PARSI NLU: A Suite of Language Understanding Challenges for Persian


1Allen Institute for AI, USA 2Google, USA 3George Washington University, USA 4UC Irvine, USA 5University of Pittsburgh, USA 6TaskRabbit, USA 7Arizona State University, USA 8UC Santa Cruz, USA 9University of Southern California, USA 10IMRSV Data Labs, Canada 11EPFL, Switzerland 12University of Illinois - Chicago, USA 13University of Maryland Baltimore County, USA 14Rutgers University, USA 15University of Pennsylvania, USA 16Expedia Inc., USA 17University of Houston, USA 18Indiana University - Bloomington, USA 19Microsoft, Canada

Abstract

Despite the progress made in recent years in addressing natural language understanding (NLU) challenges, the majority of this progress remains to be concentrated on resource-rich languages like English. This work focuses on Persian language, one of the widely spoken languages in the world, and yet there are few NLU datasets available for this language. The availability of high-quality evaluation datasets is a necessity for reliable assessment of the progress on different NLU tasks and domains. We introduce PARSI NLU, the first benchmark in Persian language that includes a range of language understanding tasks—reading comprehension, textual entailment, and so on. These datasets are collected in a multitude of ways, often involving manual annotations by native speakers. This results in over 14.5k new instances across 6 distinct NLU tasks. Additionally, we present the first results on state-of-the-art monolingual and multilingual pre-trained language models on this benchmark and compare them with human performance, which provides valuable insights into our ability to tackle natural language understanding challenges in Persian. We hope PARSI NLU fosters further research and advances in Persian language understanding.1

1https://git.io/JJuRO.

1The point of view of the authors are their own and not attributable to the company they work for.

1Introduction

In recent years, considerable progress has been made in building stronger NLU models, particularly supported by high-quality benchmarks (Bowman et al., 2015; Rajpurkar et al., 2016; Wang et al., 2019) for resourceful languages like English. However, in many other languages, such benchmarks remain scarce, unfortunately, stagnating the progress towards language understanding in these languages.

In this work, we focus on developing natural language understanding (NLU) benchmarks for Persian (also known as Farsi). This language has many attributes that make it distinct from other well-studied languages. In terms of script, Persian is similar to Semitic languages (e.g., Arabic). Linguistically, however, Persian is an Indo-European language (Masica, 1993) and thus distantly related to most of the languages of Europe as well as the northern part of the Indian subcontinent. Such attributes make Persian a unique case to study in terms of language technologies. Although Persian is a widely spoken language (Simons and Fennig, 2017), our ability to evaluate performance and measure the progress of NLU models on this language remains limited. This is mainly due to the lack of major language understanding benchmarks that can evaluate progress on a diverse range of tasks.

In this work, we present PARSI NLU, a collection of NLU challenges for Persian.² PARSI NLU contains challenges for reading comprehension, multiple-choice question-answering, textual entailment, sentiment analysis, question paraphrasing,

²We focus on the standard Iranian Persian, spoken by over 80 million people. There are other dialects of Persian spoken in other countries, e.g., Afghanistan and Tajikistan.
and machine translation (examples in Figure 1). ParsiNLU offers data for tasks that have never been explored before in the context of the Persian language. We are not aware of any publicly available dataset for Persian question answering (§3.2.2), reading comprehension (§3.2.1), and paraphrasing (§3.2.5). For the rest of the tasks, we improve at least one aspect of the existing datasets (e.g., better data construction, more comprehensive evaluation, and evaluation of less investigated genres or domains). To ensure the quality of the presented challenge tasks, we rely on the annotations from native Persian speakers or novel data collection techniques, such as search engine autocomplete (§3.2.1) and past collegiate exams (§3.2.2). To the best of our knowledge, this is the first comprehensive collection of its own, composed of a variety of Persian NLU tasks.

We conduct a collection of empirical work (§4) to establish the difficulty of ParsiNLU. We benchmark each ParsiNLU task via collecting state-of-the-art multilingual and mono-lingual language models (LMs), as well as estimating the human upper bound scores. The gap between human and machine baselines indicate the need for further research and stronger models for Persian. We hope that the release of ParsiNLU will encourage more research on Persian NLP.

2 Related Work

Cross-lingual Benchmarks. There are several recent cross-lingual benchmarks; however, almost none includes Persian: XNLI (Conneau et al., 2018) for entailment; PWNS-X (Yang et al., 2019) for paraphrasing; XCOPA (Ponti et al., 2020) for choice of plausible alternatives; and XQuAD, MLQA, TyDi, and MKQA (Artetxe et al., 2020b; Lewis et al., 2020; Clark et al., 2020a; Longpre et al., 2020) for reading comprehension. These datasets have also been integrated as part of multitask multilingual evaluation suites such as XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020). Unfortunately, the Persian portion of the former benchmark covers only two tagging tasks (POS and NER) and the latter does not cover Persian.

NLU Benchmarks for Other Languages. Benchmarks like GLUE (Wang et al., 2019) encourage development of better and stronger models on a diverse set of challenges. There have been several efforts to create GLUE-like benchmarks for other languages; for example, CLUE for Chinese (Xu et al., 2020), GLUECoS for Hindi (Khanuja et al., 2020), and RussianSuperGLUE (Shavrina et al., 2020). We view ParsiNLU in the same family of benchmarks, dedicated to the Persian language.

NLU Datasets for Persian. Prior work on creating evaluation resources for the Persian language has focused on low-level tasks in narrow domains (e.g., datasets for POS [Bijankhan, 2004], NER [Shahshahani et al., 2019], Parsing [Seraji et al., 2013]). Complementary to these efforts, we aim at providing an NLU evaluation benchmark for Persian, consisting of a wide variety of tasks. Below we mention several related works and how we build upon them.

FarsTail (Amirkhani et al., 2020) is a concurrent work on the entailment task, where the dataset is constructed semi-automatically based on existing multiple-choice exams. Different from this work, our entailment datasets are built with the annotations of native speakers of Persian and some use of machine translation (§3.2.4). Therefore, we hypothesize our construction represents a slightly different distribution than that of FarsTail.

There is a rich set of works on Persian sentiment analysis. We build upon these works and differ from them in the following manners: (a) The existing work mainly focuses on document-level sentiment identification which does not capture the nuanced judgments with respect to aspects and entities of the context (HosseinzadehBendarkheili et al., 2019; Sharami et al., 2020, inter alia). In addition to such document-level annotations, we provide aspect-level sentiment annotations (§3.2.3). (b) The majority of existing resources, such as MirasOpinion (Ashrafi Asli et al., 2020), focus on binary or ternary sentiment classes. However, our annotations contain a more granular sentiment intensity with five labels (§3.2.3). (c) Compared to the aspect-level datasets of Hosseini et al. (2018) and Ataei et al. (2019), we cover two relatively less investigated domains: food & beverages and movies, each posing new challenges for Persian sentiment analysis.

Machine translation of Persian $\rightleftharpoons$ English is one of the few tasks that has enjoyed decent attention (Tiedemann and Nygaard, 2004; Mohaghegh et al., 2010; Pilevar et al., 2011; Mohaghegh et al., 2011; Rasooli et al., 2013; Karimi et al., 2018; Kashefi, 2018; Khojasteh et al., 2020). Unfortunately, most published work for this task focus on
Figure 1: Examples of the PassNLU tasks. For each task (other than Machine Translation, which already contains English phrases) we show the English translations for ease of communication to non-Persian readers. The purple tags indicate the example category, according to their construction (explained in the main text under Section 3.2).
niche domains and datasets. Our contribution to this task is compiling a set of high-quality evaluation sets from a broad range of domains, based on the existing datasets as well as datasets introduced in this work. The hope is that this will help future work on Persian MT to evaluate their systems on a variety of domains to get a more realistic measure of machine translation.

To the best of our knowledge, this is the first work that publishes an evaluation benchmark for Persian language, promoting future studies on several NLU tasks such as question answering (§3.2.2), reading comprehension (§3.2.1), and paraphrasing (§3.2.5), among others.

3 ParsiNLU

3.1 Design Considerations

We now discuss possible design choices for constructing the dataset and the underlying reasons.

Naturally Occurring Instances. A common way of collecting data for low-resource languages has been using automated translation of the benchmark datasets of high-resource languages (Artetxe et al., 2020b; Ponti et al., 2020). This can be a poor practice, as recent investigations have shown translation artifacts in data gathered via translation of existing tasks (Artetxe et al., 2020a). It is important for any NLP dataset to reflect the natural distribution of the target language tokens and their associated cultural contexts. Therefore, one should avoid over-reliance on automatic conversion of resources from high-resource languages to minimize any unnatural instances or artifacts (Khvalchik and Malkin, 2020).

Experts Over Crowdworkers. While crowdsourcing has been the common approach for building datasets, we choose to work with a few native Persian speakers to construct the dataset. Crowdworkers are difficult to train and often generate more noisy annotations. However, expert annotators who are closely familiar with the task at hand often generate better quality annotations. Using crowdsourcing is further complicated by the fact that crowdsourcing platforms do not have an active community of Persian-speaking workers due to limited international financial transactions and crowdsourcing platforms. A study done by Pavlick et al. (2014, Table 6) shows that there are almost no crowdworkers for Persian on the Amazon Mechanical Turk platform.

3.2 Constructing ParsiNLU tasks

Examples are shown in Figure 1. We now explain the data construction of each task.

3.2.1 Reading Comprehension

We use the commonly used definition of reading-comprehension task: extracting a substring from a given context paragraph that answers a given question.

SQuAD (Rajpurkar et al., 2016) is one of the most popular reading comprehension datasets in English. Similar datasets to SQuAD are developed in other languages using varying degrees of human or semi-automatic translation techniques: KorQuAD for Korean (Lim et al., 2019), MMQA for Hindi (Gupta et al., 2018), and so on. For constructing our reading comprehension tasks, we avoid using SQuAD as a source and use a process resembling that of Kwiatkowski et al. (2019) that would lead to more natural questions.

Collecting Questions. Our efforts to translate questions from the English dataset indicated that such questions are often about topics that are not of much importance in Persian. For instance, there are many questions in SQuAD (Rajpurkar et al., 2016) about major US sports events (e.g., Superbowl, NFL) or western civilization history that might not be common among Persian speakers. Instead, we follow a pipeline that is more similar to the one introduced by Kwiatkowski et al. (2019), setting our goal to annotate answers for an existing naturalistic set of questions in Persian, as opposed to writing questions for existing paragraphs.

Unlike Kwiatkowski et al. (2019), we do not have direct access to query logs. Thus we follow the approach of Berant et al. (2013) and Khashabi et al. (2021), which relies on a query auto-completion API for collecting questions. Similarly, we use Google’s auto-completion,\(^3\) which enables us to mine a rich, yet a natural set of questions in Persian as it is reflective of popular questions posed by users of Google.

We start with a seed set of question terms (e.g., “چه کسی” [che kasi] meaning “who”, and “کجاست” [koja] meaning “where”) We bootstrap based on this set, by repeatedly querying parts of previously extracted questions, in order to discover a longer and richer set of questions. We hypothesize that such questions extracted from the auto-complete

\(^3\)http://google.com/complete/search?client=chrome&q=...
algorithm are highly reflective of popular questions posed by Persian-speaking users of Google. We filter out any results shorter than 5 tokens as they are often incomplete questions. This process yields over 50k questions.

Subsequently, we automatically filter out open-ended questions with no concrete answers (e.g., "نیچهی باری با رژی؟" [nætdʒɛ ye bɔzi bu ʒəpɔn?] meaning "What is the result of the game with Japan?"). Our filtering was guided by the observation that typically more complete questions lead to Google results that include well-established sources (such as Wikipedia). Hence, we perform this filtering by retrieving the Google search results for each question and checking if any of the top 10 search results overlap with a pre-defined list of credible websites. We keep only the questions that match this criterion.

**Annotating Paragraphs and Answers.** In this step, native speakers of Persian select a paragraph and an answer span within the paragraph that answers each of the questions. At the first step, the annotators read the question and correct any grammatical errors and typos (e.g., "اشتاسان" [otsan] is corrected to "استان" [ostan] “state”). Next, they annotate all the minimal and coherent spans that contains the answer to the question, from a paragraph obtained from a relevant web page (from the Google search results retrieved from an earlier step). Whenever possible, we annotate all valid spans as the answer (for example, "همدان" [hæmdan] and "همدان" [hæmdan e hæmdan], as shown in Figure 1). The paragraph that contains this answer is also annotated as the context of the question.

Overall, 6 native-speaker annotators annotated a collection of 1.3k question-answer-paragraph triplets (Table 2).

**Annotation Quality.** To ensure the quality of the annotations, the answers to each question were labeled by two independent annotators. Any misalignment of the answer spans or missing any valid spans were indicated as disagreements.

Such disagreements were resolved in further adjudication.

### 3.2.2 Multiple-Choice QA

Multiple-choice questions are one of the common formats for evaluation of fact-retrieval and reasoning (Richardson et al., 2013; Clark et al., 2020b).

Following prior works, we define the task as: given a natural language question, pick the correct answer among a list of multiple candidates. A key difference from reading comprehension (§3.2.1) is that the instances are open-domain (i.e., no context paragraph is provided). Hence, a model would either need to retrieve external supporting documents or have stored the necessary knowledge internally to be able to answer the question.

**Sources of Questions.** We use existing sources of multiple-choice questions, rather than annotating new ones. We collect the questions from a variety of sources: (i) The literature questions of the annual college entrance exams in Iran, for the past 15 years. These questions often involve the understanding of poetry and their implied meaning, knowledge of Persian grammar, and the history of literature. (ii) Employment exams that are expected to assess an individual’s depth in various topics (accounting, teaching, mathematics, logic, etc.). (iii) Common knowledge questions, which involve questions about topics such as basic science, history, or geography.

Most of these sources are scanned copies of the original exams in image format. We use an existing Persian OCR tool to convert the image data to a textual format. Then 4 annotators fix any mistakes made by the OCR system and convert the result into a structured format. Overall, this yields 2460 questions with an average of 4.0 candidate answers (Table 2). Additionally, the task comes with a label indicating the type of knowledge it requires: ‘literature’ (understanding of literary expressions), ‘common-knowledge’ (encyclopedic knowledge or everyday activities), and ‘math & logic’ (logical or mathematical problems). Examples from each category of questions are included in Figure 1.

**Annotation Quality.** To further examine the quality of the annotations, we randomly sampled 100 questions from the annotations and cross-checked the OCR output with the original data. We discovered that 94 of such questions exactly matched the original data, and the rest required minor modifications. We thus conclude that the annotated data is of high quality.

### 3.2.3 Aspect-Based Sentiment Analysis

Sentiment Analysis (SA) is the study of opinions (i.e., positive, negative, or neutral sentiment)
expressed in a given text (Liu, 2012). Aspect-based Sentiment Analysis (ABSA) is a more fine-grained SA that aims to extract aspects of entities mentioned in the text and determine sentiment toward these aspects (Pontiki et al., 2014). For instance, “it tastes good but it’s so expensive ...” (Figure 1) conveys positive and negative sentiments with respect to taste and price aspects of the mentioned product (entity), respectively.

Annotation Scheme. We follow the existing ABSA scheme (Pontiki et al., 2014). For every review, we do two types of annotations: (1) We assign an overall sentiment to each review, selecting from one of the following values: very-negative, negative, neutral, positive, very positive, and mixed. The mixed category indicates reviews where none of the sentiments are dominant (mix of positive and negative, or borderline cases), hence it is hard to detect the primary sentiment of a review. We also assign neutral label to reviews that express no clear sentiment toward an entity or any aspect of it. (2) We annotate pairs of \((a, s)\) where \(a\) is an aspect that belongs to a predefined set of aspects for each domain and \(s\) expresses the sentiment toward the aspect \(a\).

Collecting Reviews. At first, we collect reviews from two different domains: (1) food & beverages and (2) movies. We chose these domains since they are relatively less investigated in the existing literature (see §2 for past work). For the food & beverages category, we extracted\(^7\) reviews from the online grocery section of Digikala,\(^8\) and for the movie reviews category, we crawled reviews from Tiwall.\(^9\) Both of these websites are well known and popular websites among Persian speakers.

Defining Aspects. Following the ABSA scheme, we predefined a set of aspects for each domain. For food & beverages, we crawled Digikala and retrieved all listed aspects for product reviews in the food & beverages category. Subsequently, we manually aggregated the extracted aspects and merged those with significant semantic overlap. We also added taste/smell as a new aspect category because users frequently commented on this aspect. For movie reviews, we created an initial list of aspects based on the movie review aspects defined by Thet et al. (2010). In consultation with a movie critic, we resolved the potential overlaps among aspect categories and created a set of aspects that capture various perspectives of movie reviews. Overall, this process resulted in 6 and 7 aspects for food & beverages and movie review domains, respectively (Table 1).

After defining the sentiment aspects, we trained four native speaker annotators for the final round of annotations. This results in 2423 instances for the sentiment task (Table 2).

Annotation Quality. To measure the quality of the annotations, we randomly selected 100

\(\text{Table 1: The predefined sentiment aspects (§3.2.3).}\)

<table>
<thead>
<tr>
<th>Task</th>
<th>Attribute</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Comprehension</td>
<td># of instances</td>
<td>1300</td>
</tr>
<tr>
<td></td>
<td>avg. question length (tokens)</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>avg. paragraph length (tokens)</td>
<td>94.6</td>
</tr>
<tr>
<td></td>
<td>avg. answer length (tokens)</td>
<td>7.6</td>
</tr>
<tr>
<td>Multiple-Choice QA</td>
<td># of instances</td>
<td>2460</td>
</tr>
<tr>
<td></td>
<td>% of ’literature’ questions</td>
<td>834</td>
</tr>
<tr>
<td></td>
<td>% of ’common-knowledge’ questions</td>
<td>949</td>
</tr>
<tr>
<td></td>
<td>% of ’math &amp; logic’ questions</td>
<td>677</td>
</tr>
<tr>
<td></td>
<td>avg. # of candidates</td>
<td>4.0</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td># of instances</td>
<td>2423</td>
</tr>
<tr>
<td></td>
<td>% of ’food &amp; beverages’ reviews</td>
<td>1917</td>
</tr>
<tr>
<td></td>
<td>% of ’movie’ reviews</td>
<td>506</td>
</tr>
<tr>
<td></td>
<td>avg. length of reviews (words)</td>
<td>22.01</td>
</tr>
<tr>
<td></td>
<td># of annotated pairs of (aspect, sentiment)</td>
<td>2539</td>
</tr>
<tr>
<td>Textual Entailment</td>
<td># of instances</td>
<td>2,700</td>
</tr>
<tr>
<td></td>
<td>% of ’natural’ instances</td>
<td>1,370</td>
</tr>
<tr>
<td></td>
<td>% of ’miril’ instances</td>
<td>1,330</td>
</tr>
<tr>
<td></td>
<td>avg. length of premises (tokens)</td>
<td>23.4</td>
</tr>
<tr>
<td></td>
<td>avg. length of hypotheses (tokens)</td>
<td>11.8</td>
</tr>
<tr>
<td>Question Paraphrasing</td>
<td># of instances</td>
<td>4,644</td>
</tr>
<tr>
<td></td>
<td>% of ’natural’ instances</td>
<td>2,521</td>
</tr>
<tr>
<td></td>
<td>% of ’quip’ instances</td>
<td>2,123</td>
</tr>
<tr>
<td></td>
<td>avg. length of Q1 (tokens)</td>
<td>10.7</td>
</tr>
<tr>
<td></td>
<td>avg. length of Q2 (tokens)</td>
<td>11.0</td>
</tr>
<tr>
<td>Machine Translation</td>
<td># of instances</td>
<td>47,745</td>
</tr>
<tr>
<td></td>
<td>% of ’QP’ subset</td>
<td>489</td>
</tr>
<tr>
<td></td>
<td>% of ’Quran’ subset</td>
<td>6,236</td>
</tr>
<tr>
<td></td>
<td>% of ’Bible’ subset</td>
<td>31,020</td>
</tr>
<tr>
<td></td>
<td>% of ’Mizan’ subset (eval. only)</td>
<td>10,000</td>
</tr>
</tbody>
</table>

\(\text{Table 2: Statistics on various subsets of the dataset.}\)
samples from each domain and calculated the Inter-Annotator Agreement (IAA) using Cohen’s kappa (Cohen, 1960) on annotations elicited from two independent annotators. Based on the computed IAA values, there is a substantial agreement on sub-task 1 (0.76), and moderate agreement on sub-tasks 2 and 3 (0.49 and 0.47, respectively).

Distribution of the Labels. Here we report the distribution of the labels for this task. Figure 2 shows the distribution of the document-level sentiment labels. As expected, most reviews are associated with extreme sentiments (very positive or very negative) and a relatively small portion of them are neutral. There is also a non-negligible portion of the reviews that contains mixed sentiments (partially positive and partially negative).

3.2.4 Textual Entailment

Textual entailment (Dagan et al., 2013; Bowman et al., 2015) is typically defined as a 3-way classification to determine whether a hypothesis sentence entails, contradicts, or is neutral with respect to a given premise sentence. We construct two subsets: (i) based on available natural sentences, and (ii) based on the available English entailment dataset. The former approach yields high-quality instances, but it is a relatively slower annotation task. The latter is slightly easier, but yields less interesting instances.

Based on Natural Sentences. We start with randomly sampled raw sentences, selected from 3 different resources: Miras,10 Persian Wikipedia, and VOA corpus.11 In this random sampling process, we specifically sample sentences that contain conjunctive adverbs (e.g., “[,]” [amal] meaning “but”), along with their preceding sentences. We chose such examples as there is a higher chance that these sentences naturally contain inference relationships. We ask annotators to consider both sentences and write a premise and corresponding entailling, contradicting, and neutral sentences, whichever they deem appropriate. To minimize annotation artifacts and avoid creating an artificially easy dataset, we specifically instruct annotators to avoid using simple modifications, such as simply negating a sentence or changing a word to its synonym. For the rest of the work, we refer to this set as the ‘natural’ set.

Based on Existing Datasets. In this approach, we use existing datasets in English. We start with the MNLI dataset (Williams et al., 2018) and translate them with the publicly available Google Translate API.12 Subsequently, expert annotators carefully review and fix inaccurate translations. Furthermore, each translated document is reviewed by a native-speaker annotator to correct the translational mistakes. Our annotations show that about 66.4% of the translated documents have gone through some form of correction by our annotators. For the rest of the draft, we refer to this set as ‘mnli’.

Overall, our two-pronged construction with 6 annotators results in 2.7k entailment instances (Table 2). Examples from each collected subset are included in Figure 1.

Annotation Quality. To verify the annotation quality, we quantify the agreement of 3 independent annotators, on 150 random examples. On this subset, we observe a Fleiss Kappa (Fleiss, 1971) of 0.77, indicating a substantial inter-annotator agreement (Landis and Koch, 1977).

Distribution of the Labels. As the label distribution (Figure 3) shows, the distribution of the

---

10https://github.com/miras-tech/MirasText.  
11https://jon.dehdari.org/corpora/.  
labels across the three categories are not far from uniform distribution.

### 3.2.5 Question Paraphrasing

This task is defined as determining whether two given questions are paraphrases or not. This task has been previously used to improve downstream applications like document retrieval (Zukerman and Raskutti, 2002; Callison-Burch et al., 2006; Duboue and Chu-Carroll, 2006).

Similar to the construction of the entailment task (§3.2.4), we take two different approaches: (i) based on available natural sentences, and (ii) based on an existing English question paraphrasing dataset.

**Based on Natural Sentences.** We start with questions mined using Google auto-complete (§3.2.1) as well as an additional set of questions mined from Persian discussion forums. We create pairs of questions with high token overlap. Each pair is annotated as *paraphrase* or *not-paraphrase* by native-speakers. We drop the pair if any of the questions is incomplete. For the rest of this document, we refer to this subset as ‘natural’.

**Based on Existing Datasets.** We start with the QQP dataset, which is a dataset of English question-pairs, and translate it with Google Translate API. Later, expert annotators carefully review the translations and amend any inaccuracies. We observe that about 65.6% of the translated documents have gone through some form of correction by our annotators.

Overall, the annotations involved 4 annotators and resulted in 4682 question paraphrasing instances (Table 2). Examples from each collected subset are included in Figure 1.

**Annotation Quality.** After the annotation of the earlier steps, the examples were reviewed by another annotator familiar with the task. The disagreements were labeled and adjudicated among the annotators, in order to ensure the quality of the resulting labels.

**Distribution of the Labels.** As the label distribution shows (Figure 4), the label distributions of the two splits (‘qqp’ vs ‘natural’) are not much different.

---

We experiment with several recent LMs, to assess accuracy we use the Voices dataset (Prokopidis et al., 2016). The latter (Angelidis and Lapata, 2018; Sun et al., 2019). For the third sub-task (aspect-specific sentiment) available online. baseline performance of the state-of-the-art mono-human expert performance) and also to establish P the difficulty of the future work create comparable experiments to that of ours. We compile our training set at the union of the following datasets: (i) questions obtained from the question paraphrasing task (§3.2.5, by translating the QQP instances), (ii) the training set of the Mizan dataset (Kashefi, 2018), and (iii) the TEP dataset (Pilevar et al., 2011) and Global Voices dataset (Prokopidis et al., 2016). The latter two are not included in our evaluation set because of their noisy translations to prevent any inaccurate evaluations. Note that the Quran and Bible documents are intentionally not included in the training data, in order to measure models’ generalization to unseen documents.

4 Experiments

We experiment with several recent LMs, to assess the difficulty of the ParsNLU tasks (compared to human expert performance) and also to establish baseline performance of the state-of-the-art mono-and multilingual pre-trained models. All the baseline models used in this work are available online. Evaluation Metrics. For each task, we pick a common set of existing metrics: For reading-comprehension, we use $F1$ between gold answer and the response string (Rajpurkar et al., 2016); for question paraphrasing, textual entailment, multiple-choice question-answering, and sentiment analysis, we use accuracy. For the first two sub-tasks of sentiment analysis (document-level sentiment, aspect extraction), we use macro-$F1$. For the third sub-task (aspect-specific sentiment) we use accuracy as our target evaluation metric (Angelidis and Lapata, 2018; Sun et al., 2019). For machine translation we use SacreBLEU (Post, 2018).

Task Splits. For each task, we have provided statistics on eval, train, and dev splits in Table 3. In doing so, we have ensured that enough instances are included in our evaluation sets.

<table>
<thead>
<tr>
<th>Task</th>
<th>Train</th>
<th>Dev</th>
<th>Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Comprehension</td>
<td>600</td>
<td>125</td>
<td>575</td>
</tr>
<tr>
<td>Multiple-Choice</td>
<td>1271</td>
<td>139</td>
<td>1050</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>1894</td>
<td>235</td>
<td>294</td>
</tr>
<tr>
<td>Textual Entailment</td>
<td>756</td>
<td>271</td>
<td>1,751</td>
</tr>
<tr>
<td>Question Paraphrasing</td>
<td>1,830</td>
<td>898</td>
<td>1,916</td>
</tr>
<tr>
<td>Machine Translation</td>
<td>1.6m</td>
<td>2k</td>
<td>47k</td>
</tr>
</tbody>
</table>

Table 3: Split sizes for different tasks.

Human Performance. To have an estimate of the performance and the difficulty of the challenges, we report human performance on a random subset (100-150) of instances from each task. Similar to Wang et al. (2019), we collect annotations from three human annotators, adjudicate the inconsistencies, and evaluate it against the gold labels to estimate human performance for each task.

Models. For evaluation of our baselines, we use state-of-the-art LMs. Multilingual BERT (mBERT) (Devlin et al., 2019) is pre-trained on the masked LM task over 104 languages. Additionally, we use two specialized variants of BERT for Persian: wikiBERT\(^{17}\) (trained on Persian Wiki) and ParsBERT (Farahani et al., 2020).\(^{18}\) We also use mT5 (Xue et al., 2021), which is a multilingual variant of T5 (Raffel et al., 2020).

Model Selection. We train each model with various hyperparameters and select the best one according to their development set performance. For the BERT-based models, we fine-tune them according to the cross product of the following hyperparameters: (1) Batch sizes: \{8, 16\} for small/base models and \{1, 2\} for large models; (2) Training epochs: \{3, 7\}; (3) Learning-rates: \{3 × 10^{-5}, 5 × 10^{-5}\}. For mT5 models, we fine-tune them for 20$k steps, dumping checkpoints every 1$k step. For the translation task, we trained the models for 200$k steps since the task has much larger training data. We use 10^{-3} learning-rate.

Input/Output Encoding. We formulate question paraphrasing (§3.2.5) and entailment (§3.2.4) tasks as text classification tasks.\(^{19}\) For sentiment analysis (§3.2.3), we follow formulation of Sun et al. (2019) and encode the instances as questions per aspect. The expected output is the sentiment

\(^{16}\)Included in the repository mentioned in footnote 1.

\(^{17}\)https://github.com/TurkuNLP/wikibert.

\(^{18}\)https://github.com/hooshvare/parsbert.

\(^{19}\)https://git.io/JYTNr.
polarity of the input review with respect to the input aspect-specific question. This formulation has the benefit that it is not restricted to a particular domain and its associated set of aspects, unlike alternatives such as multiclass classification.

**Experimental Setups.** First, we fine-tune our models on Persian (our dataset). The results of this setup are listed in the top segment of Table 4.

Following recent work on generalization across languages (Artetxe et al., 2020b), we evaluate English models on our Persian benchmark. We use the commonly used English datasets to supervise mT5 on each task and evaluate the resulting model on the evaluation section ofarsiNLU. The English datasets used here are as follows: SQuAD 1.1 (Rajpurkar et al., 2016) for reading comprehension (size: 88k); the union of ARC (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), and CommonsenseQA (Talmor et al., 2019) for multiple-choice question-answering (size: 18k); SNLI (Bowman et al., 2015) for textual entailment (size: 550k); QQP for question paraphrasing (size: 350k); and the Arabic-English subset of OPUS-100 (Zhang et al., 2020) for machine translation (size: 1m). We don’t do such mixing for sentiment analysis because existing English datasets are not quite compatible with our sentiment schema. The results are reported in the middle section of Table 4.

Finally, we train models on the union of Persian and English datasets. Since English datasets tend

---

**Table 4: Evaluation of Persian-only models (top), English-only (middle), and Persian+English (bottom) models on Persian tasks. Best baseline scores are indicated in bold.**
to be much larger than Persian ones, we make sure that the batches of training data, on average, contain the same number of instances from each language. Similar treatments of task mixing have also been adopted by Khashabi et al. (2020) and Raffel et al. (2020). The results of this setup are at the bottom segment of Table 4.

4.1 Results

Below are key insights from the empirical work:

**Humans Do Well on ParsiNLU.** As shown in the last row of Table 4, the human upper-bound scores are relatively high across the board. This is an indication of a reasonable degree of consensus between the ground-truth and judgments of native speakers and hence, the quality of our dataset.

**Models Haven’t Solved ParsiNLU Yet.** The majority of the models significantly lag behind human performance. This is especially true for the mid-sized (‘large’ or smaller) models that are commonly used. It is encouraging that our largest model (mT5-XL) achieves close to human performance, for certain tasks (e.g., question paraphrasing), although this model is prohibitively large and it requires a massive amount of compute. However, even these large models still struggle for most of the remaining tasks, particularly multiple-choice QA.

**English Models Successfully Transfer to Persian.** Consistent with prior observations (Artetxe et al., 2020b), multilingual models (mT5, in this case) trained with English data show a surprising degree of generalization to other languages (to Persian, in our case). Training on English data is particularly helpful for challenges that were originally translated from English datasets (such as ‘qqp’ and ‘mnli’).

**Joint Training on English and Persian Helps.** For most of the tasks, combining Persian and English yields better results than training solely on Persian or English data.

While joint training generally helps, such combinations are not guaranteed to lead to positive gains all the times. Whether the ‘Eng + Per’ models will beat either of the Persian-only or English-only models depends on whether their strengths (large size of ‘Eng’ and distributional alignment of ‘Per’) align or go against each other. Because of this issue, the combined models are not always better than the individual models.

5 Discussion

We now discuss several limitations of the current dataset and the experiments. We then outline several directions for future work.

**Beyond Current Models.** As shown in the earlier experiments, for most of the tasks the current mid-sized models perform significantly worse than humans. This is particularly pronounced for the multiple-choice QA task where there is over a 40% gap between the model and human performance, and increasing the model size (number of parameters) shows minimal benefits.

We hypothesize that the difficulty of our multiple-choice questions (and other tasks, to some extent) for the models are partly due to the reasoning and abstraction needed to answer them. For example, the ‘literature’ questions often demand creating connection several pieces of poetry, based on abstract interpretations of their meanings. Likewise, most of the ‘math & logic’ questions require several ‘hops’ of algebraic operations to get to the final answer. We hypothesize that these challenges (multi-hop reasoning over high-level abstractions of language) cannot solely be addressed with more training data and likely require a dramatic rethinking of our architectures design. For example, the poor performance on ‘math & logic’ questions might be due to models’ inability to comprehend Persian numbers and do logical reasoning with them, a topic that is briefly studied in English (Geva et al., 2020). There might also be value in exploring multitask setups across our various tasks (Zaremoodi et al., 2018), which we delegate to the future work. We hope this benchmark will encourage more of such studies, especially in the context of the Persian language.

**Coverage of Dialects.** There are other dialects of Persian, including Dari and Tajiki dialects, that are not covered by our dataset. We acknowledge this limitation and hope the future work will create broader and more inclusive collections.

6 Conclusion

This work introduced ParsiNLU, a benchmark for high-level language understanding tasks in Persian. We present a careful set of steps that we have followed to construct each of the tasks with the help of native speakers (§3.2). We have presented human scores to establish estimated upper-bounds for each task. This is followed by evaluating
state-of-the-art models on each task and quantifying the human–machine gap (§4).

To the best of our knowledge, this is the first work that publishes a language understanding benchmark for Persian language. We hope that PARSI NLU inspires more activity in the Persian NLU tasks, as well as contributing to the latest efforts in multilingual NLU.

Acknowledgments

The authors would like to thank Alireza Nourian for providing the OCR system used in the work and the anonymous reviewers for their constructive feedback. Thanks to Google’s TensorFlow Research Cloud (TFRC) for making research TPUs available.

References


Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinxian Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A chinese language understanding evaluation benchmark. In Proceedings of COLING, pages 4762–4772.


