machine translation techniques. Then, going backwards, translations for each of these in the targeted language are assumed to be paraphrase candidates. Colin and Callison-Burch [17] used the German-English the French-English, Spanish-English, and Italian-English portions of the Europarl corpus as sources. They report an accuracy of paraphrases extracted over multiple corpora as 57.4%.

Regneri et. al extracted paraphrase fragments form paraphrase sentences. With the aim of generality, they used sentential paraphrases from four different corpora: The Microsoft Paraphrase Corpus (MSR) [18] , The Microsoft Video Description Corpus (MSVD) [19], The TACoS Corpus [20], The "House" Corpus [21]. Two annotators labelled each ordered fragment pair as paraphrases, containment, backwards containment, unrelated or invalid. The overall annotator agreement was 0.50, according to Cohen's Kappa (moderate agreement). Conflicts were resolved by a third annotator.

Agirre et al.'s [11] STS (The Semantic Textual Similarity) CORE corpus contains 2,250 pairs of headlines, machine translation evaluation sentences, and glosses (concept definitions). This corpus was annotated through crowdsourcing on 6-value scale as: 5: identical, 4: strongly related, 3: related, 2: somewhat related, 1: unrelated, 0: completely unrelated.

Bernhard et al. [22] developed QP (The Question Paraphrase Corpus) with the purpose of better understanding of how questions in social Q&A sites can be automatically analyzed and retrieved. 1000 questions and their paraphrases (in total 7434) are collected from randomly selected FAQ files in the Education category of the WikiAnswers web site. They report 80% accuracy for the task of question paraphrase retrieval.

3 Methodology

Our dataset is driven from Turkish BilCon2005 [23] news corpus which was created for the purpose of event detection and tracking. This corpus contains 209.305 news collected from five different Turkish news web sources: CNN Türk (http://www.cnnturk.com), Haber 7 (http://www.haber7.com), Milliyet Gazetesi (http://www.milliyet.com.tr), TRT (http://www.trt.net.tr), Zaman Gazetesi (http://www.zaman.com.tr) throughout the year 2005. 5872 of these news are profiled with Topic Title, Event Summary, What, Who, When, Where and other relevant data. This news is then categorized into 13 topics like natural disasters, accidents, bank, elections, and etc.

For the preliminary study we chose randomly the topic "Bank" and parsed it into sentences. After removing all the duplicates and short sentences with less than 3 words, 399 sentences are left. The average length of the sentences is 17.21words, with the shortest 3 words and the longest 74 words.

We then calculated the distance of each sentence to all other sentences with 7 different distance metrics: Chebyshev, Cosine, Euclid, Hamming, City Block, Correlation and Spearman. For each sentence, we selected two sentences with the least distance calculated by each metric to be marked by the human annotators via a user interface as shown in Fig.1 with five marking options as: Paraphrase, Entailing, Entailed, opposite, not-Paraphrase. The target statement is shown on top of the screen. Three annotators labelled each sentence in the list with a label provided via pull down menu. Annotators marked the sentence as paraphrase when they believe the sentence gives the same or very similar meaning with the target sentence. Annotators marked the sentence as entailed when they think that it can be inferred from the target statement and entailing vice versa. The final decision is made if at least two annotators marked the same choice.

4 Results

Each of 7 similarity metric considered proposes 2 candidates as paraphrase for every sentence in the set of 399 sentences. So, we have 399*14= 5586 sentences proposed as candidates. By further eliminating the same candidates proposed by different metrics we are left with 2472 sentences to be annotated by humans. Symmetric measures of the Kappa analysis for three annotators with initial "S", "T", and "U" are given in Table 2. The second row shows the symmetric agreement ratios between the annotators when the annotations are quantized to 1 (Paraphrase - Paraphrase /entailed /entailing) and 0 (Not Paraphrase - opposite /notParaphrase). The third row shows the symmetric agreement ratios between the annotators on 5 scale detail.

The sentences that are symmetrically labelled as paraphrases of each other, i.e., for given sentences A and B, if A is labelled as paraphrase of B and B is labelled as paraphrase of A, are interpreted as true paraphrases. The annotators have marked 147 (6%) pairs as paraphrases, 85 of which are bidirectional, and 62 of which are unidirectional.

Table 3 shows the percentage ratios (number of TP/(399*2) of true positives of the metrics.

Fig. 1. Annotation interface

Table 2. Kappa results for Inter-annotator agreement

Ranking	S-T	S-U	U-T
***Binary Scale*:** Paraphrase / notParaphrase	0.87	0.75	0.81
5 Scale: Paraphrase / entailed / entailing / opposite / notParaphrase	0.83	0.70	0.77

Table 3. True Positive Percentages of the Text Similarity Metrics

Sim-Metric	% of TP
Chebyshev	0,085
Cosine	0,343
Euclid	0,243
Hamming	0,243
CityBlock	0,248
Correlation	0,338
Spearman	0,333

As seen from Table 3 similarity metrics group into three. Chebyshev performs the least, in the middle performing group we have Euclid and Hamming and in the most performing group we have, Cosine, Correlation and Spearman.

Rus (2014) argues that paraphrase sentences extracted from the same news have large word overlap. Table 4 gives the word overlap ratios of the sentence pairs marked as paraphrase and not paraphrase by the annotators.

Table 4. The word overlap ratios of the sentence pairs

	Max	Min	Avg.
Paraphrase	0.98	0.0	0.49
Not paraphrase	0.86	0.0	0.12
All sentences	0.98	0.0	0.14

5 Conclusion

In this paper, we present preliminary work for the first stage, Turkish paraphrase corpus development, of a study with the ultimate goal of paraphrase recognition. The method is based on applying different text similarity metrics on the sentences driven from similar topics and choosing the sentences with the highest similarity scores as the candidates. Here we noticed that that some of the methods are not functioning well for measuring similarity at sentence level for Turkish. We will continue processing BilCon2005 Corpus, topic by topic, with improved data cleaning, pre-processing and choice of text similarity metrics.

As it is argued in [10], paraphrase sentences extracted from the same news have large word overlap which is in conflict with the definition of "expressing the same meaning with different (own) words". Most of the words in our sentence pairs overlap as in MSRP sample sentences given below.

Text A: *York had no problem with MTA's insisting the decision to shift funds had been within its legal rights.*

Text B: *York had no problem with MTA's saying the decision to shift funds was within its powers.*

Our next goal is to foster this corpus with sentences obtained from several other sources with different nature. We are planning to include sentences from different translations of the same texts, paraphrases from Turkish Language level assessment exams in order to enrich the corpus by including broader range of linguistic phenomena [10] and challenge the problem on a wider space.

Acknowledgement

This work is carried under the grant of TÜBİTAK – The Scientific and Technological Research Council of Turkey to Project No: 114E126, Using Certainty Factor Approach and Creating Paraphrase Corpus for Measuring Similarity of Short Turkish Texts.

References

1. Quirk, C., Raghaverda, U., Arul, M.: Generative models of noisy translations with applications to parallel fragment extraction. In: MT Summit XI, Copenhagen, Denmark, pp. 321–327 (2007)
2. Regneri, M., Koller, A., Pinkal, M.: Learning Script Knowledge with Web Experiments. In: the Association for Computational Linguistics (2010)
3. Wang, R., Callison-Burch, C.: Paraphrase fragment extraction from monolingual comparable corpora. In: the ACL BUCC-2011 Workshop (2011)
4. Barzilay, R., Lee, L.: Learning to paraphrase: An unsupervised approach using multiple sequence alignment. In: HLT-NAACL 2003. (2003)

5. Dolan, B., Quirk, C., Brockett, C.: Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In: COLING 2004 (2004)
6. Quirk, C., Brockett, C., Dolan, W.: Monolingual Machine Translation for Paraphrase Generation. In: the 2004 Conference on Empirical Methods in Natural Language Processing, pp.142-149 (2004)
7. McCarthy, P., McNamara, D.: The user-language paraphrase challenge. In: Special ANLP topic of the 22nd International Florida Artificial Intelligence Research Society Conference, Florida (2008)
8. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In Benno Stein, P., ed.: SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), pp.1-9 (2009)
9. Lintean, M., Vasile, R., Azevedo, R.: Automatic Detection of Student Mental Models during Prior Knowledge Activation in MetaTutor. International Journal of Artificial Intelligence in Education 21(3), 169-190. (2011)
10. Rus, V., Banjade, R., Lintean, M.: On Paraphrase Identification Corpora. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, pp.2422-2429 (2014)
11. Agirre, E., Cer, D., Diab, M., González-Agirre, A., Guo, W.: SEM 2013 shared task: Semantic Textual Similarity. In: the Second Joint Conference on Lexical and Computational Semantics, Atlanta, Georgia, USA, vol. volume 1. Association for Computational Linguistics, pp.32-43 (2013)
12. Dagan, I., Glickman, O., Magnini, B.: The PASCAL recognising textual entailment challenge. In Qui˜nonero-Candela, J. ., ed.: Machine Learning Challenges. Lecture Notes in Computer Science, vol. 3944, pp.177-190 (2006)
13. Rus, V., McCarthy, P., Lintean, M., McNamara, D., Graesser, A.: Paraphrase identification with lexico-syntactic graph subsumption. In Sutcliffe, D., ed.: the 21st International Florida Artificial Intelligence Research Society Conference, pp.201-206 (2008)
14. Androutsopoulos, I., Malakasiotis, P.: A survey of paraphrasing and textual entailment methods. Journal of Artificial Intelligence Research 38, 135-187 (2010)
15. Brockett, C., Dolan, W.: Support Vector Machines for Paraphrase Identification and Corpus Construction. In: Third International Workshop on Paraphrasing (IWP2005) (2005)
16. Barzilay, R., Kathy, M.: Extracting paraphrases from a parallel corpus. In: the 39th Annual Meeting of the Association for Computational Linguistics, Toulouse, France, pp.50-57 (2001)
17. Bannard, C., Callison-Burch, C.: Paraphrasing with Bilingual Parallel Corpora. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan, pp.597-604 (2005)
18. Dolan, W., Brockett, C.: Automatically Constructing a Corpus of Sentential Paraphrases. In: Third International Workshop on Paraphrasing (IWP2005)

19. Chen, D., Dolan, W.: Collecting Highly Parallel Data for Paraphrase Evaluation. In: the proceedings of The 49th Annual Meetings of the Association for Computational Linguistics (ACL), Portland (2011)
20. Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., Pinkal, M.: Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics 1, 25-36 (2013)
21. Regneri, M., Wang, R.: Using discourse information for paraphrase extraction. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp.916-927 (2012)
22. Bernhard, D., Gurevych, I.: Answering learners' questions by retrieving question paraphrases from social Q&A sites. In: the Third Workshop on Innovative Use of NLP for Building Educational Applications, Ohio, USA, pp.44-52 (2008)
23. Can, F., Kocberber, S., Baglioglu, O., Kardas, S., Ocalan, H., Uyar, E.: New event detection and topic tracking in Turkish. Journal of the American Society for Information Science and Technology 61(4), 802-819 (2010)