


ARTICLE

Abstractive summarization with deep reinforcement learning using semantic similarity rewards

Figen Beken Fikri¹ , Kemal Oflazer² and Berrin Yanıkoğlu^{1,3}

¹Faculty of Engineering and Natural Sciences, Sabancı University, Istanbul, Türkiye, ²Qatar Computer Science Program/Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA, and ³Center of Excellence in Data Analytics (VERIM), Sabancı University, Istanbul, Türkiye

Corresponding author: Figen Beken Fikri; Email: fbekenfikri@sabanciuniv.edu

(Received 15 August 2022; revised 12 September 2023; accepted 27 September 2023)

Abstract

Abstractive summarization is an approach to document summarization that is not limited to selecting sentences from the document but can generate new sentences as well. We address the two main challenges in abstractive summarization: how to evaluate the performance of a summarization model and what is a good training objective. We first introduce new evaluation measures based on the semantic similarity of the input and corresponding summary. The similarity scores are obtained by the fine-tuned BERTurk model using either the cross-encoder or a bi-encoder architecture. The fine-tuning is done on the Turkish Natural Language Inference and Semantic Textual Similarity benchmark datasets. We show that these measures have better correlations with human evaluations compared to Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores and BERTScore. We then introduce a deep reinforcement learning algorithm that uses the proposed semantic similarity measures as rewards, together with a mixed training objective, in order to generate more natural summaries in terms of human readability. We show that training with a mixed training objective function compared to only the maximum-likelihood objective improves similarity scores.

Keywords: Abstractive summarization; Deep reinforcement learning; Evaluation metric; Semantic textual similarity; Natural Language Inference

1. Introduction

Automatic document summarization aims to create a summary that captures the important details in a given text. It has become an important research area, with the massive amount of documents available in social media, online forums, and news articles.

There are two approaches to summarization: extractive and abstractive. Extractive summarization yields a summary by selecting parts from the given document. As such, it is guaranteed to generate grammatically correct sentences, however, the resulting summary is constrained to use sentences in the input text. In contrast, abstractive summarization constitutes a potentially more powerful approach, as it captures the semantic of the input and generates sentences to summarize it. However, the challenge in this approach lies in producing faithful summaries that are also natural and grammatically correct.

The two most popular approaches for abstractive summarization are based on supervised learning and reinforcement learning. In supervised learning models, a *teacher-forcing* algorithm is widely used to maximize the log-likelihood of the ground-truth sequence given the input text, or equivalently minimizes the negative log-likelihood loss (Bengio *et al.* 2015; Ranzato *et al.* 2016):



$$L_{ml} = - \sum_{t=1}^T \log p(y_t^* | y_1^*, \dots, y_{t-1}^*, X) \quad (1)$$

where $y^* = \{y_1^*, y_2^*, \dots, y_T^*\}$ is the ground-truth sequence for a given input X .

In reinforcement learning approaches to abstractive summarization, the idea is to learn a policy that optimizes a specific discrete metric rather than the maximum-likelihood loss. However, optimizing a model does not guarantee that the output will have better quality and readability (Liu *et al.* 2016; Paulus, Xiong, and Socher 2018). Paulus *et al.* (2018) suggested that a maximum-likelihood training objective can be helpful for the policy learning algorithm to generate more natural summaries, given the fact that it calculates the probability of a token \hat{y}_t based on the previously predicted sequence $\{\hat{y}_1, \dots, \hat{y}_{t-1}\}$ and the input sequence X . Hence, they proposed a mixed training objective, to capture human readability while optimizing the evaluation metrics; this idea is adopted in our work as well.

The most widely used evaluation metric for summarization is Recall-Oriented Understudy for Gisting Evaluation (ROUGE), which compares an automatically generated summary to a human-generated summary by considering the overlapping units, such as n-grams, word sequences, and word pairs, between them (Lin 2004). Although ROUGE is a widely used evaluation metric, it is not very suitable for the evaluation of abstractive summarization systems as it relies on superficial lexical overlap between the ground truth and the generated summaries. Furthermore, for languages with complex morphology, such as Turkish, the ROUGE metric is even less suitable. For instance, both of the following Turkish sentences have the meaning of “I want to call the embassy,” with just a single word of overlap:

Büyükelçiliği aramak istiyorum. (“I want to call the embassy.”)

Büyükelçiliğe telefon etmek istiyorum. (“I want to make a phone call to the embassy.”)

While “ara-mak” (call-INFL) is a verb that takes an object in accusative case, “telefon et-mek” (phone-INFL) 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 (highlighted with underline). Although, these sentences are semantically equivalent, considering the first one as the ground truth and second one as the generated sentence, ROUGE-1, ROUGE-2, and ROUGE-3 scores of these sentences would be 0.25, 0, and 0.25, respectively.

Another metric widely used for summarization evaluation is BLEU, which is a precision-based metric that is mainly used for automatic evaluation of machine translation (Papineni *et al.* 2002). It measures how many n-grams in the generated summary appear in the ground-truth summary, and also uses a brevity penalty to discourage generated summaries that are shorter than the ground-truth summary. Similar to ROUGE, BLEU is also a simple measure that is not fully suited for evaluating summarization quality. Furthermore, with either one of these two metrics, there is an inconsistency between the training objective and the evaluation measure. In other words, while training is done by maximum-likelihood objective, testing is done by BLEU or ROUGE metrics that are based on the number of matching n-grams between the generated summary and the corresponding ground-truth summary. On the other hand, we cannot use these two metrics as the training objective, as they are not differentiable. However, recent studies show that reinforcement learning can use non-differentiable evaluation measures (Bahdanau *et al.* 2016; Rennie *et al.* 2017; Paulus *et al.* 2018).

Based on the mentioned issues, we focus on two main challenges in abstractive summarization: How to evaluate the results and what is a good training objective. In Beken Fikri, Oflazer, and Yanikoğlu (2021), we proposed alternative evaluation measures; here we also propose a reinforcement learning framework for abstractive summarization that uses these measures.

For evaluation, we propose to use the similarity between the sentence embeddings of the generated summary and the corresponding ground-truth summary, obtained by the BERTurk

model using a bi-encoder or cross-encoder architecture. Both models are fine-tuned on the Turkish Semantic Textual Similarity benchmark (STSb-TR) (Beken Fikri *et al.* 2021) and Turkish Natural Language Inference (NLI-TR) (Budur *et al.* 2020) datasets, which are translated from the original datasets for English. We then train an abstractive summarization model using a deep reinforcement learning framework, where the proposed similarity metric is used as the reward signal.

To the best of our knowledge, this is the first study to explore large language models fine-tuned with a translated dataset to obtain similarity measures. Similar to BERTScore, these measures are based on Bidirectional Encoder Representations from Transformers (BERT) but differ by evaluating sentence-level similarity rather than token-by-token comparison.

Our paper is structured in the following way: In Section 2, we review recent studies in abstractive summarization systems. Then, we explain our methodology in Section 3. In Section 4, we describe our experiments and report our quantitative results as well as qualitative analysis. In Section 5, we discuss our findings and limitations. Finally, in Section 6, we present our conclusions.

2. Related work

Here we provide a summary of research on abstractive summarization and evaluation metrics. This section is divided into three subsections, each focusing on different aspects of the field.

2.1. Supervised learning based approaches

State-of-the-art abstractive summarization models are based on neural sequence-to-sequence models (Sutskever, Vinyals, and Le 2014). Initial work by Rush *et al.* (2015) introduced neural sequence-to-sequence abstractive sentence summarization model with an attention-based encoder and a feed-forward neural network language model decoder. Chopra, Auli, and Rush (2016) presented another abstractive summarization model with a convolutional attention-based encoder and a recurrent neural network (RNN) decoder. Nallapati *et al.* (2016) further explored RNNs for both encoder and decoder, with a novel switching generator pointer approach to address out-of-vocabulary words. In a later work, multi-sentence abstractive text summarization was addressed by See, Liu, and Manning (2017) through a hybrid pointer generator network and a coverage mechanism (Tu *et al.* 2016). Subsequent work by Gehrmann, Deng, and Rush (2018) was built upon the pointer generator summarization model by introducing content selection for relevant document portions and a bottom-up copy attention mechanism.

Recent advances in pre-trained language models have significantly enhanced text summarization as well. Liu and Lapata (2019) introduced BERTSum exploring the use of BERT in text summarization framework, using both extractive and abstractive approaches. Meanwhile, Dong *et al.* (2019) introduced the Unified Pre-trained Language Model with unidirectional, bidirectional, and sequence-to-sequence language modeling objectives. In a prominent work, Raffel *et al.* (2020) presented the text-to-text-transfer-transformer (T5) pre-trained on the new open-source dataset called the Colossal Clean Crawled Corpus (C4) they introduced. In T5, every text processing problem is considered in a text-to-text framework, and a single model can be trained with the same loss function and decoding process on different NLP tasks. Additionally, Lewis *et al.* (2020) proposed BART, a denoising autoencoder for pre-training sequence-to-sequence models. Pre-training occurs in two steps: First, an arbitrary noising function is used to corrupt the text. Then, BART is trained to reconstruct the original text. The authors observed improved summarization results. Zhang *et al.* (2020a) introduced PEGASUS, a transformer-based encoder-decoder abstractive summarization model with gap-sentence generation. Furthermore, Qi *et al.* (2020) presented ProphetNet, a sequence-to-sequence pre-training model based on transformer (Vaswani

et al. 2017), with future n-gram prediction and n-stream self-attention mechanism. Several recent works proposed abstractive summarization methods built upon existing large language models (LLMs) (e.g. GSum Dou *et al.* 2021, SimCLS Liu and Liu 2021, SeqCo Xu *et al.* 2022, and BRIO Liu *et al.* 2022).

Unlike English, abstractive summarization studies using LLMs in Turkish remain relatively underexplored. A recent noteworthy study in this direction was conducted by Baykara and Güngör (2022a) who introduced a new Turkish summarization benchmark dataset named TR-News. The authors experimented with BERTSum and pointer generator networks by incorporating linguistically-oriented tokenization techniques. Another significant contribution conducted a comprehensive analysis of BERT models, mBART (Liu *et al.* 2020), and multilingual text-to-text-transfer-transformer (mT5) model (Xue *et al.* 2021) on the Turkish text summarization and title generation tasks (Baykara and Güngör 2022b). They evaluated their models on the TR-News and MultiLingual SUMmarization (MLSum) (Scialom *et al.* 2020) Turkish datasets, focusing on ROUGE scores and presenting baseline results. Notably, the authors observed improved performance across the combined TR-News and MLSum Turkish datasets. In the scope of our study, we opted for the utilization of the mT5 model, renowned for its exceptional performance as showcased on the MLSum Turkish dataset (Baykara and Güngör 2022b).

More recently, reinforcement learning-based approaches have gained more interest in abstractive summarization, as discussed in the next section.

2.2. Reinforcement learning approaches

Paulus *et al.* (2018) introduced a neural summarization model in which ROUGE scores were optimized as a reward. The authors adopted self-critical policy gradient training algorithm (Rennie *et al.* 2017) and applied a mixed training objective function. Their method showed enhanced readability of generated summaries and suitability for longer output sequences. Policy-based reinforcement learning has been widely adopted for text summarization tasks with various rewards (e.g. Chen and Bansal 2018; Sharma *et al.* 2019) and explored with mixed training objectives (e.g. Pasunuru and Bansal 2018; Kryściński *et al.* 2018; Zhang and Bansal 2019; Scialom *et al.* 2019). In further studies, Böhm *et al.* (2019) introduced a reward function learned from human ratings, resulting in improved summarization quality and outperforming state-of-the-art systems. Similarly, Stiennon *et al.* (2020) enhanced summary quality by training models to predict human-preferred summaries, utilizing this predictive model as a reward function for reinforcement learning-based policy fine-tuning. Meanwhile, Zhang *et al.* (2020c) developed a framework to evaluate the factual correctness of a generated summary using an external information extraction system that compares it against the human reference summary. They jointly optimized factual correctness, textual overlap, and language model objectives via reinforcement learning. In another work, Laban *et al.* (2020) proposed an unsupervised abstractive summarization model that generates summaries directly from source documents by optimizing coverage, fluency, and brevity using reinforcement learning (RL)-based rewards. Yadav *et al.* (2021) introduced two novel question-aware semantic rewards for abstractive question summarization: (1) question-type identification and (2) question-focus recognition. They integrated these rewards into an encoder-decoder-based ProphetNet transformer model (Qi *et al.* 2020) by utilizing the mixed training objective proposed by Paulus *et al.* (2018). Recent studies employed reinforcement learning to generate multiple summaries with varying lengths for a given text (Hyun *et al.* 2022) and to optimize factual consistency of generated summaries (Roit *et al.* 2023).

Similar to the prevalence of supervised learning approaches in English, RL studies for abstractive summarization are abundant. However, in the context of Turkish, such studies are notably scarce. The reinforcement learning framework proposed in this paper draws parallels with Yadav *et al.* (2021) as we fine-tune BERT-based models to obtain our reward signal. Our goal is to

enhance the alignment of predicted summaries with the corresponding ground-truth summaries through a mixed training objective.

While abstractive summarization has seen a steep progress in recent years, research on how to evaluate the quality of generated summaries lagged behind. In the next section, we discuss research in evaluating the quality of generated summaries.

2.3. Evaluation metrics

A number of studies have been conducted to assess the factual correctness of the generated summaries. One notable approach, based on the idea that the source document should entail the information in a summary, was explored by Falke *et al.* (2019). Their research focused on using textual entailment to identify factual inaccuracies in generated summaries. They attempted to reduce factual errors by considering the re-ranking of alternative summaries using models trained on Natural Language Inference (NLI) datasets. However, their findings demonstrated that standard NLI models struggled to assess factual correctness. In a similar vein, Kryściński *et al.* (2020) developed a model-based approach for evaluating factual consistency in generated summaries at the document-sentence level. Meanwhile, Zhao, Cohen, and Webber (2020) addressed the problem of unsupported information in generated summaries known as “factual hallucination.” To assess faithfulness, Durmuş, He and Diab (2020) and Wang, Cho, and Lewis (2020) proposed question-answering-based frameworks.

Another relevant line of research emerged, focusing on using deep embeddings to compare generated and ground-truth texts, first in the context of machine translation. For instance, YiSi (Lo 2019) used an embedding model or cross-lingual embedding model to evaluate lexical semantic similarity using cosine embeddings. Zhang *et al.* (2020b) presented BERTScore, a metric calculating the similarity between generated and reference translations. This similarity is assessed using the power of BERT by summing cosine similarities between token embeddings. Several works proposed neural evaluation metrics by leveraging pre-trained language models (e.g. MoverScore Zhao *et al.* 2019, BLEURT Sellam, Das, and Parikh 2020, COMET Rei *et al.* 2020, Prism Thompson and Post 2020, and BARTScore Yuan, Neubig, and Liu 2021). In a more recent work, Zhao, Strube, and Eger (2023) introduced DiscoScore that uses BERT to model discourse coherence from different perspectives. Meanwhile, Eddine, Shang, and Vazirgiannis (2023) presented DATScore as an extension of BARTScore (Yuan *et al.* 2021) and data augmentation techniques.

In this study, we propose to use a BERT model with a cross-encoder or bi-encoder architecture to measure the similarity between two summaries, by training it on a translated sentence similarity dataset.

3. Methodology

We present our methodology by first describing semantic textual similarity (Section 3.1) and natural language inference (Section 3.2) tasks used in our work and the proposed approaches. We also summarize major datasets in these two areas, along with recent works on translation of these datasets into Turkish. We then describe the proposed evaluation method, which is based on fine-tuning pre-trained language models (cross-encoders and bi-encoders) to learn to predict semantic similarity between two pieces of text (Section 3.3). Finally, we describe our abstractive summarization approach via policy gradient reinforcement learning, using the proposed semantic similarity measure as a reward (Section 3.4).

It is important to highlight that our selection of models and datasets is rooted in established benchmarks for the English language. Furthermore, drawing inspiration from the pioneering work of Reimers and Gurevych (2019), we opt for models—BERTurk, mBERT, and XLM-R—that reflect a range of language-specific and cross-lingual capabilities.

Table 1. Sample translations from the Semantic Textual Similarity benchmark dataset along with the corresponding English sentences. The similarity score between two Turkish sentences are set to the similarity between the corresponding English sentences (Beken Fikri *et al.* 2021) (Section 3.1)

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

3.1. Semantic textual similarity

Semantic textual similarity aims to determine how semantically similar two pieces of text are. It has many applications in areas such as machine translation, summarization, text generation, question answering, dialog, and speech systems and has become a popular topic with the competitions organized by SemEval since 2012.

In a recent study, we translated the English Semantic Textual Similarity benchmark dataset (STSb) (Cer *et al.* 2017) into Turkish using the Google Cloud Translation API (Google Cloud 2021). The translations in the new dataset, called STSb-TR, were found to be of high quality (Beken Fikri *et al.* 2021). The STSb dataset consists of a selection of the English datasets used in SemEval STS studies between 2012 and 2017. There are 8628 sentence pairs in total (5749 train, 1500 dev, 1379 test). Scores range from 0 (no semantic similarity) to 5 (semantically equivalent) on a continuous scale. Table 1 shows some examples from the STSb dataset and their translations.

In this study, we used the translated dataset to learn the semantic similarity scores of two summaries by fine-tuning state-of-the-art pre-trained language models.

3.2. Natural language inference

Natural language inference is the study of determining whether there is an entailment, contradiction, or neutral relationship between the given premise and the hypothesis sentences. There are two major datasets in literature for natural language inference in English: the Stanford Natural Language Inference (SNLI) (Bowman *et al.* 2015) and the MultiGenre Natural Language Inference (MultiNLI) (Williams, Nangia, and Bowman 2018). The SNLI dataset has around 570k sentence pairs while the MultiNLI dataset contains around 433k sentence pairs. The MultiNLI dataset is similar to SNLI in format, but it contains a wider range of text genres.

Table 2. Example sentences illustrating the Natural Language Inference task (Budur *et al.* 2020) (Section 3.2)

	English	Turkish
Premise	Three men are sitting near an orange building with blue trim	Üç adam mavi süslemeli turuncu bir binanın yanında oturuyor
Entailment	Three males are seated near an orange building with blue trim	Üç erkek mavi süslü turuncu bir binanın yakınında oturuyor
Contradiction	Three women are standing near a yellow building with red trim	Üç kadın kırmızı süslemeli sarı bir binanın yanında duruyor
Neutral	Three males are seated near an orange house with blue trim and a blue roof	Üç erkek mavi süslü ve mavi çatılı turuncu bir evin yakınında oturuyor

Recently, SNLI and MultiNLI datasets have been translated into Turkish, called NLI-TR (Budur *et al.* 2020). Examples shown in Table 2 illustrate the relationships that are represented in the datasets, by presenting Turkish sentence pairs with their matching original English sentences.

Natural language inference is a closely related task to measuring semantic textual similarity and has been used as a pre-training step in the English STS studies, for example Reimers and Gurevych (2019, 2020). The NLI-TR dataset, which is specifically designed for Natural Language Inference tasks in Turkish, serves as a valuable resource for pre-training our models. By leveraging this dataset, we enhance the ability of our models to capture contextual and semantic information unique to the Turkish language. This pre-training step, combined with fine-tuning the STSb-TR dataset, results in improved similarity models that offer a more accurate representation of text summarization evaluation.

3.3. Predicting text similarity using fine-tuned large language models

We fine-tune pre-trained large language models, namely BERTurk, mBERT, and XLM-R, to predict the semantic similarity between two pieces of text. The trainings are done over two related tasks, using the relevant datasets translated from English: (1) The STSb-TR dataset, which is the translated version of the English STSb dataset (Beken Fikri *et al.* 2021). (2) The NLI-TR dataset, which consists of translated versions of the SNLI and MultiNLI datasets, facilitates research on natural language inference in Turkish across various text genres (Budur *et al.* 2020). The models are either only fine-tuned on the STSb-TR dataset or were first fine-tuned on the NLI-TR dataset and then on STSb-TR dataset.

The two main architectures that are fine-tuned to predict sentence pair scoring tasks are cross-encoders and bi-encoders, illustrated in Fig. 1. In the cross-encoder approach, two sentences are passed simultaneously to the transformer network, which directly computes their semantic similarity score between 0 and 1. We used the cross-encoder architecture described by Reimers and Gurevych (2019), which is trained with the binary cross-entropy loss:

$$BCE = (y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \quad (2)$$

where y is the ground-truth label and \hat{y} is the predicted label (Table 3).

The bi-encoder BERT, also called Sentence-BERT, is proposed by Reimers and Gurevych (2019). It is a modification of the pre-trained BERT network that enables to derivation of semantically meaningful fixed-size sentence embeddings. The training loss varies depending on the dataset used to train the model: the classification objective was employed during training on the NLI dataset, while the regression objective was used during the training on the STSb dataset (Reimers and Gurevych 2019).

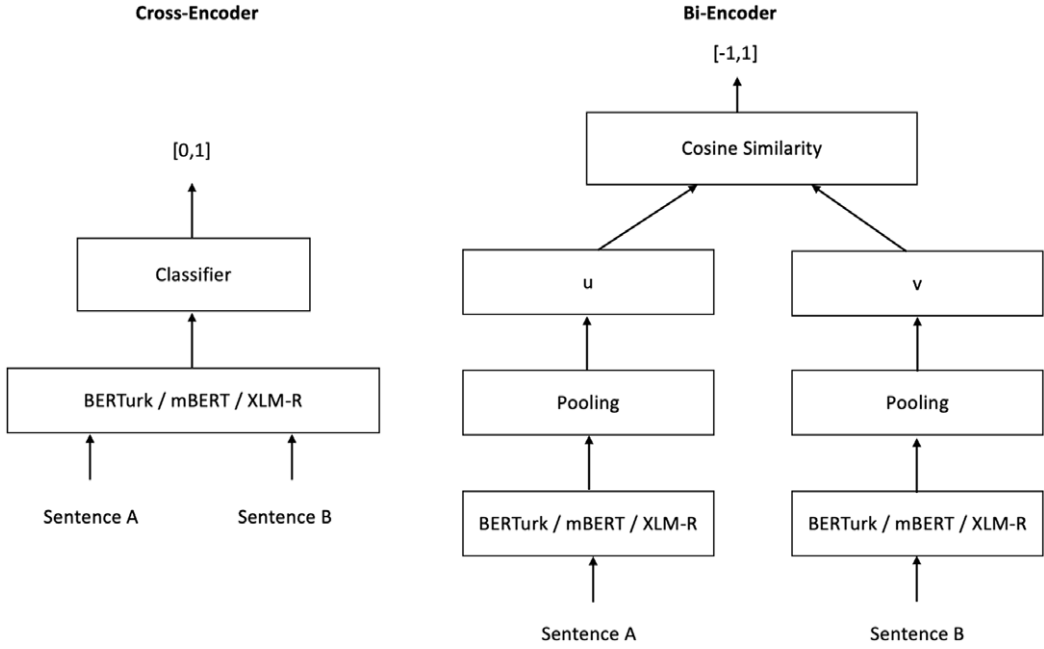


Figure 1. Cross-encoder and bi-encoder model architectures (Reimers and Gurevych 2019).

In the regression objective, the cosine similarity between two sentence embeddings is employed with the mean square error loss:

$$MSE = (1/N) \cdot \sum_{i=1}^N (y - \text{cos_sim}(u, v))^2 \tag{3}$$

where y is the ground-truth label and $\text{cos_sim}(u, v)$ is the predicted label that represents the cosine similarity of sentence embeddings u and v .

In the classification objective, the model optimizes the cross-entropy loss between the true labels and the predictions obtained with softmax, where the logits z for the n -dimensional sentence embeddings u and v of the input texts are concatenated with element-wise difference and multiplied by a trainable weight $W_t \in \mathbb{R}^{k \times 3n}$:

$$z = W_t \times [u, v, |u - v|]$$

$$CE = - \sum_{l=1}^k (y \cdot \log(\text{softmax}(z))) \tag{4}$$

The fine-tuned models, in two alternative architectures (cross-encoder and bi-encoder), are then evaluated for how well they perform in semantic similarity predictions, in the context of abstractive summarization (see Table 4).

3.4. Policy gradient reinforcement with similarity rewards

As the second major contribution of this work, we have trained an abstractive summarization model using the reinforcement learning framework, where the proposed similarity metric is used as the reward signal. In this model, we fine-tuned the mT5 model, which is a multilingual variant of the text-to-text-transfer-transformer (T5) (Raffel *et al.* 2020), that was pre-trained for 101

Table 3. Correlations between the semantic textual similarities predicted by the fine-tuned models in varying architecture and train sets, and the corresponding ground-truth similarity scores in the Semantic Textual Similarity benchmark test set. Pearson and Spearman correlations are reported as $\rho \times 100$

Model	Bi-Encoder		Cross-Encoder	
	Pearson	Spearman	Pearson	Spearman
Trained on STS				
BERTurk	81.97	81.43	83.32	82.22
mBERT	73.28	72.84	79.08	78.15
XLm-R	71.89	71.02	79.18	78.56
Trained on NLI first and then on STS				
BERTurk	82.85	83.31	85.36	84.59
mBERT	75.74	75.41	79.30	78.39
XLm-R	77.26	77.32	81.94	81.21

Table 4. Correlations between ROUGE, BERTScore, and proposed evaluation methods and the human judgments (Section 4.1.3). Pearson and Spearman correlations are reported as $\rho \times 100$

Metric	Human judgments							
	Relevance		Consistency		Fluency		Average	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
Rouge-1	42.8	43.9	28.2	32.4	21.4	20.3	36.8	37.5
Rouge-2	38.3	41.6	27.4	35.8	16.4	20.8	32.8	38.0
Rouge-L	41.8	42.0	26.3	28.9	20.2	18.6	35.2	35.1
BERTScore	45.5	45.8	25.1	22.5	24.7	19.9	37.9	38.1
Bi-Encoder BERTurk								
STS	55.4	52.8	30.3	30.0	25.6	26.7	44.3	45.9
NLI + STS	58.8	58.7	32.8	32.7	31.2	30.2	48.8	51.9
Cross-Encoder BERTurk								
STS	56.9	53.5	38.0	32.5	34.1	27.9	51.3	48.6
NLI + STS	60.0	59.2	40.0	34.2	34.6	29.3	53.5	52.1

languages, including Turkish, using a new Common Crawl-based dataset (Xue *et al.* 2021). The mT5 model decoder acts as an *agent* that interacts with the *environment* to take *actions* (predicting the next word in the sequence) based on the learned *policy* and observes as *reward* the semantic similarity of the generated and ground-truth summary.

The model is trained with the self-critical policy gradient training algorithm (Rennie *et al.* 2017) that was proposed for image captioning. This algorithm enhances the REINFORCE

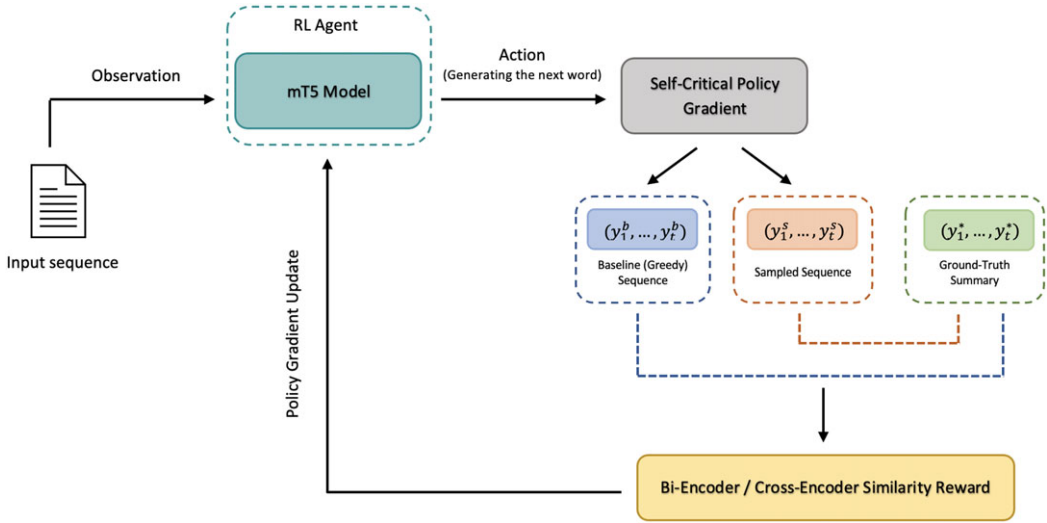


Figure 2. Self-critical policy gradient training process with bi-encoder/cross-encoder similarity rewards.

algorithm (Williams 1992), by introducing a learned baseline to enhance training stability. The baseline approximates the expected reward generated by the model itself. This approach enables models to improve their policies by comparing their performance to their self-generated sequences, addressing exposure bias issues, and enhancing overall training efficiency. Paulus *et al.* (2018) used the self-critical policy gradient training algorithm for abstractive summarization.

In this study, following a similar approach to Paulus *et al.* (2018), we generate two separate summary sequences at each training step: a sampled sequence and a baseline sequence. The sampled sequence y^s is obtained by sampling the words y_t^s from the probability distribution $p(y_t^s | y_1^s, \dots, y_{t-1}^s, X)$. The baseline sequence y^b is obtained by greedy decoding, that is by selecting the word with the highest posterior probability at each time step. We use the proposed semantic similarity ($SemSim(y^*, \hat{y})$) between a generated sequence \hat{y} and the ground-truth y^* as our reward function. Since the generated sequences have varying lengths, we normalize the log probabilities by dividing them with the sequence lengths. So, L_{rl} is defined as follows:

$$L_{rl} = (SemSim(y^*, y^b) - SemSim(y^*, y^s)) \cdot \left(\frac{1}{n} \sum_{t=1}^n \log p(y_t^s | y_1^s, \dots, y_{t-1}^s, X) \right) \tag{5}$$

In Equation (5), minimizing L_{rl} is equivalent to maximizing the conditional likelihood of the sampled sequence y^s when it receives a higher reward than the baseline sequence y^b . This optimization process encourages the model to generate sequences that are more likely to receive higher rewards compared to a baseline sequence.

The reinforcement learning process, using the bi-encoder and cross-encoder similarity rewards, is visualized in Fig. 2. The training is actually done in two stages: It starts with minimizing only the maximum-likelihood loss given in Equation (1). The second stage of the training consists of fine-tuning with a mixed training objective, which is a combination of maximum-likelihood training objective and the reinforcement training objective, as described in (Paulus *et al.* 2018):

$$L_{mixed} = \gamma L_{rl} + (1 - \gamma)L_{ml} \tag{6}$$

where L_{rl} is the reinforcement learning loss defined in Equation (5), L_{ml} is the maximum-likelihood loss defined in Equation (1), and γ is a scaling factor. Training only with L_{rl} to optimize a specific metric may not guarantee the quality and readability of the generated sequence; hence,

we have used a mixed training objective to optimize a specific measure with reinforcement training objective L_{rl} as well as to increase quality and readability of the output with maximum-likelihood training objective L_{ml} .

4. Experiments and quantitative results

We evaluate the effectiveness of the proposed similarity measures and abstractive summarization model, along with the pre-trained language models and datasets used in our experimental framework.

4.1. Evaluation of fine-tuned models for predicting semantic similarity

As explained in Section 3.3, we fine-tune different pre-trained language models for predicting the semantic similarity between two summaries. Here we describe the pre-trained models (Section 4.1.1) used in our evaluations, along with the results of two evaluations. In the first one, we evaluate the fine-tuned models in terms of the correlations of their predicted similarity scores and semantic similarities obtained in STSb-TR dataset (Section 4.1.2). In the second evaluation, we present comparison of human evaluations with the proposed and alternative evaluation measures (Section 4.1.3). Details of the Turkish summarization dataset are given in Section 4.2.1.

4.1.1. Pre-trained language models

We experimented with the pre-trained language models BERTurk, mBERT, and XLM-R for semantic similarity using both bi-encoder and cross-encoder architectures. For summarization, we used the mT5 model (Xue *et al.* 2021).

BERTurk and mBERT. BERT is a deep learning model based on pre-training bidirectional representations from unlabeled text using a masked language model pre-training objective. In contrast to unidirectional language models for pre-training, masked language model pre-training objective allows the representation to combine the left and right contexts (Devlin *et al.* 2019). In this study, we used BERTurk (Hugging Face 2021b), which is a BERT model for Turkish (Schweter 2020), and mBERT (Hugging Face 2021a) is a multilingual BERT pre-trained on the top 104 languages with the largest Wikipedia corpus (Pires, Schlinger, and Garrette 2019).

XLM-R. XLM-RoBERTa model (Hugging Face 2021c) has been pre-trained on a large filtered CommonCrawl data containing 100 languages 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.

mT5. mT5 (Xue *et al.* 2021) is a variant of the text-to-text-transfer-transformer (T5) model (Raffel *et al.* 2020) that was pre-trained for 101 languages on a new Common Crawl-based dataset. It has the same model architecture as T5 and its pre-training objective includes T5's self-supervised training, but not T5's supervised training. So, it has to be fine-tuned before applying to any downstream task like text summarization.

4.1.2. Evaluation of fine-tuned models for semantic similarity

Following the work of Reimers and Gurevych (2019), we fine-tuned our models on the NLI-TR dataset with a 3-way softmax-classifier for one epoch with the bi-encoder, and four epochs with the cross encoder. We used a batch size of 16, an Adam optimizer with learning rate of $2e-5$, and a linear learning rate warm-up over 10% of the training data. In our bi-encoder training settings,

we used the default mean pooling. Then, we fine-tuned each model on the STSb-TR dataset with 4 epochs and 10 random seeds as suggested by Reimers and Gurevych (2018, 2019). Only the XLM-R bi-encoder model was trained with 20 random seeds on STSb-TR to have at least 5 successful models.

The fine-tuned models are evaluated by calculating the Pearson and Spearman correlations between their predicted similarity scores and the gold similarity scores that are published in the original English STSb dataset (Cer *et al.* 2017). The results, presented in Table 3, show that the similarity scores predicted by our models are highly correlated with the ground truth, with the highest correlation being achieved by the BERTurk model trained with both datasets and a cross-encoder architecture (4th row of results). This shows that the proposed similarity measure, especially the BERTurk cross-encoder model, is a good option for measuring sentence similarity.

Another observation is that training the models first on the NLI-TR dataset increases the performance of the model in assessing semantic similarities. This is particularly noticeable for the XLM-R models. The BERTurk model gives very good results when trained directly on the STSb-TR dataset, but it shows even higher performance when trained on the NLI-TR dataset first. These results show the benefits of multi-task learning, in line with the literature.

In the rest of the paper, we evaluate the four best-performing models (BERTurk cross-encoder and bi-encoder models trained on STS or NLI + STS) as evaluation measures and use the two best-performing models (BERTurk cross-encoder and bi-encoder models trained on NLI + STS) for training abstractive summarization models.

4.1.3. Comparison of the proposed similarity measures with human evaluations

In addition to evaluating semantic similarities using a sentence similarity dataset (Section 4.1.2), we also evaluate their performance in evaluating generated summaries. For this, we compare the correlations of the predicted similarities with human evaluations and compare them to the correlations of ROUGE and BERTScore, for comparison.

For this evaluation, we use generated summaries by the pre-trained mT5 model (Gündeş 2021) for 50 randomly selected articles from the MLSUM Turkish dataset (see Section 4.2.1). The generated summaries were evaluated by native Turkish annotators on a scale of 1 (very bad) to 5 (very good), 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) and evaluate each criterion separately.

We then compared alternative semantic similarity-based evaluation methods, namely ROUGE scores, BERTScore, and the proposed evaluations, and computed their correlations with human judgments about the quality of the summarization. The results shown in Table 4 show that semantic similarity-based evaluation correlates with human judgments better than ROUGE and BERTScore evaluations. Furthermore, cross-encoder-based similarity measures showed higher correlations with human evaluations compared to bi-encoder-based similarity measures, as also observed in previous work (Beken Fikri *et al.* 2021).

All the correlations were significant ($p < .05$) except for the correlations between Fluency and bi-encoder BERTurk model trained on STS, BERTScore, ROUGE-L as well as correlations between BERTScore and consistency. The Pearson and Spearman correlations are also visualized in Figs. 3 and 4, respectively.

The details of the summary annotation process are as follows. The annotators' (3 university students, 1 Ph.D. student, and 1 professor) age ranged between 23 and 66 (avg. 32,8). The participants were paid a small amount for successful completion of the evaluations. They were given one text and the corresponding summary at a time. Three examples were given in the instructions to illustrate how to rate the summaries. Each participant attended only once and there were no time limit in the evaluation process. The analysis of the annotations showed that average relevance was 3.5 ± 0.8 , average consistency was 4.5 ± 0.8 , and average fluency was 4.3 ± 0.8 . Inter-annotator

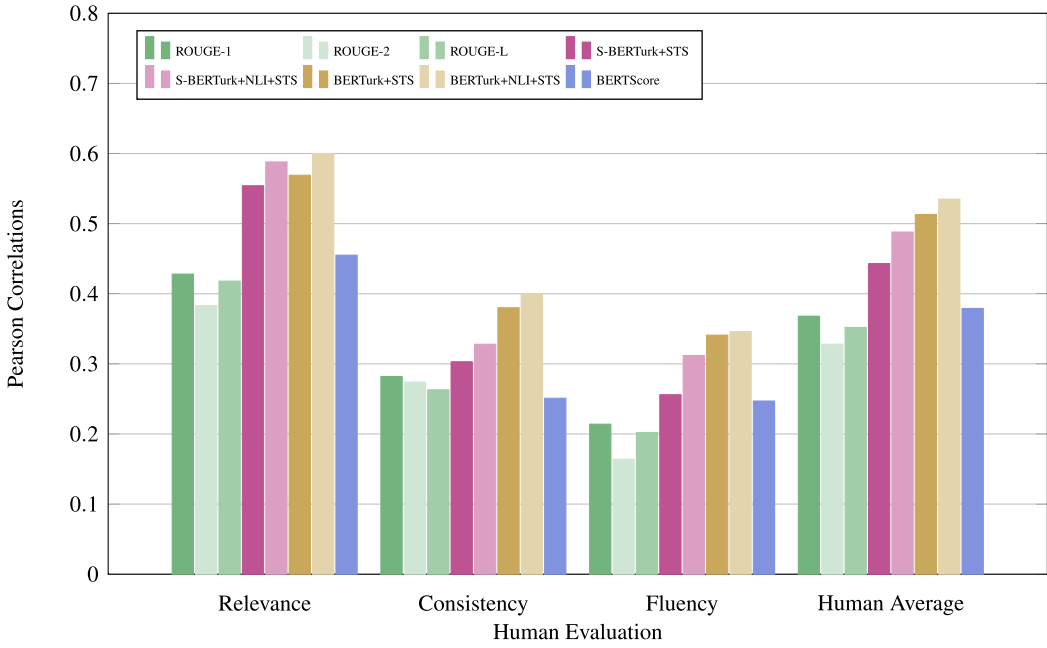


Figure 3. Pearson correlations of the evaluations with human judgments.

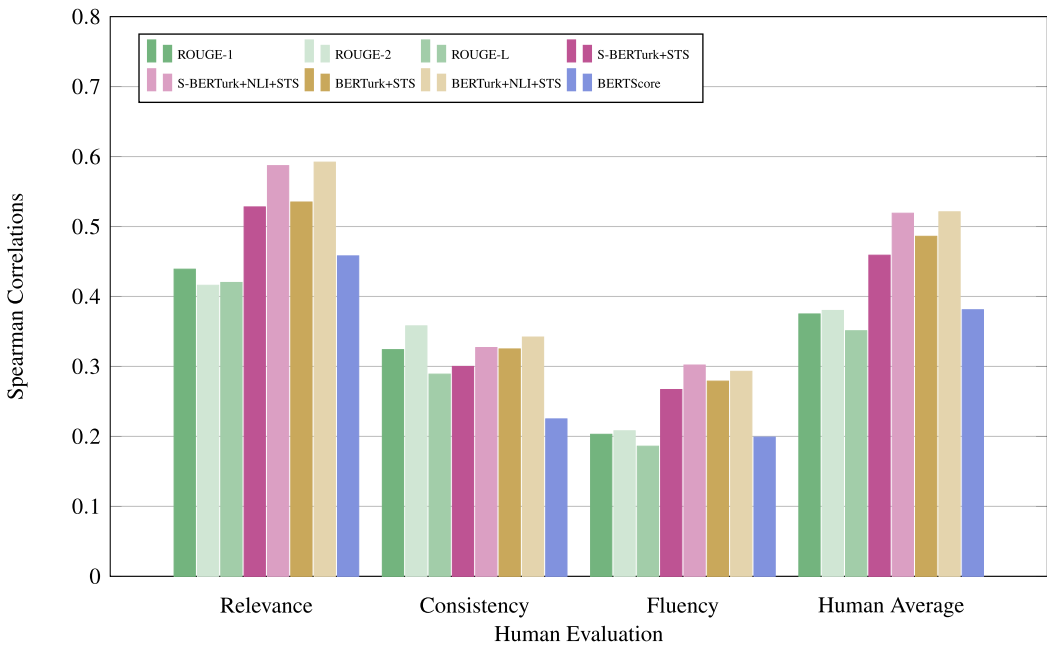


Figure 4. Spearman correlations of the evaluations with human judgments.

agreement in terms of Krippendorff's alpha scores (Krippendorff 2011) for fluency, consistency, and relevance were 0.33, 0.34, and 0.35, respectively, which shows that there is a fair agreement. Each of the three criteria (relevance, fluency, and consistency) has 50 evaluations.

4.2. Evaluation of the proposed summarization models

In this section, we compare the performance of the summarization models that are fine-tuned using the maximum-likelihood estimation (MLE)-only or the reinforcement learning (RL) paradigm with the proposed similarity measures as the reward signal. The evaluation consists of two parts. In the first part, we give different evaluation results (ROUGE, BERTScore, or proposed) for summaries generated by different models (Section 4.2.3). In the second one, we give human assessment of summary quality (Section 4.2.4).

4.2.1. Summary dataset

We use the MLSUM Turkish dataset for training and testing our summarization models. MLSUM is the first large-scale MultiLingual SUMmarization dataset that contains 1.5M+ article/summary pairs including Turkish (Scialom *et al.* 2020). The dataset was created using the same methods as the CNN/DailyMail dataset (Nallapati *et al.* 2016). They used news articles as the text source and their highlights/descriptions as the summary (Scialom *et al.* 2020).

The MLSUM Turkish subset was built by crawling the articles on Turkish website (Internet Haber) between 2010 and 2019. All articles under 50 words and summaries under 10 words were eliminated. The data was split into train, validation, and test sets, with respect to the publishing dates. The data from 2010 to 2018 was used for training, data between January and April 2019 was used for validation, and data up to December 2019 was used for test (Scialom *et al.* 2020). There are 249,277 train, 11,565 validation, and 12,775 test samples in the dataset.

4.2.2. Summarization models

In our summarization models, we used mT5 (Xue *et al.* 2021), which is a variant of T5 model (Raffel *et al.* 2020). This model was then fine-tuned on MLSUM Turkish dataset for abstractive summarization.

The details of the training process are as follows. We started fine-tuning the mT5, by minimizing the maximum-likelihood loss for 15 epochs, using a batch size of 8, accumulation steps of 4, and learning rate of $10e-5$ with Adafactor optimization and cosine annealing learning rate decay schedule (Loshchilov and Hutter 2016). The maximum news length was set to 512 and maximum summary length was determined as 120. We initialized our reinforcement learning models with the fine-tuned models that best performed on the validation set in terms of bi-encoder and cross-encoder similarity scores (Beken Fikri 2023). We continued training one more epoch with the mixed objective learning function (Equation (6)) for cross-encoder and bi-encoder similarity rewards separately. The scaling factor γ was set to 0.95. We also continued training each initial model one more epoch with only maximum-likelihood loss, to obtain comparable results. Further training was not possible due to time and computational resources.

In the continued MLE-only and reinforcement learning models, maximum news length and maximum summary length were set to 256 and 60, respectively, and batch size was 4 due to memory constraints. It should be noted that the same maximum summary length is used for target summaries in both training and testing. We utilized the scripts provided by Yadav *et al.* (2021) and adapted them for our work.

Table 5. Results of the mT5 summarization models trained on MLSUM dataset. We reported the average results for MLE-only (that best performed on validation set in terms of bi-encoder and cross-encoder similarity results) and RL training objectives on the MLSUM test set (Section 4.2). All values are scaled to 100

Model	Similarity based evaluation		ROUGE			BERTScore
	Cross encoder	Bi encoder	R-1	R-2	R-L	
<i>Bi-encoder</i>						
MLE- <i>initial</i>	58.70	66.32	33.14	22.47	31.82	76.09
MLE- <i>only</i>	59.14	66.39	33.60	22.82	32.19	75.88
MLE + RL <i>ROUGE-L</i>	59.17	66.47	33.62	22.97	32.20	75.77
MLE + RL <i>Bi-encoder</i>	59.34	66.68	33.62	22.90	32.22	75.76
<i>Cross-encoder</i>						
MLE- <i>initial</i>	58.57	66.14	33.27	22.56	31.95	76.22
MLE- <i>only</i>	59.23	66.48	33.70	22.89	32.28	75.93
MLE + RL <i>ROUGE-L</i>	59.07	66.43	33.60	22.91	32.20	75.71
MLE + RL <i>cross-encoder</i>	59.29	66.57	33.64	22.97	32.21	75.79

4.2.3. Evaluation of summarization models with different metrics

We evaluated the summaries generated by the fine-tuned mT5 models (Section 4.2.2) for input articles from the MLSUM Turkish test set. Specifically, we compare the semantic similarity, ROUGE, and BERTScore evaluation results for summarization models trained with MLE and MLE + RL with bi-encoder and cross-encoder semantic similarity rewards, as well as ROUGE-L rewards. The results shown in Table 5 indicate the average performance of the mT5 summarization models on the MLSUM test set. The given results vary significantly across the different metrics. This is expected, since each metric measures different aspects of the similarity between the ground-truth summary and the generated summary. The lowest scores are obtained with ROUGE, while the highest scores are obtained with BERTScore. However, the scores are not directly comparable among themselves and a high BERTScore does not necessarily imply that the generated summaries are of higher quality. One limitation of BERTScore is that it primarily focuses on token-level similarity, which may result in overestimated scores.

For the case of the proposed evaluation models (similarity-based evaluation), we see that training the summarization model with the RL paradigm gives the best results for both the bi-encoder or cross-encoder models (rows 4 and 8 of the results), compared to MLE-only or RL with ROUGE-L score as the reward metric. This is not the case for ROUGE or BERTScore. We thus see that improving the evaluation measure yields an improvement in overall quality of the summaries, as we already showed that the bi-encoder and cross-encoder similarity scores are better aligned with human judgments (Section 4.1.3).

4.2.4. Human evaluations for the summarization models

In addition to the quantitative analysis of the summarization models using well-known ROUGE, BERTScore, and our proposed measures, we analyzed the effectiveness of our proposed framework qualitatively as well. For this, we collected human evaluations of generated summaries in three dimensions to better assess the model outcomes. To collect human evaluations, we first identified the list of generated summaries, which were different in all fine-tuned models and from

Table 6. Human evaluations for the summarization models (higher the better). Results are shown for the bi-encoder and cross-encoder models separately and together (all). *n* is the sample size (Section 4.2.4)

	Human judgments			
	Relevance	Consistency	Fluency	Average
Bi-encoder <i>n</i> = 25				
MLE-only	3.02 ± 1.28	3.18 ± 1.38	2.99 ± 1.46	3.06 ± 1.21
MLE + RL ROUGE-L	2.93 ± 1.28	3.18 ± 1.34	2.88 ± 1.52	2.99 ± 1.22
MLE + RL Bi-encoder	3.12 ± 1.38	3.40 ± 1.38	3.10 ± 1.56	3.21 ± 1.29
Cross-encoder <i>n</i> = 25				
MLE-only	2.87 ± 1.29	3.17 ± 1.37	3.14 ± 1.45	3.06 ± 1.19
MLE + RL ROUGE-L	3.11 ± 1.25	3.38 ± 1.38	3.30 ± 1.43	3.26 ± 1.19
MLE + RL cross-encoder	3.03 ± 1.32	3.54 ± 1.34	3.30 ± 1.48	3.29 ± 1.20
All <i>n</i> = 50				
MLE-only	2.94 ± 1.29	3.18 ± 1.37	3.07 ± 1.45	3.06 ± 1.20
MLE + RL ROUGE-L	3.02 ± 1.27	3.28 ± 1.37	3.09 ± 1.49	3.13 ± 1.21
MLE + RL similarity	3.08 ± 1.35	3.47 ± 1.36	3.20 ± 1.52	3.25 ± 1.24

the actual summaries. Then, we randomly sampled 25 documents for bi-encoder models and 25 documents for cross-encoder models (50 documents in total) as human evaluation sets.

In the evaluation process, the participants were given the news article and three generated summaries from the MLE-only, MLE + RL model with ROUGE-L reward and the bi-encoder or cross-encoder reward. They were asked to rate each predicted summary on a scale of 1 (very bad) to 5 (very good) 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) and evaluate each criterion separately. The ground-truth summary was not provided during the evaluation of these three generated summaries. Overall, the participants evaluated 150 summaries and they were paid TRY500 for non-degenerate completion of the evaluations. Five native Turkish speakers participated in our study (2 undergraduate, 1 university graduate, and 2 graduate students) ages between 21 and 29 (avg. 24). Inter-annotator agreement in terms of Krippendorff's alpha scores (Krippendorff 2011) for relevance, fluency, and consistency were 0.17, 0.18, and 0.22, respectively, showing that there is a slight to fair agreement. Each of the three criteria (relevance, fluency, and consistency) has 150 evaluations.

The results of the human evaluations are presented in Table 6. We first observe that the RL model with the bi-encoder similarity reward outperformed the MLE-only model and the RL model with the ROUGE-L reward in all dimensions (relevance, consistency, fluency) and on average. Similarly, the RL model with cross-encoder similarity reward outperformed the MLE-only model and RL model with ROUGE-L reward in all but one dimension. Finally, the overall comparison (last 3 rows of Table 6) shows that the models trained with similarity rewards performed better than the MLE-only model, in all the evaluation criteria.

In Tables 7 and 8, sample articles from MLSUM Turkish test set with the reference and generated summaries are given along with the corresponding translations. In the examples, we see that the summaries generated by RL models with bi-encoder and cross-encoder similarity rewards are

Table 7. Sample article with the reference and generated summaries from MLE-only model and MLE + RL models with ROUGE-L and bi-encoder similarity rewards, respectively

Article

Bursa'nın İznik ilçesi Abdulvahap Mahallesi'ndeki evlerinden sabah saatlerinde tarlalarına çalışmaya giden vatandaşlar, yol kenarındaki bahçede, ceviz ağacına asılı 2 ölü sincap gördü. Hemen polis ile belediyeyi arayan vatandaşlar duruma tepki gösterdi. Olay yerine gelen ekipler, bacaklarından iple ağaca asılan ölü sincapları, ağaçtan indirdi. Polis, sincapları öldürüp, ağaca asan kişi ya da kişileri tespit etmek için soruşturma başlattı

Citizens, who went to work in their fields in the morning hours from their homes in the Abdulvahap District of Bursa's Iznik district, saw 2 dead squirrels hanging on a walnut tree in the garden by the roadside. Citizens who immediately called the police and the municipality reacted to the situation. The teams who came to the scene took the dead squirrels, which were hung from the tree with a rope from their legs, from the tree. The police launched an investigation to identify the person or persons who killed the squirrels and hung them on the tree

Reference summary

BURSA'nın İznik ilçesinde, öldürüldükten sonra bacaklarından ceviz ağacına iple asılmış 2 sincap, bulundu. Polis, sincapları öldürüp ağaca asanları arıyor

After being killed, 2 squirrels were found hanging from a walnut tree by a rope in the Iznik district of BURSA. The police are looking for those who killed the squirrels and hung them on the tree

Generated summary (MLE-only)

Bursa'nın İznik ilçesi Abdulvahap Mahallesi'ndeki evlerinden sabah saatlerinde tarlalarına çalışmaya giden vatandaşlar, ceviz ağacına asılı 2 ölü sincap gördü

Citizens, who went to work in their fields in the morning hours from their homes in the Abdulvahap District of Iznik district of Bursa, saw 2 dead squirrels hanging on a walnut tree

Semantic similarity scores BERTurk+NLI + STS (Cross-Encoder/ Bi-Encoder): 69.53/ 63.03

Generated summary (MLE + RL ROUGE-L)

Bursa'nın İznik ilçesi Abdulvahap Mahallesi'ndeki evlerinden sabah saatlerinde tarlalarına çalışmaya giden vatandaşlar, yol kenarındaki bahçede, ceviz ağacına asılı 2 ölü sincapları, ağaçtan indirdi. Polis, sincapları . . .

Bursa'nın İznik ilçesi Abdulvahap Mahallesi'ndeki evlerinden sabah saatlerinde tarlalarına çalışmaya giden vatandaşlar, yol kenarındaki bahçede, ceviz ağacına asılı 2 ölü sincapları, ağaçtan indirdi. Polis, sincapları . . .

Semantic similarity scores BERTurk+NLI + STS (Cross-Encoder/ Bi-Encoder): 70.12/ 64.67

Generated summary (MLE + RL Bi-Encoder)

Bursa'nın İznik ilçesinde, yol kenarındaki bahçede, ceviz ağacına asılı 2 ölü sincap bulundu. Polis, bacaklarından iple ağaca asılan 2 ölü sincapları öldürüp, ağaca asan kişi ya da kişileri te . . .

Two dead squirrels were found hanging from a walnut tree in the garden by the road in Iznik district of Bursa. The police identified . . . the person or persons who killed 2 dead squirrels hanging from the tree with a rope from their legs and hanged them on the tree

Semantic similarity scores BERTurk+NLI + STS (Cross-Encoder/ Bi-Encoder): 81.02/ 89.62

more similar to the reference summaries, compared to the summaries generated by ROUGE-L reward and MLE-only models, in terms of semantic similarity scores. It should be noted that the reference or generated summaries could be truncated as we restricted the maximum summary length in our experiments.

Table 8. Sample article with the reference and generated summaries from MLE-only model and MLE + RL models with ROUGE-L and cross-encoder similarity rewards, respectively

Article

Olay, başkent Tiflis'teki bir teknoloji mağazasında meydana geldi. 32 yaşında olduğu belirtilen bir kişi, mağazaya girerek, bir kadın çalışanı bıçakla rehin aldı. Soyguncu, mağazayı soymaya çalıştı. Güvenlik kamerasına yansıyan görüntülerde, şüphelinin kadını koluyla boynundan kavrayarak tuttuğu ve elindeki bıçağı sallayarak diğer çalışanları tehdit ettiği görüldü. Polis ekiplerinin geldiğini gören şüpheli, kadını bırakarak ellerini havaya kaldırdı ve teslim oldu. Gözaltına alınan şüphelinin 7 ila 10 yıl arası hapis cezasına çarptırılabilceği belirtiliyor

The incident occurred in a technology store in the capital, Tbilisi. A 32-year-old man entered the store and took a female employee hostage with a knife. The robber tried to rob the store. In the footage reflected on the security camera, it was seen that the suspect grabbed the woman by the neck with his arm and threatened other employees by waving the knife in his hand. The suspect, who saw the police teams coming, left the woman, raised his hands in the air and surrendered. It is stated that the suspect, who was taken into custody, could be sentenced to 7 to 10 years in prison

Reference summary

Gürcistan'da soygun amacıyla bir teknoloji mağazasına giren kişi, kadın çalışanı bıçakla rehin aldı. Polisi görünce kadını bırakıp teslim olan şüpheli, gözaltına alınırken, olay ise saniye s. . .

A person who entered a technology store in Georgia with the intention of robbery took the female employee hostage with a knife. The suspect, who left the woman and surrendered when he saw the police, was taken into custody, while the incident was second by s. . .

Generated summary (MLE-only)

İtalya'nın başkenti Tiflis'te bir teknoloji mağazasına girerek, bir kadın çalışanı bıçakla rehin aldı. Olayda 3 kişi gözaltına alındı

He broke into a technology store in Tbilisi, the capital of Italy, and took a female employee hostage with a knife. Three people were detained in the incident

Semantic similarity scores BERTurk+NLI + STS (Cross-Encoder/ Bi-Encoder): 65.10/ 68.69

Generated summary (MLE + RL ROUGE-L)

ANTALYA'nın Tiflis ilçesinde bir teknoloji mağazasına girerek, bir kadın çalışanı bıçakla rehin aldı. Polis ekiplerinin geldiğini gören şüpheli, kadını bırakarak ellerini havaya kaldırdı.

He entered a technology store in the Tbilisi district of ANTALYA and took a female employee hostage with a knife. The suspect, who saw the police teams coming, raised his hands in the air, leaving the woman.

Semantic similarity scores BERTurk+NLI + STS (Cross-Encoder/ Bi-Encoder): 72.69/ 85.93

Generated summary (MLE + RL Cross-Encoder)

İtalya'nın başkenti Tiflis'teki bir teknoloji mağazasına girerek, bir kadın çalışanı bıçakla rehin aldı. Polis ekiplerinin geldiğini gören şüpheli, kadını bırakarak ellerini havaya kaldırdı ve teslim oldu

He broke into a technology store in Tbilisi, the capital of Italy, and took a female employee hostage with a knife. The suspect, who saw the police teams coming, left the woman, raised his hands in the air and surrendered

Semantic similarity scores BERTurk+NLI + STS (Cross-Encoder/ Bi-Encoder): 72.70/ 87.21

5. Discussion

5.1. Findings and contributions

Our work introduced novel BERT-based semantic similarity evaluation measures to assess the quality of abstractive summaries. We showed that the proposed similarity models have higher correlations with human evaluations, compared to ROUGE scores and BERTScore. By observing

that our proposed evaluation measures have high correlations with human preferences, we used the similarity measures as rewards in a reinforcement learning framework. This unique approach offers a novel way to guide the model toward generating summaries that better align with human preferences and offer semantic coherence. We show the effectiveness of our strategy through the outcomes of ablation experiments. Our results show that training by optimizing semantic similarity scores instead of ROUGE scores or MLE-only training yields better summarization results, in terms of both semantic similarity scores and human evaluations.

We have shown that for low-resource languages such as Turkish, the translated STSb dataset works well for fine-tuning BERT models for predicting semantic similarity. In addition, the two-step fine-tuning, first on the natural language inference task using the NLI-TR dataset and then with STSb-TR, has shown to be more successful compared to only using the most related dataset (STSb-TR). This is in line with other results in literature, where multi-task learning helps with generalization. The introduction of the NLI-TR dataset, with its extensive training samples, added a significant volume of diverse linguistic patterns and contextual information to our model. Furthermore, the NLI-TR dataset is designed to capture entailment and contradiction relationships between sentences. This dataset-specific feature corresponds to the nature of semantic similarity tasks, where measuring the degree of similarity often involves analyzing the contextual alignment and relatedness between sentences. Hence, the improvement we observed in semantic similarity model outcomes can be attributed to the transfer of knowledge acquired during NLI-TR fine-tuning to the STSb dataset.

5.2. Limitations

BERT-based similarity models have been shown to be effective in various natural language processing tasks; however, they may struggle with detecting minor differences between compared sentences. In our study, BERT models were fine-tuned on semantic textual similarity benchmark dataset, which has sentence pairs along with their human annotated similarity scores, including sentence pairs with small lexical differences but high semantic similarity. However, the dataset size is also rather small, with a limited scope and biased toward specific topics, genres, or domains. We applied a pre-training step on the NLI-TR dataset successfully for training our semantic similarity models, to remedy this size limitation to some extent.

In our summarization experiments, we used the mT5 model, which is a multilingual variant of the T5 architecture. However, unlike its predecessor, mT5 lacks the benefit of supervised pre-training, needing fine-tuning before it can be efficiently applied to downstream tasks. This fine-tuning process, involving reinforcement learning, requires a significant amount of computational time and resources, often requiring an extensive duration. The substantial time required for fine-tuning limited the ability to investigate a wide range of hyper-parameter settings.

We acknowledge certain limitations in our study related to the human evaluations. Our human evaluations were conducted in two phases: (1) Correlating human judgments with our proposed similarity models, and (2) comparing MLE-only and RL models with both ROUGE and our similarity rewards. While the first phase involved evaluating 50 article/summary pairs, the second phase required participants to assess 150 summaries across 50 articles, resulting in a higher cognitive workload. One of the main limitations was the level of subjectivity involved in the human evaluations. Together with the cognitive load, the subjectivity was more prominent in the second phase, resulting in lower inter-annotator agreement. We provided a larger compensation to participants to assure their motivation and involvement, within budgetary constraints.

Automatic summarization systems are relatively novel but have the potential to be used widely and increase productivity. On the other hand, the dataset used is a news dataset, which could be biased in many ways, such as the represented viewpoints. The model trained on this data might pick up on these biases in its generated summaries.

6. Summary and conclusion

In this work, we focused on two main issues in abstractive summarization: how to evaluate the results and what is a good training objective. We presented semantic similarity-based summarization evaluation measures and a reinforcement learning framework with the semantic similarity rewards.

We proposed evaluation measures using similarity scores obtained by fine-tuning the BERTurk model using cross-encoder and bi-encoder model architectures on NLI-TR (Budur *et al.* 2020) and STSb-TR (Beken Fikri *et al.* 2021) datasets. We showed that the proposed evaluation measures have better correlations with human evaluations compared to ROUGE scores, according to both Pearson and Spearman correlations. We further showed that using bi-encoder and cross-encoder similarities as rewards improved the model results in terms of the proposed evaluation measures, as well as BERTScore and ROUGE scores. Our qualitative analyses demonstrated that the proposed models can generate summaries that are more similar to the ground truth, as compared to MLE-only models and RL models with ROUGE rewards.

It is worth mentioning that our rewards are not model-dependent in our reinforcement learning framework and can be explored in other downstream sequence-to-sequence tasks like paraphrase generation, text simplification, and semantic search. Also, the suggested framework can be applied to other languages following the described methodology.

Acknowledgments. None.

Competing interests. The authors declare none.

References

- Bahdanau D., Brakel P., Xu K., Goyal A., Lowe R., Pineau J., Courville A. and Bengio Y. (2016). An actor-critic algorithm for sequence prediction. arXiv preprint arXiv: 1607.07086.
- Baykara B. and Güngör T. (2022a). Abstractive text summarization and new large-scale datasets for agglutinative languages Turkish and Hungarian. *Language Resources and Evaluation* 56(3), pp. 973–1007.
- Baykara B. and Güngör T. (2022b). Turkish abstractive text summarization using pretrained sequence-to-sequence models. *Natural Language Engineering*, 29(5), pp. 1275–1304.
- Beken Fikri F. (2023). *Abstractive Summarization with Semantically-Driven Evaluation and Reinforcement Learning*. PhD Thesis, Sabancı University.
- Beken Fikri F., Oflazer K. and Yanıkoğlu B. (2021). *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 (ACL), pp. 24–33.
- Bengio S., Vinyals O., Jaitly N. and Shazeer N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems* 28.
- Böhm F., Gao Y., Meyer C.M., Shapira O., Dagan I. and Gurevych I. (2019). *Better rewards yield better summaries: Learning to summarise without references*. 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)*, pp. 3110–3120.
- Bowman S., Angeli G., Potts C. and Manning C.D. (2015). *A large annotated corpus for learning natural language inference*. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 632–642.
- Budur E., Özçelik R., Güngör T. and Potts C. (2020). *Data and representation for Turkish natural language inference*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8253–8267.
- Cer D., Diab M., Agirre E., Lopez-Gazpio I. and Specia L. (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)*, pp. 1–14.
- Chen Y.-C. and Bansal M. (2018). *Fast abstractive summarization with reinforce-selected sentence rewriting*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pp. 675–686.
- Chopra S., Auli M. and Rush A.M. (2016). *Abstractive sentence summarization with attentive recurrent neural networks*. In *Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies*, pp. 93–98.
- Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave É., Ott M., Zettlemoyer L. and Stoyanov V. (2020). *Unsupervised cross-lingual representation learning at scale*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451.

- Devlin J., Chang M.-W., Lee K. and Toutanova K. (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 (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186.
- Dong L., Yang N., Wang W., Wei F., Liu X., Wang Y., Gao J., Zhou M. and Hon H.-W. (2019). Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems* 32.
- Dou Z.-Y., Liu P., Hayashi H., Jiang Z. and Neubig G. (2021). Gsum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4830–4842.
- Durmuş E., He H. and Diab M. (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 (ACL)*, pp. 5055–5070.
- Eddine M.K., Shang G. and Vazirgiannis M. (2023). DATScore: Evaluating translation with data augmented translations. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 912–922.
- Falke T., Ribeiro L.F., Utama P.A., Dagan I. and Gurevych I. (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 (ACL)*, pp. 2214–2220.
- Gehrmann S., Deng Y. and Rush A.M. (2018). Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4098–4109.
- Gündeş Z. (2021). mt5-small based Turkish summarization system (accessed 17 March 2021). Available at <https://huggingface.co/ozcangundes/mt5-small-turkish-summarization>.
- Google Cloud (2021). Translating text (basic) (accessed 19 February 2021). Available at <https://cloud.google.com/translate/docs/basic/translating-text>.
- Hugging Face (2021a). BERT multilingual base model (cased) (accessed 3 May 2021). Available at <https://huggingface.co/bert-base-multilingual-cased>.
- Hugging Face (2021b). dbmdz Turkish BERT model (cased) (accessed 3 May 2021). Available at <https://huggingface.co/dbmdz/bert-base-turkish-cased>.
- Hugging Face (2021c). XLM-RoBERTa (base-sized model) (accessed 3 May 2021). Available at <https://huggingface.co/xlm-roberta-base>.
- Hyun D., Wang X., Park C., Xie X. and Yu H. (2022). Generating multiple-length summaries via reinforcement learning for unsupervised sentence summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2939–2951.
- Krippendorff K. (2011). Computing krippendorff's alpha-reliability.
- Kryściński W., McCann B., Xiong C. and Socher R. (2020). Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332–9346.
- Kryściński W., Paulus R., Xiong C. and Socher R. (2018). Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1808–1817.
- Laban P., Hsi A., Canny J. and Hearst M.A. (2020). The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5135–5150.
- Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V. and Zettlemoyer L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880.
- Lin C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization Branches Out, Post Conference Workshop of ACL 2004*.
- Liu C.-W., Lowe R., Serban I.V., Noseworthy M., Charlin L. and Pineau J. (2016). How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2122–2132.
- Liu Y., Gu J., Goyal N., Li X., Edunov S., Ghazvininejad M., Lewis M. and Zettlemoyer L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics* 8, 726–742.
- Liu Y. and Lapata M. (2019). 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)*, pp. 3730–3740.
- Liu Y. and Liu P. (2021). Simcls: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 1065–1072.
- Liu Y., Liu P., Radev D. and Neubig G. (2022). Brio: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2890–2903.
- Lo C.K. (2019). YiSi-A unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *WMT 2019*, 507.

- Loshchilov I. and Hutter F. (2016). SGDR: Stochastic gradient descent with warm restarts. arXiv preprint arXiv: 1608.03983.
- Nallapati R., Zhou B., dos Santos C., Gülçehre Ç. and Xiang B. (2016). *Abstractive text summarization using sequence-to-sequence RNNs and beyond*. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290.
- Papineni K., Roukos S., Ward T. and Zhu W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318.
- Pasunuru R. and Bansal M. (2018). Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 2 (Short Papers)*, pp. 646–653.
- Paulus R., Xiong C. and Socher R. (2018). A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations (ICLR)*.
- Pires T., Schlinger E. and Garrette D. (2019). How multilingual is multilingual BERT?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001.
- Qi W., Yan Y., Gong Y., Liu D., Duan N., Chen J., Zhang R. and Zhou M. (2020). ProphetNet: Predicting future n-gram for sequence-to-sequence pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2401–2410.
- Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W. and Liu P.J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21, 1–67.
- Ranzato M., Chopra S., Auli M. and Zaremba W. (2016). Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations (ICLR)*.
- Rei R., Stewart C., Farinha A.C. and Lavie A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics (ACL), pp. 2685–2702, Online.
- Reimers N. and Gurevych I. (2018). Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. arXiv preprint arXiv: 1803.09578.
- Reimers N. and Gurevych I. (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)*, pp. 3982–3992.
- Reimers N. and Gurevych I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4512–4525.
- Rennie S.J., Marcheret E., Mroueh Y., Ross J. and Goel V. (2017). Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, pp. 1179–1195.
- Roit P., Ferret J., Shani L., Aharoni R., Cideron G., Dadashi R., Geist M., Girgin S., Husenot L., Keller O., Momchev N., Ramos Garea S., Stanczyk P., Vieillard N., Bachem O., Elidan G., Hassidim A., Pietquin O. and Szepeski I. (2023). Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics, pp. 6252–6272.
- Rush A.M., Harvard S., Chopra S. and Weston J. (2015). A neural attention model for sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Schweter S. (2020). BERTurk-BERT models for Turkish, 3770924, Online. Available at <https://doi.org/10.5281/zenodo>.
- Scialom T., Dray P.-A., Lamprier S., Piwowarski B. and Staiano J. (2020). MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8051–8067.
- Scialom T., Lamprier S., Piwowarski B. and Staiano J. (2019). Answers unite! Unsupervised metrics for reinforced summarization models. 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)*, pp. 3246–3256.
- See A., Liu P.J. and Manning C.D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pp. 1073–1083.
- Sellam T., Das D. and Parikh A. (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics (ACL), pp. 7881–7892, Online.
- Sharma E., Huang L., Hu Z. and Wang L. (2019). An entity-driven framework for abstractive summarization. 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)*, pp. 3280–3291.
- Stiennon N., Ouyang L., Wu J., Ziegler D.M., Lowe R., Voss C., Radford A., Amodei D. and Christiano P. (2020). Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 3008–3021.
- Sutskever I., Vinyals O. and Le Q.V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*.
- Thompson B. and Post M. (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)*, pp. 90–121.

- Tu Z., Lu Z., Liu Y., Liu X. and Li H.** (2016). *Modeling coverage for neural machine translation*. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pp. 76–85.
- Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł. and Polosukhin I.** (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* 30.
- Wang A., Cho K. and Lewis M.** (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 (ACL)*, pp. 5008–5020.
- Williams A., Nangia N. and Bowman S.** (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 (NAACL): Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122.
- Williams R.J.** (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 8(3), 229–256.
- Xu S., Zhang X., Wu Y. and Wei F.** (2022). Sequence level contrastive learning for text summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 11556–11565.
- Xue L., Constant N., Roberts A., Kale M., Al-Rfou R., Siddhant A., Barua A. and Raffel C.** (2021). *mT5: A massively multilingual pre-trained text-to-text transformer*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies*, pp. 483–498.
- Yadav S., Gupta D., Abacha A.B. and Demner-Fushman D.** (2021). Reinforcement learning for abstractive question summarization with question-aware semantic rewards. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP) (Volume 2: Short Papers)*, pp. 249–255.
- Yuan W., Neubig G. and Liu P.** (2021). BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems* 34, pp. 27263–27277.
- Zhang J., Zhao Y., Saleh M. and Liu P.** (2020a). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*. PMLR, pp. 11328–11339.
- Zhang S. and Bansal M.** (2019). Addressing semantic drift in question generation for semi-supervised question answering. 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)*, pp. 2495–2509.
- Zhang T., Kishore V., Wu F., Weinberger K.Q. and Artzi Y.** (2020b). BERTScore: Evaluating text generation with BERT. In *Proceedings of International Conference on Learning Representations*.
- Zhang Y., Merck D., Tsai E., Manning C.D. and Langlotz C.** (2020c). Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5108–5120.
- Zhao W., Peyrard M., Liu F., Gao Y., Meyer C.M. and Eger S.** (2019). Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. 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)*, pp. 563–578.
- Zhao W., Strube M. and Eger S.** (2023). DiscoScore: Evaluating text generation with bert and discourse coherence. In *Proceedings of the 17th Conference of the Association for Computational Linguistics*, pp. 3847–3865.
- Zhao Z., Cohen S.B. and Webber B.** (2020). Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2237–2249.