Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1603–1612
Marseille, 20-25 June 2022

BasqueGLUE: A Natural Language Understanding Benchmark for Basque

Gorka Urbizu[1], Iñaki San Vicente[1], Xabier Saralegi[1], Rodrigo Agerri[2], Aitor Soroa[2]

[1]Elhuyar Foundation
[2]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{g.urbizu, i.sanvicente, x.saralegi}@elhuyar.eus
{rodrigo.agerri, a.soroa}@ehu.eus

Abstract

Natural Language Understanding (NLU) technology has improved significantly over the last few years and multitask benchmarks such as GLUE are key to evaluate this improvement in a robust and general way. These benchmarks take into account a wide and diverse set of NLU tasks that require some form of language understanding, beyond the detection of superficial, textual clues. However, they are costly to develop and language-dependent, and therefore they are only available for a small number of languages. In this paper, we present BasqueGLUE, the first NLU benchmark for Basque, a less-resourced language, which has been elaborated from previously existing datasets and following similar criteria to those used for the construction of GLUE and SuperGLUE. We also report the evaluation of two state-of-the-art language models for Basque on BasqueGLUE, thus providing a strong baseline to compare upon. BasqueGLUE is freely available under an open license.

Keywords: Neural Language Models, Natural Language Understanding, Less-Resourced Languages

1. Introduction

The Transformer architecture (Vaswani et al., 2017) has led to a family of powerful neural language models (LM) that can scale up to billions of parameters, and which are pre-trained using enormous quantities of text (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019). These models have achieved remarkable results in many Natural Language Understanding (NLU) tasks by following the transfer learning approach, where the LMs are first pre-trained over vast amounts of unannotated text, and then fine-tuned to a certain downstream task using a comparatively much smaller amount of annotated data.

Experimentation and evaluation with neural language models in NLU have been greatly facilitated by the existence of evaluation frameworks such as GLUE (Wang et al., 2018) or SuperGLUE (Wang et al., 2019), that assess the capacity of the models to understand text beyond the detection of superficial clues. Those benchmarks provide a unified and robust evaluation framework for evaluating NLU systems that is not exclusive to a single task, genre, or dataset. Initially, most NLU assessment frameworks were focused on English, but gradually such resources are being published for other languages (Xu et al., 2020; Le et al., 2020; Wilie et al., 2020; Kakwani et al., 2020; Shavrina et al., 2020; Park et al., 2021).

This paper presents BasqueGLUE[1], the first NLU evaluation framework for Basque. We believe that BasqueGLUE is a significant contribution towards developing NLU tools in Basque, which we believe will facilitate the technological advance for the Basque language. In order to create BasqueGLUE we took as a reference the GLUE and SuperGLUE frameworks. When possible, we re-used existing datasets for Basque, adapting them to the corresponding task formats if necessary. Additionally, BasqueGLUE also includes six new datasets that have not been published before. In total, BasqueGLUE consists of nine Basque NLU tasks and covers a wide range of tasks with different difficulties across several domains. As with the original GLUE benchmark, the training data for the tasks vary in size, which allows to measure the performance of how the models transfer the knowledge across tasks.

In addition to the evaluation framework, we also provide an evaluation of the existing state of the art transformer models for Basque, namely BERTeus (Agerri et al., 2020), and ElhBERTeu[2], a new BERT model trained with more data and which has been developed specifically for this paper.

In the following section we describe related work both on the use of neural language models to address NLU tasks and on the construction of NLU benchmarks. In Section 3 we explain the construction process of BasqueGLUE. In Section 4 we report the results obtained on BasqueGLUE by two monolingual Basque language models, one of them pre-trained for this work. We finish with some concluding remarks and future work.

[1]https://github.com/Elhuyar/BasqueGLUE

[2]https://huggingface.co/elh-eus/ElhBERTeu

2. Related Work

The marked improvement provided by neural language models in the field of Natural Language Processing (NLP) in the recent years has shown the weakness and limitations of many datasets. Moreover, for optimal performance, NLU technology should be able to robustly process language independently of specific tasks and datasets (Wang et al., 2018).

In order to evaluate these new Transformer-based language models, benchmarks for general NLU like GLUE (Wang et al., 2018) were created by collecting diverse tasks with varying degrees of difficulty and training data, and putting them together using a unified format.

Still, recent models have surpassed human performance on those benchmarks within a year of their release, even though it is known that we are yet far from solving NLU (Kiela et al., 2021). In response to this, there have been several efforts to create harder tasks, with popular examples those of SuperGLUE (Wang et al., 2019), GEM (Gehrmann et al., 2021) or BIG-Bench (BIG-bench, 2021).

All these benchmarks mentioned above are in English, and therefore they cannot be used to evaluate NLU in other languages. Although there are some cross-lingual benchmarks such as XGLUE (Liang et al., 2020) or XTREME (Hu et al., 2020), training datasets released for them are in English only.

Taking these issues into consideration, several monolingual benchmarks for other major languages are flourishing: CLUE for Chinese (Xu et al., 2020), FLUE for French (Le et al., 2020), a Russian variant of SuperGLUE (Shavrina et al., 2020), ALUE for Arabic (Seelawi et al., 2021), or the Korean KLUE (Park et al., 2021). Only few of those datasets are focused on less-resourced languages, namely, an Indonesian variant (Wilie et al., 2020), IndicGLUE for Indic languages (Kakwani et al., 2020), or CLUB for Catalan (Rodriguez-Penagos et al., 2021).

2.1. Language Models

Since the publication of BERT (Devlin et al., 2019), a Masked Language Model (MLM) based on the Transformer architecture (Vaswani et al., 2017), most of the NLP tasks have shifted towards the approach of fine-tuning big pre-trained language models on task-specific labelled data.

Since then, many new variants of MLM have emerged. Among those encoder-based MLMs, we can find an optimized version of BERT called RoBERTa (Liu et al., 2019), models like Electra (Clark et al., 2020) and SpanBERT (Joshi et al., 2020), which change the masking method or Albert (Lan et al., 2019), which decreases the size of the model by employing shared weights among the layers.

Meanwhile, other models for language generation like GPT2 (Radford et al., 2019) and GPT3 (Brown et al., 2020), based on auto-regressive decoders, and models for sequence-to-sequence tasks based on a MLM approach over the original transformer encoder-decoder architecture, like BART (Lewis et al., 2020), T5 (Raffel et al., 2020) and byT5 (Xue et al., 2021a) have also shown good performance on NLU tasks.

Witnessing the fast-paced progress in NLP for English and a few other high resource languages, multilingual models like mBERT (Devlin et al., 2019), mBART (Liu et al., 2020) or mT5 (Xue et al., 2021b), have made possible to use these pre-trained LMs for cross-lingual transfer-learning in low-resource languages, where the availability of manually labelled training data is perennially scarce.

The first Transformer based language model able to process Basque text was multilingual BERT (mBERT) (Devlin et al., 2019), a BERT model trained using the largest 104 Wikipedias. Agerri et al. (2020) showed the limitations of this model and introduced BERTeus, the first monolingual model for Basque, achieving state-of-the-art results in four tasks; ixambert, a multilingual BERT model released by Otegi et al. (2020), which was pre-trained on Basque, Spanish and English, obtains better results than mBERT for the Chatbot related intent classification task (López de Lacalle et al., 2021). Although a few other models have been released [3], to the best of our knowledge, no proper evaluation has been published.

3. Design of BasqueGLUE

BasqueGLUE follows the design principles of the English GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks. BasqueGLUE is built around nine Basque NLU tasks, which cover a wide range of dataset sizes as well as various difficulties across several domains. In BasqueGLUE, performance is evaluated by a single automatic metric. The datasets, including the test sets, are publicly available, along with the code to run the unified evaluation.

The tasks were selected among the datasets already available for Basque NLU, aiming to comply, whenever possible, with the following criteria used to build SuperGLUE (Wang et al., 2019):

- Task substance: tasks should test a system's ability to understand and reason about texts in Basque.
- Task difficulty: tasks should be beyond the scope of current state-of-the-art systems, but solvable by most college-educated Basque native speakers.
- Evaluability: tasks must have an automatic performance metric that correlates well to human judgments of output quality.
- Public data: in SuperGLUE, tasks are required to have existing public training data in order to minimize the risks involved in newly-created

[3] https://huggingface.co/models?language=eu&sort=downloads

Corpus	Train	Dev	Test	Task	Metric	Domain
NERC$_{id}$	51,539	12,936	35,855	NERC	F1	News
NERC$_{ood}$	64,475	14,945	14,462			News, Wikipedia
FMTODeu$_{intent}$	3,418	1,904	1,087	Intent classification	F1	Dialog system
FMTODeu$_{slot}$	19,652	10,791	5,633	Slot filling	F1	Dialog system
BHTCv2	8,585	1,857	1,854	Topic classification	F1	News
BEC2016eu	6,078	1,302	1,302	Sentiment analysis	F1	Twitter
VaxxStance	864	206	312	Stance detection	MF1*	Twitter
QNLI$_{eu}$	1,764	230	238	QA/NLI	Acc	Wikipedia
WiC$_{eu}$	408,559	600	1,400	WSD	Acc	Wordnet
EpecKorrefBin	986	320	587	Coreference resolution	Acc	News

Table 1: The 9 tasks included in BasqueGLUE. NERC$_{id}$ stands for NERC in-domain, while NERC$_{ood}$ stands for NERC out-of-domain. Acc refers to accuracy, while F1 refers to micro-average F1-score. The metric used for VaxxStance is the macro-average F1-score of two classes: FAVOR and AGAINST.

datasets. Nevertheless, in this new benchmark for Basque, we decided to include some new datasets, which have been created recently for other purposes, or have been built from previously annotated and well-known datasets, due to the relatively low number of annotated resources available for Basque. We made this decision with the aim of including more diverse tasks to be able to evaluate better the NLU capabilities of the models.

- Task format: we opted for tasks that have relatively simple input and output formats to avoid pushing researchers into creating complex task-specific model architectures.

- License: task data must be available under licences that allow use and redistribution for research purposes.

3.1. Selected Tasks

In this section we describe the tasks included in BasqueGLUE and their corresponding datasets. Table 1 presents a summary of them, including their sizes, the metrics used for evaluation and the type of data sources. Table 2 introduces an illustrative example for each of the tasks in BasqueGLUE, with its corresponding English translation.

3.1.1. Named Entity Recognition

The first task we have included in BasqueGLUE is the well-known NLP task of Named Entity Recognition and Classification (NERC), a sequence labelling task.

The NERC dataset is divided into two subtasks: in-domain NERC and out-of-domain NERC. As for the performance metric, we use the average of the F1 values from both subtasks.

For the in-domain NERC subtask (NERC$_{id}$), EIEC (Alegria et al., 2004) was the previous standard dataset for Basque NERC, used also to evaluate BERTeus (Agerri et al., 2020). EIEC is composed of texts from news sources and it is annotated following the BIO annotation scheme over four categories: person, organization, location, and miscellaneous. Aiming for a larger dataset, we merged EIEC with a new NERC dataset containing texts from the Basque newspaper Naiz[4], which was annotated with the same guidelines used in EIEC. For this benchmark, we merged both datasets and created train, development and test sets, so examples from both sources are present in all splits. In the out of domain NERC subtask (NERC$_{ood}$), the training set comprises data from the news domain, whereas the test set contains data from Wikipedia articles. Specifically, the training set is obtained by joining the training and development sets of the in-domain datasets. For the out-of-domain evaluation, a new dataset has been created from Wikipedia, which was annotated together with the data from Naiz, following the same guidelines. The dataset from Wikipedia is split into development and test sets for the NERC$_{ood}$ BasqueGLUE task.

3.1.2. Intent Classification (FMTODeu$_{intent}$)

The next task included in the benchmark is intent classification, an NLU task in the field of dialogue systems that aims to identify the intent that the user denotes in a sentence, among a set of predefined classes. Thus, it is approached as a multi-class sequence classification task. The dataset we selected for the task is the Facebook Multilingual Task Oriented Dataset for Basque or FMTODeu (López de Lacalle et al., 2020). The examples are annotated with one of 12 different intent classes corresponding to *alarm*, *reminder* or *weather* related actions. We will name the dataset FMTODeu$_{intent}$ in this benchmark, in order to differentiate it from the slot filling task also included in the dataset (see section 3.1.3). We maintain the original train/dev/test partitions. Micro F1-score is used for evaluation.

3.1.3. Slot Filling (FMTODeu$_{slot}$)

The third task is slot filling, that comes also from the field of dialogue systems and it is usually performed

[4] https://naiz.eus

in conjunction with the intent classification task. The objective is to identify entities associated with intents expressed in sentences formulated by the user. FMTODeu (López de Lacalle et al., 2020) includes this entity annotations, and thus it is straightforward to use. The task is a sequence labelling task similar to NERC, following BIO annotation scheme over 11 categories. We will name the dataset FMTODeu$_{slot}$, and same as for intent classification, we maintain the original train/dev/test partition. Micro F1-score is used as performance metric.

3.1.4. Topic Classification (BHTCv2)

Topic classification is another multi-class sequence classification task. The dataset we provide here is based on the BHTC dataset (Agerri et al., 2020). It contains news headlines (brief article descriptions) from the Basque weekly newspaper Argia[5]. News are classified according to twelve thematic categories. This dataset was previously used in the evaluation of BERTeus (Agerri et al., 2020). While the examples in original BHTC were lower-cased and had punctuation their removed, we decided to keep the original texts for BasqueGLUE. We also removed several duplicate examples from train and examples that were not in Basque. From now on, we will call this dataset BHTCv2. Same as for intent classification, performance is measured using micro F1 score.

3.1.5. Sentiment Analysis (BEC)

Sentiment Analysis is a well known task present in most NLU benchmarks. The aim is to correctly classify the polarity of the given texts, among positive, neutral or negative classes in our case.

The Basque Election Campaign 2016 Opinion Dataset (BEC2016eu) is a new dataset for the task of sentiment analysis, a sequence classification task, which contains tweets about the campaign for the Basque elections from 2016. The crawling was carried out during the election campaign period (2016/09/09-2016/09/23), by monitoring the main parties and their respective candidates. The tweets were manually annotated as positive, negative or neutral. The metric we propose for this task and dataset is micro F1 score.

3.1.6. Stance Detection (VaxxStance)

Stance Detection (SD) is one of the tasks within the universe of Fake News detection, also approached as a sequence classification task. The aim of it is to detect stance in social media on a very controversial and trendy topic. The task is to determine whether a given tweet expresses an AGAINST, FAVOR or NEUTRAL stance towards the topic.

The VaxxStance (Agerri et al., 2021) dataset is included in BasqueGLUE. It deals with tweets regarding the antivaxxers movement. The dataset did not include a development set, and thus we split the original training data creating new training and development sets as a result. The split was done randomly, in an effort to provide development data and make the scores obtained in the new benchmark more consistent, for a fair comparison among the models being evaluated. Following the original VaxxStance track and the literature of SD shared tasks we measure the performance of the systems by means of macro-average F1 score (MF1) of two classes: FAVOR and AGAINST (although all three classes, including NEUTRAL, need to be included for training and prediction).

3.1.7. QNLI

In order to include a QA task in the benchmark, we adapted the QA dataset ElkarHizketak (Otegi et al., 2020), a low resource conversational Question Answering (QA) dataset for Basque created by native speaker volunteers. The dataset is built on top of Wikipedia sections about popular people and organizations, and it contains around 400 dialogues and 1600 question and answer pairs.

We adapted this dataset into a sentence-pair binary classification tasks, following the design of QNLI for English (Wang et al., 2019). We form a pair with each question and each sentence in the corresponding context. Afterwards, we filter out the negative samples with the lowest lexical overlap between the question and the sentence, until we are left with a balanced dataset. As evaluation metric, we follow the English QNLI design and use accuracy.

3.1.8. WiC

Word in Context or WiC (Pilehvar and Camacho-Collados, 2019) is a word sense disambiguation (WSD) task included in the English SuperGLUE benchmark. It is designed as a particular form of sentence pair binary classification. Given two text snippets and a polysemous word that appears in both of them (the span of the word is marked in both snippets), the task is to determine whether the word has the same sense in both sentences. Performance is evaluated using accuracy.

We generated a dataset taking EPEC-EuSemcor (Pociello et al., 2011) corpus as a starting point. EPEC-EuSemcor is a sense-tagged corpus for Basque. This corpus comprises a set of occurrences of nouns which have been annotated with Basque WordNet v1.6 senses (Pociello et al., 2011). It contains 42,615 occurrences of nouns manually annotated, corresponding to the 407 most frequent Basque nouns. We only made use of the occurrences from context sentences of 10-50 words in length.

The Basque WiC dataset follows the design of the English WiC dataset (Pilehvar and Camacho-Collados, 2019), pairing context sentences for each noun to create the instances of the classification task.

We adopted the same strategy as in the English dataset to boost the clarity of the dataset and removed all pairs whose senses were first degree connections in the WordNet semantic graph (including sister senses) and those which belong to the same supersense. For

[5] https://www.argia.eus

NERC

Tokens: Helburuetako bat McLareni eta Ferrariri aurre egitea izango du taldeak .
Labels: O O B-ORG O B-ORG O O O O O O
One of the objectives that will have the team is to confront McLaren and Ferrari.

Intent Classification (FMTODeu$_{intent}$)

Text: Alarma ezarri gaurko 6:00etan
Set the alarm today at 6:00am
Intent: `alarm/set_alarm`

Slot Filling (FMTODeu$_{slot}$)

Tokens: Euria egingo du gaur ?
Labels: `B-weather/attribute` O O `B-datetime` O
Is it going to rain today?

Topic Classification (BHTCv2)

Text: Gurasotasun baimena eta seme-alabak zaintzeko baimena lau hilabetera luzatzeko proposamena egitea onartu du Europako Batzordeak. Proposamenak aldaketa handia ekarriko luke Hego Euskal Herrian, lau asteetara luzatu berri baita baimen hori.
The European Commission has approved to make the proposal of extending paternity leave to four months. The proposal would represent an important change in Hego Euskal Herria, as it has been extended recently to four weeks.
Topic: `Gizartea`
society

Sentiment Analysis (BEC)

Text: Mezu txoro, patetiko eta lotsagarri hori ongi hartuko duenik badela uste du PSEk.
PSE thinks there are people who will respond positively to that crazy, pathetic and shameful message.
Polarity: `Negative`

Stance Detection (VaxxStance)

Text: Gure nagusiak babestuko dituen txertoa martxan da. Zor genien. Gaur mundua apur bat hobeagoa da. #OsasunPublikoarenGaraipena #GureGaraipena
The vaccine that will protect our elderly people is on it's way. We owned them. Today the world is a little bit better. #TheVictoryOfPublicHealthcare #OurVictory
Stance: `FAVOR`

QNLI

Question: "Irrintziaren oihartzunak" dokumentalaz gain, zein best lan egin ditu zinema arloan?
Aside from the documentary "Irrintziaren ohiartzunak", in what other projects has she worked on in the field of cinema?
Sentence: "Irrintziaren oihartzunak" du lehen filma zuzendari eta gidoilari gisa.
"Irrintziaren oihartzunak" is her first film as a director and scriptwriter.
NLI: `not_entailment`

WiC

Sentence1: Asterix, zazpi egunen segida asmatu zuen galiarra .
Asterix, the Gaul who invented the 7 days week.
Sentence2: Etxeko landareek sasoi aktiboan tenperatura epelak behar dituzte : egunez 25 C ingurukoak .
House plants need warm temperatures during active season: around 25C in daylight .
Same_sense: `False`

Coreference (EpecKorrefBin)

Text: Birmoldaketan daudenen artean Katalunia , Madril , Hego Euskal Herria , Aragoi , Balear irlak eta Errioxa aurkitzen dira . Horien artean , Hego Euskal Herriak 47.870 milioi pezeta jasoko ditu .
Among those under reorganization are Catalonia, Madrid, Southern Basque Country, Aragon, Balearic islands and Rioja . Among them, the Southern Basque Country will receive 47,870 million pesetas.
Coreference: `True`

Table 2: Examples for each task in BasqueGLUE. Underlined text is specially marked in the datasets. Text in monospaced font represents the expected model output. English translation of the input text in Basque is provided in italics.

this pruning we used a mapping from Wordnet v1.6 to Wordnet v3.0, as it is customary for other datasets related to WSD (Raganato et al., 2017).
We enforced the constraint of not having repeated contextual sentences across the test and development sets. Thus, we set apart 1,400 and 600 instances for the test and development sets, respectively.
The remaining pairs whose context sentences did not overlap with the test and development sets formed our training data (where we allow the repetition of the context sentences, to increase the training data size). Lastly, we ensured that all the splits are balanced for positive and negative examples.

3.1.9. Coreference Resolution (EpecKorrefBin)

The last task we selected for our BasqueGLUE benchmark is coreference resolution, for which there is already available a dataset for Basque, EPEC-KORREF (Soraluze et al., 2012). However, coreference resolution involves clustering mentions into entities, without any easy format to approach the whole task. Thus, we decided to convert it into a binary classification problem, adopting the same format used for the Winograd Schema Challenge (WSC) task (Levesque et al., 2012). In this new task, the model has to predict whether two mentions from a text, which can be pronouns, nouns or noun phrases, are referring to the same entity. We adapted EPEC-KORREF to this task, and we renamed this dataset as EpecKorrefBin. We limited mention pairs to those in the same sentence or consecutive sentences.
To create negative examples, we selected mention pairs that included one pronoun or have the same mention type (e.g. both being proper nouns for locations). Then, in an attempt to make the task more challenging, we filtered out the most similar mention pairs among positive examples and also those which were most different from the negative ones. As string similarity measures, we used Levenshtein distance for positive examples and token set ratio for negative ones. Finally, we ensured that all the splits are balanced for positive and negative examples.

4. Evaluation

4.1. Baselines

We have compared two models for implementing the baselines. One of the models is BERTeus (Agerri et al., 2020), a BERT model for Basque trained over 224M token texts from news sources and Wikipedia. The second model is another BERT model we have trained with a bigger monolingual corpus collected recently, which we will call ElhBERTeu[6].
To train ElhBERTeu, we increased the BMC corpus (Agerri et al., 2020) used for BERTeus. Specifically, data from 2020 and 2021 from the same sources in BMC was added, as well as novel news sources and texts from other domains, such as science (both academic and divulgative), literature or subtitles. More details about the corpora used and their sizes are shown in Table 3. Texts from news sources were oversampled (duplicated) as done during the training of BERTeus. In total 575M tokens were used for pre-training ElhBERTeu.

Domain	Size
News	2 * 224M
Wikipedia	40M
Science	58M
Literature	24M
Others	7M
Total	575M

Table 3: Corpora used to pre-train ElhBERTeu. Size in tokens. Text from news sources has been oversampled (duplicated).

ElhBERTeu was trained following the design decisions for BERTeus (Agerri et al., 2020). The tokenizer and the hyper-parameter settings remained the same, with the only difference being that the full pre-training of the model (1M steps) was performed with a sequence length of 512 on a v3-8 TPU.
For implementing the baselines we fine-tune both models separately for each task, setting the initial learning rate at 3e-5 and using a maximum of 10 epochs to choose the best performing model over the development set. For all the tasks except WiC and coreference, we use the transformers library (Wolf et al., 2020) and fine-tuned following the standard procedures for sequence labelling and text classification tasks. We concatenate the sentences with a [SEP] token in the input for sentence-pair format tasks.
Regarding WiC and coreference tasks, we fine-tune and evaluate using the jiants toolkit (Phang et al., 2020), which was employed for SuperGLUE. For WiC, we concatenate the representation of the marked word to the [CLS] representation. For coreference resolution, we approached the task as a span-based task, and adopt the implementation that jiants provides for the WSC task.

4.2. Results

Table 4 shows the results for both models. In the case of NERC, we only show the average of the F1 scores for NERC$_{id}$ and NERC$_{ood}$[7], as mentioned in section 3.1.1. This is done to avoid a task having more importance than the rest in the final average score.
If we look at the tasks individually, we obtain mixed results. BERTeus scores higher for coreference while ElhBERTeu scores stand out in F$_{slot}$, WiC and spe-

[6]The model will be publicly available. Reference will be added to the final version of the paper

[7]Specifically, BERTeus scores 81.92 F1 on NERC, average of 86.40 F1 on NERC$_{id}$ and 77.44 F1 on NERC$_{ood}$. ElhBERTeu scores 82.30 F1 on NERC, average of 86.81 F1 on NERC$_{id}$ and 77.78 F1 on NERC$_{ood}$.

Model	AVG	NERC F1	F_{intent} F1	F_{slot} F1	BHTC F1	BEC F1	Vaxx MF1	QNLI acc	WiC acc	coref acc
BERTeus	73.23	81.92	82.52	74.34	78.26	69.43	59.30	74.26	70.71	68.31
ElhBERTeu	73.71	82.30	82.24	75.64	78.05	69.89	63.81	73.84	71.71	65.93

Table 4: Results of the models we evaluate on the BasqueGLUE language understanding benchmark for Basque. The model BERTeus is the BERT model presented by Agerri et al. (2020). The score obtained at NERC is the average of the in domain and out of domain subtasks. F_{intent} and F_{slot} refer to FMTODeu$_{intent}$ and FMTODeu$_{slot}$ respectively. BHTC refers to BHTCv2, BEC is BEC2016eu, Vaxx refers to VaxxStance, and coref refers to EpecKorrefBin.

cially VaxxStance. For the rest of the datasets performances are very similar, with BERTeus coming first in FMOTD$_{intent}$, BHTCv2 and QNLI, and ElhBERTeu wining in NERC and BEC. Overall, scoring a half-point higher on the average of the benchmark, we can conclude that ElhBERTeu is better at NLU tasks according to BasqueGLUE.

As for the adequacy of the benchmark, results show that there is a clear room for improvement in all the tasks, proving the datasets in BasqueGLUE are indeed challenging for the current state-of-the-art language models. GLUE was published with an average-score of 70 for the best baseline model, while SuperGLUE reported an average-score of 74.6 for the best baseline model, similar to the average score of 73.71 obtained by ElhBERTeu in BasqueGLUE.

5. Conclusions

This paper introduces BasqueGLUE, the first Natural Language Understanding Benchmark for Basque, which will be useful to evaluate the NLU ability of large pre-trained language models in a robust and general way. The benchmark contains a wide and diverse set of NLU tasks that require some form of language understanding. The datasets for the benchmark has been elaborated from previously existing datasets, adapting some of them to simpler task formats, following similar criteria to those used for the construction of GLUE and SuperGLUE.

We also present the evaluation, on this new benchmark, of two BERT language models trained for Basque, which we are finally able to compare exhaustively at NLU, and conclude that the best model among these two, according to the results in BasqueGLUE, is ElhBERTeu.

BasqueGLUE is freely available and released under an open license, including the scripts for conducting a unified evaluation. ElhBERTeu is also available at Huggingface.

6. Acknowledgements

This work has been partially funded by the Basque Government (DeepText, (Elkartek grant no. KK-2020/00088)) and the IARPA BETTER Program contract No. 2019-19051600006 (ODNI, IARPA). We also acknowledge the support of Googles's TFRC program. Rodrigo Agerri is funded by the RYC-2017–23647 fellowship and by the ANTIDOTE - EU CHIST-ERA project (PCI2020-120717-2) funded by the Agencia Estatal de Investigación through the INT-Acciones de Programación Conjunta Internacional (MINECO) 2020 call.

7. Bibliographical References

Agerri, R., San Vicente, I., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., and Agirre, E. (2020). Give your text representation models some love: the case for basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Aremu, A., Bosselut, A., Chandu, K. R., Clinciu, M.-A., Das, D., Dhole, K., Du, W., Durmus, E., Dušek, O., Emezue, C. C., Gangal, V., Garbacea, C., Hashimoto, T., Hou, Y., Jernite, Y., Jhamtani, H., Ji, Y., Jolly, S., Kale, M., Kumar, D., Ladhak, F., Madaan, A., Maddela, M., Mahajan, K., Mahamood, S., Majumder, B. P., Martins, P. H., McMillan-Major, A., Mille, S., van Miltenburg, E., Nadeem, M., Narayan, S., Nikolaev, V., Niyongabo Rubungo, A., Osei, S., Parikh, A., Perez-Beltrachini, L., Rao, N. R., Raunak, V., Rodriguez, J. D., Santhanam, S., Sedoc, J., Sellam, T., Shaikh, S., Shimorina, A., Sobrevilla Cabezudo, M. A., Strobelt, H., Subramani, N., Xu, W., Yang, D., Yerukola, A., and Zhou, J. (2021). The GEM benchmark: Natural language generation, its evaluation and metrics.

In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online, August. Association for Computational Linguistics.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. M., and Kumar, P. (2020). IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online, November. Association for Computational Linguistics.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., et al. (2021). Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May. European Language Resources Association.

Levesque, H., Davis, E., and Morgenstern, L. (2012). The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Liang, Y., Duan, N., Gong, Y., Wu, N., Guo, F., Qi, W., Gong, M., Shou, L., Jiang, D., Cao, G., Fan, X., Zhang, R., Agrawal, R., Cui, E., Wei, S., Bharti, T., Qiao, Y., Chen, J.-H., Wu, W., Liu, S., Yang, F., Campos, D., Majumder, R., and Zhou, M. (2020). Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, abs/2004.01401.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

López de Lacalle, M., Saralegi, X., and López, I. (2021). Reducing annotation effort for cross-lingual transfer learning: The case of nlu for basque. In *Proceedings of The Workshop on Mixed-Initiative ConveRsatiOnal Systems (MICROS) at ECIR 2021*.

Otegi, A., Agirre, A., Campos, J. A., Soroa, A., and Agirre, E. (2020). Conversational question answering in low resource scenarios: A dataset and case study for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 436–442.

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J. Y., Park, J., Song, C., Kim, J., Song, Y., Oh, T., Lee, J., Oh, J., Lyu, S., Jeong, Y., Lee, I., Seo, S., Lee, D., Kim, H., Lee, M., Jang, S., Do, S., Kim, S., Lim, K., Lee, J., Park, K., Shin, J., Kim, S., Park, L., Oh, A., Ha, J.-W., and Cho, K. (2021). KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Pilehvar, M. T. and Camacho-Collados, J. (2019). Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Raganato, A., Camacho-Collados, J., and Navigli, R. (2017). Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110. Association for Computational Linguistics.

Rodriguez-Penagos, C., Armentano-Oller, C., Villegas, M., Melero, M., Gonzalez, A., Bonet, O. d. G., and Pio, C. C. (2021). The catalan language club. *arXiv preprint arXiv:2112.01894*.

Seelawi, H., Tuffaha, I., Gzawi, M., Farhan, W., Talafha, B., Badawi, R., Sober, Z., Al-Dweik, O., Freihat, A. A., and Al-Natsheh, H. (2021). ALUE: Arabic language understanding evaluation. In *Proceed-*

ings of the Sixth Arabic Natural Language Processing Workshop, pages 173–184, Kyiv, Ukraine (Virtual), April. Association for Computational Linguistics.

Shavrina, T., Fenogenova, A., Anton, E., Shevelev, D., Artemova, E., Malykh, V., Mikhailov, V., Tikhonova, M., Chertok, A., and Evlampiev, A. (2020). RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online, November. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *EMNLP 2018*, page 353.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., and Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China, December. Association for Computational Linguistics.

Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., Xu, Y., Sun, K., Yu, D., Yu, C., Tian, Y., Dong, Q., Liu, W., Shi, B., Cui, Y., Li, J., Zeng, J., Wang, R., Xie, W., Li, Y., Patterson, Y., Tian, Z., Zhang, Y., Zhou, H., Liu, S., Zhao, Z., Zhao, Q., Yue, C., Zhang, X., Yang, Z., Richardson, K., and Lan, Z. (2020). CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., and Raffel, C. (2021a). Byt5: Towards a token-free future with pre-trained byte-to-byte models. *arXiv preprint arXiv:2105.13626*.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021b). 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.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv preprint arXiv:1906.08237*.

8. Language Resource References

Agerri, R., San Vicente, I., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., and Agirre, E. (2020). Give your text representation models some love: the case for basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4781–4788.

Agerri, R., Centeno, R., Espinosa, M., de Landa, J. F., and Rodrigo, A. (2021). Vaxxstance@ iberlef 2021: Overview of the task on going beyond text in cross-lingual stance detection. *Procesamiento del Lenguaje Natural*, 67:173–181.

Alegria, I., Arregi, O., Balza, I., Ezeiza, N., Fernandez, I., and Urizar, R. (2004). Design and development of a named entity recognizer for an agglutinative language. In *First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition.*

BIG-bench. (2021). Beyond the imitation game benchmark (big-bench). `https://github.com/google/BIG-bench`.

López de Lacalle, M., Saralegi, X., and San Vicente, I. (2020). Building a task-oriented dialog system for languages with no training data: the case for basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2796–2802.

Otegi, A., Agirre, A., Campos, J. A., Soroa, A., and Agirre, E. (2020). Conversational question answering in low resource scenarios: A dataset and case study for basque. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 436–442.

Phang, J., Yeres, P., Swanson, J., Liu, H., Tenney, I. F., Htut, P. M., Vania, C., Wang, A., and Bowman, S. R. (2020). `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Pociello, E., Agirre, E., and Aldezabal, I. (2011). Methodology and construction of the basque wordnet. *Language Resources and Evaluation. Springer. Volume 45, Issue 2, pp 121-142. ISSN 1574-020X. DOI 10.1007/s10579-010-9131-y. official*.

Soraluze, A., Arregi, O., Arregi, X., Ceberio, K., and De Ilarraza, A. D. (2012). Mention detection: First steps in the development of a basque coreference resolution system. In *KONVENS*, pages 128–136.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma,

C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.