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 Mahabadi**[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, [2]Google, [3]George Washington U., [4]UC Irvine, [5]U. of Pittsburgh, [6]TaskRabbit, [7]Arizona State U., [8]UC Santa Cruz
[9]U. of Southern California, [10]IMRSV Data Labs, [11]EPFL, [12]U. of Illinois - Chicago, [13]U. of Maryland Baltimore County
[14]Rutgers U., [15]U. of Pennsylvania, [16]Expedia Group, [17]U. of Houston, [18]Indiana U. - Bloomington, [19]Microsoft

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 rich 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 high-level tasks — *Reading Comprehension*, *Textual Entailment*, etc. 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. Besides, we present the first results on state-of-the-art monolingual and multi-lingual 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 a collection of NLU tasks for Persian. This language has many attributes that make it distinct from other well-studied languages. In terms of script, Persian is similar to the 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 among the top 25 widely spoken languages (Simons and Fennig, 2017), our ability to evaluate performance and measure 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*, and *machine translation* (examples in Fig. 1). PARSINLU offers data for tasks that have never been explored before. 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

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

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

[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.

comprehensive evaluation, evaluation of less investigated genres or domains, etc.). 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 auto-complete (§3.2.1), past collegiate exams (§3.2.2), etc. 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 multi-lingual and monolingual 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 machine solvers for Persian. We will make the PARSINLU benchmark and the implementation of the models publicly available for the common good and encouraging research on Persian NLP — among other resource-scarce languages.

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, XQuAD, MLQA, TyDI and MKQA (Artetxe et al., 2019; Lewis et al., 2019; Clark et al., 2020; Longpre et al., 2020) for reading comprehension. These datasets have also been integrated as part of multi-task multilingual evaluation suites such as the 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 contain a Persian subset.

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 Russian-GLUE.[3] 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 Persian language has been limited to single task evaluation in a narrow domain or low-level tasks (e.g., datasets for POS (Bijankhan, 2004), NER (Shahshahani et al., 2019), Parsing (Seraji et al., 2013)). In contrast, we aim at providing a general 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), an *entailment* task, is a concurrent work, which is constructed semi-automatically based on multiple-choice exams. While we incorporate this dataset as one of our evaluation sets (§3.2.4), we introduce additional entailment datasets constructed differently.

The existing work on Persian *sentiment analysis* are mainly focused on *document-level* sentiment identification which ignores the nuanced judgements with respect to granular elements in the context (HosseinzadehBendarkheili et al., 2019; Sharami et al., 2020, inter alia). Moreover, the majority of such resources, such as MirasOpinion (Ashrafi Asli et al., 2020), not only lack intensity labels, but also are limited to binary or ternary sentiment classes. To overcome such limitations, we provide *aspect-level* sentiment annotations along with sentiment intensity labels (§3.2.3). The works closest to ours are Hosseini et al. (2018); Ataei et al. (2019). The data provided by Hosseini et al. (2018) comes with aspect-level labels, however, it is limited to digital products domain. Ataei et al. (2019) use a rather large set of aspects but lack any significant quality analysis. Moreover, they do not include granular sentiment labels. We address these limitations by providing high-quality annotations for aspect-based sentiment, equipped with sentiment intensity labels (§3.2.3). 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., 2017; Kashefi, 2018; Khojasteh et al., 2020). Unfortu-

[3]https://russiansuperglue.com/leaderboard/

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)

movie

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

Textual Entailment:

پیش_فرض: 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

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

Premise: Poor people in more than a couple 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:

بزگترین 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
Answer: Asia

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
Answer: candidate 2

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 ببند ** این چنین مگذار ما را یا رهاکن یا ببند
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*
Answer: candidate 3

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 souther 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 می کند، 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 for each task are included in PARSINLU. For tasks other than Machine Translation, we show the English translations for non-Persian readers. The purple tags indicate the type of the instance, according to their construction (explained in the corresponding subsection under Section 3.2).

nately, most published work in this domain perform evaluations on limited datasets and domains.

Our contribution to this task is compiling a set of high-quality evaluation sets from a broad range of domains. 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. Additionally, we contribute to the existing resources by introducing new training and evaluation sets for MT.

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.

Quality, over quantity. Contrary to the recent trend that emphasized on dataset sizes (Bowman et al., 2015; Rajpurkar et al., 2016), we opt for the *quality* of the dataset over its *size*. The field has seen an increasing number of approaches that rely *less* on the supervised annotated data and *more* on unsupervised pre-training (Radford et al., 2019; Brown et al., 2020). We expect this trend to only continue and models become much more efficient with respect to the amount of labeled data.

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., 2019; 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., 2020). It is important for any NLP dataset to reflect the *natural* distribution of the target language tokens and its 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 Galkin, 2020).

Experts, over crowdworkers. While crowdsourcing has been the common approach for building datasets, we choose to work with few native Persian speakers to construct the dataset. Crowdworkers are difficult to train and often generate more noisy annotations. However, expert annotators that 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 access restrictions to crowdsourcing platforms. A study done by Pavlick et al. (2014, Table 6) rates Persian (among few other languages) with almost no crowd-workers.

3.2 Constructing PARSINLU tasks

We provide separate explanations for the construction of each task. Examples are shown in Fig. 1.

3.2.1 Reading Comprehension

We use the commonly used definition of reading-comprehension task: generating an answer, given a *question* and a context *paragraph*.

SQuAD (Rajpurkar et al., 2016) is one of 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), etc. For constructing our reading comprehension tasks, we avoid using SQuAD as a source and employ a process resembling that of Kwiatkowski et al. (2019) that would lead to more natural questions.

Collecting questions. Our efforts to translate questions from 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 sport 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 follow the approach of Berant et al. (2013) which relies on a query auto-completion API for collecting questions. Similarly, we use Google's auto-completion[4] which enables us to mine a rich, yet 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 kasɪ] meaning "who", "کجا" [kojɑ] meaning "where", etc.) 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. Such questions extracted from the auto-complete 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 in 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 "The results 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[5] for each question and checking if any of the top 10 search results overlap with a pre-defined list of credible websites.[6] 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 answer each of the questions. At the first step, the annotators read the question and correct any grammatical errors and typos (e.g., "اتسان" is corrected to "استان" [ostɑn] "state"). Next, they annotate *the shortest coherent span* 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. The paragraph that contains this answer is also annotated as the context of the question.

Overall, 6 native-speaker annotators annotated a collection of 575 question-answer-paragraph triplets (Table 2).[7]

Annotation quality. To verify the quality of the annotations, one more annotator went over the existing annotations in the evaluation set. The disagreements were resolved in further adjudication.

3.2.2 Multiple-Choice QA

Multiple-choice questions are widely adopted for teaching purposes and, within NLP, are one of the common formats for evaluation of fact-retrieval and reasoning (Clark et al., 2019; Talmor et al., 2019). 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 supporting documents from an external source, or have stored the necessary knowledge internally to be able to answer such QAs.

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 understanding poetry and their implied meaning, knowledge of Persian grammar, and the history of literature. (ii) Employment exams that are expected to assess 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 the above sources are scanned copies of the original exams in image format. We use an existing Persian OCR tool to convert the image data to textual format.[8] Then 4 annotators fix any mistakes made by the OCR system and convert the result into a structured format. Overall, this yields *1970* 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 Fig. 1.

Annotation quality. To further examine the quality of the annotations, we randomly sampled

[4]http://google.com/complete/search?client=chrome&q=...
[5]https://github.com/MarioVilas/googlesearch
[6]fa.wikipedia.org, bbcpersian.com, etc.
[7]We will release a bigger set in our revision.
[8]https://www.sobhe.ir/alefba/

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. This shows high quality of the annotated data.

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, such as a review (Liu, 2012). Applications of SA include tasks such as market prediction, product review assessment, gauging public opinion about socio-political matters, etc.

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 ..."* (Fig. 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 the past work). For the *food & beverages* category, we extracted[9] reviews from the online grocery section of Digikala,[10] and for the *movie* reviews category, we crawled reviews from Tiwall,[11]. Both of these websites are well-known and popular websites among Persian speakers.

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

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

Defining aspects. Following 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 since 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. 2423 instances for the sentiment task (Table 2) were annotated accordingly.

Annotation quality. To measure the quality of the annotations, we randomly selected 100 samples from each domain and calculated the Inter-Annotator Agreement (IAA) using Cohen's kappa (Cohen, 1960) on annotations elicited from two independent annotators. We computed IAA for three sub-tasks including: (1) overall sentiment, (2) aspect annotation, and (3) (aspect, sentiment) pair annotation. Details of IAA are reported in Appendix B. Overall, there is a *substantial* agreement on sub-task 1, and *moderate* agreement on sub-tasks 2 and 3.

[9]https://github.com/rajabzz/digikala-crawler

[10]https://www.digikala.com/

[11]https://www.tiwall.com/

Task	Attribute	Statistic
Reading Comprehension	# of instances	575
	avg. question length (tokens)	6.3
	avg. paragraph length (tokens)	94.6
	avg. answer length (tokens)	7.6
Multiple-Choice QA	# of instances	1970
	% of 'literature' questions	517
	% of 'common-knowledge' questions	949
	% of 'math & logic' questions	504
	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.

3.2.4 Textual Entailment

Textual Entailment (TE; Dagan et al., 2013) and its newer variant, Natural Language Inference (NLI; Bowman et al., 2015), are typically defined as a 3-way classification task where the goal is 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 available English query-paraphrasing dataset. The former approach yields high quality instances, however, 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,[12] Persian Wikipedia and VOA corpus.[13] In this random sampling process, we specifically sample sentences that contain conjunctive adverbs (e.g, "امّا" [amɑ] meaning "but", etc.), 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 a 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 MNLI dataset (Williams et al., 2018) and translate them with the publicly available Google Translate API.[14] Afterwards, 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 Fig. 1.

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

3.2.5 Question Paraphrasing

For a given pair of natural-language questions, one must determine whether they are paraphrases or not. Paraphrasing has a broad range of applications (Callison-Burch et al., 2006; Kriz et al., 2018) and, in particular, query-paraphrasing can be used to improve document retrieval (Zukerman and Raskutti, 2002; 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 available English question paraphrasing dataset.

Based on natural sentences. We start with questions mined using Google auto-complete

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

Task	Train	Dev	Eval
Reading Comprehension	-	-	575
Multiple-Choice	830	90	1050
Sentiment Analysis	1894	235	294
Textual Entailment	756	271	1751
Question Paraphrasing	1830	898	1916
Machine Translation	-	-	47,745

Table 3: Split sizes for different tasks.

(§3.2.1) as well as an additional set of questions mined from Persian discussion forums.[15] We create pairs of questions with high token overlap. Each pair is annotated by native-speaker annotator as *paraphrase* or *not-paraphrase*. We drop the pair if any of the questions is incomplete. For the rest of this document, we call this subset 'natural' set.

Based on existing datasets. In this approach, we use existing datasets in English. We start with QQP dataset[16] and translate them with Google Translate API. Afterwards, expert annotators carefully re-annotate the result of the translations to fix any inaccuracies. Our annotations show 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 Fig. 1.

Annotation quality. After the annotation is done in the earlier step, they are reviewed by another annotator. Disagreements are adjudicated to ensure the quality of the samples.

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. 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 followings: (i) *Quran* which has been translated into many languages, including English and Persian (Tiedemann and Nygaard, 2004). We use several English/Persian translations of Quran by different translators to create high-quality evaluation sets, with 10 gold standard translations for each direction. Having multiple gold standards is particularly helpful for automatic evaluation of machine-translation since such metrics work best when provided with several gold standards (Gupta et al., 2019). (ii) similarly, we use Persian and English versions of *Bible*[17] as another evaluation set. (iii) *QP*: using 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 evaluation set that contains 47,745 sentences, from 4 different domains (Table 2.) For an automatic translation system to work well on our evaluation set, it needs to generalize to a variety of unseen distribution.

While our main contribution here is providing a more comprehensive *evaluation* of machine-translation, we also provide training and 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) translation instances obtained from the training subset of our question paraphrasing task (§3.2.5), (ii) the training set of *Mizan* dataset (Kashefi, 2018), (iii) TEP dataset (Pilevar et al., 2011) and Global Voices dataset (Prokopidis et al., 2016). The latter two are not included in our evaluation set due to their noisy translations, which could lead to inaccurate evaluations. Needless to say, participants in this task are encouraged to use any resources at their disposal to train stronger models, as long as the used resources do not overlap with our proposed evaluation data.

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

[15] http://javabkoo.com/

[16] https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

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

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 our data	mBERT (base)	-	31.1	28.6	27.6	71.8	60.5	80.4	75.3
	WikiBERT (base)	-	34.0	31.4	32.1	76.2	62.8	80.0	75.5
	ParsBERT (base)	-	**35.4**	29.5	27.1	76.7	62.2	79.4	72.0
	mT5 (small)	-	30.3	24.9	32.0	51.9	51.0	75.2	72.0
	mT5 (base)	-	29.7	27.2	**36.3**	57.8	59.9	79.1	75.1
	mT5 (large)	-	31.7	28.7	34.6	69.1	71.6	84.6	76.6
	mT5 (XL)	-	30.0	27.0	34.3	**77.2**	74.5	88.6	80.3
fine-tuned on English	training data	SQuAD ($88k$)	ARC+OpenBookQA+ComQA ($18k$)			SNLI ($550k$)		QQP ($350k$)	
	mT5 (small)	28.6	17.7	28.9	30.2	45.1	55.6	73.5	75.1
	mT5 (base)	43.0	16.8	29.2	30	44.4	43.3	83.2	81.8
	mT5 (large)	60.1	23.7	**34.3**	29.4	46.5	54.9	88.1	86.6
	mT5 (XL)	**65.5**	23.7	32.4	27.7	66.2	**77.8**	**89.2**	**87.0**
	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 (EN → FA)				Machine Translation (FA → EN)			
	Subtask →	food	movies	food	movies	food	movies	quran	bible	qp	mizan	quran	bible	qp	mizan
trained on our data	mBERT (base)	55.15	48.64	87.08	73.24	53.92	34.65	-	-	-	-	-	-	-	-
	WikiBERT (base)	51.98	58.48	91.88	77.98	56.54	41.58	-	-	-	-	-	-	-	-
	ParsBERT (base)	59.07	56.84	91.06	76.82	53.92	37.62	-	-	-	-	-	-	-	-
	mT5 (small)	54.6	49.4	86.4	78.6	52.4	40.6	6.7	1.9	11.1	6.1	14.5	2.3	13.9	9.7
	mT5 (base)	56.6	52.9	88.6	80.5	52.9	46.5	8.7	2.0	16.5	7.3	17.3	2.3	24.9	10.9
	mT5 (large)	**62.9**	**72.5**	**92.2**	85.0	**58.1**	53.5	10.0	2.1	15.8	8.9	20.4	2.5	28.5	**12.8**
	mT5 (XL)	-	70.6	-	**85.8**	-	**54.5**	**13.1**	**2.2**	**20.7**	**9.3**	**28.2**	**2.6**	29.2	8.7
fine-tuned on English	training data	-						-				OPUS-100 ($1m$)			
	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
	Human	88.4	90.3	93.1	91.6	71.0	61.6	-	-	-	-	-	-	-	-

Table 4: Results of evaluation on PARSINLU tasks by fine-tuning models on Persian dataset (§4.1) and English datasets (§4.2). Best baseline scores are indicated as **bold**.

mono- and multi-lingual pre-trained models.

All the baseline models used in this work are available online.[18]

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); we use *accuracy* for question paraphrasing, textual-entailment, multiple-choice question-answering and sentiment analysis. For the first two sub-tasks of sentiment analysis (sentence-level sentiment, aspect extraction), we use macro-F1. For the third sub-task (aspect-specific sentiment) we use accuracy as our target evaluation metric. For machine-translation we use *Sacre-BLEU* (Post, 2018).

Task splits. For each task, we have provided eval, train, and dev splits in Table 3. In doing so, we have ensured that enough instances are included in our evaluation sets. For the future revisions of our work, we are planning to add more instances to our train and dev splits.[19]

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. Inspired by the the setup used by 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. In evaluation of our baselines, we use state-of-the-art LMs. Multilingual BERT (mBERT) (Devlin et al., 2019) is pretrained on the masked LM task over 104 languages. Additionally, we use two specialized variants of BERT for Persian: wikiBERT[20] which is trained on Persian Wikipedia and ParsBERT (Farahani et al., 2020).[21] Finally, we use mT5 (Xue et al., 2020), which is a multilingual variant of the T5 architec-

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

[19]For the current release, the reading comprehension comes with evaluation set only.

[20]https://github.com/TurkuNLP/wikibert

[21]https://github.com/hooshvare/parsbert

ture.[22]

Input/output encoding. We formulate question paraphrasing (§3.2.5) and entailment (§3.2.4) tasks as text classification tasks.[23]

For the ABSA task (§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 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 multi-class classification.

4.1 Experiment: task-specific fine-tuning

For each task (except reading comprehension; cf., footnote 19) we fine-tune models on each task's training set and perform model selection on their corresponding held-out development sets (details in Appendix A). The results of fine-tuning experiment are shown in Table 4.

Humans do well on PARSINLU. As it can be seen in the last row of Table 4, our human upper-bound scores are relatively high across the board. This indicates that, to a reasonable degree, there is a consensus between the ground-truth labels and judgments of average native speakers.

Models struggle on PARSINLU. The majority of the models significantly lag behind human performance. This point is especially true for the mid-sized models that are commonly used, where the margin ranges from 13.2% (in query-paraphrasing) to 54% (in question-answering).

Not surprisingly, larger models perform better across all the tasks at the cost of massive size. It is encouraging that our largest model (mT5-XL) achieves close to human performance, for certain tasks (e.g., question paraphrasing), however, this model is prohibitively large and it requires a massive amount of compute. The mT5-XL model is especially close to human performance on question paraphrasing task, likely because of the inherent simplicity of this task (relatively short inputs, binary output). However, even these large models still struggle for most of the remaining tasks, particularly multiple-choice QA.

[22]Our experiments does not contain mT5-XXL since it was not released at the time of this publication.

[23]https://github.com/huggingface/transformers/tree/master/examples/text-classification

4.2 Experiment: zero-shot evaluation

One recently emerged experimental setup deals with the generalization across languages (e.g., evaluating a system trained on English data or another language) (Artetxe et al., 2019).

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, the union of ARC (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018) and CommonsenseQA (Talmor et al., 2019) for multiple-choice question-answering, SNLI (Bowman et al., 2015) for textual entailment, QQP[24] for question paraphrasing, and Arabic-English subset of OPUS-100 (Zhang et al., 2020) for machine translation. The results are reported in the bottom-half of Table 4.

English models successfully transfer to Persian. Consistent with the prior observations made in the field (Artetxe et al., 2019), multilingual models (mT5, in this case) trained with English data show a surprising degree of generalization to other languages (to Persian, in our case).

One potential caveat of cross-lingual transfer models is that they generally demand more supervision than training with target language instances. In other words, if collecting labeled data for a target language is not a bottleneck, supervising models with language-specific data remains a more effective approach.

5 Discussion

We discuss several limitations about the current dataset, experiments and outline several directions for future work.

Beyond current models. As shown before, the current mid-sized models perform significantly worse than humans. Even the largest models remain quite weak on tasks like multiple-choice QA.

Our conjecture is that more training data will lead to slight improvements in performance, however, it is not clear that it necessarily leads to better generalization to, for example, other variants of question-answering in Persian.

It might be possible to address the limitations of the current models with better designs. For exam-

[24]See footnote 16.

ple, the poor performance on 'math & logic' questions might be due to model's inability to comprehend Persian numbers and do logical reasoning with them, a topic that is briefly studied in English (Geva et al., 2020). We hope this will encourage more of such works, especially in the context of Persian language.

Coverage of dialects. In addition to Iranian Persian, there are other dialects of Persian spoken by millions of people, including Afghani and Tajiki dialects. 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 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-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.

Acknowledgement

The authors would like to thank Alireza Nourian for providing the OCR system used in the work. 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*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. *arXiv preprint arXiv:2004.04721*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

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 The 12th Language Resources and Evaluation Conference*, pages 2855–2861, Marseille, France. European Language Resources Association.

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 the 2013 conference on empirical methods in natural language processing*, 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 *EMNLP*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa:

A benchmark for information-seeking question answering in typologically diverse languages. *arXiv preprint arXiv:2003.05002*.

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, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, et al. 2019. From'f'to'a'on the ny regents science exams: An overview of the aristo project. *arXiv preprint arXiv:1909.01958*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

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

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

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 the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 33–36.

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.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. *arXiv preprint arXiv:2004.04487*.

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 the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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 the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391.

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 *2019 27th Iranian Conference on Electrical Engineering (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 generalization. *arXiv preprint arXiv:2003.11080*.

Akbar Karimi, Ebrahim Ansari, and Bahram Sadeghi Bigham. 2017. Extracting an english-Persian parallel corpus from comparable corpora. *arXiv preprint arXiv:1711.00681*.

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 the 58th Annual Meeting of the Association for Computational Linguistics*.

Hadi Abdi Khojasteh, Ebrahim Ansari, and Mahdi Bohlouli. 2020. LSCP: Enhanced large scale colloquial Persian language understanding. *arXiv preprint arXiv:2003.06499*.

Maria Khvalchik and Mikhail Galkin. 2020. El departamento de nosotros: How machine translated corpora affects language models in mrc tasks. *arXiv preprint arXiv:2007.01955*.

Reno Kriz, Eleni Miltsakaki, Marianna Apidianaki, and Chris Callison-Burch. 2018. Simplification using paraphrases and context-based lexical substitution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 207–217.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. MLQA: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*.

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.

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 *EMNLP*.

Mahsa Mohaghegh, Abdolhossein Sarrafzadeh, and Tom Moir. 2010. Improved language modeling for English-Persian statistical machine translation. In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*, pages 75–82, Beijing, China. Coling 2010 Organizing Committee.

Mahsa Mohaghegh, Abdolhossein Sarrafzadeh, and Tom Moir. 2011. Improving Persian-English statistical machine translation:experiments in domain adaptation. In *Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP)*, pages 9–15, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

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.

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.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

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 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. Parallel global voices: a collection of multilingual corpora with citizen media stories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 900–905.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.

Mohammad Sadegh Rasooli, Ahmed El Kholy, and Nizar Habash. 2013. Orthographic and morphological processing for Persian-to-English statistical machine translation. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1047–1051, Nagoya, Japan. Asian Federation of Natural Language Processing.

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

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*.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Tun Thura Thet, Jin-Cheon Na, and Christopher SG 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 the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference.

In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Liang Xu, Xuanwei Zhang, Lu Li, Hai Hu, Chenjie Cao, Weitang Liu, Junyi Li, Yudong Li, Kai Sun, Yechen Xu, et al. 2020. CLUE: A Chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3678–3683.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Ingrid Zukerman and Bhavani Raskutti. 2002. Lexical query paraphrasing for document retrieval. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

A Model selection

For the BERT-based models, we fine-tune them according to the cross product of the following hyper-parameters:

- Batch sizes: $\{8, 16\}$ for small/base models and $\{1, 2\}$ for large models.
- Training epochs: $\{3, 7\}$
- Learning-rates: $\{3 \times 10^{-5}, 5 \times 10^{-5}\}$

For text-to-text architectures (T5 and BART), we fine-tune them for $20k$ steps, dumping checkpoints every $1k$ step. We use 10^{-3} learning-rate.

B ABSA Inter-Annotator Agreement

To measure the quality of the annotation, we randomly selected 100 samples from each domain and calculated the Inter-Annotator Agreement (IAA) for pairs of annotators using Cohen's kappa (Cohen, 1960).

The results of IAA for our sentiment analysis tasks (§3.2.3) are shown in Table 5. We report agreements for three sub-tasks: (1) overall sentiment, (2) aspect annotation, and (3) aspect-sentiment annotation. Agreements for sub-task 1 is based on the percentage of the instances that receive the same label by our annotators. Similarly, for sub-task 2, the agreement is based on the percentage of the cases that annotators have annotated the same aspects, from the set of pre-defined aspects for each domain (Table 1). For sub-task 3, we check whether there is agreement based on *both* the aspect and its associated sentiment.

Sub-task #	Task name	Food	Movie
1	Overall sentiment	0.72	0.81
2	Aspect annotation	0.49	0.49
3	Aspect-Sentiment annotation	0.48	0.47

Table 5: Inter-Annotator Agreement for different sentiment analysis annotation tasks.

Overall, there is a *substantial* agreement on sub-task 1, *moderate* agreement on sub-tasks 2 and 3, which indicate a reasonable level of quality for our data. Note that lower agreement scores for sub-tasks 2 and 3 is, at least partly, due to the challenging nature of aspect and aspect-specific sentiment annotations.