Morphosyntactic Evaluation for Text Summarization in Morphologically Rich Languages: A Case Study for Turkish

Batuhan Baykara(✉) and Tunga Güngör

Department of Computer Engineering, Boğaziçi University,
Bebek, 34342 Istanbul, Turkey
{batuhan.baykara,gungort}@boun.edu.tr

Abstract. The evaluation strategy used in text summarization is critical in assessing the relevancy between system summaries and reference summaries. Most of the current evaluation metrics such as ROUGE and METEOR are based on n-gram exact matching strategy. However, this strategy cannot capture the orthographical variations in abstractive summaries and is highly restrictive especially for languages with rich morphology that make use of affixation extensively. In this paper, we propose several variants of the evaluation metrics that take into account morphosyntactic properties of the words. We make a correlation analysis between each of the proposed approaches and the human judgments on a manually annotated dataset that we introduce in this study. The results show that using morphosyntactic tokenization in evaluation metrics outperforms the commonly used evaluation strategy in text summarization.

Keywords: Text summarization · Morphologically rich languages · Text summarization evaluation

1 Introduction

Large volumes of textual data have become available since the emergence of the Web. It becomes gradually more challenging to digest the vast amount of information that exists in sources such as websites, news, blogs, books, scientific papers, and social media. Hence, text summarization has emerged as a popular field of study in the past few decades which aims to simplify and make more efficient the process of obtaining relevant piece of information.

Text summarization can be defined as automatically obtaining brief, fluent, and salient piece of text from a much longer and more detailed input text. The two main approaches to text summarization are extractive text summarization and abstractive text summarization. Extractive summarization aims to summarize a given input by directly copying the most relevant sentences or phrases without any modification according to some criteria and ordering them. Abstractive summarization, on the other hand, aims to automatically generate

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
E. Métais et al. (Eds.): NLDB 2023, LNCS 13913, pp. 201–214, 2023.
https://doi.org/10.1007/978-3-031-35320-8_14

new phrases and sentences based on the given input and incorporate them in the output summary.

Evaluation of summarization methods is critical to assess and benchmark their performance. The main objective of evaluation is to observe how well the output summary is able to reflect the reference summaries. The commonly used evaluation methods in summarization such as ROUGE [17] and METEOR [3] are based on n-gram matching strategy. For instance, ROUGE computes the number of overlapping word n-grams between the reference and system summaries in their exact (surface) forms. While the exact matching strategy is not an issue for extractive summarization where the words are directly copied, it poses a problem for abstractive summarization where the generated summaries can contain words in different forms. In the abstractive case, this strategy is very strict especially for morphologically rich languages in which the words are subject to extensive affixation and thus carry syntactic features. It severely punishes the words that have even a slight change in their forms. Hence, taking the morphosyntactic structure of these morphologically rich languages into account is important for the evaluation of text summarization.

In this paper, we introduce several variants of the commonly used evaluation metrics that take into account the morphosyntactic properties of the language. As a case study for Turkish, we train state-of-the-art text summarization models mT5 [31] and BERTurk-cased [27] on the TR-News dataset [4]. The summaries generated by the models are evaluated with the proposed metrics using the reference summaries. In order to make comparisons between the evaluation metrics, we perform correlation analysis to see how well the score obtained with each metric correlates with the human score for each system summary-reference summary pair. Turkish is a low-resource language and it is challenging to find manually annotated data in text summarization. Hence, for correlation analysis, we annotate human relevancy judgements for a randomly sampled subset of the TR-News dataset and we make this data publicly available[1]. Correlation analysis is performed using the annotated human judgements to compare the performance of the proposed morphosyntactic evaluation methods as well as other popular evaluation methods.

2 Related Work

Text summarization studies in Turkish have been mostly limited to extractive approaches. A rule-based system is introduced by Altan [2] tailored to the economics domain. Çığır et al. [7] and Kartal and Kutlu [13] use classical sentence features such as position, term frequency, and title similarity to extract sentences and use these features in machine learning algorithms. Özsoy et al. [21] propose variations to the commonly applied latent semantic analysis (LSA) and Güran et al. [12] utilize non-negative matrix factorization method. Nuzumlalı and Özgür [19] study fixed-length word truncation and lemmatization for Turkish multi-document summarization.

[1] https://github.com/batubayk/news_datasets.

Recently, large-scale text summarization datasets such as MLSum [28] and TR-News [4] have been released which enabled research in abstractive summarization in Turkish. The abstractive studies are currently very limited and they mostly utilize sequence-to-sequence (Seq2Seq) architectures. Scialom et al. [28] make use of the commonly used pointer-generator model [29] and the unified pretrained language model (UniLM) proposed by Dong et al. [10]. Baykara and Güngör [4] follow a morphological adaptation of the pointer-generator algorithm and also experiment with Turkish specific BERT models following the strategy proposed by Liu and Lapata [18]. In a later study, Baykara and Güngör [5] use multilingual pretrained Seq2Seq models mBART and mT5 as well as several monolingual Turkish BERT models in a BERT2BERT architecture. They obtain state-of-the-art results in both TR-News and MLSum datasets.

Most of the evaluation methods used in text summarization and other NLP tasks are more suitable for well-studied languages such as English. ROUGE [17] is the most commonly applied evaluation method in text summarization which basically calculates the overlapping number of word n-grams. Although initially proposed for machine translation, METEOR [3] is also used in text summarization evaluation. METEOR follows the n-gram based matching strategy which builds upon the BLEU metric [22] by modifying the precision and recall computations and replacing them with a weighted F-score based on mapping unigrams and a penalty function for incorrect word order. Recently, neural evaluation methods have been introduced which aim to capture semantic relatedness. These metrics usually utilize embeddings at word level such as Word mover distance (WMD) [15] or sentence level such as Sentence mover distance (SMD) [8]. BERTScore [32] makes use of the BERT model [9] to compute a cosine similarity score between the given reference and system summaries.

There has been very limited research in summarization evaluation for Turkish which has different morphology and syntax compared to English. Most of the studies make use of common metrics such as ROUGE and METEOR [21,28]. Recently, Beken Fikri et al. [6] utilized various semantic similarity metrics including BERTScore to semantically evaluate Turkish summaries on the MLSum dataset. In another work [30], the BLEU+ metric was proposed as an extension to the BLEU metric by incorporating morphology and Wordnet into the evaluation process for machine translation.

3 Overview of Turkish Morphology

Turkish is an agglutinative language which makes use of suffixation extensively. A root word can take several suffixes in a predefined order as dictated by the morphotactics of the language. It is common to find words affixed with 5–6 suffixes. During the affixation process, the words are also subject to a number of morphophonemic rules such as vowel harmony, elisions, or insertions. There are two types of suffixes as inflectional suffixes and derivational suffixes. The inflectional suffixes do not alter the core meaning of a word whereas the derivational suffixes can change the meaning or the part-of-speech.

Table 1. Morphological analysis of an example sentence.

Input	Morphological Analysis
tutsağı	[tutsak:Noun] tutsağ:Noun+A3sg+ı:Acc
serbest	[serbest:Adj] serbest:Adj
bıraktılar	[bırakmak:Verb] bırak:Verb+tı:Past+lar:A3pl

Table 1 shows the disambiguated morphological analysis of the sentence *tutsağı serbest bıraktılar* (*they released the prisoner*) as an example. The square bracket shows the root and its part-of-speech, which is followed by the suffixes attached to the root and the morphological features employed during the derivation[2].

4 Methodology

In this section, we explain the proposed methods that are based on the morphosyntactic features of Turkish and the evaluation metrics used in the study.

4.1 Morphosyntactic Variations

While comparing a system summary and a reference summary, the evaluation metrics used in text summarization use either the surface forms or the lemma or stem forms of the words. As stated in Sect. 1, the former approach is too restrictive and misses matches of the inflected forms of the same words, whereas the latter approach is too flexible and allows matches of all derivations of the same root which causes semantically distant words to match. In this work, we propose and analyze several other alternatives in between these two extreme cases based on morphosyntactic properties of the language. The obtained system and reference summaries are preprocessed according to the details of each proposed method before being passed to the evaluation metrics (ROUGE, METEOR, etc.). The implementation of the evaluation metrics are not changed. The proposed methods can easily be adapted to other morphologically rich languages in the case of readily available morphological analyzer tools.

Table 2 gives the list of the methods used to process the words before applying the evaluation metrics and shows the result of each one for the example sentence depicted in Table 1. The Surface method leaves the words in their written forms, while the Lemma (Stem) method strips off the suffixes and takes the lemma (stem) forms of the words. The lemma and stem forms are obtained using the Zemberek library [1] which applies morphological analysis and disambiguation processes. For the Lemma and Stem methods, in addition to their bare forms,

[2] The morphological features used in the example are as follows: Acc = accusative, A3pl = third person plural number/person agreement, A3sg = third person singular number/person agreement, Past = past tense.

Table 2. Proposed methods based on morphosyntactic variations of words.

Method	Processed Text
Surface	tutsağı serbest bıraktılar
Lemma	tutsak serbest bırak
Stem	tutsağ serbest bırak
Lemma and all suffixes	tutsak ##ı serbest bırak ##tı ##lar
Lemma and combined suffixes	tutsak ##ı serbest bırak ##tılar
Lemma and last suffix	tutsak ##ı serbest bırak ##lar
Lemma and all suffixes with Surface	tutsağı##tutsak tutsağı##ı serbest##serbest bıraktılar##bırak bıraktılar##tı bıraktılar##lar
Lemma and combined suffixes with Surface	tutsağı##tutsak tutsağı##ı serbest##serbest bıraktılar##bırak bıraktılar##tılar
Lemma and last suffix with Surface	tutsağı##tutsak tutsağı##ı serbest##serbest bıraktılar##bırak bıraktılar##lar

six different variations based on different usages of the suffixes are employed. The suffixes used in these variations are also obtained from the morphological parse by the Zemberek library. Only the variations of the Lemma method are shown in the table to save space; the same forms are also applied to the Stem method. The methods are explained below.

Surface: The text is only lower-cased and punctuations are removed. All the other methods also perform the same cleaning and lower-casing operations. For Turkish, this is the default evaluation strategy for all the metrics.

Lemma: The text is lemmatized and the lemma forms of the words are used.

Stem: The text is stemmed and the stem forms of the words are used.

Lemma and all Suffixes: The text is lemmatized and the suffixes are extracted. The lemma and each suffix of a word are considered as separate tokens.

Lemma and Combined Suffixes: The text is lemmatized and the suffixes are extracted. The suffixes are concatenated as a single item. The lemma and the concatenated suffixes of a word are considered as separate tokens.

Lemma and Last Suffix: The text is lemmatized and the suffixes are extracted. The lemma and the last suffix of a word are considered as separate tokens.

The last three methods above split the lemma and the suffixes and use them as individual tokens. This may cause the same tokens obtained from different words to match mistakenly. For instance, if the system summary contains the word *tutsağı* (*the prisoner*) (the accusative form of *tutsak* (*prisoner*)) and the reference summary contains the word *gardiyanı* (*the guardian*) (the accusative form of *gardiyan* (*guardian*)), the morphological parse will output the suffix 'ı' for both of them. The evaluation metric (e.g. ROUGE-1) will match these two suffixes (tokens) although they belong to different words. To prevent such cases,

we devise another variation of these three methods where the surface form of the word is prefixed to each token generated from the word as explained below.

Lemma and all Suffixes with Surface: The text is lemmatized and the suffixes are extracted. The surface form of a word is added as a prefix to the lemma and each of the suffixes of the word. The lemma and each suffix of the word are then considered as separate tokens.

Lemma and Combined Suffixes with Surface: The text is lemmatized and the suffixes are extracted. The suffixes are concatenated as a single item. The surface form of a word is added as a prefix to the lemma and the concatenated suffixes of the word. The lemma and the concatenated suffixes of the word are then considered as separate tokens.

Lemma and Last Suffix with Surface: The text is lemmatized and the suffixes are extracted. The surface form of a word is added as a prefix to the lemma and the last suffix of the word. The lemma and the last suffix of the word are then considered as separate tokens.

4.2 Evaluation Metrics

We use five different metrics for comparing system summaries and reference summaries. We apply the morphosyntactic variations to the summaries and then score the performance using these metrics. In this way, we make a detailed analysis related to which combinations of evaluation metrics and morphosyntactic tokenizations correlate well with human judgments. We explain below each metric briefly.

ROUGE [17] is a recall-oriented metric which is commonly used in text summarization evaluation. ROUGE-N computes the number of overlapping n-grams between the system and reference summaries while ROUGE-L considers the longest common sub-sequence matches.

METEOR [3] is another commonly used metric in text summarization [14, 28]. It is based on unigram matches and makes use of both unigram precision and unigram recall. Word order is also taken into account via the concept of chunk.

BLEU [22] is a precision-oriented metric originally proposed for machine translation evaluation. It uses a modified version of n-gram precision and takes into account both the common words in the summaries and also the word order by the use of higher order n-grams. Although not common as ROUGE, BLEU is also used in text summarization evaluation as an additional metric [11,23].

BERTScore [32] is a recent metric proposed to measure the performance of text generation systems. It extracts contextual embeddings of the words in the system and reference summaries using the BERT model and then computes pairwise cosine similarity between the words of the summaries.

chrF [24] is an evaluation metric initially proposed for machine translation. The F-score of character n-gram matches are calculated between system output and references. It takes into account the morphosyntax since the method is based on character n-grams.

In this work, we make use of the Huggingface's `evaluate` library[3] for all the metrics explained above. We use the monolingual BERTurk-cased [27] model for computing the BERTScore values.

5 Dataset, Models, and Annotations

In this section, we first explain the dataset and the models used for the text summarization experiments. We then give the details of the annotation process where the summaries output by the models are manually scored with respect to the reference summaries. The human judgment scores will be used in Sect. 6 to observe the goodness of the proposed morphosyntactic methods.

5.1 Dataset

We use the **TR-News** [4] dataset for the experiments. TR-News is a large-scale Turkish summarization dataset that consists of news articles. It contains 277,573, 14,610, and 15,379 articles, respectively, for train, validation, and test sets.

5.2 Models

In this work, we use two state-of-the-art abstractive Seq2Seq summarization models. The models are trained on the TR-News dataset and used to generate the system summaries of a sample set of documents to compare with the corresponding reference summaries.

mT5 [31] is the multilingual variant of the T5 model [25] and closely follows its model architecture with some minor modifications. The main idea behind the T5 model is to approach each text-related task as a text-to-text problem where the system receives a text sequence as input and outputs another text sequence.

BERTurk-cased [27] is a bidirectional transformer network pretrained on a large corpus. It is an encoder-only model used mostly for feature extraction. However, Rothe et al. [26] proposed constructing a Seq2Seq model by leveraging model checkpoints and initializing both the encoder and the decoder parts by making several modifications to the model structure. Consequently, we constructed a BERT2BERT model using BERTurk-cased and finetuned it on abstractive text summarization.

The maximum encoder length for mT5 and BERTurk-cased are set to, respectively, 768 and 512, whereas the maximum decoder length is set to 128. The learning rate for the mT5 model is 1e−3 and for the BERTurk-cased model 5e−5. An effective batch size of 32 is used for both models. The models are finetuned for a maximum of 10 epochs where early stopping with patience 2 is employed based on the validation loss.

[3] https://github.com/huggingface/evaluate.

Table 3. Average scores and inter-annotator agreement scores for the models. In the first row, the averages of the two annotators are separated by the/sign.

	BERTurk-cased	mT5
Avg. annotator score	5.86/6.22	6.00/5.88
Pearson correlation	0.85	0.88
Krippendorff's alpha	0.84	0.87
Cohen's Kappa coefficient	0.44	0.25

5.3 Human Judgment Annotations

In order to observe which morphosyntactic tokenizations and automatic summarization metrics perform well in evaluating the performance of text summarization systems for morphologically rich languages, we need a sample dataset consisting of documents, system summaries, reference summaries, and relevancy scores between the system and reference summaries. For this purpose, we randomly sampled 50 articles from the test set of the TR-news dataset. For each article, the system summary output by the model is given a manual score indicating its relevancy with the corresponding reference summary. This is done for the mT5 model and the BERTurk-cased model separately. The relevancy scores are annotated by two native Turkish speakers with graduate degrees. An annotator is shown the system summary and the reference summary for an article without showing the original document and is requested to give a score. We decided to keep the annotation process simple by giving a single score to each system summary-reference summary pair covering the overall semantic relevancy of the summaries instead of scoring different aspects (adequacy, fluency, style, etc.) separately. The scores range from 1 (completely irrelevant) to 10 (completely relevant).

Table 3 shows the average scores of the annotators and the inter-annotator agreement scores. The averages of the two annotators are close to each other for both models. The Pearson correlation and Krippendorf's alpha values being around 0.80–0.90 indicate that there is a strong agreement in the annotators' scores. We also present the Cohen's Kappa coefficient as a measure of agreement between the annotators. The values of 0.44 and 0.25 signal, respectively, moderate agreement and fair agreement between the scores [16]. Since the Cohen's Kappa coefficient is mostly suitable for measuring agreement in categorical values rather than quantitative values as in our case, the results should be approached with caution.

Table 4. Pearson correlation results of the morphosyntactic methods with prefix tokens for the BERTurk-cased summarization model. Bold and underline denote, respectively, the best score and the second-best score for a column.

BERTurk-cased	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BLEU	BERTScore	chrF
Surface	0.770	0.723	0.750	0.736	0.649	**0.800**	0.789
Lemma with Surface	**0.802**	**0.730**	**0.768**	**0.807**	**0.776**	0.766	**0.804**
Stem with Surface	0.792	0.728	0.759	0.802	0.773	0.763	0.801
Lemma and all suffixes with Surface	0.773	0.712	0.743	0.796	0.765	0.760	0.793
Stem and all suffixes with Surface	0.768	0.712	0.740	0.794	0.764	0.760	0.791
Lemma and combined suffixes with Surface	0.774	0.718	0.747	0.796	0.771	0.768	0.797
Stem and combined suffixes with Surface	0.767	0.718	0.741	0.794	0.770	0.767	0.794
Lemma and last suffix with Surface	0.781	0.718	0.749	0.798	**0.776**	0.766	0.795
Stem and last suffix with Surface	0.774	0.718	0.743	0.798	**0.776**	0.766	0.792

Table 5. Pearson correlation results of the morphosyntactic methods with prefix tokens for the mT5 summarization model. Bold and underline denote, respectively, the best score and the second-best score for a column.

mT5	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BLEU	BERTScore	chrF
Surface	0.682	0.648	0.693	0.697	0.591	0.693	0.718
Lemma with Surface	**0.701**	**0.669**	**0.709**	0.753	0.719	0.682	**0.739**
Stem with Surface	0.688	0.665	0.700	0.742	0.714	0.674	0.734
Lemma and all suffixes with Surface	0.699	0.658	0.700	**0.771**	**0.730**	**0.694**	0.733
Stem and all suffixes with Surface	0.693	0.658	0.698	0.767	0.728	0.690	0.731
Lemma and combined suffixes with Surface	0.685	0.653	0.693	0.750	0.714	0.690	0.738
Stem and combined suffixes with Surface	0.677	0.653	0.688	0.745	0.712	0.687	0.734
Lemma and last suffix with Surface	0.692	0.653	0.699	0.749	0.712	0.674	0.734
Stem and last suffix with Surface	0.684	0.653	0.693	0.743	0.710	0.671	0.730

6 Correlation Analysis

In this work, we mainly aim at observing the correlation between the human evaluations and the automatic evaluations for the system generated summaries. For each of the proposed morphosyntactic tokenization methods (Sect. 4.1), we first apply the method to the system and reference summaries of a document and obtain the tokenized forms of the words in the summaries. We then evaluate the similarity of the tokenized system and reference summaries with each of the standard metrics (Sect. 4.2). Finally, we compute the Pearson correlation

between the human score (average of the two annotators) given to the reference summary-system summary pair (Sect. 5.3) and the metric score calculated based on that morphosyntactic tokenization.

In this way, we make a detailed analysis of the morphosyntactic tokenization method and text summarization metric combinations. The results are shown in Tables 4 and 5. For the ROUGE metric, we include the results for the ROUGE-1, ROUGE-2, and ROUGE-L variants that are commonly used in the literature. For the tokenization methods that include suffixes, we show only the results with the surface forms of the words prefixed to the tokens (*with Surface*). The results without the prefixed tokens are given in the Appendix. Interestingly, the methods that do not use the prefix forms correlate better with the human judgments, although they tend to produce incorrect matches as shown in Sect. 4.1.

We observe that the Lemma method mostly yields the best results for the summaries generated by the BERTurk-cased model. The Lemma method is followed by the Stem method. These results indicate that simply taking the root of the words in the form of lemma or stem before applying the evaluation metrics is sufficient instead of more complex tokenizations. One exception is the BERTScore metric which works best with the surface forms of the words. This may be regarded as an expected behavior since BERTScore is a semantically-oriented evaluation approach while the others are mostly syntactically-oriented metrics. Hence, when fed with the surface forms, BERTScore can capture the similarities between different orthographical forms of the words.

The summaries generated by the mT5 model follow a similar pattern in ROUGE evaluations. The Lemma method and the Stem method yield high correlations with human scores. On the other hand, the other three metrics correlate better with human judgments when suffixes are also incorporated as tokens into the evaluation process in addition to the lemma or stem form. The BERTScore metric again shows a good performance when used with the Surface method.

We observe a significant difference between the correlation scores of the BERTurk-cased model and the mT5 model. The higher correlation results of the BERTurk-cased model indicate that summaries with better quality are generated. This may be attributed to the fact that BERTurk-cased is a monolingual model unlike the multilingual mT5 model and this distinction might have enabled it to produce summaries with better and more relevant context.

The high correlation ratios obtained with the Lemma tokenization approach may partly be attributed to the success of the Zemberek morphological tool. Zemberek has a high performance in morphological analysis and morphological disambiguation for Turkish [1]. When the Lemma and Stem methods are compared, we see that the Lemma method outperforms the Stem method for both models and for all evaluation metrics. This is the case for both the bare forms of these two methods and their variations. The tokenization methods where the last suffixes are used follow the top-ranking Lemma and Stem methods in BERTurk-cased evaluations, whereas they fall behind the tokenization variations with all suffixes in mT5 evaluations. The motivation behind the last suffix strategy is that the last suffix is considered as one of the most informative morphemes in

Turkish [20]. We see that this simple strategy is on par with those that use information of all the suffixes.

Finally, comparing the five text summarization evaluation metrics shows that METEOR yields the best correlation results for both models followed by the chrF metric. Although the underlying tokenization method that yields the best performance is different in the two models (Lemma for BERTurk-cased and Lemma with all suffixes in mT5), we can conclude that the METEOR metric applied to lemmatized system and reference summaries seems as the best metric for text summarization evaluation. This is an interesting result considering that ROUGE is the most commonly used evaluation metric in text summarization.

It should be noted that the Surface method corresponds to the approach used in the evaluation tools for these metrics. That is, the ROUGE, METEOR, BLEU, chrF, and BERTScore tools used in the literature mostly follow a simple strategy and work on the surface forms of the words. However, Tables 4 and 5 show that other strategies such as using the lemma form or using the lemma form combined with the suffixes nearly always outperform this default strategy. This indicates that employing morphosyntactic tokenization processes during evaluation increases correlation with human judgments and thus contributes to the evaluation process.

7 Conclusion

In this study, we introduced various morphosyntactic methods that can be used in text summarization evaluation. We trained state-of-the-art text summarization models on the TR-News dataset. The models were used to generate the system summaries of a set of documents sampled from the test set of TR-News. The relevancy of the system summaries and the reference summaries were manually scored and correlation analysis was performed between the manual scores and the scores produced by the morphosyntactic methods. The correlation analysis revealed that making use of morphosyntactic methods in evaluation metrics outperforms the default strategy of using the surface form for Turkish. We make the manually annotated evaluation dataset publicly available to alleviate the resource scarcity problem in Turkish. We believe that this study will contribute to focus on the importance of preprocessing in evaluation in this area.

Appendix

The correlation results of the morphosyntactic tokenization methods without the prefix tokens are shown in Tables 6 and 7.

Table 6. Pearson correlation results of the morphosyntactic methods without prefix tokens for the BERTurk-cased summarization model. Bold and underline denote, respectively, the best score and the second-best score for a column.

BERTurk-cased	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BLEU	BERTScore	chrF
Surface	0.770	0.723	0.750	0.736	0.649	**0.800**	0.789
Lemma	**0.831**	**0.744**	**0.795**	**0.809**	0.671	0.775	0.797
Stem	0.815	0.738	0.777	0.799	0.668	0.768	0.791
Lemma and all suffixes	0.796	0.737	0.762	0.783	0.768	0.746	**0.798**
Stem and all suffixes	0.789	0.736	0.757	0.779	0.766	0.745	0.794
Lemma and combined suffixes	0.798	0.727	0.769	0.793	0.763	0.752	0.794
Stem and combined suffixes	0.789	0.725	0.758	0.789	0.759	0.753	0.789
Lemma and last suffix	0.807	0.733	0.769	0.789	**0.773**	0.756	0.793
Stem and last suffix	0.795	0.732	0.757	0.784	0.768	0.757	0.788

Table 7. Pearson correlation results of the morphosyntactic methods without prefix tokens for the mT5 summarization model. Bold and underline denote, respectively, the best score and the second-best score for a column.

mT5	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BLEU	BERTScore	chrF
Surface	0.682	0.648	0.693	0.697	0.591	**0.693**	0.718
Lemma	**0.713**	**0.677**	**0.708**	0.737	0.602	0.682	0.723
Stem	0.696	0.659	0.693	0.716	0.594	0.675	0.714
Lemma and all suffixes	0.702	0.648	0.691	0.730	**0.701**	0.671	0.719
Stem and all suffixes	0.693	0.642	0.688	0.721	0.695	0.666	0.714
Lemma and combined suffixes	0.691	0.652	0.690	**0.748**	0.678	0.687	**0.727**
Stem and combined suffixes	0.680	0.643	0.679	0.737	0.669	0.690	0.720
Lemma and last suffix	0.700	0.656	0.702	0.741	0.678	0.656	0.718
Stem and last suffix	0.688	0.647	0.690	0.730	0.669	0.652	0.710

References

1. Akın, A.A., Akın, M.D.: Zemberek, an open source NLP framework for Turkic languages. Structure **10**, 1–5 (2007)
2. Altan, Z.: A Turkish automatic text summarization system. In: IASTED International Conference on (2004)
3. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan, pp. 65–72. Association for Computational Linguistics (2005)
4. Baykara, B., Güngör, T.: Abstractive text summarization and new large-scale datasets for agglutinative languages Turkish and Hungarian. Lang. Resour. Eval. **56**, 1–35 (2022)
5. Baykara, B., Güngör, T.: Turkish abstractive text summarization using pretrained sequence-to-sequence models. Nat. Lang. Eng. 1–30 (2022)

6. Beken Fikri, F., Oflazer, K., Yanikoglu, B.: Semantic similarity based evaluation for abstractive news summarization. In: Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021). Association for Computational Linguistics, Online (2021)
7. Çığır, C., Kutlu, M., Çiçekli, İ.: Generic text summarization for Turkish. In: ISCIS, pp. 224–229. IEEE (2009)
8. Clark, E., Celikyilmaz, A., Smith, N.A.: Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 2748–2760. Association for Computational Linguistics (2019)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics (2019)
10. Dong, L., et al.: Unified language model pre-training for natural language understanding and generation. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. (2019)
11. Graham, Y.: Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In: EMNLP (2015)
12. Güran, A., Bayazit, N.G., Bekar, E.: Automatic summarization of Turkish documents using non-negative matrix factorization. In: 2011 International Symposium on Innovations in Intelligent Systems and Applications, pp. 480–484. IEEE (2011)
13. Kartal, Y.S., Kutlu, M.: Machine learning based text summarization for Turkish news. In: 2020 28th Signal Processing and Communications Applications Conference (SIU), pp. 1–4. IEEE (2020)
14. Koupaee, M., Wang, W.Y.: WikiHow: a large scale text summarization dataset (2018)
15. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, Lille, France, vol. 37, pp. 957–966. PMLR (2015)
16. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–74 (1977)
17. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, Barcelona, Spain, pp. 74–81. Association for Computational Linguistics (2004)
18. Liu, Y., Lapata, M.: Text summarization with pretrained encoders. 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), Hong Kong, China, pp. 3730–3740. Association for Computational Linguistics (2019)
19. Nuzumlalı, M.Y., Özgür, A.: Analyzing stemming approaches for Turkish multi-document summarization. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 702–706. Association for Computational Linguistics (2014)
20. Oflazer, K., Say, B., Hakkani-Tür, D.Z., Tür, G.: Building a Turkish Treebank. In: Abeillé, A. (ed.) Treebanks. Text, Speech and Language Technology, vol. 20, pp.

261–277. Springer, Dordrecht (2003). https://doi.org/10.1007/978-94-010-0201-1_15
21. Özsoy, M.G., Çiçekli, İ., Alpaslan, F.N.: Text summarization of Turkish texts using latent semantic analysis. In: Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010, pp. 869–876. Association for Computational Linguistics, USA (2010)
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318. Association for Computational Linguistics (2002)
23. Parida, S., Motlícek, P.: Abstract text summarization: a low resource challenge. In: EMNLP (2019)
24. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, pp. 392–395. Association for Computational Linguistics (2015). https://doi.org/10.18653/v1/W15-3049, https://aclanthology.org/W15-3049
25. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020)
26. Rothe, S., Narayan, S., Severyn, A.: Leveraging pre-trained checkpoints for sequence generation tasks. Trans. Assoc. Comput. Linguist. **8**, 264–280 (2020)
27. Schweter, S.: BERTurk - BERT models for Turkish (2020). https://doi.org/10.5281/zenodo.3770924
28. Scialom, T., Dray, P.A., Lamprier, S., Piwowarski, B., Staiano, J.: MLSUM: the multilingual summarization corpus. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8051–8067. Association for Computational Linguistics, Online (2020)
29. See, A., Liu, P.J., Manning, C.D.: Get to the point: Summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, pp. 1073–1083. Association for Computational Linguistics (2017)
30. Tantuğ, A.C., Oflazer, K., El-Kahlout, I.D.: BLEU+: a tool for fine-grained BLEU computation. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco. European Language Resources Association (ELRA) (2008)
31. Xue, L., et al.: mT5: a massively multilingual pre-trained text-to-text transformer. ArXiv abs/2010.11934 (2021)
32. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: evaluating text generation with BERT (2019)