MUKAYESE: Turkish NLP Strikes Back

Ali Safaya[†,*] **Emirhan Kurtuluş**[‡] **Arda Göktoğan**[§] **Deniz Yuret**[†]

[†] KUIS AI Center, Koç University [†] Computer Engineering Department, Koç University
[‡] Cağaloğlu Anadolu Lisesi, Istanbul [§] Computer Engineering Department, Bilkent University

Abstract

Having sufficient resources for language X lifts it from the *under-resourced* languages class, but not necessarily from the *under-researched* class. In this paper, we address the problem of the absence of organized benchmarks in the Turkish language. We demonstrate that languages such as Turkish are left behind the state-of-the-art in NLP applications. As a solution, we present MUKAYESE, a set of NLP benchmarks for the Turkish language that contains several NLP tasks. We work on one or more datasets for each benchmark and present two or more baselines. Moreover, we present four new benchmarking datasets in Turkish for language modeling, sentence segmentation, and spell checking. All datasets and baselines are available under: https://github.com/alisafaya/mukayese

1 Introduction

Although some human languages, such as Turkish, are not classified as under-resourced languages, only a few research communities are working on them (Joshi et al., 2020). As a result, they are left behind in developing state-of-the-art systems due to the lack of organized benchmarks and baselines. In this study, we aim to address this gap for the Turkish language with MUKAYESE (Turkish word for "comparison/benchmarking"), an extensive set of datasets and benchmarks for several Turkish NLP tasks.

We survey several tasks in Turkish NLP and observe an absence of organized benchmarks and research. We demonstrate how the lack of benchmarks affects under-studied languages such as Turkish and how it can keep the state of research behind the state-of-the-art of NLP. We accomplish this by presenting state-of-the-art baselines that outperform previous work significantly. We believe that MUKAYESE will set a basis for boosting NLP research for Turkish. Therefore, we encourage research communities from other under-studied languages to follow a similar path.

In our work on MUKAYESE, we study seven NLP tasks in the Turkish language. We evaluate available datasets in Turkish for these tasks and describe the process of creating four new datasets for tasks that do not have accessible datasets. Furthermore, in addition to evaluating existing methods, we provide at least two baseline models/methods per task. More details are enlisted in Table 1.

Our overall contribution to Turkish NLP can be summarized as the following: (a) Set of seven organized benchmarks for NLP. (b) Four new datasets in Turkish for language modeling, sentence segmentation, as well as spellchecking and correction. (c) Dataset splits for fair benchmarking. (d) Several replicable baselines for each task. (e) Benchmarking state-of-the-art methods on Turkish.

Moreover, Mukayese is a part of the Turkish Data Depository (TDD) project[1]. The main goal of TDD is collecting and organizing Turkish Natural Language Processing NLP resources and providing a research basis for Turkish NLP.

The rest of the paper is organized as follows: We review similar efforts in Section 2. Then, we advert to benchmarks and NLP in Section 3. Next, we give a background on the Turkish language resources in Section 4. We explain the approach we follow for each task in Section 5, and we provide dataset details, evaluation results, and explain the baselines for each task in Section 6.

2 Related Work

In this section, we discuss efforts similar to ours. We give an overview of efforts on building multilingual benchmarks, and we mention some of the monolingual benchmarks as well.

Corresponding author: asafaya19@ku.edu.tr

[1] https://tdd.ai

Findings of the Association for Computational Linguistics: ACL 2022, pages 846 - 863

Task	Datasets	Metrics	Baselines
Language Modeling	- **trnews-64** - **trwiki-67**	- Bits-per-char - Perplexity	- Adapt. Trans. - SHA-RNN
Machine Translation	- Wmt-16 - MuST-C	- BLEU	- ConvS2S - Transformer - mBART50
Named-Entity Recognition	- WikiAnn - Milliyet-Ner	- CoNLL F1	- Bilstm-Crf - Bert - Bert-Crf
Sentence Segmentation	- **trseg-41**	- Segment F1-Score	- spaCy - Punkt - Ersatz
Spellchecking & Correction	- **trspell-10**	- F1-Score - Accuracy	- Zemberek - Hunspell
Summarization	- Mlsum	- Rouge-L - Meteor	- Transformer - mBart50 - mT5
Text Classification	- Offenseval - *News-Cat*	- F1-Score	- Bilstm - CNN Text - Bert

Table 1: List of the NLP Tasks we work on for the Turkish language in Mukayese. We list the datasets, metrics, and baselines we use for each task. New datasets presented in this paper are marked in **bold**, and ones for which we present train/test splits are marked in *italic*.

There exist various endeavors at building multilingual benchmarks. One example for this is Xtreme (Hu et al., 2020), a multilingual benchmark containing 40 different languages and nine different tasks. These tasks include Classification, Named Entity Recognition (NER), and Question Answering (QA). However, most of these datasets are created by translating existing English datasets manually or automatically. Therefore, they have limitations and cannot be utilized to build a research basis in a specific language.

There are several benchmarks for NLP tasks for both low-resource and high-resource languages when it comes to monolingual benchmarks. Duh et al. (2020) proposes a benchmark for two low-resourced African languages on Neural Machine Translation (NMT), namely Somali and Swahili. Similarly, there are efforts to build benchmarks for high-resource but under-studied languages such as Alue benchmark for Arabic (Seelawi et al., 2021), and Klej benchmark for Polish (Rybak et al., 2020). Both benchmarks focus on Natural Language Understanding (NLU). Most of these benchmarks have public leaderboards to disseminate studies in NLP for their languages.

While most of previous benchmarks focus on one task such as NLU or NMT, Mukayese covers a comprehensive set of NLP tasks with seven different benchmarks on a variety of tasks. The reasoning behind this is to catalyze the research of Turkish NLP, and encourage research in all NLP applications.

3 Benchmarks and NLP

Following the research on NLP over the years, we can observe how datasets and benchmarks are fundamental. In this section, we discuss the importance of benchmarks for the progress of NLP.

Benchmarks are very essential for measuring the progress of NLP. For instance, the SQuAD dataset (Rajpurkar et al., 2016) is used to examine the progress of English Question Answering, and GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019) provide benchmarks for English Language Understanding.

Such progress has been enabled by the existence of benchmarks, which allowed for fair and meaningful comparison, and showed if there is room for improvement. In addition, organized benchmarks and datasets enable the research community to make progress with minimal amount of domain knowledge.

This is especially important when it comes to languages with fewer speakers, and research communities are more likely to contribute when such organized tasks are presented (Martínez-Plumed

et al., 2021). Thus, this is essential if we want to include other communities in the development of under-resourced and under-studied languages.

However, there are several things to keep in mind when dealing with benchmarks and leaderboards. Such leaderboards should be created transparently, and the results need to be evaluated with all factors taken into account. Some of these factors are model size, energy efficiency, and generalization (Linzen, 2020). Otherwise, we can run into the risk of these leaderboards resulting in inefficient and non-robust models. Ethayarajh and Jurafsky (2020) describe a few limitations of current leaderboards and suggest practices to mitigate these limitations.

We take these practices into account and present the benchmarks of MUKAYESE. We provide more details about our methodology in Section 5.

4 Background on Turkish

The Turkish language has distinctive characteristics compared to well-studied languages in the literature, such as English, Spanish, and German. Due to its agglutinative morphological nature, Turkish nouns can produce more than 100 inflected forms, while verbs can produce even more (Oflazer and Saraçlar, 2018). Therefore benchmarks designed for English are not necessarily applicable for Turkish.

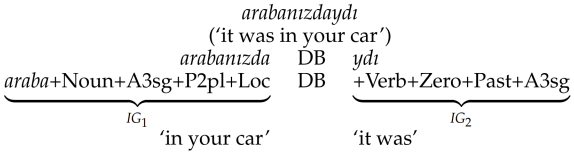

Figure 1: An example of a word constituting multiple inflectional groups (Eryiğit et al., 2008).

Unlike many other languages, a single word can constitute multiple different inflectional groups. An example is displayed in Figure 1. We provide more details on the features of the Turkish language in Appendix A.

There are several attempts at constructing comprehensive sets of resources and evaluation for Turkish. Sak et al. (2008) introduced a morphological parser, and a morphological disambiguator accompanied by a web corpus. More recently, Eryiğit (2014) proposed an online Turkish NLP Pipeline, which includes Normalization, Tokenization, Morphological Analysis, NER, and Syntactic Parsing.

However, among previously proposed methods and datasets, none are presented in a comparative way. This study aims to make a comprehensive inventory of different tools, corpora, and evaluation measures for the Turkish language. Such inventory may be used for researchers and practitioners who are looking for tools and datasets for Turkish NLP.

5 Methodology

In MUKAYESE, we focus on under-researched tasks of NLP in the Turkish language. After defining the task and assessing its importance, we construct the following three key elements for each benchmark:

Datasets are the first element to consider when it comes to a benchmark. We define the minimum requirements of a benchmark dataset as follows: (i) accessible with reasonable size. (ii) Satisfactory quality. (iii) A publicly shareable, compliant applicable regulations (GDPR licensing).

We chose the dataset sizes in a task-specific manner, unless used in a few-shot setting, benchmarks with small datasets will lack generalizability, and models trained on them might suffer from overfitting. On the other hand, training models on enormous datasets might be costly and inefficient (Ethayarajh and Jurafsky, 2020).

Another feature to assess is the quality of the dataset. A manually annotated dataset with a low Interannotator Agreement (IAA) rate is not suitable for benchmarking. Moreover, to build a generalizable benchmark, we need to consider using a dataset representing the general domain. For instance, sentence segmentation methods of editorial texts do not work on user-generated content such as social media posts, as we show in Subsection 6.4.

Metrics are the second element of benchmarks. We need to decide on one or more evaluation metrics to evaluate and compare methodologies. In order to do so, we have to answer the following questions: (a) Does this metric measure what our task aims to do? (b) How well does it correlate with human judgment? (c) Are there any issues/bugs to consider in these metrics? (For example, using accuracy to measure performance on an unbalanced set does not give a representative idea of model performance).

Baselines are the final element of benchmarking. In order to characterize the performance characteristics of different methodologies, it is bet-

ter to diversify our baselines as much as possible. For instance, we can compare pretrained vs. non-pretrained approaches, rule-based systems vs. trained systems, or unsupervised vs. supervised models.

6 Tasks

We provide benchmarks in the form of Datasets, Metrics, Baselines triplets for each of the following NLP tasks:

6.1 Language Modeling

Auto-regressive language modeling is a generative process, which focuses on modeling the probability $P(X)$ of a text sequence of n tokens, where $X = (x_1, x_2, ..., x_n)$, and $P(X) = \prod_{i=1}^{n} P(x_i|x_{<i})$. This type of language modeling is known as Auto-regressive (AR) or causal language modeling. The main objective of the model is to learn to estimate the probability of a given text sequence.

In our work, we focus on neural approaches for this task (Bengio et al., 2003), where we present two new benchmarking corpora for AR language modeling and report the results of two different baseline models.

Datasets We present two different corpora for AR language modeling, namely TRNEWS-64 and TRWIKI-67, along with their train/validation/test splits. These corpora are presented in a similar fashion to enwik8 (Hutter, 2006) and WikiText (Merity et al., 2017) English corpora. We provide statistics about these corpora in Table 10 in Appendix C.1.

TRWIKI-67 is a language modeling corpus that contains 67 million words of raw Turkish Wikipedia articles. We extracted this corpus from a recent Turkish Wikipedia dump[2] using WikiExtractor (Attardi, 2015). Additionally, further preprocessing was applied to get rid of the redundant text. Only the articles' raw text and titles were kept and presented in their cased format (with no upper/lower case transformations).

Due to the agglutinative nature of the Turkish language, most of the words are derived by combining one or more suffixes with one of the roots (Oflazer and Saraçlar, 2018). To make use of this attribute of the Turkish language, we train a sentencepiece unigram model (Kudo, 2018) with a vocabulary size of 32K, only using the training split of the corpus. Although we advise using the tokenized version of this corpus to encourage reproducibility, we provide a raw version of this corpus that can be utilized as a benchmark for language modeling tasks on a character, subword, or word level.

TRNEWS-64 is another language modeling corpus that contain 64 million words of news columns and articles that was retrieved from TS Timeline Corpus (Sezer, 2017). It can be utilized as a benchmark for modeling long-range dependencies in the Turkish language, as it contains relatively long documents (See Table 10). This corpus consists of a mix of news articles collected from different journals about various domains and topics. Since TRNEWS-64 is intended for language modeling on character level, articles were lightly pre-processed, and no further tokenization was applied. We provide more details about TRNEWS-64 and TRWIKI-67 in Appendix C.1.

Metrics Language models are trained on minimizing the negative log-likelihood (NLL) of the training set, and their performance is measured based on how well they can generalize on the test set:

$$\text{NLL}(X_{test}) = -\frac{1}{n}\sum_{i=1}^{n} log\ p_\theta(x_i|x_{test} < i) \quad (1)$$

Word or sub-word level language models are evaluated using the word perplexity (PPL) metric, a derivative of NLL. On the other hand, character language models are evaluated using entropy-based Bits-per-character (BPC) metric, which is also another derivative of NLL (Huyen, 2019). We consider PPL for the evaluation of models on TRWIKI-67, and BPC for TRNEWS-64. Note that lower is better for both metrics.

We note that PPL needs to be computed with the same count of tokens, otherwise it needs to be normalized (See Appendix C.1). Moreover, models considered to be evaluated using either of these corpora, are meant to have no training data other than that corpus' training split.

Baselines We consider two baseline models of different families. The first one is Single Headed Attention - RNN (SHA-RNN) (Merity, 2019), which is a Recurrent Neural Network-based language model, and the second is Adaptive Transformer (ADAP.TRANS) (Sukhbaatar et al., 2019), which is based on Transformer architecture (Vaswani et al., 2017). We choose these models for two main reasons. First, we want to compare

[2]https://dumps.wikimedia.org/trwiki/20210720/: accessed on 20 July 2021.

	TRWIKI-67		TRNEWS-64	
	#PARAM	PPL	#PARAM	BPC
ADAP.TRANS	92M	14.64	38M	1.024
SHA-RNN	87M	**12.54**	53M	**0.938**

Table 2: Results of language modeling baseline models, with their no of parameters. Perplexity (PPL) is reported for TRWIKI-67, and Bits-per-char (BPC) for TRNEWS-64, on their test sets.

models from different families (RNNs vs. Transformers). Second, compared to their counterparts such as (Lei, 2021; Dai et al., 2019), these models represent the state-of-the-art when it comes to the ratio of performance to the training cost and the number of parameters. For more details on the training refer to Appendix C.1.

In Table 2, we provide the results of these models, which we train and evaluate separately on TRWIKI-67 and TRNEWS-64 corpora (See Table 10 for more details on the splits of each corpus).

Note that even though we follow the same architectural settings for character-level and subword-level modeling, different tokenization algorithms of TRWIKI-67 (subword-level) and TRNEWS-64 (character-level) lead to different vocabulary sizes, which leads to a difference in the number of parameters.

Unlike the case for the English language (Merity, 2019), SHA-RNN performed better than Adaptive Transformer for both of the presented Turkish corpora. This implies the necessity of establishing such benchmarks for other languages as well. We leave investigating this feature for future research.

6.2 Machine Translation

Machine translation is the problem of translating a piece of text from one language to another. Over the years, neural machine translation models have become dominant, especially in low resource settings, benefiting from transfer learning (Zoph et al., 2016). In this work, we focus on evaluating neural machine translation models for translation between English and Turkish languages. We provide the results of three different baselines on two datasets.

Datasets The first dataset we evaluate is the Turkish-English subset of WMT-16[3], it consists of manually translated Turkish-English sentence pairs. The second one is the Turkish-English subset of Multilingual Speech Translation Corpus (MUST-C) (Di Gangi et al., 2019). For details on the split refer to Table 12 in Appendix C.3.

Metrics We evaluate our models on the relevant test sets for translation in both directions. We utilize BLEU Score (Papineni et al., 2002) for the assessment of translation quality.

	WMT-16		MUST-C	
	tr-en	en-tr	tr-en	en-tr
from scratch				
Stahlberg et al. (2018)	19.17	13.61	-	-
CONVS2S (180M)	13.22	12.78	21.79	13.3
TRANS. (58M)	17.29	15.72	27.01	15.52
pre-trained				
MBART50 (680M)	**24.17**	**18.54**	**32.97**	**19.61**

Table 3: BLEU scores of machine translation baselines. Results are provided for translations in both directions.

Baselines In this task, we train three different models. First, we train a TRANSFORMER (Vaswani et al., 2017) with the same settings for the encoder and the decoder parts, where we use 6 layers, with 4 attention heads each, and hidden size of 512. Second, we utilize the Convolutional sequence-to-sequence CONVS2S model (Gehring et al., 2017) following the same settings. The last model is mBART 50 (Tang et al., 2020), a multilingual model pre-trained on 50 different languages, which we fine-tune for each dataset separately.

In Table 3, we present BLEU score of the models on each translation dataset in both directions. The benefit of pre-training can be seen in the case of MBART50, where it outperforms the counterparts that we train from scratch. Additionally, we compare our work to the results reported by Stahlberg et al. (2018) on WMT-16. Their model is based on fusing language model decoding into seq2seq model with dot-attention (Luong et al., 2015).

6.3 Named-Entity Recognition (NER)

We include the Named-Entity Recognition (NER) task in our set of benchmarks, as it has an essential role in NLP applications. In this task, words representing named-entities are detected in the text input and assigned one of the predefined named-entity classes such as *Person* or *Location* (Chinchor and Robinson, 1998). We benchmark three different models on two NER datasets for Turkish and compare our work with previous work.

Datasets The first dataset we use is MILLIYET-NER (Tür et al., 2003), which is a set of manually, annotated news articles from the Turkish Milliyet

[3]http://www.statmt.org/wmt16/

news resource[4]. The second is the Turkish subset of the semi-automatically annotated Cross-lingual NER dataset WIKIANN or (PAN-X) (Pan et al., 2017), which consists of Turkish Wikipedia articles. Both datasets have three entity classes as shown in Table 11 in Appendix C.2.

Metrics Following previous work on Turkish NER (Yeniterzi, 2011; Şeker and Eryiğit, 2012), we report the CoNLL F-1 metric (Tjong Kim Sang, 2002) to assess our NER baselines. CoNLL F-1 counts a named entity as correct, only if it is an exact match of the corresponding entity in the ground truth.

	MILLIYET	WIKIANN
(Yeniterzi, 2011)	91.56	-
(Şeker and Eryiğit, 2012)	91.94	-
(Güngör et al., 2018)	93.37	-
BILSTM-CRF	95.54	**93.8**
BERTURK	95.31	92.82
BERTURK-CRF	**96.48**	93.07

Table 4: Evaluation results (CoNLL F_1) of NER models on test sets.

Baselines We train three different baseline models for this task. One with no pre-trained embeddings, which utilizes bi-directional Long Short Term Memory with Conditional Random Fields (BILSTM-CRF) (Panchendrarajan and Amaresan, 2018). The remaining two models employ pre-trained representations from BERT (Devlin et al., 2019). In one of the models, we investigate the benefit of adding a CRF layer on top of BERT. As for the pre-trained BERT model, we use BERTURK base, which is pre-trained on a large Turkish corpus (Schweter, 2020).

In Table 4, we provide the evaluation results (CoNLL F_1) for the three baselines on both datasets' test sets. Additionally, we compare our results with previous work of (Yeniterzi, 2011; Şeker and Eryiğit, 2012; Güngör et al., 2018) on the MILLIYET-NER dataset. We note that CoNLL F_1 of human performance on Turkish NER is expected to be in the range of 98-99% (Tür et al., 2003).

6.4 Sentence Segmentation

Sentence segmentation is the task of detecting sentence boundaries in a given article. Despite its fundamental place in the NLP pipelines, sentence segmentation attracts little interest. Common approaches are rule-based systems that rely on cues such as punctuation marks and capital letters (Jurafsky and Martin, 2018).

Datasets We present TRSEG-41, a new sentence segmentation dataset for Turkish. This dataset consists of 300 sampled scientific abstracts from (Özturk et al., 2014), 300 curated news articles from TRNEWS-64, and a set of 10K tweets. For the scientific abstracts, our sampling rationale is to maximize the number of abbreviations that reduce the accuracy of the rule-based approaches. As for the news subset, we maximize the length of documents and the number of proper nouns. In the Twitter subset, we balance the number of multi/single sentence tweets, and preprocess the tweets by replacing all URLs with `http://some.url`, and all user mentions with `@user`.

A single annotator labels the sentence boundaries of the data samples. We present two dataset splits, one for training and development and one for testing and benchmarking. The statistics of the splits can be found in Table 13 in Appendix C.4.

Applying sentence segmentation to user-generated content such as social media posts or comments can be quite challenging. To simulate such difficult cases and expose the weaknesses of rule-based methods, we create another version of **TRSEG-41** where we artificially corrupt the boundaries of sentences. This is done by randomly converting sentences to lowercase or uppercase with 50% probability, or by removing all punctuation marks with 50% probability.

Metrics Our evaluation procedure is based on the metrics F1 score, Precision, Recall for each segment. Unlike (Wicks and Post, 2021), we evaluate our models on the entire test set, without removing sentences with ambiguous boundaries. Furthermore, in order to highlight the gap in performance, we cross-evaluate our systems on the original and corrupted set.

	F1-SCORE	PRECISION	RECALL
SPACY	0.74 / 0.37	0.76 / 0.48	0.72 / 0.30
Training (Original)			
ERSATZ	**0.89** / **0.40**	**0.98** / 0.51	0.81 / **0.33**
PUNKT	0.87 / 0.39	0.88 / **0.52**	**0.86** / 0.32
Training (Corrupted)			
ERSATZ	**0.88** / **0.40**	**0.97** / **0.51**	0.81 / **0.33**
PUNKT	0.85 / 0.39	0.86 / 0.50	**0.84** / 0.31

Table 5: Results of sentence segmentation baselines. Metrics are reported for both corrupted and clean versions of the test set in the ORIGINAL / CORRUPTED format.

[4]https://www.milliyet.com.tr/

Baselines For this task, we employ three methods as baseline models. ERSATZ, a context-based approach that relies on supervised training (Wicks and Post, 2021), the unsupervised PUNKT tokenizer (Kiss and Strunk, 2006), and SPACY Sentencizer tool (Montani et al., 2021). While ERSATZ utilizes the Transformer (Vaswani et al., 2017) architecture, spaCy Sentencizer is a rule-based sentence boundary detector, whereas Punkt Tokenizer relies on an unsupervised training approach.

We experiment with these models on four different training and testing set combinations, where we train using the original and corrupted training sets separately and test on both test sets. Results are presented in Table 5. In all settings, SPACY SENTENCIZER is outperformed by its trained counterparts. Among the baselines, ERSATZ performed the best. Our experiments show that deep learning models are more robust to corruption in the data.

Please refer to Appendix C.4 for dataset creation process and samples, and an analysis on the behaviour of our baselines.

6.5 Spellchecking and Correction

Spellcheckers are among the most widely used NLP tools. The basic task is to check for misspellings in an input and suggest a set of corrections. Different methods can be employed for error correction, such as looking up words that minimize the edit distance from a dictionary or utilizing probabilistic models with N-grams to suggest the most likely correct word based on the context (Jurafsky and Martin, 2018). Due to the complexity of the Turkish Morphology, it is possible to derive over a hundred of words from one verb (Oflazer and Saraçlar, 2018). This makes the spellchecking task quite challenging. Hence, we focus on contextless (single word) spellchecking and correction as a start, and leave in-context spellchecking for future work.

We present a new benchmarking dataset for contextless spellcheckers and a computationally efficient and accurate dictionary for Turkish.

Datasets We present **TRSPELL-10**, a dataset of 10K words, for benchmarking spellchecking and correction. The dataset consists of tuples of input and correct (gold) words.

To create this dataset, we randomly sample 8500 Turkish words from the TS Corpus Word List (Sezer, 2013, 2017). We create artificial misspellings by applying random insertions, deletions, and substitutions on 65% of the words, where we apply at most two operations on the same word. The remaining 35% of the words are unchanged. Moreover, we add 1K random foreign words, and 500 randomly generated word-like character sequences.

As a quality check of these artificial misspellings, given a list of corrupted words, we ask our annotators to provide us a list of suggestions up to 10 suggestions per word. Their suggestion lists had the gold output 91% of the time.

Metrics We evaluate spellcheckers' ability to detect misspellings using the macro-averaged F1-Score metric. Additionally, we evaluate their spell correction accuracy (SCA) based on the suggestions provided for misspelled words.

	SCA	F1
HUNSPELL-TR (Zafer, 2017)	25.52	86.52
ZEMBEREK (Akın and Akın, 2007)	62.12	96.56
OUR HUNSPELL	**71.72**	**99.62**

Table 6: Spell correction accuracy (SCA) and macro-averaged F1 scores of spellchecking methods on TRSPELL-10.

Baselines We take advantage of the agglutinative nature of the Turkish language by developing a Hunspell-based (Trón et al., 2005) dictionary for Turkish. Using a list of 4M words we filter from Web crawls and Turkish corpora, we optimize the splits that minimize the size of the root dictionary and the affix list.

We compare this dictionary to HUNSPELL-TR (Zafer, 2017) another Hunspell-based Turkish dictionary, and to ZEMBEREK spellchecker (Akın and Akın, 2007), which is designed based on morphological features of the Turkish language. As shown in Table 6, our dictionary surpasses other baselines in terms of both error correction accuracy and error detection ability.

For dataset creation process and samples, please refer to Appendix C.5.

6.6 Summarization

Abstractive text summarization is the task of generating a short description (summary) of an article (longer text). Formally, given a sequence of tokens (input article) $X = (x_1, x_2, ..., x_n)$ and its summary $Y = (y_1, y_2, ..., y_m)$, the main task is to model the conditional probability: $P(Y|X) = \prod_{i=1}^{m} P(y_i|y_{<i}, X)$.

For this task, we work on the Multi-lingual Summarization (MLSUM) dataset (Scialom et al., 2020) and present state-of-the-art summarization results for Turkish.

Datasets MLSUM is a multi-lingual dataset for abstractive summarization. This dataset consists of a large set of crawled news articles with their abstracts in multiple languages. We focus on the Turkish subset of MLSUM.

We removed 4378 duplicated instances and 12 overlapping instances among the splits while assessing the dataset's quality. Further details in Appendix C.6.

Metrics To assess the quality of the generated summaries, we use the N-gram co-occurrence-based ROUGE-L (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) metrics. We report two different results for each model, one on the original, and one for the cleaned set.

	ROUGE-L	METEOR
(Scialom et al., 2020)	32.90/ –	26.30/ –
TRBART (120M)	35.54/35.08	26.47/25.81
MBART50 (680M)	39.21/38.47	30.84/30.36
MT5-BASE (220M)	**39.92/38.76**	**31.72/31.47**

Table 7: Evaluation of different models on MLSUM test set along with their no of parameters. Metrics are calculated for both (Original/Cleaned) test sets.

Baselines As a baseline model for summarization, we present TRBART, a Seq2Seq Transformer (Vaswani et al., 2017) trained following the configuration of BART Base (Lewis et al., 2020), which is a state-of-the-art model for abstractive summarization in English.

Moreover, we fine-tune two different pre-trained models. The first model is Multilingual BART (MBART50) (Tang et al., 2020), which is pre-trained on data from 50 different languages. The second model is Multilingual Text to Text Transformer (MT5-BASE) (Xue et al., 2021). As shown in Table 7, all models perform better than the best proposed baseline (Scialom et al., 2020), which follows the UniLM architecture (Dong et al., 2019).

6.7 Text Classification

Text classification can be utilized in several applications such as sentiment analysis or topic identification. In this task we take a sequence of text as an input, and output a probability distribution over the given classes. In our work on Turkish we benchmark three models on two datasets from different domains.

	OFFENSEVAL	NEWS-CAT	Avg.
BILSTM	0.747	0.808	0.777
CNN-TEXT	0.751	0.883	0.817
BERTURK	**0.823**	**0.944**	**0.883**

Table 8: Evaluation results (macro averaged F1-Score) of our baseline models for text classification task. The last column represent the average F1-scores of each model.

Datasets We work on the news categorization (NEWS-CAT) dataset (Amasyalı and Yıldırım, 2004). In this dataset, news articles are labeled with one of the following five categories *health, sports, economy, politics, magazine*. There is no splits provided in the original work for NEWS-CAT dataset. Hence we shuffle the dataset and construct our own splits in a stratified way, keeping the class distribution balanced across splits. We use 750 samples for training, 150 samples for validation, and 250 samples for testing. More details on the dataset can be found in Appendix C.7.

Since no information about the quality of annotation or Inter-annotator Agreement (IAA) rates were provided in for NEWS-CAT (Amasyalı and Yıldırım, 2004), we applied a quality assessment by re-annotating the test set. We asked three annotators to label the documents of test set with one of the given five categories. The annotators agreed with the gold annotation with an average IAA rate of FLEISS $\kappa = 0.88$.

The second dataset is the corpus of Offensive Speech Identification in Social media (OFFENSEVAL) (Çöltekin, 2020). This dataset was collected from Twitter, where the tweets are annotated for offensive speech with *offensive*, or *non-offensive* labels. We choose these datasets for benchmarking since they vary in domain and average article length.

Metrics We use the macro averaged F1-Score to account for the imbalance in classes within the datasets.

Baselines We measure the performance of three deep learning models—one with pre-training and two without pre-training. The pre-trained model is the BERT (Devlin et al., 2019) based Turkish pre-trained (BERTURK) model (Schweter, 2020). The remaining two models employ randomly initialized word embeddings of size 256. In one of them we use two layers of Bidirectional

LSTM (BILSTM) (Hochreiter and Schmidhuber, 1997) with a hidden size of 256. In the other model (CNN-TEXT), we use Convolutional Neural Networks for Sentence Classification (Kim, 2014) with 32 filters instead of 2.

Looking at F1 scores in Table 8, we can observe the advantage of pre-trained BERTURK model over BILSTM and CNN-TEXT.

7 Conclusion

We believe that while some languages such as Turkish do not fall under the definition of under-resourced languages, they attract relatively little research interest as a result of the lack of organized benchmarks and baselines. To address this problem, we presented MUKAYESE, a comprehensive set of benchmarks along with corresponding baselines for seven different tasks: Language Modeling, Machine Translation, Named Entity Recognition, Sentence Segmentation, Spell Checking and Correction, Summarization, and Text Classification, as well as four new benchmarking datasets in Turkish for Language Modeling, Sentence Segmentation, and Spell Checking and Correction. For future work, the same methodology can be followed to include more tasks such as Dependency Parsing, Morphological Analysis, coreference resolution. We hope that MUKAYESE encourages more researchers to get involved in the development of Turkish NLP, and it sets an example and leads to an increase in efforts on under-researched languages.

Acknowledgements

We thank Buse Çarık, Reyyan Yeniterzi, and Taner Sezer for their helpful discussions and feedback. Ali Safaya was supported by KUIS AI Center fellowship. Moreover, parts of the results reported in this paper were performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Ahmet Afsin Akın and Mehmet Dündar Akın. 2007. Zemberek, an open source nlp framework for turkic languages. 10:1–5.

MF Amasyalı and T Yıldırım. 2004. Otomatik haber metinleri sınıflandırma. In *2004 12th Signal Processing and Communications Applications Conference (SIU)*, pages 224–226. IEEE.

Giusepppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

N. Chinchor and P. Robinson. 1998. Appendix E: MUC-7 named entity task definition (version 3.5). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Çağrı Çöltekin. 2020. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184, Marseille, France. European Language Resources Association.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

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

Mattia A Di Gangi, Matteo Negri, and Marco Turchi. 2019. One-to-many multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 585–592. IEEE.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kevin Duh, Paul McNamee, Matt Post, and Brian Thompson. 2020. Benchmarking neural and statistical machine translation on low-resource African languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2667–2675, Marseille, France. European Language Resources Association.

Gülşen Eryiğit. 2014. ITU Turkish NLP web service. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–4, Gothenburg, Sweden. Association for Computational Linguistics.

Gülşen Eryiğit, Joakim Nivre, and Kemal Oflazer. 2008. Dependency parsing of turkish. *Computational Linguistics*, 34(3):357–389.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1243–1252. JMLR.org.

Onur Güngör, Suzan Üsküdarlı, and Tunga Güngör. 2018. Recurrent neural networks for turkish named entity recognition. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Marcus Hutter. 2006. The human knowledge compression prize.

Chip Huyen. 2019. Evaluation metrics for language modeling. *The Gradient*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Daniel Jurafsky and James H Martin. 2018. Speech and language processing (draft). *https://web. stanford. edu/~ jurafsky/slp3*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Tao Lei. 2021. When attention meets fast recurrence: Training language models with reduced compute. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7633–7648, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Fernando Martínez-Plumed, Pablo Barredo, Seán Ó hÉigeartaigh, and José Hernández-Orallo. 2021. Research community dynamics behind popular ai benchmarks. *Nature Machine Intelligence*, 3(7):581–589.

Stephen Merity. 2019. Single headed attention rnn: Stop thinking with your head. *Computing Research Repository*, arXiv:1911.11423. Version 2.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Tomas Mikolov, Anoop Deoras, Stefan Kombrink, Lukas Burget, and Jan "Honza" Cernocky. 2011. Empirical evaluation and combination of advanced language modeling techniques. In *Interspeech*. ISCA.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Sofie Van Landeghem, Adriane Boyd, Henning Peters, Paul O'Leary McCann, Maxim Samsonov, Jim Geovedi, Jim O'Regan, György Orosz, Duygu Altinok, Søren Lind Kristiansen, Roman, Explosion Bot, Leander Fiedler, Grégory Howard, Wannaphong Phatthiyaphaibun, Yohei Tamura, Sam Bozek, murat, Mark Amery, Björn Böing, Pradeep Kumar Tippa, Leif Uwe Vogelsang, Bram Vanroy, Ramanan Balakrishnan, Vadim Mazaev, and GregDubbin. 2021. explosion/spaCy: v3.2.0: Registered scoring functions, Doc input, floret vectors and more.

Kemal Oflazer and Murat Saraçlar, editors. 2018. *Turkish Natural Language Processing*. Springer International Publishing.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Seçil Özturk, Bülent Sankur, Tunga Gungör, Mustafa Berkay Yilmaz, Bilge Köroğlu, Onur Ağin, Mustafa İşbilen, Çağdaş Ulaş, and Mehmet Ahat. 2014. Turkish labeled text corpus. In *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, pages 1395–1398. IEEE.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. Bidirectional LSTM-CRF for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: Comprehensive benchmark for Polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, Online. Association for Computational Linguistics.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2008. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *Advances in Natural Language Processing*, pages 417–427, Berlin, Heidelberg. Springer Berlin Heidelberg.

Stefan Schweter. 2020. Berturk - bert models for turkish.

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

Haitham Seelawi, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. 2021. ALUE: Arabic language understanding evaluation. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 173–184, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Gökhan Akın Şeker and Gülşen Eryiğit. 2012. Initial explorations on using CRFs for Turkish named entity recognition. In *Proceedings of COLING 2012*, pages 2459–2474, Mumbai, India. The COLING 2012 Organizing Committee.

Taner Sezer. 2013. Ts corpus: Herkes İçin türkçe derlem. *Proceedings 27th National Linguistics Conference*, pages 217–225.

Taner Sezer. 2017. Ts corpus project: An online turkish dictionary and ts diy corpus. *European Journal of Language and Literature*, 9(1):18–24.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *Computing Research Repository*, arXiv:1909.08053. Version 4.

Felix Stahlberg, James Cross, and Veselin Stoyanov. 2018. Simple fusion: Return of the language model. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 204–211, Brussels, Belgium. Association for Computational Linguistics.

Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. 2019. Adaptive attention span in transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 331–335, Florence, Italy. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *Computing Research Repository*, arXiv:2008.00401.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Viktor Trón, Györgi Gyepesi, Péter Halácsky, András Kornai, László Németh, and Dániel Varga. 2005. Hunmorph: Open source word analysis. In *Proceedings of Workshop on Software*, pages 77–85, Ann Arbor, Michigan. Association for Computational Linguistics.

Gökhan Tür, Dilek Hakkani-TüR, and Kemal Oflazer. 2003. A statistical information extraction system for turkish. *Natural Language Engineering*, 9(2):181–210.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, page 6000–6010.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32, page 3266–3280. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3995–4007, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 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: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Reyyan Yeniterzi. 2011. Exploiting morphology in Turkish named entity recognition system. In *Proceedings of the ACL 2011 Student Session*, pages 105–110, Portland, OR, USA. Association for Computational Linguistics.

Harun Reşit Zafer. 2017. hunspell-tr. `https://github.com/hrzafer/hunspell-tr`.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Turkish Language

Even though in formal language the Subject-Object-Verb order is predominantly used, Turkish is a free-order language, meaning that words can freely change order depending on the context without changing the meaning but only the accentuation. The English sentence "I am going to school." can

be translated into Turkish as "Ben okula gidiyorum." where all 6 permutations of the words are valid and meaningful:

- Ben okula gidiyorum
- Ben gidiyorum okula
- Gidiyorum ben okula
- Gidiyorum okula ben
- Okula gidiyorum ben
- Okula ben gidiyorum

In Turkish, morphologically ambiguous words are common in running texts. Depending on the context, the same word can have varying morphological features. For instance, the word "masalı" can correspond to the following:

masal+Noun+A3sg+Pnon+Acc(=the story)
masal+Noun+A3sg+P3sg+Nom(=his story)
masa+Noun+A3sg+Pnon+Nom^DB+Adj+With(=with tables)

Given all these language features, Turkish language needs special attention by the research community, and we cannot assume that methods with good performance on English would yield good results on Turkish.

B Computational Costs and Implementations

We utilize NVIDIA TESLA V100 GPUS with 32GBs of memory for training our baselines. In Table 9, we depict approximate estimations of the training time for each of our compute-intensive baselines.

The implementations of TRANSFORMER (Vaswani et al., 2017), and CONVS2S (Gehring et al., 2017) are based on the open-source library Fairseq (Ott et al., 2019). We use the Flair library (Akbik et al., 2019) for the BERT-CRF model in the Named Entity Recognition task. The remaining deep learning models used as our baselines are either implemented using the Huggingface Transformers library (Wolf et al., 2020). All reported experiments and implementations of deep learning models are performed using PyTorch (Paszke et al., 2019).

C Datasets and Baselines

C.1 Language Modeling

We provide some samples from TRWIKI-67 and TRNEWS-64 corpora in Table 16. These corpora are presented with minimal pre-processing. We remove non-Turkish characters and redundant texts such as category lists and tables from TRWIKI-67. Sentences and words are counted based on `sent_tokenize`, `word_tokenize` methods of NLTK (Bird et al., 2009).

Model	Dataset	GPU Hr	Batch S.
LANGUAGE MODELING			
SHA-RNN	trwiki-67	30	16
SHA-RNN	trnews-64	24	32
Adap. Transformer	trwiki-67	72	16
Adap. Transformer	trnews-64	56	16
MACHINE TRANSLATION			
ConvS2S	Wmt-16	12x2	4000*
ConvS2S	MuST-C	11x2	4000*
Transformer	Wmt-16	8x2	4096*
Transformer	MuST-C	7x2	4096*
mBART50	Wmt-16	24x2	2
mBART50	MuST-C	22x2	2
SUMMARIZATION			
Transformer	Mlsum	12	4
mBART50	Mlsum	51	2
mT5-Base	Mlsum	38	2

Table 9: Computational costs per models. * Fairseq uses dynamic batching, so we report max number of tokens per batch.

	#articles	#words	#tokens	avg.sent
TRWIKI-67				
Training	374K	63.5M	139M	12.8
Validation	10K	1.7M	4M	13.3
Test	10K	1.7M	4M	12.9
Total	394K	67M	147M	12.8
TRNEWS-64				
Training	140K	59.7M	421M	23
Validation	5K	2.1M	15M	22.8
Test	5K	2.1M	15M	22.9
Total	150K	64M	450M	23

Table 10: Statistics about TRWIKI-67 and TRNEWS-64 corpus splits. The column *avg. sents* refers to the average number of sentences per article. Tokens are characters for TRNEWS-64 and sentencepiece tokens for TRWIKI-67.

We follow the same architectures proposed by (Merity, 2019; Sukhbaatar et al., 2019). The only difference in architecture is based on vocabulary size due to difference in training data. For training, we use vocabulary size of 32K sentencepiece for TRWIKI-67, and 124 for TRNEWS-64 which includes the Turkish alphabet with punctuation and some other common characters. We train both models until no improvement over the validation

set, then following the original implementation we lower the learning-rate, dividing it by 10 and run until no improvement on the validation set again.

C.1.1 Normalizing perplexity

The Perplexity metric is defined as the exponent of the average entropy over a corpus (Mikolov et al., 2011):

$$\text{PPL}(X_{test}) = exp(-\frac{1}{N}\sum_{i=1}^{n} log\ p_\theta(x_i|x_{test}< i)) \quad (2)$$

where N is the original number of tokens in X_{test}, and n is the number of tokens of X_{test} when tokenized using a certain tokenization algorithm. Depending on what tokenization is used, N might or might not be equal to n. To accommodate this issue, N should always be the same when calculating perplexity for different models (Shoeybi et al., 2019).

C.2 Named Entity Recognition (NER)

We provide statistics about dataset splits for both MILLIYET-NER and WIKIANN in Table 11.

	Training	Validation	Test
WIKIANN			
Location	9679	5014	4914
Organization	7970	4129	4154
Person	8833	4374	4519
Total words	149786	75930	75731
MILLIYET-NER			
Location	8821	942	1126
Organization	8316	842	873
Person	13290	1400	1603
Total words	419996	45532	49595

Table 11: Distribution of Named entities over classes in MILLIYET-NER and WIKIANN datasets.

C.3 Machine Translation

We utilize two datasets for Machine Translation, WMT-16 dataset, which was presented at the first Conference of Machine Translation (WMT), and MuST-C dataset. This corpus was extracted from movies and TV shows subtitles. Statistics of both datasets are presented in Table 12.

C.4 Sentence Segmentation

In this section, we provide additional information for our Sentence Segmentation 6.4 Benchmark.

In both clean and corrupted training cases, ErSatz and Punkt are trained with all subsets. Following the authors, our baseline model ErSatz is trained without changing the original architecture with a vocabulary size of 500, left and right context of 5 for 100 epochs using early stopping. We use the NLTK (Bird et al., 2009) implementation of the Punkt tokenizer (Kiss and Strunk, 2006) for both training and testing purposes. The spaCy tokenizer (Montani et al., 2021) is used with the default settings provided by the library.

	#Sentences	#Words
Turkish		
MUST-C	236K / 1.3K / 2K	3.4M / 19K / 33K
WMT-16	205K / 1K / 3K	3.6M / 14K / 44K
English		
MUST-C	236K / 1K/ 2K	4.6M / 26K / 45K
WMT-16	205K / 1K / 3K	4.4M / 19K / 58K

Table 12: Statistics of machine translation datasets. Each cell represents the (Train / Validation / Test) values of the datasets in the corresponding row. WMT-16 and MUST-C refer to Turkish-English subsets.

	#Articles	#Sentences	#Words
News	300	6K	102K
Tweets	10K	28K	242K
Abstracts	300	6K	112K
Total	10.6K	40K	456K

Table 13: Statistics of TRSEG-41 dataset.

Table 18 provides examples from each subset of the TRSEG-41 dataset along with their corrupted versions. The dataset is annotated by a single human. The reason for maximizing the number of abbreviations and proper nouns is that rule-based methods are designed to be sensitive to local language features such as periods and capital letters. In editorial texts, sentence segmentation can achieve high success. Therefore, we apply automated random corruption process as described in Section 6.4. The rationale behind this is to eliminate the aforementioned context for rule-based approaches and to promote learning methods.

Table 19 shows examples of the results of our baselines. The results show that while the models are able to perform successful sentence segmentation on clean editorial text, they experience an evident drop in performance on corrupted versions.

C.4.1 F1-Score

In this benchmark, we compare the performances of the models via F1-Score. For sentence segmentation, we define F1-Score as the accuracy measure of the position of the dots in a given piece of text among spaced tokens. This means that for a para-

graph containing N words, all words are separated as distinct tokens, leaving $N-1$ locations to place the dots as separators. Our measure is based on the correctness of the placed dots in this given setting. We calculate F1-Score in the following way:

$$\frac{TP}{TP + 1/2(FP + FN)} \tag{3}$$

where TP is true positive rate, FP is false positive rate, and FN is false negative rate. Our calculation is based on the *Scorer* submodule of the spaCy library.

C.5 Spellchecking and Correction

In this section, we provide a detailed description of the spellchecking dataset with the statistics about the word set and corruption methods.

The dataset consist of 10K words it total, and includes pairs of gold and corrupted words. 8500 words are randomly sampled from TS Corpus Word List (Sezer, 2013, 2017), 1K random words are included from foreign language and 500 randomly generated word-like character sequences are added.

For 70% of the sampled Turkish words, we apply one corruption with 70% probability, two corruptions with 25% probability and three corruptions with 5% probability. The following corruption methods with their probability distribution is applied for a single corruption:

- For a probability of 1/2, the word is asciified.
- For a probability of 1/6, a random character in the word is substituted by another character sampled from a distribution simulating the placement of keys in standard Turkish-Qwerty keyboards.
- For a probability of 1/6, a random character is inserted into the word sampled from a distribution simulating the placement of keys in standard Turkish-Qwerty keyboards.
- For a probability of 1/6, a random character is deleted from a word sampled from a distribution simulating the placement of keys in standard Turkish-Qwerty keyboards.

The remaining 30% of the words are uncorrupted, therefore their gold and input versions are same. For evaluating against inserted foreign words and randomly generated character sequences where no gold output exists, we use an empty string as the gold output.

C.6 Summarization

We remove these instances from the dataset for a more accurate evaluation and evaluate our models on both the original and the cleaned sets. In Table 14, we provide some statistics about both sets, before and after the deduplication.

	Original	Cleaned
Avg. article length	259.1	258.4
Avg. summary length	18.5	18.3
Splits		
Training	249277	246490
Validation	11565	10852
Test	12775	11897
Total	273617	269239

Table 14: Statistics of the Turkish subset of MLSUM. The number of samples is provided for each split before and after the deduplication.

We provide summaries predicted by our models in Table 17.

C.7 Text Classification

In table 15, we provided statistics about both of the datasets we used for text classification task.

	OFFENSEVAL	NEWS-CAT
Avg. #words	8.5	227.3
#Classes	2	5
Splits		
Training	28000	750
Validation	3277	150
Test	3515	250
Total	34792	1150

Table 15: Statistics of NEWS-CAT and OFFENSEVAL dataset splits.

TRWIKI-67
== NGC 1710 == NGC 1710, Yeni Genel Katalog'da yer alan bir galaksidir. Gökyüzünde Aslan takımyıldızı yönünde bulunur. E-S0 tipi bir merceksi, eliptik galaksidir. Amerikan astronom Francis Leavenworth tarafından 1885 yılında 66,04 cm (26 inç) çaplı mercekli tip bir teleskopla keşfedilmiştir. == Şenol Gürşan == Şenol Gürşan, (d. 17 Ekim 1964, Pınarhisar, Kırklareli) Türk avukat ve siyasetçi. İstanbul Üniversitesi Hukuk Fakültesi'ni bitirmiş ve serbest avukat olarak çalışmıştır. Kırklareli İlim Yayma Cemiyeti Kuruculuğu ve Başkanlığı görevlerinde bulunmuştur. 2009 yılında Adalet ve Kalkınma Partisi Kırklareli il yönetim kurulu üyesi olmuş, TBMM 24. dönem AK Parti Kırklareli milletvekili, Türkiye-Polonya Dostluk Grubu Başkanı ve TBMM KİT Komisyonu Sözcüsü olmuştur. Gelecek Partisi Kurucular Kurulu üyesi olup aynı zamanda partinin genel sekreteridir. İyi düzeyde Almanca bilen Gürşan, evli ve 2 çocuk babasıdır.
TRNEWS-64
Dolar dün 2.5075 liraya kadar çıkarak rekor kırmasının ordından bugün 2.49 - 2.50 lira aralığında hareket etti. Cari işlemler açığının beklentilere paralel gelmesinin de etkisiyle 2.4820 liraya kadar çekilen dolar, daha sonra gelens alımlarla 2.5085'e çıkarak rekorunu tazeledi. ABD para birimi daha sonra 2.5050 - 2.5070 düzeylerinde hareket ederken, euro da 2.8380 lira düzeylerine çıktı ve yarı yarıya euro ve dolardan oluşan döviz sepeti de 2.63 düzeyinin üstüne çıktı. DW Türkçe Servisi'nin aktardığına göre, 'Aghet' (Ağıt) konserinin Almanya'nın İstanbul Başkonsolosluğu'ndaki temsiline Cumhurbaşkanı Recep Tayyip Erdoğan da davet edildi. Alman haber ajansı dpa'nın haberinde, Erdoğan'ın yanı sıra Başbakan Binali Yıldırım, Dışişleri Bakanı Mevlüt Çavuşoğlu ile Kültür ve Turizm Bakanı Nabi Avcı'nın da davetliler arasında olduğu belirtildi. Habere göre, gönderilen davetiyelerde etkinlikte 'Türk ve Ermeni geçmişlerindeki yaralar' ile ifade ve sanat özgürlüğünün ele alınacağı ifade edildi. Dresden Senfoni Orkestrası tarafından hazırlanan 'Aghet' konseri, İstanbul Başkonsolosluğu'nda 13 Kasım'da gerçekleştirilecek. Etkinlikte ayrıca Türk-Ermeni-Alman Dostluk Derneği'nin kurulması planlanıyor.

Table 16: Text samples from TRWIKI-67 and TRNEWS-64 corpora.

INPUT
Bursa İnegöl ilçesi Deydinler Mahallesi'nde yaşayan Erdoğan Bitirim evde gördüğü yılanı elleriyle yakalayıp doğaya saldı. Havaların sıcak olmasıyla birlikte son günlerde sayıları artan yılanlar vatandaşları tedirgin ediyor. Erdoğan Bitirim evinde yakaladığı yılanı doğaya salarken o anları cep telefonuyla kayıt altına aldı. Bitirim, yılana her hangi bir zarar vermediği belirterek, "Çok hızlı ve seri hareket ediyordu. Birkaç kez bana saldırmaya kalktı ama ben onu yakaladım. Yakaladığım yılanı zarar vermeden doğa saldım. Yaklaşık 1 metre boyunda bir yılandı" dedi.
REFERENCE
Bursa'nın İnegöl ilçesinde bir vatandaş evinde eliyle yakaladığı yılanı doğaya saldı.
TRBART
bursa'nın inegöl ilçesinde yaşayan erdoğan bitirim, yılanı elleriyle yakalayıp doğaya saldı.
MBART50
BURSA'nın İnegöl ilçesinde yaşayan Erdoğan Bitirim, evde gördüğü yılanı elleriyle yakalayıp doğaya saldı.
MT5-BASE
Bursa'nın İnegöl ilçesinde yaşayan Erdoğan Bitirim evinde yakaladığı yılanı doğaya saldı.

Table 17: Example of summaries generated by the three baselines for a sample from the test set of MLSUM.

Clean Abstract Sample:

Bu çalışmanın amacı, bayan ve erkek voleybolcular ile güreşçilerin statik, yaylanarak, düşerek ve tekrarlı sıçrama performanslarını karşılaştırmaktır.
Bu çalışmaya Yaşar Doğu Beden Eğitimi ve Spor Yüksekokulunda okuyan 2. ve 3. Ligde mücadele eden 20 bayan voleybolcu, 20 erkek voleybolcu ile Milli 20 erkek güreşçi gönüllü olarak katılmıştır.
Bayan voleybolcuların yaş ortalaması 21.15 yıl, voleybolcu erkeklerin 20.80 yıl ve güreşçilerin 20.60 yıldır.
Bütün denekler statik sıçrama, yaylanarak sıçrama, düşerek sıçrama ve tekrarlı sıçrama yapmışlardır.
Sıçrama değerlerinin belirlenmesi, New Test Power Timer System 300 Series aleti kullanılarak yapılmıştır.
Ayrıca çalışmaya katılan sporcuların, beden kitle indeksi (BKİ), esneklik ve vücut yağ yüzdesi değerleri ölçülmüştür.
Üç grup arasında fark olup olmadığına bakmak amacıyla Kruskal Vallis testi, ikili karşılaştırmalarda Mann Whitney U testi kullanılmıştır.
Sporcuların karşılaştırıldığında güreşçi erkeklerin voleybolcu erkelerden daha esnek oldukları görülmüştür.
boy, vücut ağırlığı, BKI, vücut yağ yüzdesi arasında anlamlı derecede farklılık bulunmuştur.
Voleybolcu erkeklerin Düşerek, Statik, Yaylanarak ve Tekrarlı sıçrama yükseklikleri ve güçleri voleybolcu bayanlardan ve güreşçilerden yüksek bulunmuştur.
Güreşçilerin ise statik ve yaylanarak sıçrama yükseklikleri ve güçleri bayan voleybolculardan daha yüksek bulunmuştur.
Erkek voleybolcuların sıçrama değerlerinin güreşçilerden yüksek çıkması yapılan spor branşı ile ilgilidir.
Voleybolcu bayanların sıçrama performansının güreşçi erkeklerden daha iyi olması beklenirken cinsiyet faktörünün bu durumun önüne geçtiği görülmüştür.
Sonuç olarak, yapılan spor branşının ve cinsiyetin sıçrama performansı üzerinde önemli etkisinin olduğu görülmüştür.

Corrupted Abstract Sample:

BU ÇALIŞMANIN AMACI BAYAN VE ERKEK VOLEYBOLCULAR İLE GÜREŞÇİLERİN STATİK YAYLANARAK DÜŞEREK VE TEKRARLI SIÇRAMA PERFORMANSLARINI KARŞILAŞTIRMAKTIR
bu çalışmaya yaşar doğu beden eğitimi ve spor yüksekokulunda okuyan 2. ve 3. ligde mücadele eden 20 bayan voleybolcu, 20 erkek voleybolcu ile milli 20 erkek güreşçi gönüllü olarak katılmıştır.
Bayan voleybolcuların yaş ortalaması 21.15 yıl, voleybolcu erkeklerin 20.80 yıl ve güreşçilerin 20.60 yıldır.
Bütün denekler statik sıçrama yaylanarak sıçrama düşerek sıçrama ve tekrarlı sıçrama yapmışlardır
Sıçrama değerlerinin belirlenmesi, New Test Power Timer System 300 Series aleti kullanılarak yapılmıştır.
Ayrıca çalışmaya katılan sporcuların beden kitle indeksi BKİ esneklik ve vücut yağ yüzdesi değerleri ölçülmüştür
üç grup arasında fark olup olmadığına bakmak amacıyla kruskal vallis testi, ikili karşılaştırmalarda mann whitney u testi kullanılmıştır.
Sporcuların karşılaştırıldığında güreşçi erkeklerin voleybolcu erkelerden daha esnek oldukları görülmüştür.
boy, vücut ağırlığı, BKI, vücut yağ yüzdesi arasında anlamlı derecede farklılık bulunmuştur.
Voleybolcu erkeklerin Düşerek Statik Yaylanarak ve Tekrarlı sıçrama yükseklikleri ve güçleri voleybolcu bayanlardan ve güreşçilerden yüksek bulunmuştur
Güreşçilerin ise statik ve yaylanarak sıçrama yükseklikleri ve güçleri bayan voleybolculardan daha yüksek bulunmuştur
Erkek voleybolcuların sıçrama değerlerinin güreşçilerden yüksek çıkması yapılan spor branşı ile ilgilidir
VOLEYBOLCU BAYANLARIN SIÇRAMA PERFORMANSININ GÜREŞÇİ ERKEKLERDEN DAHA İYİ OLMASI BEKLENİRKEN CİNSİYET FAKTÖRÜNÜN BU DURUMUN ÖNÜNE GEÇTİĞİ GÖRÜLMÜŞTÜR
Sonuç olarak, yapılan spor branşının ve cinsiyetin sıçrama performansı üzerinde önemli etkisinin olduğu görülmüştür.

Clean Tweet Sample:

@user @user o kulların açılmasını 1 gün erteledi, çünkü alın size müjde veriyorum diyecek.
başka bişey yok, işler çığırından çıkmış

Corrupted Tweet Sample:

@user @user O KULLARIN AÇILMASINI 1 GÜN ERTELEDİ ÇÜNKÜ ALIN SİZE MÜJDE VERİYORUM DİYECEK
BAŞKA BİŞEY YOK İŞLER ÇIĞIRINDAN ÇIKMIŞ

Clean News Sample:

Kanal 2 televizyonunda, İsrail'in tanınmış gazeteci ve analistlerinden Ehud Yaari ile birlikte konuk olan Lieberman'a, "Bakanlığı döneminde hem Türkiye hem de Mısır'dan büyükelçilerin kovulduğuna" işaret edilerek, gazetede yayımlanan haberin doğru olup olmadığı soruldu.
Haberin doğru olmadığını söyleyen lieberman'a, bu kez, "Peki (Dışişleri'nde) böyle şeyler konuştunuz mu?" sorusu yöneltildi.
Lieberman, bu soruya, "Dışişleri Bakanlığı'nda her gün yüzlerce fikir tartışma konusu edilir" yanıtını verdi.
Bunun üzerine, Ehud Yaari, "Açıkça söyleyin, PKK terör örgütüne silah vs. sağlama gibi, yardım etme konusu konuşuldu mu?" diyerek, sorusunu yineledi.
Lieberman, soruya bu kez "Hayır, kesinlikle konuşulmadı" karşılığını verdi.
Lieberman, Palmer Komisyonu raporunun "Mavi Marmara" baskını ile ilgili olarak İsrail'in eyleminin ve Gazze'ye ablukanın haklı olduğunu açıkça ortaya koyduğunu da ifade etti.
Lieberman, Türkiye ile ilişkilerin normalleştirilmesinin yeniden sağlanacağı ve Türkiye'nin, böyle bir normalleşmenin çıkarına olacağını göreceği umudunda olduğunu da kaydetti.
"Alevlerin seviyesini düşürmeye çalışıyoruz" İsrail Başbakanı Binyamin Netanyahu da Türkiye ile yaşanan krizin kendi seçimleri olmadığını öne sürdü.
Türkiye ile ilişkilerin daha da kötüye gitmesini önlemeye çalıştıklarını savunan Netanyahu, halihazırda, iki ülke arasındaki "Alevlerin seviyesini düşürmeye" uğraştıklarını belirterek, "Umarım bu gerginlik, eğer karşı taraf da isterse, sona erdirilecektir" diye konuştu.

Corrupted News Sample:

Kanal 2 televizyonunda, İsrail'in tanınmış gazeteci ve analistlerinden Ehud Yaari ile birlikte konuk olan Lieberman'a, "Bakanlığı döneminde hem Türkiye hem de Mısır'dan büyükelçilerin kovulduğuna" işaret edilerek, gazetede yayımlanan haberin doğru olup olmadığı soruldu.
Haberin doğru olmadığını söyleyen liebermana bu kez Peki Dışişlerinde böyle şeyler konuştunuz mu sorusu yöneltildi
Lieberman, bu soruya, "Dışişleri Bakanlığı'nda her gün yüzlerce fikir tartışma konusu edilir" yanıtını verdi.
Bunun üzerine Ehud Yaari Açıkça söyleyin PKK terör örgütüne silah vs sağlama gibi yardım etme konusu konuşuldu mu diyerek sorusunu yineledi
LIEBERMAN, SORUYA BU KEZ "HAYIR, KESİNLİKLE KONUŞULMADI" KARŞILIĞINI VERDİ.
lieberman palmer komisyonu raporunun mavi marmara baskını ile ilgili olarak israilin eyleminin ve gazzeye ablukanın haklı olduğunu açıkça ortaya koyduğunu da ifade etti
lieberman, türkiye ile ilişkilerin normalleştirilmesinin yeniden sağlanacağı ve türkiye'nin, böyle bir normalleşmenin çıkarına olacağını göreceği umudunda olduğunu da kaydetti.
Alevlerin seviyesini düşürmeye çalışıyoruz İsrail Başbakanı Binyamin Netanyahu da Türkiye ile yaşanan krizin kendi seçimleri olmadığını öne sürdü
Türkiye ile ilişkilerin daha da kötüye gitmesini önlemeye çalıştıklarını savunan Netanyahu halihazırda iki ülke arasındaki Alevlerin seviyesini düşürmeye uğraştıklarını belirterek Umarım bu gerginlik eğer karşı taraf da isterse sona erdirilecektir diye konuştu

Table 18: A sample from each of the abstracts, news, and tweets test subsets of TRSEG-41. **Clean** means the unedited and uncorrupted version of the data. **Corrupted** is the corrupted version of this abstract as specified in Section 6.4. The annotation of each sample is denoted by line-separation.

Punkt Tokenizer Corrupted Tweet Sample Output:
@user @user O KULLARIN AÇILMASINI 1 GÜN ERTELEDI ÇÜNKÜ ALIN SIZE MÜJDE VERIYORUM DIYECEK BAŞKA BIŞEY YOK IŞLER ÇIĞIRINDAN ÇIKMIŞ
spaCy Tokenizer Corrupted Tweet Sample Output:
@user @user O KULLARIN AÇILMASINI 1 GÜN ERTELEDI ÇÜNKÜ ALIN SIZE MÜJDE VERIYORUM DIYECEK BAŞKA BIŞEY YOK IŞLER ÇIĞIRINDAN ÇIKMIŞ
ErSatz Tokenizer Corrupted Tweet Sample Output:
@user @user O KULLARIN AÇILMASINI 1 GÜN ERTELEDI ÇÜNKÜ ALIN SIZE MÜJDE VERIYORUM DIYECEK BAŞKA BIŞEY YOK IŞLER ÇIĞIRINDAN ÇIKMIŞ
Punkt Tokenizer Corrupted News Sample Output:
Kanal 2 televizyonunda, İsrail'in tanınmış gazeteci ve analistlerinden Ehud Yaari ile birlikte konuk olan Lieberman'a, "Bakanlığı döneminde hem Türkiye hem de Mısır'dan büyükelçilerin kovulduğuna" işaret edilerek, gazetede yayımlanan haberin doğru olup olmadığı soruldu. Haberin doğru olmadığını söyleyen liebermana bu kez Peki Dışişlerinde böyle şeyler konuştunuz mu sorusu yöneltildi Lieberman, bu soruya, "Dışişleri Bakanlığı'nda her gün yüzlerce fikir tartışma konusu edilir" yanıtını verdi. Bunun üzerine Ehud Yaari Açıkça söyleyin PKK terör örgütüne silah vs sağlama gibi yardım etme konusu konuşuldu mu diyerek sorusunu yineledi LIEBERMAN, SORUYA BU KEZ "HAYIR, KESİNLİKLE KONUŞULMADI" KARŞILIĞINI VERDİ. lieberman palmer komisyonu raporunun mavi marmara baskını ile ilgili olarak imidrulesrailin eyleminin ve gazzeye ablukanın haklı olduğunu açıkça ortaya koyduğunu da ifade etti lieberman, türkiye ile ilişkilerin normalleştirilmesinin yeniden sağlanacağı ve türkiye'nin, böyle bir normalleşmenin çıkarına olacağını göreceği umudunda olduğunu da kaydetti. Alevlerin seviyesini düşürmeye çalışıyoruz İsrail Başbakanı Binyamin Netanyahu da Türkiye ile yaşanan krizin kendi seçimleri olmadığını öne sürdü Türkiye ile ilişkilerin daha da kötüye gitmesini önlemeye çalıştıklarını savunan Netanyahu halihazırda iki ülke arasındaki Alevlerin seviyesini düşürmeye uğraştıklarını belirterek Umarım bu gerginlik eğer karşı taraf da isterse sona erdirilecektir diye konuştu
spaCy Tokenizer Corrupted News Sample Output:
Kanal 2 televizyonunda, İsrail'in tanınmış gazeteci ve analistlerinden Ehud Yaari ile birlikte konuk olan Lieberman'a, "Bakanlığı döneminde hem Türkiye hem de Mısır'dan büyükelçilerin kovulduğuna" işaret edilerek, gazetede yayımlanan haberin doğru olup olmadığı soruldu. Haberin doğru olmadığını söyleyen liebermana bu kez Peki Dışişlerinde böyle şeyler konuştunuz mu sorusu yöneltildi Lieberman, bu soruya, "Dışişleri Bakanlığı'nda her gün yüzlerce fikir tartışma konusu edilir" yanıtını verdi. Bunun üzerine Ehud Yaari Açıkça söyleyin PKK terör örgütüne silah vs sağlama gibi yardım etme konusu konuşuldu mu diyerek sorusunu yineledi LIEBERMAN, SORUYA BU KEZ "HAYIR, KESİNLİKLE KONUŞULMADI" KARŞILIĞINI VERDİ. lieberman palmer komisyonu raporunun mavi marmara baskını ile ilgili olarak israilin eyleminin ve gazzeye ablukanın haklı olduğunu açıkça ortaya koyduğunu da ifade etti lieberman, türkiye ile ilişkilerin normalleştirilmesinin yeniden sağlanacağı ve türkiye'nin, böyle bir normalleşmenin çıkarına olacağını göreceği umudunda olduğunu da kaydetti. Alevlerin seviyesini düşürmeye çalışıyoruz İsrail Başbakanı Binyamin Netanyahu da Türkiye ile yaşanan krizin kendi seçimleri olmadığını öne sürdü Türkiye ile ilişkilerin daha da kötüye gitmesini önlemeye çalıştıklarını savunan Netanyahu halihazırda iki ülke arasındaki Alevlerin seviyesini düşürmeye uğraştıklarını belirterek Umarım bu gerginlik eğer karşı taraf da isterse sona erdirilecektir diye konuştu
ErSatz Tokenizer Corrupted News Sample Output:
Kanal 2 televizyonunda, İsrail'in tanınmış gazeteci ve analistlerinden Ehud Yaari ile birlikte konuk olan Lieberman'a, "Bakanlığı döneminde hem Türkiye hem de Mısır'dan büyükelçilerin kovulduğuna" işaret edilerek, gazetede yayımlanan haberin doğru olup olmadığı soruldu. Haberin doğru olmadığını söyleyen liebermana bu kez Peki Dışişlerinde böyle şeyler konuştunuz mu sorusu yöneltildi Lieberman, bu soruya, "Dışişleri Bakanlığı'nda her gün yüzlerce fikir tartışma konusu edilir" yanıtını verdi. Bunun üzerine Ehud Yaari Açıkça söyleyin PKK terör örgütüne silah vs sağlama gibi yardım etme konusu konuşuldu mu diyerek sorusunu yineledi LIEBERMAN, SORUYA BU KEZ "HAYIR, KESİNLİKLE KONUŞULMADI" KARŞILIĞINI VERDİ. lieberman palmer komisyonu raporunun mavi marmara baskını ile ilgili olarak israilin eyleminin ve gazzeye ablukanın haklı olduğunu açıkça ortaya koyduğunu da ifade etti lieberman, türkiye ile ilişkilerin normalleştirilmesinin yeniden sağlanacağı ve türkiye'nin, böyle bir normalleşmenin çıkarına olacağını göreceği umudunda olduğunu da kaydetti. Alevlerin seviyesini düşürmeye çalışıyoruz İsrail Başbakanı Binyamin Netanyahu da Türkiye ile yaşanan krizin kendi seçimleri olmadığını öne sürdü Türkiye ile ilişkilerin daha da kötüye gitmesini önlemeye çalıştıklarını savunan Netanyahu halihazırda iki ülke arasındaki Alevlerin seviyesini düşürmeye uğraştıklarını belirterek Umarım bu gerginlik eğer karşı taraf da isterse sona erdirilecektir diye konuştu

Table 19: Predictions of the proposed ErSatz, Punkt, and spaCy baselines. ErSatz and Punkt are trained on the **Clean** version of the TRSEG-41 training set. The listed predictions are for the samples provided in Table 18.