PARSINLU: A Suite of Language Understanding Challenges for Persian

Daniel Khashabi[1] **Arman Cohan**[1] **Siamak Shakeri**[2] **Pedram Hosseini**[3] **Pouya Pezeshkpour**[4] **Malihe Alikhani**[5] **Moin Aminnaseri**[6] **Marzieh Bitaab**[7] **Faeze Brahman**[8] **Sarik Ghazarian**[9] **Mozhdeh Gheini**[9] **Arman Kabiri**[10] **Rabeeh Karimi Mahabagdi**[11] **Omid Memarrast**[12] **Ahmadreza Mosallanezhad**[7] **Erfan Noury**[13] **Shahab Raji**[14] **Mohammad Sadegh Rasooli**[15] **Sepideh Sadeghi**[2] **Erfan Sadeqi Azer**[2] **Niloofar Safi Samghabadi**[16] **Mahsa Shafaei**[17] **Saber Sheybani**[18] **Ali Tazarv**[4] **Yadollah Yaghoobzadeh**[19]

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

1 Introduction

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 PARSINLU, a collection of NLU challenges for Persian.[2] PARSINLU contains challenges for *reading comprehension*, *multiple-choice question-answering*, *textual entailment*, *sentiment analysis*, *question paraphrasing*,

[1]https://git.io/JIuRO.

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

[2]We 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.

Transactions of the Association for Computational Linguistics, vol. 9, pp. 1147–1162, 2021. https://doi.org/10.1162/tacl_a_00419
Action Editor: Mark Johnson. Submission batch: 3/21; Revision batch: 6/21; Published 10/2021.

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 $\rightleftarrows$ 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

Question Paraphrasing:

سوال ۱: کدام شهرهای ایران در وضعیت سفید کرونا هستند؟

Q1:Which cities in Iran are in white zones for corona?

سوال ۲: چه شهرهایی در وضعیت قرمز کرونا هستند؟

Q2: What cities are red zones of corona?

Answer: not a paraphrase pair

natural

سوال ۱: چگونه می‌توانم لایک‌هایم را از تمام صفحات فیس‌بوک حذف کنم؟

Q1:How can I unlike all Facebook pages?

سوال ۲: اگر من عکس شخصی را در فیس بوک لایک کنم و بلافاصله آنلایک کنم، آیا فیس بوک به ایشان اطلاع رسانی می کند که من را فاش کند؟

Q2: If I like someone's picture in Facebook and then unlike it will there be any notification in Facebook disclosing me?

Answer: not a paraphrase pair

quora

Sentiment Analysis:

نظر: طعم خوبی داره اما حتی در شگفت انگیز هم خیلی گرونه، حدودا دو برابر قیمت گوشت گرم رو داره.

بار معنایی: منفی

نشانه گذاری: (طعم: مثبت) (قیمت/ارزش خرید: خیلی منفی)

Review: it tastes good but it's so expensive even with a special offer. It's almost double the price of fresh meat.

Review sentiment score: negative (-1)

Aspect annotation: (taste: positive), (price: very negative)

product

نظر: یک فیلم با بازی‌های بد و شخصیتهای درست پرداخته نشده، و روایت نه چندان جالب... اصلا دیدنش رو توصیه نمیکنم.

بار معنایی: خیلی منفی

نشانه گذاری: (بازی/شخصیت پردازی: منفی) (داستان/روایت: منفی)

Review: a movie with poor acting and underdeveloped characters. Not an interesting plot ... I don't recommend seeing this movie at all.

Review sentiment score: very negative (-2)

Aspect annotation: (performance/acting: negative), (narrative: negative)

movie

Textual Entailment:

پیش‌فرض: پس از آن از بیم دستگیری توسط صدام، متواری شد و یارانش نیز یا از او برائت جستند و برخی مخفی شدند.

Premise: He then fled for fear of being captured by Saddam, and his allies also left him and some went into hiding.

فرضیه: از بیم دستگیری، صدام متواری شد و مخفی گشت.

Hypothesis: For the fear of being captured, Saddam fled and went into hiding..

Answer: contradicts

natural

پیش‌فرض: افراد فقیر در بیش از دو شهرستان در آتلانتا از کمک‌های حقوقی آتلانتا کمک می گیرند.

Premise: Poor people in more than two counties in Atlanta receive help from the Atlanta Legal Aid.

فرضیه: کمک‌های حقوقی آتلانتا در پنج ایالت کلان شهری آتلانتا خدمات دولتی را به افراد فقیر ارائه می‌دهد.

Hypothesis: Atlanta Legal Aid provides civil services to poor people in five metro Atlanta counties.

Answer: entails

mnli

Multiple-Choice QA:

بزرگترین قاره ی جهان کدام است؟

۱) آسیا ✓ ۲) اروپا ۳) آمریکا ۴) آفریقا

What is the largest continent in the world?

✓1) Asia 2) Europe 3) Americas 4) Africa

common knowledge

نجاری روزی یک صندلی و شاگردش در سه روز یک صندلی می‌سازد اگر نجار و شاگردش باهم کار کنند ۱۲ صندلی را چند روزه می‌سازند؟

۱) ۱۲ ۲) ✓ ۹ ۳) ۸ ۴) ۶

A carpenter makes a chair a day and his student makes a chair in three days If a carpenter and his student work together, how many days will they make 12 chairs?

1) 12 ✓2) 9 3) 8 4) 6

math&logic

بیت: «امیدوار بود آدمی به خیر کسان ** مرا به خیر تو امید نیست، شر مرسان» با کدام بیت تناسب مفهومی دارد؟

۱) گاهی از خاک درت مرهم به زخم ما بیند ** این چنین مگذار ما را یا رهاکن یا بیند

۲) مرا وصال نباید همان امید خوش است ** نه هر که رفت، رسید و نه هر که کشت، درود

✓ ۳) سنان جور بر دل ریش کم زن ** چو مرهم می‌نسازی نیش کم زن

۴) از پیش کسی کار کسی نگشاید ** امید به کردگار می‌باید داشت

The verse "A man hoped for the good of others. I do not hope for your good, just don't act like an evil." is closer to the meaning of which one?

*1) Sometimes put ointment on my wounds ** Don't stay passive; either leave or close the wound*

*2) I might not reach to her, but sure I have have the hope ** Not everyone who leaves, arrives and not everyone who plants, harvests*

*✓3) Your spear of pain keeps wounding me ** if you're not going to put any ointment on this pain, don't make it worse*

*4) No one will solve your problems ** Better maintain trust in God*

literature

Reading Comprehension:

سوال: نهاوند جزء کدام استان است؟

Question: Nahavand is part of which province?

پاراگراف: نهاوند یا (ناون) شهری در غرب ایران است. این شهر در جنوب غربی استان همدان قرار گرفته و مرکز شهرستان نهاوند است. نهاوند دارای ۷۲٬۲۱۸ نفر جمعیت است ...

Paragraph: Nahavand (Navan) is a city in western Iran. This city is located in the southern part of Hamedan province and it is the capital city of Nahavand. Nahavand has a population of 72,218, ...

پاسخ: همدان، استان همدان

Answer: Hamedan; Hamedan province

Machine Translation:

آن کسانی که به جهان غیب ایمان آرند و نماز به پا دارند و از هر چه روزیشان کردیم به فقیران انفاق کنند.

Who believe in the Unseen, are steadfast in prayer, and spend out of what We have provided for them.

quran

پیچیده در دایی سفید به نشان توبه با لباس مبدل عشق برادرانه در کارخانه‌ها و مجالس قانونگذاری راه می‌رود؛ پیشنهاد کمک می‌کند، اما طالب قدرت است.

shrouds herself in white and walks penitentially disguised as brotherly love through factories and parliaments; offers help, but desires power.

mizan

Figure 1: Examples of the ParsiNLU 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 crowdworkers 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 kası] meaning ''who'', and ''کجا'' [kojɑ] 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., ''نتیجه‌ی بازی با ژاپن؟'' [nætɪdʒe ye bɑzɪ bɑ ʒɑpon?] 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[4] for each question and checking if any of the top 10 search results overlap with a pre-defined list of credible websites.[5] 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., ''اتسان'' [otsɑn] is corrected to ''استان'' [ostɑn] ''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'' [hæmedɑn] and ''همدان'' [ostɑn e hæmedɑn], 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.[6] 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)

[4] https://github.com/MarioVilas/googlesearch.

[5] fa.wikipedia.org, bbcpersian.com, etc.

[6] https://www.sobhe.ir/alefba/.

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

Food & beverages aspects	Movie review aspects
ارزش خرید و قیمت - purchase value/price	موسیقی - music
بسته بندی و نگهداری - packaging	صدا - sound
حمل و نقل و ارسال - delivery	کارگردانی و تدوین - directing
کیفیت و تازگی محصول - product quality	داستان و فیلمنامه - story/screenplay
ارزش غذایی - nutritional value	بازی و شخصیت پردازی - acting/performance
طعم، مزه و بو - taste/smell	فیلمبرداری و دوربین - cinematography
	صحنه و جلوه های بصری - scene

Table 1: The predefined sentiment aspects (§3.2.3).

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

Table 2: Statistics on various subsets of the dataset.

[7] https://github.com/rajabzz/digikala-crawler.
[8] https://www.digikala.com/.
[9] https://www.tiwall.com/.

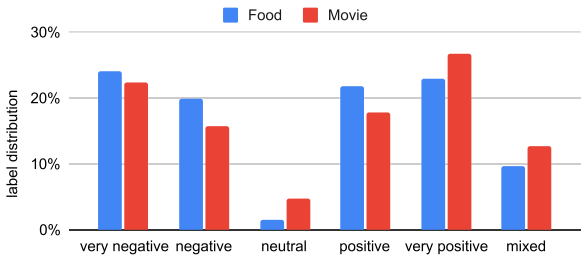

Figure 2: The distribution of the overall sentiment labels (document-level).

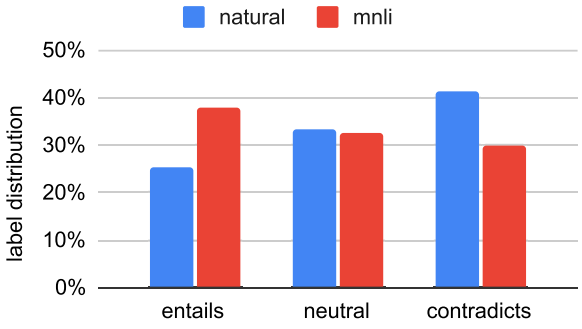

Figure 3: The distribution of the labels for the entailment task.

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, ''امّا'' [amɑ] 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 entailing, 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

[10] https://github.com/miras-tech/MirasText.
[11] https://jon.dehdari.org/corpora/.
[12] https://cloud.google.com/translate.

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 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.[13] 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,[14] 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 annotators 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.

[13] http://javabkoo.com/.

[14] https://www.kaggle.com/c/quora-question-pairs.

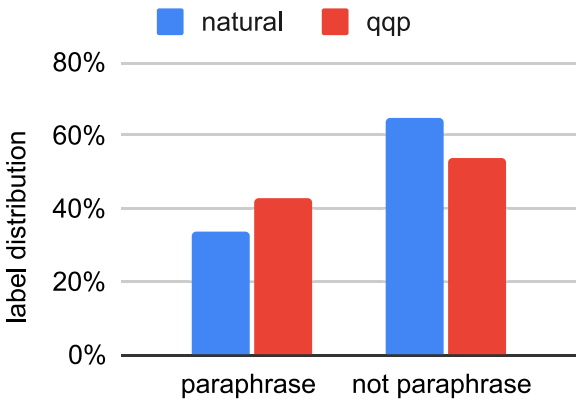

Figure 4: Label distribution for the query paraphrasing task.

3.2.6 Machine Translation

We consider the task of translating a given English sentence into Persian, and vice versa.

This task is one of the few for which several resources are available in the literature (Kashefi, 2018; Prokopidis et al., 2016; Pilevar et al., 2011). One major limitation is that there is no widely adopted comprehensive assessment of this task: Most of the works are often limited to narrow domains, and the generalization across different styles of text is rarely studied. Our contribution is to put together a collection of evaluation sets, from various domains to encourage a more holistic *evaluation* set.

Our proposed evaluation sets consist of the following: (i) *Quran*: The Quran has been translated into many languages, including English and Persian (Tiedemann and Nygaard, 2004). We use several different translations of the Quran to create high-quality evaluation sets (10 gold standard translations for each direction). Having multiple gold standards is particularly helpful for the automatic evaluation of machine translation since such metrics work best when provided with several gold standards (Gupta et al., 2019). (ii) *Bible*: Similarly, we use Persian and English versions of the Bible[15] as another evaluation set. (iii) *QQP*: We use the data obtained in the construction of question paraphrasing task (§3.2.5) to create an evaluation set for translating language questions. (iv) *Mizan*: We use the evaluation subset of the Mizan corpus (Kashefi, 2018), which is acquired based on a manual alignment of famous literary works and their published Persian translations. Overall, the combination of these four high-quality subsets yields an

[15] https://github.com/christos-c/bible-corpus.

evaluation set that contains $47k$ sentences, from 4 different domains (Table 2).

While our main contribution here is providing a more comprehensive *evaluation* of machine translation, we also provide training/dev sets to let 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 PARSINLU 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.[16]

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.

[16]Included in the repository mentioned in footnote 1.

Task	Train	Dev	Eval
Reading Comprehension	600	125	575
Multiple-Choice	1271	139	1050
Sentiment Analysis	1894	235	294
Textual Entailment	756	271	1,751
Question Paraphrasing	1,830	898	1,916
Machine Translation	$1.6m$	$2k$	$47k$

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 developement 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: $\{3 \times 10^{-5}, 5 \times 10^{-5}\}$. For mT5 models, we fine-tune them for $20k$ steps, dumping checkpoints every $1k$ step. For the translation task, we trained the models for $200k$ 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

[17]https://github.com/TurkuNLP/wikibert.
[18]https://github.com/hooshvare/parsbert.
[19]https://git.io/JYTNr.

Setup	Model ↓ - Task →	Reading Comprehension	Multiple-Choice Question Answering			Textual Entailment		Question Paraphrasing	
	Subtask →	all	literature	com-know	math & logic	natural	mnli	natural	qqp
trained on Persian	mBERT (base)	49.0	30.1.	28.7	33.8	48.7	51.6	80.4	75.3
	WikiBERT (base)	39.2	36.9	30.2	34.1	52.8	52.6	80.0	75.5
	ParsBERT (base)	40.7	33.4	28.6	32.5	51.8	53.9	79.4	72.0
	mT5 (small)	30.9	33.7	23.7	**39.1**	51.9	51.0	75.2	72.0
	mT5 (base)	42.6	34.0	24.0	36.9	57.8	59.9	79.1	75.1
	mT5 (large)	49.2	32.6	27.1	38.9	69.1	71.6	84.6	76.6
	mT5 (XL)	70.4	33.7	27.7	38.9	**77.2**	74.5	88.6	80.3
trained on English	mT5 (small)	33.0	20.9	25.7	28.9	45.1	55.6	73.5	75.1
	mT5 (base)	53.4	23.4	23.4	24.3	44.4	43.3	83.2	81.8
	mT5 (large)	67.4	27.4	33.1	25.4	46.5	54.9	88.1	86.6
	mT5 (XL)	68.2	28.3	**38.6**	22.0	66.2	77.8	**89.2**	**87.0**
trained on Per + Eng	mT5 (small)	45.3	30.9	24.9	36.6	53.3	56.2	77.9	71.3
	mT5 (base)	63.9	32.3	24.0	37.7	57.8	63.9	80.2	73.4
	mT5 (large)	73.6	30.6	28.9	38.6	70.9	72.5	85.3	78.9
	mT5 (XL)	**74.7**	**38.0**	33.7	38.0	75.5	**78.7**	88.2	80.3
	Human	86.2	80.0	85.0	85.0	87.1	90.2	92.3	88.4

Setup	Model ↓ - Task →	Sentiment (sentence sent.)		Sentiment (aspect ext.)		Sentiment (aspect sent.)		Machine Translation (Eng → Per)				Machine Translation (Per → Eng)			
	Subtask →	food	movies	food	movies	food	movies	quran	bible	qqp	mizan	quran	bible	qqp	mizan
trained on our data	mBERT (base)	55.2	48.6	87.1	73.24	53.9	34.7	–	–	–	–	–	–	–	–
	WikiBERT (base)	52.0	58.5	91.9	78.0	56.5	41.6	–	–	–	–	–	–	–	–
	ParsBERT (base)	59.1	56.8	91.1	76.8	53.9	37.6	–	–	–	–	–	–	–	–
	mT5 (small)	54.6	49.4	86.4	78.6	52.4	40.6	10.2	2.1	22.2	8.4	20.6	2.5	22.9	14.6
	mT5 (base)	56.6	52.9	88.6	80.5	52.9	46.5	11.4	2.1	27.3	9.4	22.8	2.5	34.6	14.9
	mT5 (large)	62.9	**72.5**	**92.2**	85.0	58.1	53.5	11.9	2.1	**24.8**	10.6	24.7	2.4	35.1	16.4
	mT5 (XL)	**63.1**	70.6	92.0	**85.8**	**58.9**	**54.5**	**13.5**	**2.2**	20.0	**11.0**	30.0	**2.6**	33.7	**19.3**
trained on English	mT5 (small)	–	–	–	–	–	–	–	–	–	–	6.6	1.9	7.7	3.7
	mT5 (base)	–	–	–	–	–	–	–	–	–	–	11.5	2.1	14.0	5.7
	mT5 (large)	–	–	–	–	–	–	–	–	–	–	20.2	2.3	21.0	7.4
	mT5 (XL)	–	–	–	–	–	–	–	–	–	–	25.6	2.3	30.7	9.7
trained on Per + Eng	mT5 (small)	–	–	–	–	–	–	–	–	–	–	19.2	2.5	25.6	12.1
	mT5 (base)	–	–	–	–	–	–	–	–	–	–	24.1	2.4	36.0	14.8
	mT5 (large)	–	–	–	–	–	–	–	–	–	–	29.9	2.6	36.5	18.1
	mT5 (XL)	–	–	–	–	–	–	–	–	–	–	**33.4**	2.6	**41.0**	18.2
	Human	88.4	90.3	93.1	91.6	71.0	61.6	–	–	–	–	–	–	–	–

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**.

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 of ParsiNLU. 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[20] 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

[20] See footnote 14.

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 PARSINLU 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

Hossein Amirkhani, Mohammad Azari Jafari, Azadeh Amirak, Zohreh Pourjafari, Soroush Faridan Jahromi, and Zeinab Kouhkan. 2020. Farstail: A Persian natural language inference dataset. *arXiv preprint arXiv:2009.08820*.

Stefanos Angelidis and Mirella Lapata. 2018. Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31. https://doi.org/10.1162/tacl_a_00002

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. In *Proceedings of EMNLP*, pages 7674–7684. https://doi.org/10.18653/v1/2020.emnlp-main.618

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. On the cross-lingual transferability of monolingual representations. In *Proceedings of ACL*, pages 4623–4637. https://doi.org/10.18653/v1/2020.acl-main.421

Seyed Arad Ashrafi Asli, Behnam Sabeti, Zahra Majdabadi, Preni Golazizian, Reza Fahmi, and Omid Momenzadeh. 2020. Optimizing annotation effort using active learning strategies: A sentiment analysis case study in Persian. In *Proceedings of LREC*, pages 2855–2861.

Taha Shangipour Ataei, Kamyar Darvishi, Soroush Javdan, Behrouz Minaei-Bidgoli, and Sauleh Eetemadi. 2019. Pars-absa: An aspect-based sentiment analysis dataset for Persian. *arXiv preprint arXiv:1908.01815*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of EMNLP*, pages 1533–1544.

Mahmood Bijankhan. 2004. The role of the corpus in writing a grammar: An introduction to a software. *Iranian Journal of Linguistics*, 19(2):48–67.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*. https://doi.org/10.18653/v1/D15-1075

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of NAACL*, pages 17–24. https://doi.org/10.3115/1220835.1220838

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020a. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470. https://doi.org/10.1162/tacl_a_00317

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Peter Clark, Oren Etzioni, Tushar Khot, Daniel Khashabi, Bhavana Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, Sumithra Bhakthavatsalam, Dirk Groeneveld, Michal Guerquin, and Michael Schmitz. 2020b. From 'F' to 'A' on the NY Regents Science Exams: An overview of the Aristo Project. *AI Magazine*, 41(4):39–53. https://doi.org/10.1609/aimag.v41i4.5304

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46. https://doi.org/10.1177/001316446002000104

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 EMNLP*, pages 2475–2485. https://doi.org/10.18653/v1/D18-1269

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing Textual Entailment: Models and Applications*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00509ED1V01Y201305HLT023

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Pablo Duboue and Jennifer Chu-Carroll. 2006. Answering the question you wish they had asked: The impact of paraphrasing for question answering. In *Proceedings of NAACL*, pages 33–36. https://doi.org/10.3115/1614049.1614058

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020. Parsbert: Transformer-based model for Persian language understanding. *arXiv preprint arXiv:2005.12515*.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378. https://doi.org/10.1037/h0031619

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of ACL*, pages 946–958. https://doi.org/10.18653/v1/2020.acl-main.89

Deepak Gupta, Surabhi Kumari, Asif Ekbal, and Pushpak Bhattacharyya. 2018. MMQA: A multi-domain multi-lingual question-answering framework for English and Hindi. In *Proceedings of LREC*.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey P. Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of SIGDIAL*, pages 379–391. https://doi.org/10.18653/v1/W19-5944

Pedram Hosseini, Ali Ahmadian Ramaki, Hassan Maleki, Mansoureh Anvari, and Seyed Abolghasem Mirroshandel. 2018. Sentipers: A sentiment analysis corpus for Persian. *arXiv preprint arXiv:1801.07737*.

Fatemeh HosseinzadehBendarkheili, Rezvan MohammadiBaghmolaei, and Ali Ahmadi. 2019. Product quality assessment using opinion mining in Persian online shopping. In *Proceedings of ICEE*, pages 1917–1921. IEEE.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of ICML*, pages 4411–4421. PMLR.

Akbar Karimi, Ebrahim Ansari, and Bahram Sadeghi Bigham. 2018. Extracting an English-Persian parallel corpus from comparable corpora. In *Proceedings of LREC*.

Omid Kashefi. 2018. Mizan: A large Persian-English parallel corpus. *arXiv preprint arXiv:1801.02107*.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. GLUECoS: An evaluation benchmark for code-switched NLP. In *Proceedings of ACL*. https://doi.org/10.18653/v1/2020.acl-main.329

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing format boundaries with a single QA system. In *Proceedings of EMNLP (Findings)*, pages 1896–1907. https://doi.org/10.18653/v1/2020.findings-emnlp.171

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. GooAQ: Open question

answering with diverse answer types. *arXiv preprint arXiv:2104.08727*.

Hadi Abdi Khojasteh, Ebrahim Ansari, and Mahdi Bohlouli. 2020. LSCP: Enhanced large scale colloquial Persian language understanding. In *Proceedings of LREC*, pages 6323–6327.

Maria Khvalchik and Mikhail Malkin. 2020. Departamento de nosotros: How machine translated corpora affects language models in mrc tasks. In *Proceedings of the Workshop on Hybrid Intelligence for Natural Language Processing Tasks (HI4NLP 2020) co-located with 24th European Conference on Artificial Intelligence (ECAI 2020): Santiago de Compostela, Spain, August 29, 2020*, pages 29–33. CEUR Workshop Proceedings.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466. https://doi.org/10.1162/tacl_a_00276

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 159–174. https://doi.org/10.2307/2529310

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of ACL*, pages 7315–7330. https://doi.org/10.18653/v1/2020.acl-main.653

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of EMNLP*, pages 6008–6018. https://doi.org/10.18653/v1/2020.emnlp-main.484

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1.0: Korean QA dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *arXiv preprint arXiv:2007.15207*.

Colin P. Masica. 1993. *The Indo-aryan Languages*. Cambridge University Press.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of EMNLP*. https://doi.org/10.18653/v1/D18-1260

Mahsa Mohaghegh, Abdolhossein Sarrafzadeh, and Tom Moir. 2010. Improved language modeling for English-Persian statistical machine translation. In *Proceedings of the Workshop on Syntax and Structure in Statistical Translation*, pages 75–82.

Mahsa Mohaghegh, Abdolhossein Sarrafzadeh, and Tom Moir. 2011. Improving Persian-English statistical machine translation:experiments in domain adaptation. In *Proceedings of the Workshop on South Southeast Asian Natural Language Processing (WSSANLP)*, pages 9–15.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92. https://doi.org/10.1162/tacl_a_00167

Mohammad Taher Pilevar, Heshaam Faili, and Abdol Hamid Pilevar. 2011. TEP: Tehran English-Persian parallel corpus. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 68–79.

Springer. https://doi.org/10.1007/978-3-642-19437-5_6

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of EMNLP*, pages 2362–2376.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35. https://doi.org/10.3115/v1/S14-2004

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of WMT*, pages 186–191. https://doi.org/10.18653/v1/W18-6319

Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. Parallel global voices: A collection of multilingual corpora with citizen media stories. In *Proceedings of LREC*, pages 900–905.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQUAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*. https://doi.org/10.18653/v1/D16-1264

Mohammad Sadegh Rasooli, Ahmed El Kholy, and Nizar Habash. 2013. Orthographic and morphological processing for Persian-to-English statistical machine translation. In *Proceedings of IJCNLP*, pages 1047–1051.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of EMNLP*, pages 193–203.

Mojgan Seraji, Carina Jahani, Beáta Megyesi, and Joakim Nivre. 2013. Uppsala Persian dependency treebank annotation guidelines. Technical report, Uppsala University.

Mahsa Sadat Shahshahani, Mahdi Mohseni, Azadeh Shakery, and Heshaam Faili. 2019. Payma: A tagged corpus of Persian named entities. *Signal and Data Processing*, 16(1):91–110. https://doi.org/10.29252/jsdp.16.1.91

Javad PourMostafa Roshan Sharami, Parsa Abbasi Sarabestani, and Seyed Abolghasem Mirroshandel. 2020. Deepsentipers: Novel deep learning models trained over proposed augmented Persian sentiment corpus. *arXiv preprint arXiv:2004.05328*.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. Russiansuperglue: A Russian language understanding evaluation benchmark. In *Proceedings of EMNLP*, pages 4717–4726. https://doi.org/10.18653/v1/2020.emnlp-main.381

Gary F. Simons and Charles D. Fennig. 2017. *Ethnologue: Languages of Asia*. sil International Dallas.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of NAACL*, pages 380–385.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL*, pages 4149–4158.

Tun Thura Thet, Jin-Cheon Na, and Christopher S. G. Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6):823–848.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: http://logos.uio.no/opus. In *Proceedings of LREC*. https://doi.org/10.1177/0165551510388123

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of NourIPS*, pages 3266–3280.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL*, pages 1112–1122. https://doi.org/10.18653/v1/N18-1101

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, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A chinese language understanding evaluation benchmark. In *Proceedings of COLING*, pages 4762–4772.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. MT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of EMNLP-IJCNLP*, pages 3678–3683. https://doi.org/10.18653/v1/D19-1382

Poorya Zaremoodi, Wray Buntine, and Gholamreza Haffari. 2018. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 656–661. https://doi.org/10.18653/v1/P18-2104

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of ACL*, pages 1628–1639. https://doi.org/10.18653/v1/2020.acl-main.148

Ingrid Zukerman and Bhavani Raskutti. 2002. Lexical query paraphrasing for document retrieval. In *Proceedings of COLING*. https://doi.org/10.3115/1072228.1072389