Semantic Similarity Based Evaluation for Abstractive News Summarization

Figen Beken Fikri[1], Kemal Oflazer[2] , Berrin Yanıkoğlu[1]

[1]Department of Computer Science and Engineering, Sabancı University, Istanbul, Turkey
[2]Department of Computer Science, Carnegie Mellon University - Qatar, Doha, Qatar
[1]{*fbekenfikri,berrin*}*@sabanciuniv.edu*, [2]*ko@andrew.cmu.edu*

Abstract

ROUGE is a widely used evaluation metric in text summarization. However, it is not suitable for the evaluation of abstractive summarization systems as it relies on lexical overlap between the gold standard and the generated summaries. This limitation becomes more apparent for agglutinative languages with very large vocabularies and high type/token ratios. In this paper, we present semantic similarity models for Turkish and apply them as evaluation metrics for an abstractive summarization task. To achieve this, we translated the English STSb dataset into Turkish and presented the first semantic textual similarity dataset for Turkish. We showed that our best similarity models have better alignment with average human judgments compared to ROUGE in both Pearson and Spearman correlations.

1 Introduction

Automatic document summarization aims to produce a summary that conveys the salient information in the given text(s). Automatic summarizers provide reduction in the size of the text, as well as, combine and cluster different sources of information, while preserving the informational content. There are two approaches to summarization: extractive and abstractive. Extractive summarization yields a summary by extracting important phrases or sentences from the document. In contrast, abstractive summarization provides a much more human-like summary by capturing the internal semantic meaning and generating new sentences.

ROUGE is a widely used evaluation metric in text summarization. It compares the system summary with the human generated summary or summaries, by considering the overlapping units such as n-gram, word sequences and word pairs (Lin, 2004). However, in abstractive summarization systems, the generated summary does not necessarily contain the same words in the gold standard summary. On the contrary, an abstractive summarization model is expected to generate new words that may not even appear in the source. For agglutinative languages, the ineffectiveness of ROUGE metric becomes more apparent. For instance, both of the following sentences has the meaning "I want to call the embassy":

Büyükelçiliği aramak istiyorum.

Büyükelçiliğe telefon etmek istiyorum.

While, "aramak" is a verb that takes an object in accusative case, "telefon etmek" is a compound verb in Turkish and the equivalent of the accusative object in the first sentence is realized with a noun in dative case (as highlighted with underlines). Although, these sentences are semantically equivalent, ROUGE-1, ROUGE-2 and ROUGE-3 scores of these sentences are 0.25, 0, and 0.25 respectively.

In this paper, we present a semantic similarity model which can be applied to abstractive summarization as a semantic evaluation metric. To this end, we translated the English Semantic Textual Similarity benchmark (STSb) dataset (Cer et al., 2017) into Turkish and presented the first semantic textual similarity dataset for Turkish as well. STSb dataset is a selection of data from English STS shared tasks between 2012 and 2017. These datasets have been widely used for sentence level similarity and semantic representations research (Cer et al., 2017).

We also leveraged the NLI-TR dataset that has been presented recently for Turkish natural language inference task (Budur et al., 2020). The NLI-TR dataset combines the translated Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and MultiGenre Natural Language Inference (MultiNLI) (Williams et al., 2018) datasets.

Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), pages 24–33

Our paper is structured in the following way: In section 2, we explain recent studies and evaluation metrics. In section 3, we explain natural language inference and semantic textual similarity. We present our STSb Turkish dataset and translation quality. In section 4, we present our experiments for semantic textual similarity. In section 5, we present the experiments for summarization. We applied our best performing four semantic similarity models as evaluation metrics to the summarization results. In section 6, we present our results both qualitatively and quantitatively by comparing the semantic similarity and ROUGE scores with human judgments in Pearson and Spearman correlations.

2 Related Work

The most widely used evaluation metric for summarization is ROUGE which compares the system summary with the human generated summary or summaries by considering the overlapping units such as n-gram, word sequences and word pairs (Lin, 2004). Recently, there has been a range of studies focusing on the evaluation of factual correctness in the generated summaries. Falke et al. (2019) has studied whether textual entailment can be used to detect factual errors in generated summaries based on the idea that the source document should entail the information in a summary. The authors investigated whether factual errors can be reduced by reranking the alternative summaries using models trained on NLI datasets. They found that out-of-the-box NLI models do not perform well on the task of factual correctness. Kryscinski et al. (2020) proposed a model-based approach on the document-sentence level for verifying factual consistency in generated summaries. Zhao et al. (2020) addressed the problem of unsupported information in the generated summaries known as factual hallucination. Durmus et al. (2020) and Wang et al. (2020) suggested question answering based methods to evaluate the faithfullness of the generated summaries.

In addition to the studies focusing on summarization evaluation, there are some recently proposed metrics to evaluate generated text with the gold standard. Zhang et al. (2019) proposed BERTScore that uses BERT (Devlin et al., 2019) to compute a similarity score between the generated and reference text. Several recent works proposed new evaluaiton metrics for machine translation (BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020), YiSi (Lo, 2019), Prism (Thompson and Post, 2020)).

3 Methodology

3.1 Natural Language Inference

Natural language inference is the study of determining whether there is an entailment, a contradiction or a neutral relationship between a hypothesis and a given premise. There are two major corpora in literature for natural language inference in English. These are Stanford Natural Language Inference (SNLI) (Bowman et al., 2015) and MultiGenre Natural Language Inference (MultiNLI) (Williams et al., 2018) datasets. The SNLI corpus is about 570k sentence pairs while the MultiNLI corpus is about 433k sentence pairs. The MultiNLI corpus is in the same format as SNLI, but with more varied text genres. Recently, these corpora have been translated into Turkish (Budur et al., 2020). In this study, we used the NLI-TR dataset.[1]

3.2 Semantic Textual Similarity

Semantic textual similarity aims to determine how similar two pieces of texts are. There are many application areas such as machine translation, summarization, text generation, question answering, dialogue and speech systems. It has become a remarkable area with the competitions organized by SemEval since 2012.

Semantic textual similarity studies are very common in English, and are based on datasets that are annotated and given similarity scores by human annotators. However, annotation is costly and time consuming. Recently, with the increase of success in machine translation and the development of multi-language models, it has become possible to use datasets by translating them from one language to another, e.g., Isbister and Sahlgren (2020), Budur et al. (2020).

In this study, we use the English STS Benchmark (STSb) dataset (Cer et al., 2017) that we translated into Turkish using the Google Cloud Translation API.[2,3] The STSb dataset consists of all the English datasets used in SemEval STS studies between 2012 and 2017. It consists of 8628 sentence pairs (5749 train, 1500 dev, 1379 test), (see Table 3

[1] NLI-TR dataset consists of the translations of SNLI and MultiNLI data sets available on GitHub: https://github.com/boun-tabi/NLI-TR

[2] https://cloud.google.com/translate/docs/basic/translating-text

[3] https://github.com/verimsu/STSb-TR

Sentence 1	Sentence 2	Similarity Score
Adam ata biniyor. (The man is riding a horse.)	Bir adam ata biniyor. (A man is riding on a horse.)	5.0
Bir kız uçurtma uçuruyor. (A girl is flying a kite.)	Koşan bir kız uçurtma uçuruyor. (A girl running is flying a kite.)	4.0
Bir adam gitar çalıyor. (A man is playing a guitar.)	Bir adam şarkı söylüyor ve gitar çalıyor. (A man is singing and playing a guitar.)	3.6
Bir adam gitar çalıyor. (A man is playing a guitar.)	Bir kız gitar çalıyor. (A girl is playing a guitar.)	2.8
Bir bebek kaplan bir topla oynuyor. (A baby tiger is playing with a ball.)	Bir bebek bir oyuncak bebekle oynuyor. (A baby is playing with a doll.)	1.6
Bir kadın dans ediyor. (A woman is dancing.)	Bir adam konuşuyor. (A man is talking.)	0.0

Table 1: Sample translations from STSb-TR dataset and the corresponding labels taken from the English dataset. Original English sentences are given in parenthesis.

for details). In this dataset, each sentence pair was annotated by crowdsourcing and assigned a semantic similarity score. Five scores were collected for each pair and gold scores were generated by taking the median value of these scores (Agirre et al., 2016). Scores range from 0 (no semantic similarity) to 5 (semantically equivalent) on a continuous scale. Some examples from the STS dataset and their translations are given in Table 1.

Here, we apply various state-of-the-art models on the translated dataset, and the best performing four models are used for semantic similarity based evaluation metric for the task of abstractive summarization.

3.3 Translation Quality

It is possible to encounter some translation errors in the translated texts. The most striking mistakes are related to expressions that are not used in Turkish. For instance, the sentence in S1 is translated as T1; however, a more appropriate translation would be C1, as ”sitting” is translated differently for inanimate subjects.

S1: Old green bottle sitting on a table.

T1: Bir masada oturan eski yeşil şişe.

C1: Bir masada duran eski yeşil şişe.

Another typical error is possessive agreement mismatch. For example, the sentence S2 is translated as T2 but the correct translation would be C2.

S2: Group of people sitting at table of restaurant.

T2: Bir grup insan restoran masada oturuyor.

C2: Bir grup insan restoran masasında oturuyor.

In this paper, we assumed that such translation errors will not cause a major problem in our similarity models. In order to verify our assumption, we tested the quality of translations by selecting 50 sentence pairs (100 sentences) randomly, considering the percentage of the categories in the dataset. So, 6, 19 and 25 pairs chosen from forum, caption and news categories respectively. These sentences were translated by three native Turkish speakers who are fluent in English. We evaluated quality of the system translations with the three references using BLEU (Papineni et al., 2002) score. We used the SacreBLEU[4] tool (Post, 2018) version 1.5.1 and found BLEU score as 60.21 which shows that our system translations can be considered as very high quality translations (Google). Therefore, no changes have been made to the translations.

Table 2 shows vocabulary size (cased and uncased), type/token ratio, average word length and average sentence length values for English and Turkish datasets.[5]

[4] https://github.com/mjpost/sacrebleu

[5] Only the punctuation marks around the word and at the end of sentences were deleted.

Language	Vocab Size (Cased)	Vocab Size (Uncased)	Type/Token Ratio	Avg Word Length	Avg Sentence Length
English	18,736	16,225	0.09	4.62	10.15
Turkish	29,461	26,649	0.19	6.20	8.26

Table 2: English and Turkish STSb dataset statistics. Vocab size is the word count and type/token ratio is the number of different words divided by the total number of words. Word length is the amount of characters in the word and sentence length is the number of words in a sentence.

	Train	Dev	Test	Total
News	3,299	500	500	4,299
Caption	2,000	625	625	3,250
Forum	450	375	254	1,079
Total	5,749	1,500	1,379	8,628

Table 3: STSb dataset statistics in terms of number of sentence pairs.

4 Experiments for Semantic Textual Similarity

In order to assess the semantic similarity between a pair of texts, there are two main model structures: 1) Sentence representation models that try to map a sentence to a fixed-sized real-value vectors called sentence embeddings. 2) Cross-encoders that directly compute the semantic similarity score of a sentence pair.

In this paper, we experimented with state-of-the-art sentence representation models that are applicable to Turkish (language-specific and multilingual models) and BERT cross-encoders. In sentence representation models, we obtained the semantic similarity scores using cosine similarity. All models were tested on the STSb-TR test dataset.

4.1 Sentence Representation Models

We experimented with LASER, LaBSE, MUSE, BERT, XLM-R and Sentence-BERT models as explained below.

LASER Language-Agnostic SEntence Representations (LASER) is a language model based on the BiLSTM encoder trained on parallel data targeting translation. The model has been trained in 93 languages, including Turkish.[6] In this study, Turkish sentence embeddings were computed using a pre-trained LASER model.

LaBSE Language-agnostic BERT Sentence Embedding (LaBSE) is a BERT variant masked and trained on multilingual data for translation language modeling. The model produces language-independent sentence embeddings for 109 languages, including Turkish (Feng et al., 2020). Similar to the LASER model, Turkish sentence embeddings were computed using a pre-trained LaBSE model.

MUSE Multilingual Universal Sentence Encoder (MUSE) model is a sentence embedding model trained on multiple languages at the same time. The model creates a common semantic embedding area for a total of 16 languages, including Turkish (Yang et al., 2020). In this study, CNN[7] and Transformer[8] models that are shared publicly in TensorFlow Hub are used.

BERT Bidirectional Encoder Representations from Transformers (BERT) is designed to pre-train deep bi-directional representations from unlabeled text by conditioning together in both left and right context on all layers (Devlin et al., 2019). In this study, BERTurk[9] and M-BERT[10] (Pires et al., 2019) models were used. Sentence embeddings were obtained by averaging the BERT embeddings.[11] In addition, the models were integrated into the Siamese network that we explained in section 4.1.

XLM-R RoBERTa Transformer model[12] has been trained on a large multilingual data using a multilingual masked language modeling goal (Conneau et al., 2020). In this study, we used the model to compute sentence embeddings similar to BERT models. We also integrated it into the Siamese network used in Sentence-BERT.

[6] https://github.com/facebookresearch/LASER

[7] https://tfhub.dev/google/universal-sentence-encoder-multilingual/3

[8] https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3

[9] https://huggingface.co/dbmdz/bert-base-turkish-cased

[10] https://huggingface.co/bert-base-multilingual-cased

[11] The output of the CLS vectors yields significantly lower results compared to the results obtained.

[12] https://huggingface.co/xlm-roberta-base

Sentence-BERT Sentence-BERT (SBERT) (also called Bi-Encoder BERT) is a modification of pre-trained BERT network (or other transformer models) using Siamese and ternary network structures (Reimers and Gurevych, 2019). The model derives close fixed-size sentence embedding in vector space for semantically similar sentences. The training loss function differs depending on the dataset the model was trained on. During the training on the NLI dataset, the classification objective function was used; whereas during the training on the STSb dataset, the regression objective function was used (Reimers and Gurevych, 2019).

The classification objective function concatenates the sentence embeddings by element-wise difference and multiplies by a trainable weight. The model optimizes the cross entropy loss:

$$o = \text{softmax}(W_t(u, v, |u - v|)), W_t \epsilon R^{3n \times k}$$

where n is the size of the sentence embedding, and k is the number of labels.

In the regression objective function, the cosine similarity between two sentence embeddings, optimize the models for mean square error loss.

4.2 Cross-Encoders

We adopted cross-encoder architecture as explained in Reimers and Gurevych (2019). In the cross-encoder, both sentences are passed to the network and a similarity score between 0 and 1 obtained; no sentence embeddings are produced.[13] We experimented with BERTurk, M-BERT, and XLM-R with training on NLI-TR and STSb-TR datasets.

4.3 Results for Semantic Textual Similarity

All models were individually trained on NLI-TR and STSb-TR training datasets. Also, the models trained on the NLI-TR dataset were fine-tuned on the STSb-TR dataset. All models were then tested on the STSb-TR test dataset.

We trained/fine-tuned the models on STSb-TR dataset with 4 epochs and 10 random seeds[14] as suggested by Reimers and Gurevych (2018; 2019). Then, we reported the average test results of 5 successful models that perform best on the validation set. The models were evaluated by calculating the Spearman and Pearson correlations between the estimated similarity scores and the gold labels. Table 4 shows the results as ρ x 100. According to the results, training the models first on the NLI-TR dataset increases the model performance. This is particularly noticeable for the XLM-R models. The BERTurk model also gives very good results when trained directly on the STSb-TR dataset. Here, we observe that the existing multilingual LASER, LaBSE, MUSE models without any training for semantic textual similarity, give very good results. Compared to these models, the performance of BERT models without training are quite low. The best results were obtained by training the BERTurk model on the NLI-TR dataset first, and then on the STSb-TR dataset.

Model	Pearson	Spearman
Not trained for STS		
Avg. BERTurk embeddings	54.48	55.23
Avg. M-BERT embeddings	50.44	50.43
Avg. XLM-R embeddings	20.22	41.81
LASER	69.86	70.18
LaBSE	72.24	71.74
MUSE-CNN	71.09	69.91
MUSE-Transformer	**76.32**	**74.84**
Trained on STS		
BERTurk + STS	**83.32**	**82.22**
M-BERT + STS	79.08	78.15
XLM-R + STS	79.18	78.56
S-BERTurk + STS	81.97	81.43
S-M-BERT + STS	73.28	72.84
S-XLM-R + STS	71.89	71.02
Trained on NLI + STS		
BERTurk + NLI + STS	**85.36**	**84.59**
M-BERT + NLI + STS	79.30	78.39
XLM-R + NLI + STS	81.94	81.21
S-BERTurk + NLI + STS	82.85	83.31
S-M-BERT + NLI + STS	75.74	75.41
S-XLM-R + NLI + STS	77.26	77.32

Table 4: Experiment results for semantic textual similarity. BERTurk, M-BERT and XLM-R are cross-encoder models. S-BERTurk, S-M-BERT and S-XLM-R are bi-encoder models. Pearson and Spearman correlations were reported as ρ x 100.

5 Experiments for Summarization

To investigate the effectiveness of our semantic similarity models for summarization evaluation, we computed the correlations of ROUGE scores and our best performing four similarity models with human judgments for a state-of-the-art abstractive model. We reported semantic similarity scores for extractive baselines as well in order to observe their alignmnet with the ROUGE scores.

[13] https://www.sbert.net/examples/applications/cross-encoder/README.html

[14] Only S-XLM-R + STS was trained with 20 random seeds to have at least 5 successful models.

	Cross-Encoder		Bi-Encoder		ROUGE			Other Metrics
Model	NLI+STS	STS	NLI+STS	STS	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Lead-1	52.11	55.71	59.18	61.67	26.56	17.31	25.31	73.72
Lead-3	**60.78**	**61.86**	**69.72**	**71.01**	30.04	18.90	28.83	74.15
mT5	59.00	61.03	66.43	68.29	**33.22**	**22.44**	**31.90**	**75.90**

Table 5: Results of the summarization models on MLSUM dataset. The values under Cross-Encoder are the average similarity scores predicted by the models; whereas, the values under Bi-Encoder are the average cosine similarities of sentence embeddings computed by these models. All the values were scaled to 100.

	Relevance		Consistency		Fluency		Human Average	
Metric	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
Rouge-1	42.79	43.87	28.18	32.36	21.40	20.30	36.79	37.51
Rouge-2	38.26	41.63	27.39	**35.78**	16.43	20.83	32.76	38.02
Rouge-L	41.83	41.95	26.29	28.85	20.17	18.63	35.15	35.11
BERTScore	45.49	45.75	25.14	22.47	24.74	19.85	37.88	38.07
S-BERTurk+STS	55.44	52.82	30.25	30.04	25.63	26.70	44.26	45.86
S-BERTurk+NLI+STS	58.77	58.72	32.80	32.67	31.24	**30.17**	48.80	51.85
BERTurk+STS	56.87	53.54	38.02	32.46	34.10	27.88	51.32	48.59
BERTurk+NLI+STS	**59.98**	**59.17**	**39.95**	34.24	**34.62**	29.31	**53.54**	**52.10**

Table 6: Pearson and Spearman correlations of ROUGE, BERTScore and proposed evaluation metrics with human judgments.

5.1 Dataset

MLSUM is the first large-scale MultiLingual SUMmarization dataset which contains 1.5M+ article/summary pairs including Turkish (Scialom et al., 2020). The authors compiled the dataset following the same methodology of CNN/DailyMail dataset. They considered news articles as the text input and their paired highlights/description as the summary. Turkish dataset was created from Internet Haber[15] by crawling archived articles between 2010 and 2019. All the articles shorter than 50 words or summaries shorter than 10 words were discarded. The data was split into train, validation and test sets, with respect to the publication dates. The data from 2010 to 2018 was used for training; data between January-April 2019 was used for validation; and data up to December 2019 was used for test (Scialom et al., 2020). In this study, we obtained the Turkish dataset from HuggingFace collection.[16] The dataset consists of 249,277 train, 11,565 validation, and 12,775 test samples.

5.2 Models

We experimented on MLSUM Turkish dataset with extractive baselines Lead-1 and Lead-3 and a state-of-the-art abstractive model mT5 described below.

Lead-1 We selected the first sentence of the source text as a summary.

Lead-3 We selected the first three sentences of the source text as a summary, based on the observation that the leading three sentences are a strong baseline for summarization (Nallapati et al., 2017; Sharma et al., 2019).

mT5 Multilingual T5 (mT5) (Xue et al., 2020) is a variant of T5 model (Raffel et al., 2020) that was pre-trained for 101 languages including Turkish on a new Common Crawl-based dataset. For Turkish summarization, we used mT5 model fine-tuned on MLSUM dataset available on HuggingFace.[17] The model was trained with 10 epochs, 8 batch size and 10e-4 learning rate. The max news length was 784 and max summary length was determined as 64.[18]

5.3 Evaluations

We evaluated the summarization models using semantic similarity-based evaluation, ROUGE scores, and human judgments. All the values were scaled to 100.

Semantic Similarity Evaluations We used the best performing four semantic similarity models to evaluate the summarization models. The values under Cross-Encoder are the average similarity scores predicted by the models; whereas, the values under Bi-Encoder are the average cosine similarities of sentence embeddings computed by these models.

[15] www.internethaber.com

[16] https://github.com/huggingface/datasets/tree/master/datasets/mlsum

[17] https://huggingface.co/ozcangundes/mt5-small-turkish-summarization

[18] During inference, we set max summary length to 120.

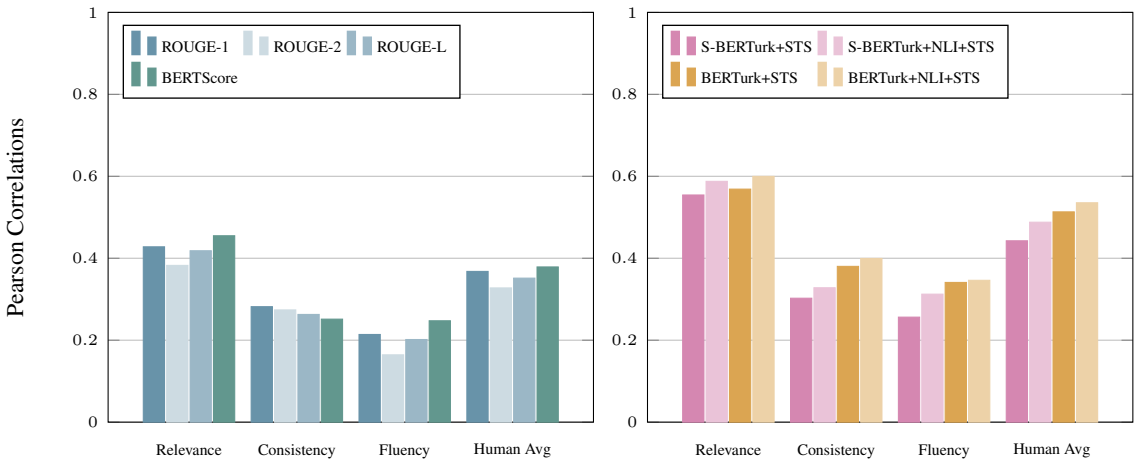

Figure 1: Pearson correlations between different evaluation metrics and human evaluations.

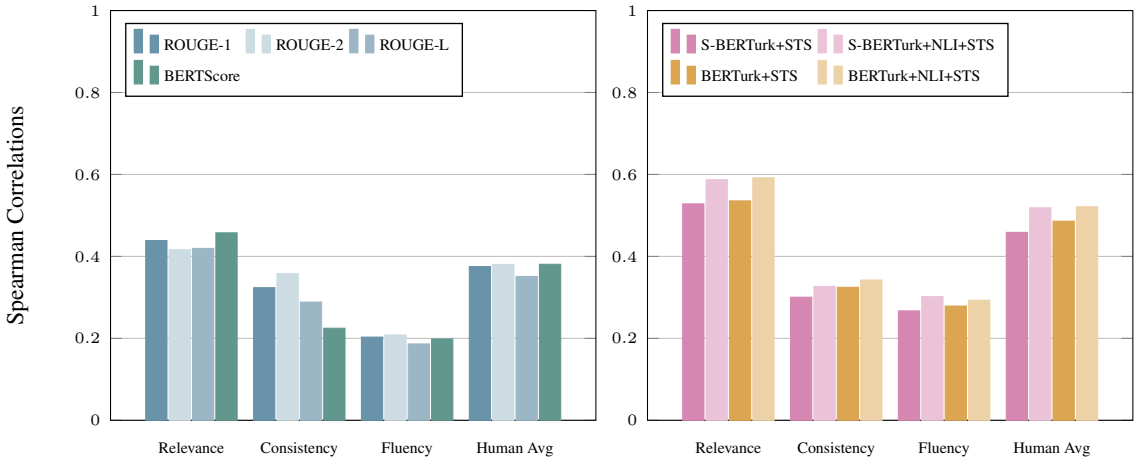

Figure 2: Spearman correlations between different evaluation metrics and human evaluations.

ROUGE We reported F1 scores for ROUGE-1, ROUGE-2 and ROUGE-L. ROUGE scores were computed using rouge package version 0.3.1.[19]

BERTScore We reported F1 score for BERTScore (Zhang et al., 2019).

Human Evaluations Human evaluations were conducted to show the effectiveness of our semantic similarity based evaluation metric. We randomly selected 50 articles from the test set with their predicted summaries via mT5 model. Following the work of Fabbri et al. (2021), we asked native Turkish annotators to rate each predicted summary in terms of relevance (selection of important content from the source), consistency (the factual alignment between the summary and the summarized source) and fluency (the quality of individual sentences) in the range of 1 (very bad) to 5 (very good). Overall, 5 annotators (3 university students, 1 Ph.D. student, and 1 professor) evaluated the summaries. Average relevance was 3.50 ± 0.78, average consistency was 4.45 ± 0.83, and average fluency was 4.34 ± 0.77.

6 Results

Quantitative Analysis We computed Pearson and Spearman correlations of human judgments with semantic similarity and ROUGE scores. Correlation values can be seen in Table 6 and are visualized in Figure 1 and Figure 2.[20] The results show that, our cross-encoder models have significantly better correlations with relevance, consistency, fluency, and human average. The correlations are higher compared to the bi-encoder models. This

[19]This is the package and version that the authors of MLSUM reported: https://github.com/recitalAI/MLSUM.

[20]All the correlations were significant ($p<.05$) except for the correlations between Fluency and S-BERTurk+STS, BERTScore, ROUGE-L as well as correlations between BERTScore and Consistency.

Article-1
Seattle şehrinin merkezinde meydana gelen olayda, Kanadalı olduğu belirtilen adam, *bir otomobilden söktüğü sunroof camıyla bölgede bulunan araçların ön camlarını* parçaladı. Araçların kaputlarına da çıkan adam, çevredeki birçok araca maddi hasar verdi. Sonrasında, çevrede bulunan otopark görevlisi adama müdahale etmek istedi. Elindeki cam tavanla bu sefer görevliye saldıran adam, *çevredeki diğer insanların müdahalesiyle etkisiz hale getirildi.* Olay yerine gelen polis, adamı gözaltına alırken; adamın uyuşturucu etkisi altında olduğu bildirildi.
Reference Summary
ABD'de bir adam, elindeki sunroof camıyla otomobillerin ön camlarını parçaladı. Adama müdahale etmek isteyen park görevlisi de adamın saldırısına uğradı.
Generated Summary
ABD'de bir otomobilden söktüğü sunroof camıyla bölgede bulunan araçların ön camlarını parçalayan adam, çevredeki diğer insanların müdahalesiyle etkisiz hale getirildi.
ROUGE-(1/2/L): 30.00, 10.53, 25.00
Semantic Similarity Scores BERTurk+NLI+STS (Cross Encoder / Bi-Encoder): 73.67 / 74.35
Human Evaluations (relevance / consistency / fluency / avg): 3.81 / 4.36 / 4.36 / 4.18

Article-2
Yangın, Salihli-Köprübaşı yolu Taytan Mahallesi Çaldırlık mevkisinde meydana geldi. Edinilen bilgiye göre, *seyir* halinde ilerleyen Servet Durmuş idaresindeki 43 HE 737 plakalı *otomobilin motor bölümünde yangın çıktı.* Alevlerin büyümesiyle birlikte otomobil ateş topuna döndü. Sürücü Durmuş hemen itfaiye ekiplerine haber verirken olay yerine gelen Manisa Büyükşehir Belediyesi Salihli İtfaiye Amirliği ekipleri yangına müdahale etti. Söndürülen otomobil kullanılamaz hale geldi. Yangınla ilgili soruşturma başlatıldı.
Reference Summary
Manisa'nın Salihli ilçesinde seyir halinde ilerleyen otomobil alevlere teslim oldu.
Generated Summary
Manisa'da seyir halindeki otomobilin motor bölümünde yangın çıktı.
ROUGE-(1/2/L): 11.11 / 0 / 11.11
Semantic Similarity Scores BERTurk+NLI+STS (Cross Encoder / Bi-Encoder): 76.16 / 81.75
Human Evaluations (relevance / consistency / fluency / avg): 4.0 / 4.8 / 5.0 / 4.6

Table 7: Example articles from MLSUM Turkish test dataset with their reference and generated summaries. The words that appear in both reference and generated summary are in blue, while the semantically similar words are in red. The italic text pieces in the article appear in the generated summary.

also shows that predicted similarity scores are more reliable than computed cosine similarities.

While the main idea of this paper is to evaluate abstractive summarization, we also showed that an extractive Lead-3 baseline yields better semantic similarity scores compared to the abstractive mT5 although it outperforms the extractive baselines in terms of BERTScore and ROUGE scores.

Qualitative Analysis We analyzed the effectiveness of our proposed metrics qualitatively as well. In Table 7, we show two example articles. In the first one, there are some overlapping words between two sentences and they share semantically similar information in the following parts: "*ABD'de bir adam, elindeki sunroof camıyla otomobillerin ön camlarını parçaladı*" and "*ABD'de bir otomobilden söktüğü sunroof camıyla bölgede bulunan araçların ön camlarını parçalayan adam*". So, we can say that both ROUGE and semantic similarity scores can be acceptable for this example. On the other hand, the second example is more critical as it has only one overlapping word between the reference and generated summary; however, there is a high semantic similarity between them and the predicted summary has high human evaluation scores. Our proposed metrics can capture this but apparently ROUGE cannot.

7 Conclusion

In this study, we presented the first Turkish semantic textual similarity corpus, called STSb-TR, by translating the original English STSb dataset via machine translation. We showed that the dataset has high quality translations and does not require costly human annotation. We applied state-of-the-art models to the STSb-TR dataset, and used the best performing four models as evaluation metrics for the text summarization task. We used natural language inference (NLI) models and observed that we can improve our semantic similarity models. We found high correlations between human judgments and our models, compared to BERTScore and ROUGE scores. Our qualitative analyses showed that the proposed models can capture the semantic similarity of reference and predicted summaries which cannot be caught by ROUGE scores. We conclude that our models can be applied as evaluation metric to abstractive summarization in Turkish.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Emrah Budur, Rıza Özçelik, Tunga Güngör, and Christopher Potts. 2020. Data and representation for Turkish natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8253–8267.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*.

Google. Evaluating models — automl translation documentation — google cloud.

Tim Isbister and Magnus Sahlgren. 2020. Why not simply translate? a first Swedish evaluation benchmark for semantic similarity. *arXiv preprint arXiv:2009.03116*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chi-kiu Lo. 2019. Yisi-a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Nils Reimers and Iryna Gurevych. 2018. Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. *arXiv preprint arXiv:1803.09578*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3973–3983.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, et al. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Zheng Zhao, Shay B Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. *arXiv preprint arXiv:2009.13312*.