Data and Representation for Turkish Natural Language Inference

Emrah Budur,[1,2] **Rıza Özçelik,**[2] **Tunga Güngör,**[2] **and Christopher Potts**[3]

[1]Garanti BBVA Technology [2]Boğaziçi University [3]Stanford University

{emrah.budur, riza.ozcelik, gungort}@boun.edu.tr,
cgpotts@stanford.edu

Abstract

Large annotated datasets in NLP are overwhelmingly in English. This is an obstacle to progress in other languages. Unfortunately, obtaining new annotated resources for each task in each language would be prohibitively expensive. At the same time, commercial machine translation systems are now robust. Can we leverage these systems to translate English-language datasets automatically? In this paper, we offer a positive response for natural language inference (NLI) in Turkish. We translated two large English NLI datasets into Turkish and had a team of experts validate their translation quality and fidelity to the original labels. Using these datasets, we address core issues of representation for Turkish NLI. We find that in-language embeddings are essential and that morphological parsing can be avoided where the training set is large. Finally, we show that models trained on our machine-translated datasets are successful on human-translated evaluation sets. We share all code, models, and data publicly.

1 Introduction

Many tasks in natural language processing have been transformed by the introduction of very large annotated datasets. Prominent examples include paraphrase (Ganitkevitch et al., 2013), parsing (Nivre et al., 2016), question answering (Rajpurkar et al., 2016), machine translation (MT; Bojar et al., 2014), and natural language inference (NLI; Bowman et al., 2015; Williams et al., 2018a).

Unfortunately, outside of parsing and MT, these datasets tend to be in English. This is not only an obstacle to progress on other languages, but it also limits the field of NLP itself: English is generally not a representative example of the world's languages when it comes to morphology, syntax, or spelling conventions and other kinds of standardization (Munro, 2012), so it's risky to assume that models and results for English will generalize to other languages.

A natural response to these gaps in our dataset coverage might be to launch new annotation efforts for multiple languages. However, this would likely be prohibitively expensive. For example, based on the costs of SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018a), we estimate that each large dataset for NLI would cost upwards of US $50,000 if created completely from scratch.

At the same time, commercial MT systems have improved dramatically in recent years (Wu et al., 2016; Johnson et al., 2017; Hieber et al., 2017, 2018; Tomasello, 2019; Hieber et al., 2020). They now offer high-quality translations between hundreds of language pairs. This raises the question: can we use these MT systems to translate English-language datasets and use the translated versions to drive more genuinely multilingual development in NLP? In this paper, we offer evidence that the answer is "yes".

Using Amazon Translate, we translated SNLI and MultiNLI from English into Turkish to create the first large Turkish NLI data sets, NLI-TR, at a tiny fraction of the cost of creating them from scratch. Turkish is an interesting challenge in this context since it is very different from English, most notably in its very free word order and complex morphology. A word in Turkish bears morpho-syntactic properties in the sense that phrases formed of several words in languages like English can be expressed with a single word form.

In our validation phase (Section 3), a team of Turkish–English bilingual speakers assessed the quality of a large sample of the translations in NLI-TR. They found the quality to be very high, which suggests that translated datasets can provide a foundation for NLI research on a resource-constrained language, even if it has significantly different characteristics from English.

We then use these datasets to study the roles of pre-trained language models and morphological parsing in successful NLI systems for Turkish (Section 4). For these experiments, we fit classifiers on top of pre-trained BERT parameters (Devlin et al., 2019) and compare the original BERT-base release, the multilingual BERT embeddings released by the BERT team, and the Turkish BERT (BERTurk) embeddings of Schweter (2020). We find BERTurk to be superior to the others for NLI-TR.

Morphological parsing is a natural preprocessing step for Turkish due to its complex morphology. Thus, we assess the use of three morphological parsers as the second case study: Zemberek (Akın and Akın, 2007), BOUN parser (Sak et al., 2011), and Turkish Morphology (Öztürel et al., 2019). We find that the parsers help where training data is sparse, but the need for a parser disappears as the training data increases. This is a striking finding: one might expect that Turkish would require morphological parsing given its complex word-formation processes. It might be regarded as welcome news, though, since the parsers are expensive to run. In Section 4.2, we report on some new optimizations of existing tools to make the relevant parsing jobs feasible, but we would still like to avoid these steps if possible, and it seems that we can for NLI.

Finally, we investigate how models trained on the machine translated datasets perform on the human translations from XNLI (Conneau et al., 2018). We find that machine translated and human translated sentences yield similar results, suggesting that it is safe to apply models trained on machine-translated datasets to human-written sentences.

2 Related Work

Early in the development of textual entailment tasks, Mehdad et al. (2010) argued for multilingual versions of them. This led to subsequent explorations of a variety of techniques, including crowdsourcing translations (Negri and Mehdad, 2010; Negri et al., 2011), relying on parallel corpora to support reasoning across languages (Mehdad et al., 2011), and automatically translating datasets using MT systems (Mehdad et al., 2010; Real et al., 2018; Rodrigues et al., 2020). This research informed SemEval tasks in 2012 (Negri et al., 2012) and 2013 (Negri et al., 2013) followed by ASSIN 1 (Fonseca et al., 2016) and 2 (Real et al., 2020) shared tasks exploring the viability of multilingual NLI. From the perspective of present-day NLI models, these datasets are very small, but they could be used productively as challenge problems.

More recently, Conneau et al. (2018) reinvigorated work on multilingual NLI with their XNLI dataset. XNLI provides expert-translated evaluation sets from English into 14 other languages, including Turkish. Though they are valuable resources to push NLI research beyond English, test sets alone are insufficient for in-language training on target languages, which is likely to lower the performance of the resulting systems.

Although it was not the main focus of the XNLI effort, Conneau et al. (2018) distributed machine translations of MultiNLI into other languages, including Turkish, which we call MultiNLI-TR$^{\text{XNLI}}$ in this paper. The translations helped them form a strong baseline for their cross-lingual models, which proved superior in their assessments. However, the quality of the translations is crucial, as the authors note. Our hope for NLI-TR is that it supports effective in-language training.

XNLI's primary focus on test sets rather than training is justified by a wide body of recent results on cross-lingual transfer learning. Multilingual embeddings (embeddings trained on multilingual corpora) have played an important role in these developments. The BERT team (Devlin et al., 2019) released multilingual embeddings and demonstrated their value using XNLI. At the same time, BERT models have been released for a variety of individual languages (see Wolf et al., 2019) and specialized domains (Alsentzer et al., 2019; Lee et al., 2020). While we might expect the language- and domain-specific embeddings to be superior for the kind of data they were trained on, the multilingual versions might be more efficient in large-scale deployments in diverse environments. Balancing these trade-offs is challenging. Here, we offer some insight into these trade-offs for Turkish.

Turkish is a morphologically-rich language in which new word forms are freely created using suffixation. Several morphological parsers (Akın and Akın, 2007; Öztürel et al., 2019; Sak et al., 2009) and morphological disambiguation systems (Akın and Akın, 2007; Sak et al., 2011) have been developed for Turkish. The state-of-the-art morphological analyzers can parse with success rates around 95%. We use three of these parsers in this work to evaluate the role of morphology in NLI systems (Section 4.2).

3 Creating and Validating NLI-TR

3.1 English NLI Datasets

We translated the Stanford Natural Language Inference Corpus (SNLI; Bowman et al., 2015) and the Multi-Genre Natural Language Inference Corpus (MultiNLI; Williams et al., 2018b) to create labeled NLI datasets for Turkish, NLI-TR.

SNLI contains ≈570K semantically related English sentence pairs. The semantic relations are entailment, contradiction, and neutral. The premise sentences for SNLI are image captions from the Flickr30K corpus (Young et al., 2014), and the hypothesis sentences were written by crowdworkers. SNLI texts are mostly short and structurally simple. We translated SNLI while respecting the train, development (dev), and test splits.

MultiNLI comprises ≈433K sentence pairs in English, and the pairs have the same semantic relations as SNLI. However, MultiNLI spans a broader range of genres, including travel guides, fiction, dialogue, and journalism. As a result, the texts are generally more complex than SNLI. In addition, MultiNLI contains *matched* and *mismatched* dev and test sets, where the sentences in the former set are from the same sources as the training set, whereas the latter consists of texts from different genres than those found in the training set. We translated the training set and both dev sets for NLI-TR.

3.2 Automatic Translation Effort

As we noted in Section 1, Turkish is a resource-constrained language with few labeled data sets compared to English. Furthermore, Turkish has a fundamentally different grammar from English that could hinder transfer-learning approaches. These facts motivate our effort to translate SNLI and MultiNLI from English to Turkish. We employ an automatic MT system and hope that it will deliver high-quality translations that we can use for NLI research and system development in Turkish.

We used Amazon Translate, a commercial neural machine translation service. Translation of all folds of SNLI and MultiNLI cost just US $2K (vs. the ≈US $100K we would expect for replicating these two datasets from scratch) and five days with no parallelization. We refer to the translated datasets as SNLI-TR and MultiNLI-TR, and collectively as NLI-TR. Translation examples are provided in Table 1. We publicly share NLI-TR.[1]

SNLI-TR and MultiNLI-TR are different from SNLI and MultiNLI in terms of token counts and vocabulary sizes. Table 2 illustrates these features before and after translation. For each fold in each dataset, translation decreased the number of tokens in the corpus, but it increased the vocabulary sizes drastically, in both the cased and uncased versions. Both of these differences are expected: many multiword expressions in English are translated into individual words due to the agglutinating nature of Turkish. For instance, the four-word English expression "when in your home" can be translated to the single word "evinizdeyken".

Table 2 also reflects the complexity difference between SNLI and MultiNLI that we noted in Section 3.1. Though SNLI contains more sentence pairs than MultiNLI, it has fewer tokens and a smaller vocabulary.

3.3 Translation Quality Assurance

Two major risks arise when using MT systems to translate NLI datasets. First, the translation quality might be low. Second, even if the individual sentences are translated correctly, the nature of the mapping from the source to the target language might affect the semantic relations between sentences. For example, English has the words "boy" and "girl" to refer to male and female children, and both those words can be translated to a gender-neutral Turkish word "çocuk". Now, consider a premise sentence "A boy is running" and its contradiction pair "A girl is running". Both sentences can be translated fluently into the same Turkish sentence, "Çocuk koşuyor", which changes the semantic relation from contradiction to entailment.

Thus, to determine the viability of NLI-TR as a tool for NLI research, we must assess both translation quality and the consistency of the NLI labels. To do this, we assembled a team of ten Turkish–English bilingual speakers who were familiar with the NLI task and were either MSc. candidates or graduates in a relevant field.

For expert evaluation, we grouped the translations into example sets of four sentences as in Table 1, where the first sentence (premise) is semantically related to the rest (hypotheses). We distributed the sets to the experts so that each set (and sentence) was examined by five randomly chosen experts and each expert co-examined approx-

[1] https://github.com/boun-tabi/NLI-TR

		English	Turkish
SNLI	**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.

Table 1: Sample translations from SNLI into NLI-TR. Each premise is associated with a hypothesis from each of the three NLI categories. Table 7 in our supplementary materials provides MultiNLI examples.

		English			Turkish		
Dataset	**Fold**	**Token Count**	**Vocab Size (Cased)**	**Vocab Size (Uncased)**	**Token Count**	**Vocab Size (Cased)**	**Vocab Size (Uncased)**
SNLI	**Train**	5900366	38565	32696	4298183	78786	66599
	Dev	120900	6664	6224	88668	11455	10176
	Test	120776	6811	6340	88533	11547	10259
MultiNLI	**Train**	6356136	81937	66082	4397213	216590	187053
	Matched Dev	161152	14493	12659	112192	27554	24872
	Mismatched Dev	170692	12847	11264	119691	26326	23941

Table 2: Comparative statistics for the English and Turkish NLI datasets. The Turkish translations have larger vocabularies and lower token counts due to the highly agglutinating morphology of Turkish as compared to English.

imately the same number of sets with each other expert. Each expert evaluated the translation by (i) grading the translation quality between 1 and 5 (inclusive; 5 the best) and (ii) checking if the translation altered the semantic relation. We distributed an annotation guide[2] to the team to standardize the criteria. In total, 500 example sets (2,000 translated sentences) were examined by five experts, yielding 10,000 annotations.

We use the average translation score of the annotations to estimate translation quality. For label consistency, there are two comparisons we can make, since we have five new annotations per example. The *annotation-level* analysis compares each new annotation with the gold label on the original English example. The *majority-level* analysis compares only the majority label (if any) of the five new annotations with the English gold label. The annotation-level analysis is more stringent, whereas the majority-level analysis directly connects with how we expect NLI-TR to be most commonly used. Table 3 reports these analyses for SNLI and MultiNLI. The results are extremely reassuring. First, average translation quality is near 5 (ceiling) for all the splits. Second, annotation-level label consistency is over 90% and majority-level label consistency is over 95%, indicating that the linguistic differences between English and Turkish are not a major issue for preserving NLI labels.

To assess the reliability of the translation quality scores, we calculated the Intra-Class Correlation (ICC; McGraw and Wong 1996). ICC is frequently adopted in medical studies to assess ordinal annotations provided by experts randomly drawn from a team. Its assumptions align well with our evaluation scheme. We obtained an ICC of 0.8426, which suggests excellent agreement (Cicchetti, 1994; Hallgren, 2012).

We also computed Krippendorff's alpha (Krippendorff, 1970), which is an inter-annotator agreement metric used more commonly in NLP. This metric is suitable for both nominal and ordinal annotations involving multiple annotators. We calculated intercoder reliability of the ordinally-scaled translation quality score as 0.47. Our annotation-level label consistency yielded a score of 0.78 whereas our majority-level label consistency resulted in a score of 0.99. In contrary to the perfect agreement in the majority-level label consistency, the Krippendorff's alpha values of annotation-level

[2] https://github.com/boun-tabi/NLI-TR

Dataset	Fold	Translation Quality	Annotation-level Label Consistency	Majority-level Label Consistency
SNLI-TR	Train	4.55 (0.78)	92.62%	98.67%
	Dev	4.46 (0.90)	90.53%	95.33%
	Test	4.45 (0.86)	87.87%	94.00%
MultiNLI-TR	Train	4.56 (0.80)	89.96%	96.22%
	Matched Dev	4.42 (0.86)	88.53%	95.33%
	Mismatched Dev	4.49 (0.82)	92.53%	98.00%
	All	4.51 (0.82)	90.72%	96.73%

Table 3: Translation quality and label consistency of the translations in SNLI-TR and MultiNLI-TR based on expert judgements. For the quality ratings (1–5), we report mean and standard deviation (in parentheses). For label consistency, we report the percentage of labels in SNLI-TR and MultiNLI-TR judged consistent with the original label, both in annotation- and sentence-level.

labels and translation quality scores suggest less overall agreement than our ICC values do, but they are still acceptable, and ICC is arguably the more appropriate metric for our study. Krippendorff's alpha is generally used for large, diverse annotation teams, and its penalties for disagreements are known to be harsh.

Overall, it seems that the very high estimates of translation quality and label consistency of NLI-TR are trustworthy, and only a small percentage of premise–hypothesis have inconsistent semantic labels between their original and translated forms. Still, we would like to better understand why inconsistencies do arise. To this end, we inspected all 49 label-inconsistent pairs in our annotations. We find that low translation quality is the leading source of such errors, which further emphasizes how essential it is to work with high-quality translations.

Of the label-inconsistent pairs with good translations, we find that about 20 probably trace to differing perspectives on how to apply the NLI annotation guidelines. Relatedly, Conneau et al. (2018) find that NLI labels often cannot be completely recovered by different annotators even with no sentence modifications.

Finally, we did find one example of label inconsistency that traces to a subtle difference between the English and Turkish lexicons. In this example, the premise "Your speeches are inflammatory" was translated to Turkish as "Konuşmalarınız çok kışkırtıcı", which can be back-translated as "Your speeches are provocative", while its entailment hypothesis "Your speeches upset people" was translated as "Konuşmaların insanları üzüyor", equivalent to "Your speeches make people sad". Annotators agreed that both of these translations are of maximum quality, but also stated that the Turkish pair should be labeled neutral. As bilingual speakers, we feel that this is essentially correct; the relevant English and Turkish adjectives are subtly different in ways that affect the NLI label. However, such examples seem to be rare and so pose minimal risk for conducting research using NLI-TR.

4 Experiments

4.1 Case Study I: Comparing BERT models on Turkish NLI Datasets

The arrival of pre-trained model-sharing hubs (e.g., Tensorflow Hub,[3] PyTorch Hub,[4] and Hugging Face Hub[5]) has democratized access to Transformer-based models (Vaswani et al., 2017), which are mostly in English. Combined with the abundance of labeled English datasets for fine-tuning, this has increased the performance gap between English and resource-constrained languages.

Here, we use NLI-TR to analyze the effects of pretraining Transformer-based models. We compare three BERT models trained on different corpora by fine-tuning them on NLI-TR. The results quantify the importance of having high-quality, language-specific resources.

4.1.1 Experimental Settings

We compared cased BERT-English (BERT-En), BERT-Multi, and BERTurk (Schweter, 2020). BERT-En is the original BERT-base model released by Devlin et al. (2019), which used an English-only

[3] https://github.com/tensorflow/hub
[4] https://pytorch.org/hub
[5] https://huggingface.co/models

corpus for training. BERT-Multi was released by the BERT team as well, and was trained on a corpus containing texts from 104 languages, including Turkish. Schweter's BERTurk also uses the same model architecture and is trained on a Turkish corpus (≈30GB).

We fine-tuned each model on train folds of NLI-TR separately and fixed the maximum sequence length to 128 for all experiments. Similarly, we used a common learning rate of 2×10^{-5} and batch size of 8 with no gradient accumulation. We fine-tuned each model for 3 epochs using HuggingFace's Transformers Library (Wolf et al., 2019). We evaluated the models on the test set of SNLI-TR and the *matched* and *mismatched* dev splits of MultiNLI-TR. Table 4 reports the accuracy of each model on the evaluation sets.

4.1.2 Results

Table 4 demonstrates that NLI-TR can be used to train high quality Turkish NLI models. We observe that every model performed better on the dev and test folds of SNLI-TR than the dev folds of MultiNLI-TR, which is an expected outcome given the greater complexity of MultiNLI compared to SNLI. The translation effort seems to have preserved this fundamental difference between the two datasets.

In addition, BERTurk, which was trained on a Turkish corpus, achieved the highest accuracy, and BERT-Multi, which used a smaller Turkish corpus, was ranked the second, consistently on every evaluation fold. The ranking emphasizes the importance of having a Turkish corpus for pre-training.

4.2 Case Study II: Comparing Morphological Parsers on Turkish NLI Datasets

In this case study, we use NLI-TR to compare three morphological parsers with regular tokenization. We train a BERT model from scratch utilizing each approach for pretraining and use NLI-TR for fine-tuning. This leads to the striking result that morphology adds additional information where training data is sparse, but its importance shrinks as the dataset grows larger.

4.2.1 Experimental Settings

Morphological Parsers We use Zemberek (Akın and Akın, 2007), BOUN (Sak et al., 2011), and Turkish Morphology (Öztürel et al., 2019) as parsers and compare them with an approach that does not do morphological parsing.

Zemberek is a mainstream Turkish NLP library used in research (Büyük, 2020; Kuyumcu et al., 2019; Özer et al., 2018; Can, 2017; Dehkharghani et al., 2016; Gulcehre et al., 2015) and applications such as iOS 12.2 and Open Office. It has 67,755 entries in its lexicon and uses a rule-based parser. BOUN implements the Turkish morphology rules described by Oflazer (1994) with a Finite State Transducer, and its lexicon has 55,278 entries. Finally, Turkish Morphology is an OpenFST-based (Allauzen et al., 2007) morphological parser that was recently released by Google Research and uses a lexicon with 47,202 entries.

Out of the box, Zemberek and BOUN can parse 398K and 51K tokens per minute respectively, whereas Turkish Morphology can process only 1K tokens. We sped up Turkish Morphology to parse 11 times more tokens per minute by implementing a dynamic programming wrapper (Bellman, 1952) that increased the cache hit ratio to 89.9%. This technique is already used by Zemberek.

Pretraining To conduct a wide range of experiments on a limited budget, we opted to use one-tenth (≈4GB, 500M tokens) of the Turkish corpus used by BERTurk (Schweter, 2020) to pretrain BERT models. We analyzed each token morphologically using Zemberek, BOUN, and Turkish Morphology and trained a BERT model using the stems of the tokens only. For the model that does not utilize morphological information, we used tokens as they are. We used the `BertWordPieceTokenizer` class of HuggingFace Tokenizers[6] with the same set of parameters for each model.

We trained each model on a single Tesla V100 GPU of an NVIDIA DGX-1 system, allocating 128GB memory for 1 day. We split the dataset into 30 equal shards for parallel processing, where each shard comprises 1M sentences, and shuffled the shards prior to training to reduce the adverse effects of variance across the sentence styles in the different shards (Goodfellow et al., 2016). We used an effective batch size of 128 with gradient accumulation to address memory limitations.

Fine-tuning We fine-tuned each model on NLI-TR with the same setting as in Section 4.1, with the exception that we trained for only 1 epoch. We measured the accuracy on the evaluation sets

[6] https://github.com/huggingface/tokenizers

Model Name	SNLI-TR		MultiNLI-TR	
	Dev	Test	Matched Dev	Mismatched Dev
BERT-En	81.83%	82.09%	69.98%	70.56%
BERT-Multi	85.37%	85.12%	75.97%	76.34%
BERTurk	**87.28%**	**87.04%**	**79.58%**	**80.87%**

Table 4: Accuracy results for the publicly available cased BERT models on NLI-TR. BERTurk performed the best in all three evaluations, highlighting the value of language-specific resources for NLI.

	SNLI-TR		MultiNLI-TR	
	Dev	Test	Matched Dev	Mismatched Dev
No Parser	76.50%	76.59%	58.24%	60.01%
Zemberek	76.47%	76.71%	59.01%	60.44%
BOUN Parser	**76.64%**	**76.89%**	59.99%	61.29%
Turkish Morphology	76.00%	76.36%	**60.13%**	**62.00%**

Table 5: Accuracy results for different morphology approaches on NLI-TR. To facilitate running many experiments, these results are for pretraining on just one-tenth of the Turkish corpus used by BERTurk and fine-tuning on NLI-TR for just one epoch.

with an interval of 1,000 training steps to observe the effect of morphological parsing as the dataset grew. Figure 1 reports the accuracy of all models with respect to fine-tuning steps on NLI-TR development sets, and Table 5 shows the final accuracies.

4.2.2 Results

Figure 1 suggests that morphological parsing is beneficial where the training set is small, but its importance largely disappears for large training sets. This is reflected also in the final results in Table 5. We relate this to the fact that BERT models create contextual embeddings of both word and subword tokens (Kudo, 2018; Kudo and Richardson, 2018; Sennrich et al., 2016). Given a sufficiently large dataset, BERT models can approximate the effects of morphological parsing even for Turkish, a morphologically-rich language.

The trends are not uniform for SNLI-TR and MultiNLI-TR. For SNLI-TR, all three models display a similar learning curve, with a slight edge for Zemberek early on. For MultiNLI-TR, models with morphological parsers are more differentiated. However, all three converge to similar performance at the end of training on both datasets (Table 5).

In light of these findings, we suggest avoiding the use of morphological parsers for Turkish NLI where the training set is large, since the benefits of such parsers are generally not enough to offset the cost of running them.

4.3 Case Study III: Evaluating NLI-TR on Human-Translated Sentences

Thus far, we have used NLI-TR for both training and assessment. One might worry that machine-translated test sets are not reliable tools for measuring how models will perform on examples written by humans. In this section, we address this concern using the Turkish dev and test portions of XNLI, which were translated entirely by humans. The models we assess on XNLI are those from our first case study as well as models trained on a different machine-translated training dataset, MultiNLI-TR$^{\text{XNLI}}$. Overall, we find that performance on XNLI is consistently very similar to performance on NLI-TR.

4.3.1 Datasets

MultiNLI-TR$^{\text{XNLI}}$ was created to investigate the performance of cross-lingual sentence embeddings compared to in-language ones (Conneau et al., 2018). It provides machine translations of only the MultiNLI training set, so we report comparisons with just the corresponding section of NLI-TR, and we train models only on these two training sets.

4.3.2 Models

We used the BERT models from Case Study I (Section 4.1) for evaluation. We fine-tuned each model on the training sets of MultiNLI-TR and MultiNLI-TR$^{\text{XNLI}}$ separately, following the same

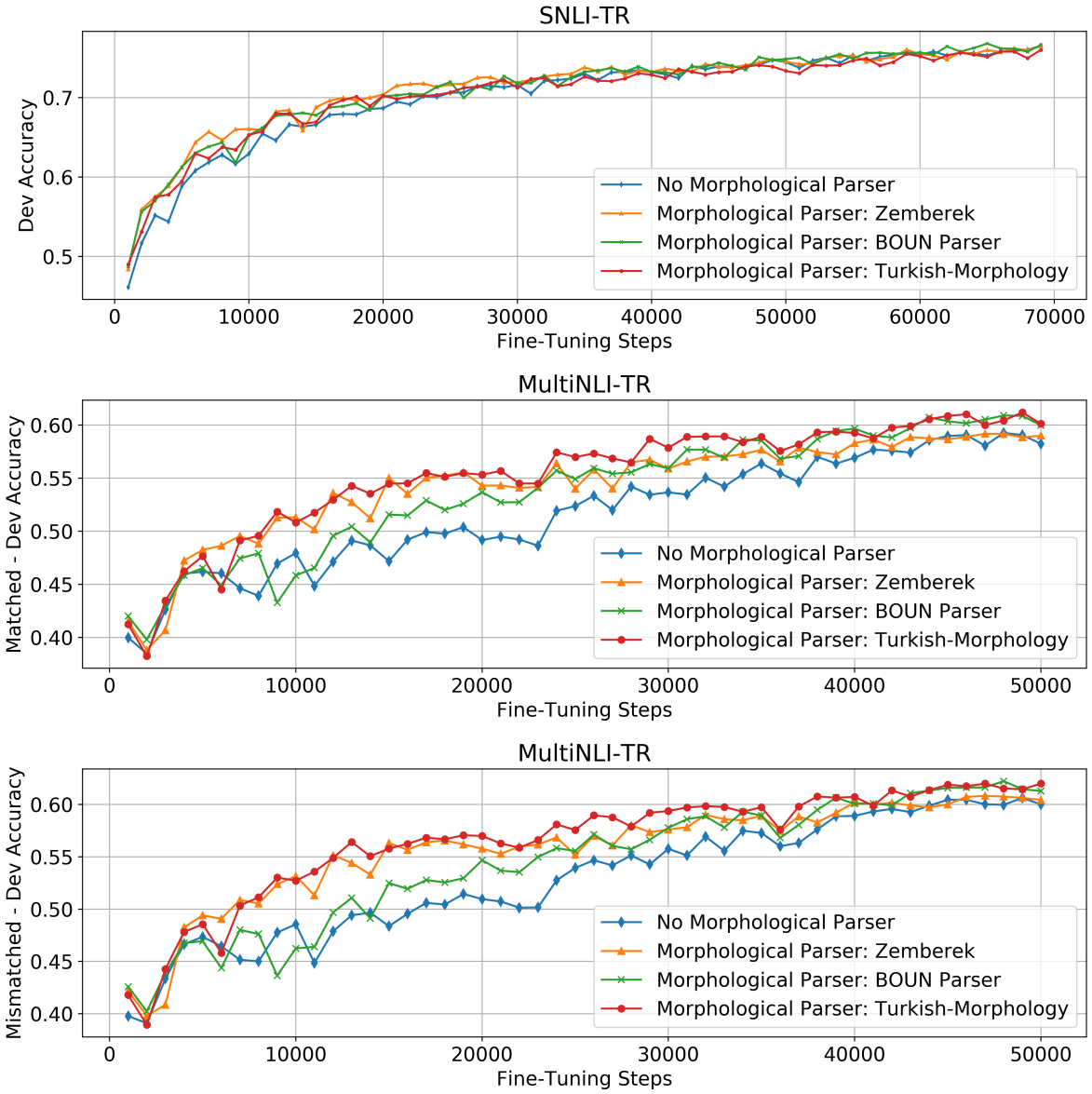

Figure 1: Development set accuracy for the three morphological parsers and a model without morphological parsing. The x-axis tracks the size of the training set. We find that morphological parsing is generally helpful in early rounds, when the training set is very small, but that its importance diminishes as the training set increases. These effects are especially clear for the two MultiNLI-TR dev sets.

fine-tuning steps as in Section 4.1, and computed their accuracy on XNLI-Dev and XNLI-Test.

4.3.3 Results

Table 6 provides the results of the experiments. All three models consistently achieve higher accuracy on XNLI-Dev and XNLI-Test when fine-tuned with MultiNLI-TR, but the performance difference is modest. Table 6 also illustrates that BERTurk, backed by a Turkish-only training corpus, outperforms the other two models on all eight evaluations. Its performance is followed by BERT-Multi, which is trained on a corpus with texts in multiple languages, including Turkish. The same result was also shown in Case Study I using the evaluation splits of NLI-TR. Therefore, machine-translated MultiNLI-TR and human-translated XNLI display similar characteristics across evaluations, which lends further credence to our claim that MT can help provide a viable path to robust Turkish NLI.

To better understand how in-language pretraining (BERTurk) helps, we investigated the 57 hypotheses from XNLI-Dev and XNLI-test where BERTurk was successful and BERT-Multi was

	MultiNLI-TR		MultiNLI-TR^{XNLI}	
Model Name	**XNLI-Dev**	**XNLI-Test**	**XNLI-Dev**	**XNLI-Test**
BERT-En	66.99%	67.74%	65.66%	65.71%
BERT-Multi	73.82%	72.95%	71.61%	71.20%
BERTurk	**76.75%**	**78.72%**	**76.43%**	**76.43%**

Table 6: Accuracy results comparing NLI-TR with another machine translated dataset. NLI-TR performed better, but the gap is modest, suggesting that both datasets have value for Turkish NLI. Figure 2 in our supplementary materials provides full learning curves. The results are very similar to those of Table 4 for MultiNLI, in overall quality and in the ranking of models.

not. For these sentences, we observed that the BERT-Multi tokenizer was often unable to segment the words into meaningful Turkish subword units, most likely due to its training on a multilingual corpus. For instance, BERT-Multi often could not segment the suffix "-me/ma", which negates a verb in Turkish, and thus bears crucial semantics for many contradiction examples (Gururangan et al., 2018). This shows that in-language training is essential not only for good vector representations but also for effective tokenization.

We also hypothesize that subtle lexical distinctions are another factor in the performance difference between BERTurk and BERT-Multi. For example, though BERT-Multi successfully identified the semantic relations created by frequent pairs such as "hiç" ('any') and "hepsi" ('all'), it missed many other distinctions like these. We propose that this is due to the more limited vocabulary of BERT-Multi for Turkish and the more robust word representations in BERTurk.

In addition to manual inspection, we computationally analyzed the pairs where BERT-Multi was unsuccessful and BERTurk was successful. We computed the frequency of each semantic class in the BERT-Multi predictions for these sentences and observed that the neutral class is the most common. This perhaps reflects the fact that neutral is the default choice where the model cannot robustly identify a semantic relation.

5 Conclusion

We created and released the first large Turkish NLI dataset, NLI-TR, by machine translating SNLI and MultiNLI. Though English and Turkish have very different grammars and thus stress-test automatic approaches, our team of experts judged the translations to be of very high quality and to preserve the original NLI labels consistently. These results suggest that MT can help address the paucity of datasets for Turkish NLI. We release code, models, and data publicly for further research.

We also used NLI-TR to investigate central issues in Turkish NLI. First, we used NLI-TR to analyze the effects of in-language pretraining. Second, we compared three morphological parsers for Turkish with simpler tokenization schemes. We found that a Turkish-only pretraining regime can enhance Turkish models significantly, and that morphological parsing is arguably worth its costs only when the training dataset is small. In our final case study, we returned to the general issue of translation quality, but now from the perspective of developing NLI systems. We showed that models trained on MultiNLI-TR perform well on the expert-translated test set from XNLI.

On the basis of these findings, we argue that MT can be more widely adopted for advancing NLP studies on resource-constrained languages. Though language-dependent tasks like dependency parsing are challenging to translate, MT can efficiently transfer large and expensive-to-create labeled datasets from English to other languages in many NLP tasks, including text classification, question answering, and text summarization. In addition, MT will presumably get cheaper, faster, and better over time, thereby further strengthening our core claims.

Acknowledgments

This research was supported by the AWS Cloud Credits for Research Program (formerly AWS Research Grants). E.Budur is thankful for the support provided by The Scientific and Technological Research Council of Turkey (TÜBİTAK) and Council of Higher Education (YÖK) under BİDEB 2214/A and 100/2000 graduate research scholarship programs, respectively. R.Özçelik gratefully

acknowledges the graduate research scholarship by TÜBİTAK under BİDEB 2211/A program.

The authors gratefully acknowledge that the computational parts of this study has been mostly performed at Boğaziçi TETAM DGX-1 GPU Cluster and partially carried out at TÜBİTAK ULAKBİM High Performance and Grid Computing Center (TRUBA resources) and Stanford Research Computing Center (FarmShare).

We thank Alara Dirik, Almira Bağlar, Berfu Büyüköz, Berna Erden, Fatih Mehmet Güler, Gökçe Uludoğan, Gözde Aslantaş, Havva Yüksel, Melih Barsbey, Melike Esma İlter, Murat Karademir, Ramazan Pala, Selen Parlar, Tuğçe Ulutuğ, Utku Yavuz for their annotation support and vital contributions. We are grateful also to Stefan Schweter and Kemal Oflazer for sharing the dataset that BERTurk was trained on and Omar Khattab, Dallas Card, Yiwei Luo, and many more researchers including the anonymous reviewers for their valuable advice, discussion and insightful comments.

References

Ahmet Afşın Akın and Mehmet Dündar Akın. 2007. Zemberek, an open source NLP framework for Turkic languages. *Structure*, 10:1–5. https://github.com/ahmetaa/zemberek-nlp.

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. OpenFst: A general and efficient weighted finite-state transducer library. In *Implementation and Application of Automata*, pages 11–23, Berlin, Heidelberg. Springer Berlin Heidelberg.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Richard Bellman. 1952. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, 38(8):716–719.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

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

Osman Büyük. 2020. Context-dependent sequence-to-sequence Turkish spelling correction. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(4):1–16.

Burcu Can. 2017. Unsupervised learning of allomorphs in Turkish. *Turkish Journal of Electrical Engineering & Computer Sciences*, 25(4):3253–3260.

Domenic V Cicchetti. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4):284.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Rahim Dehkharghani, Yucel Saygin, Berrin Yanikoglu, and Kemal Oflazer. 2016. SentiTurkNet: a Turkish polarity lexicon for sentiment analysis. *Language Resources and Evaluation*, 50(3):667–685.

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.

Erick Rocha Fonseca, Leandro Borges dos Santos, Marcelo Criscuolo, and Sandra Maria Aluísio. 2016. Visao geral da avaliaçao de similaridade semântica e inferência textual. *Linguamática*, 8(2):3–13.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares,

Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

Felix Hieber, Tobias Domhan, Michael Denkowski, and David Vilar. 2020. Sockeye 2: A toolkit for neural machine translation. In *European Association for Machine Translation*.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 200–207, Boston, MA. Association for Machine Translation in the Americas.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70.

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.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Birol Kuyumcu, Cüneyt Aksakallı, and Selman Delil. 2019. An automated new approach in fast text classification (FastText): A case study for Turkish text classification without pre-processing. In *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*, NLPIR 2019, page 1–4, New York, NY, USA. Association for Computing Machinery.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Kenneth O McGraw and Seok P Wong. 1996. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2010. Towards cross-lingual textual entailment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 321–324, Los Angeles, California. Association for Computational Linguistics.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2011. Using bilingual parallel corpora for cross-lingual textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1336–1345, Portland, Oregon, USA. Association for Computational Linguistics.

Rob Munro. 2012. *Processing Short Message Communications in Low-Resource Languages*. Ph.D. thesis, Stanford University, Stanford, CA.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 670–679, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2012. Semeval-2012 task 8: Cross-lingual textual entailment for content synchronization. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 399–407, Montréal, Canada. Association for Computational Linguistics.

Matteo Negri, Alessandro Marchetti, Yashar Mehdad, Luisa Bentivogli, and Danilo Giampiccolo. 2013. Semeval-2013 task 8: Cross-lingual textual entailment for content synchronization. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh*

International Workshop on Semantic Evaluation (SemEval 2013), pages 25–33, Atlanta, Georgia, USA. Association for Computational Linguistics.

Matteo Negri and Yashar Mehdad. 2010. Creating a bilingual entailment corpus through translations with mechanical turk: $100 for a 10-day rush. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 212–216, Los Angeles. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Kemal Oflazer. 1994. Two-level description of Turkish morphology. *Literary and linguistic computing*, 9(2):137–148.

Zeynep Özer, İlyas Özer, and Oğuz Fındık. 2018. Diacritic restoration of Turkish tweets with word2vec. *Engineering Science and Technology, an International Journal*, 21(6):1120–1127.

Adnan Öztürel, Tolga Kayadelen, and Işın Demirşahin. 2019. A syntactically expressive morphological analyzer for Turkish. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 65–75.

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.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. The ASSIN 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer.

Livy Real, Ana Rodrigues, Andressa Vieira e Silva, Beatriz Albiero, Bruna Thalenberg, Bruno Guide, Cindy Silva, Guilherme de Oliveira Lima, Igor CS Câmara, Miloš Stanojević, et al. 2018. SICK-BR: a Portuguese corpus for inference. In *International Conference on Computational Processing of the Portuguese Language*, pages 303–312. Springer.

Ruan Chaves Rodrigues, Jéssica Rodrigues da Silva, Pedro Vitor Quinta de Castro, Nádia Félix Felipe da Silva, and Anderson da Silva Soares. 2020. Multilingual transformer ensembles for Portuguese natural language tasks. In *Proceedings of the ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*, CEUR Workshop Proceedings, pages 27–38. CEUR-WS.org.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2009. A stochastic finite-state morphological parser for Turkish. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 273–276, Suntec, Singapore. Association for Computational Linguistics.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2011. Resources for Turkish morphological processing. *Language resources and evaluation*, 45(2):249–261.

Stefan Schweter. 2020. BERTurk - BERT models for Turkish.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Laura Tomasello. 2019. *Neural Machine Translation and Artificial Intelligence: What Is Left for the Human Translator?* Ph.D. thesis, University of Padua.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Appendices

		English	Turkish
SNLI	**Premise**	Several people are on stage preparing for a show.	Birkaç kişi sahnede gösteri için hazırlanıyor.
	Entailment	People are setting up for a show.	İnsanlar bir gösteri için hazırlanıyor.
	Contradiction	A house is being demolished.	Bir ev yıkılıyor.
	Neutral	A crew is getting ready for a rock concert.	Bir ekip rock konseri için hazırlanıyor.
MultiNLI	**Premise**	All rooms have color TV, alarm clock/radio, en-suite bathrooms, real hangers, and shower massage.	Tüm odalarda renkli TV, çalar saat/radyo, en-suite banyo, gerçek askılar ve duş masajı vardır.
	Entailment	All rooms also contain a ceiling fan and outlets for electronics.	Tüm odalarda ayrıca tavan vantilatörü ve elektronik prizler bulunmaktadır.
	Contradiction	You will not find a TV or alarm clock in any of the rooms.	Odaların hiçbirinde TV veya çalar saat bulunmamaktadır.
	Neutral	Color TVs, alarms, and hangers can be found in all rooms.	Tüm odalarda renkli TV'ler, alarmlar ve askılar bulunur.

Table 7: Sample translations from SNLI and MultiNLI into NLI-TR. Each premise is associated with a hypothesis from each of the three NLI categories.

	SNLI	**MultiNLI**	
Model Name	**Test**	**Matched Dev**	**Mismatched Dev**
BERT-En	**90.13%**	**83.16%**	**83.95%**
BERT-Multi	89.02%	81.74%	82.13%
BERTurk	85.84%	75.16%	75.60%

Table 8: Accuracy of the cased models in Table 4 trained on SNLI and MultiNLI. We used the same fine-tuning and evaluation procedures. BERT-En ranked the first and BERT-Multi ranked the second, emphasizing the importance of in-language training one-more time as in Section 4.1.

.

	MultiNLI-TR		**MultiNLI-TR$^{\text{XNLI}}$**	
Model Name	**XNLI-Dev-TR**	**XNLI-Test-TR**	**XNLI-Dev-TR**	**XNLI-Test-TR**
BERT-En	70.11%	70.11%	68.42%	67.70%
BERT-Multi	75.85%	74.79%	74.69%	73.77%
BERTurk	**80.11%**	**79.52%**	**79.95%**	**78.40%**

Table 9: Accuracy results of the models in Table 6 for machine translated XNLI. The outcomes agree with the ones in Section 4.3, suggesting that machine translated sentences can be used to evaluate Turkish NLI models. Here we note that, XNLI-Dev-TR, XNLI-Test-TR and MultiNLI-TR are translated with the same MT service, whereas MultiNLI-TR$^{\text{XNLI}}$ used a different one. Though this might result in a positive bias for MultiNLI-TR models, we report the accuracy of MultiNLI-TR$^{\text{XNLI}}$ models as well for the sake of completeness.

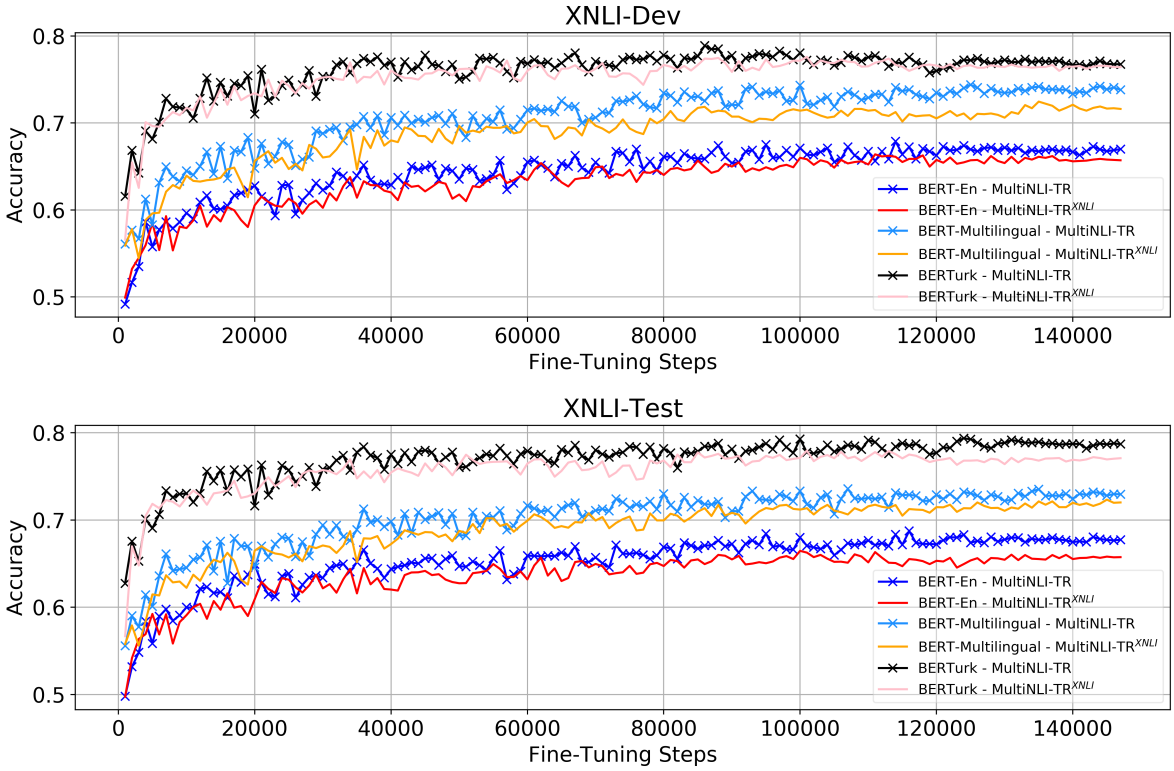

Figure 2: XNLI-Dev and XNLI-Test accuracy of three transformer models trained on MultiNLI-TR and MultiNLI-TR$^{\text{XNLI}}$. The x-axis tracks the training set size. We find that models trained on MultiNLI-TR are superior to their MultiNLI-TR$^{\text{XNLI}}$ counterparts from the start of the training until the end.

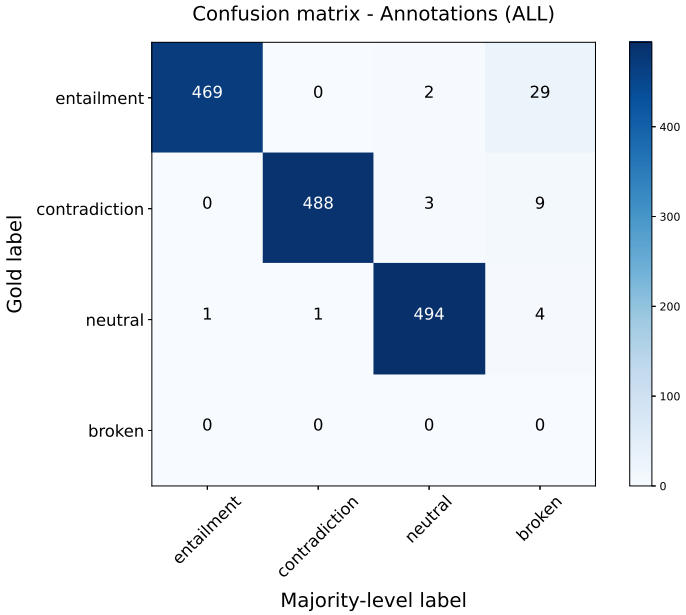

Figure 3: Confusion matrix for the majority-level label consistency results of all annotations in Table 3. The label "broken" corresponds to the pairs which have either major translation error or no majority-level label.