BasqueGLUE: A Natural Language Understanding Benchmark for Basque

Gorka Urbizu¹, Iñaki San Vicente¹, Xabier Saralegi¹, Rodrigo Agerri² and Aitor Soroa²

¹Elhuyar Foundation
²HiTZ Center - Ixa, University of the Basque Country UPV/EHU

g.urbizu@elhuyar.eus
BasqueGLUE:

1. Introduction
2. Design
   2.1. Methodology
   2.2. Tasks
   2.3. Summary
3. Evaluation
   3.1. Models
   3.2. Results
4. Conclusions
Introduction
INTRODUCTION

Motivation

- Transformer $\rightarrow$ Scalable Language Models (LM)
- Remarkable results on NLU tasks, via transfer learning:
  - Pretrain on huge unannotated text corpus
  - Finetune on downstream tasks using small annotated datasets
INTRODUCTION

Motivation

- Transformer $\rightarrow$ Big Language Models (LM)
- Remarkable results on NLU tasks, via transfer learning:
  - Pretrain on huge unannotated text corpus
  - Finetune on downstream tasks using small annotated datasets
- Benchmarks such as GLUE are key to evaluate this improvement
- However, they are only available for a small number of languages:
  - Costly to develop
  - Language-dependent
Main contributions

- BasqueGLUE: the first NLU Benchmark for Basque
  - A less-resourced language
  - Spoken by less than 1M people

- Evaluation of two LMs for Basque on BasqueGLUE, providing a strong baseline

- BasqueGLUE is freely available at https://github.com/Elhuyar/BasqueGLUE
BasqueGLUE follows the design of GLUE and SuperGLUE (Wang et al., 2018; 2019). The tasks were selected following SuperGLUE’s criteria: Task substance, Difficulty, Evaluability, Public Data, Task format & License.

BasqueGLUE is built around 9 Basque NLU tasks:
- several domains
- various difficulties
- a wide range of dataset sizes

Performance is evaluated by a single automatic metric.

The datasets are publicly available.
## Tasks

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Task</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>NERC$_{id}$</td>
<td>NERC</td>
<td>News</td>
</tr>
<tr>
<td>NERC$_{ood}$</td>
<td></td>
<td>News, Wikipedia</td>
</tr>
<tr>
<td>FMTODEu$_{intent}$</td>
<td>Intent classification</td>
<td>Dialog system</td>
</tr>
<tr>
<td>FMTODEu$_{slot}$</td>
<td>Slot filling</td>
<td>Dialog system</td>
</tr>
<tr>
<td>BHTCv2</td>
<td>Topic classification</td>
<td>News</td>
</tr>
<tr>
<td>BEC2016eu</td>
<td>Sentiment analysis</td>
<td>Twitter</td>
</tr>
<tr>
<td>VaxxStance</td>
<td>Stance detection</td>
<td>Twitter</td>
</tr>
<tr>
<td>QNLI$_{eu}$</td>
<td>QA/NLI</td>
<td>Wikipedia</td>
</tr>
<tr>
<td>WiC$_{eu}$</td>
<td>WSD</td>
<td>Wordnet</td>
</tr>
<tr>
<td>EpecKorrefBin</td>
<td>Coreference resolution</td>
<td>News</td>
</tr>
</tbody>
</table>
Sequence labeling task

2 subtasks:
- In-domain (News -> News)
  - EIEC (Alegria et al., 2004) + Naiz
- Out-of-domain (News -> Wiki)
  - EIEC + Naiz -> Wikipedia

Metric: AVG of the F1 of both subtasks

**NERC (NERC\textsubscript{id}, NERC\textsubscript{ood})**

<table>
<thead>
<tr>
<th>Tokens:</th>
<th>Helburuetako</th>
<th>bat</th>
<th>McLareni</th>
<th>eta</th>
<th>Ferrariri</th>
<th>aurre</th>
<th>egitea</th>
<th>izango</th>
<th>du</th>
<th>taldeak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labels:</td>
<td>O</td>
<td>O</td>
<td>B-ORG</td>
<td>O</td>
<td>B-ORG</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>

One of the objectives that will have the team is to confront McLaren and Ferrari.
Intent Classification ($\text{FMTODeu}_{\text{intent}}$)

- A NLU task in the field of dialogue systems that aims to identify the intent of users
- A multi-class sequence classification task

- Facebook Multilingual Task Oriented Dataset for Basque (FMTODeu):
  - (López de Lacalle et al., 2020)
  - 12 different intent classes: alarm, reminder or weather related actions.

- Metric: Micro F1

<table>
<thead>
<tr>
<th>Intent Classification (FMTODeu_{intent})</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text:</strong> Alarma ezarri gaurko 6:00etan</td>
</tr>
<tr>
<td><em>Set the alarm today at 6:00am</em></td>
</tr>
<tr>
<td><strong>Intent:</strong> alarm/set_alarm</td>
</tr>
</tbody>
</table>
Slot filling (FMTODEu_{slot})

- Another task from dialogue systems, to identify entities associated with user’s intents
- Sequence labeling task

- FMTODEu dataset:
  - BIO annotation over 11 categories

- Metric: Micro F1

<table>
<thead>
<tr>
<th>Tokens:</th>
<th>Euria</th>
<th>egingo</th>
<th>du</th>
<th>gaur</th>
<th>?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labels:</td>
<td>B-weather/attribute</td>
<td>0</td>
<td>0</td>
<td>B-datetime</td>
<td>0</td>
</tr>
</tbody>
</table>

Is it going to rain today?
Topic classification (BHTCv2)

- A multi-class sequence classification task
- The dataset is based on the BHTC (Agerri et al., 2020)
  - Contains news headlines from the Argia Basque magazine
  - 12 categories
- Metric: Micro F1

---

**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, lauasteetara 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**
DESIGN

Sentiment analysis (BEC2016eu)

- Sentiment Analysis is a common text classification task in NLU benchmarks

- The Basque Election Campaign 2016 Opinion Dataset (BEC2016eu):
  - Contains tweets about the campaign for the Basque elections from 2016
  - Classify the polarity of a text (Positive/Neutral/Negative)

- Metric: Micro F1

---

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)

- Stance Detection is a sequence classification task from Fake News detection
- To detect stance in social media on hot topics.
- VaxxStance dataset (Agerri et al., 2021):
  - Tweets regarding the antivaxxers movement
  - expresses a stance towards the topic (AGAINST, FAVOR or NEUTRAL)
- Metric: Macro F1 of two classes: FAVOR and AGAINST
**DESIGN**

**Question Answering (QNLI<sub>eu</sub>)**

- A sentence-pair binary classification task as done for QNLI for English (Wang et al., 2019)
- Adapted from the ElkarHizketak QA dataset (Otegi et al., 2020)
- We form a pair with each question and each sentence in the corresponding context
- The task: whether the sentence contains the answer of the question or not
- Metric: Accuracy

---

**QNLI**

**Question:** “Irrintziaren oihartzunak” dokumentalaz gain, zein best lan egin ditu zinema arloan?

*Aside from the documentary “Irrintziaren ohiartxunak”, 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
DESIGN

**Word in Context (WiC\textsubscript{eu})**

- WiC is a word sense disambiguation task
- Given two words in context the task is to determine whether they have the same sense
- Generated a new dataset from EPEC-EuSemcor dataset (Pociello et al., 2011)
- Binary span classification task, following the design of WiC for English
- Metric: Accuracy

---

**WiC**

**Sentence 1:** Asterix, zazpi \underline{egunen} segida asmatu zuen galiarra.

* Asterix, the Gaul who invented the \underline{7 days week}.

**Sentence 2:** Etxeko landareek sasoi aktiboan temperatura epelak behar dituzte: \underline{egunez 25 C ingurukoak}.

* House plants need warm temperatures during active season: around \underline{25C in daylight}.

**Same sense:** False
Coreference Resolution (EpecKorrefBin)

- We simplified this clustering problem into a binary span classification task
- Given a text and two mentions in it, whether they refer to the same entity or not
- Adapted EPEC-KORREF dataset (Soraluze et al., 2012) into WSC format

- Metric: Accuracy

---

Coreference (EpecKorrefBin)


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
## Summary

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

Table 1: The 9 tasks included in BasqueGLUE. NERC<sub>id</sub> stands for NERC in-domain, while NERC<sub>ood</sub> 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.
EVALUATION

Models

- Berteus (Agerri et al. 2020)
- ElhBERTeu:
  - A BERT base model.
  - Similar to Berteus
  - 50% bigger and more diverse training corpus
  - [https://huggingface.co/elh-eus/ElhBERTeu](https://huggingface.co/elh-eus/ElhBERTeu)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>2 * 224M</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>40M</td>
</tr>
<tr>
<td>Science</td>
<td>58M</td>
</tr>
<tr>
<td>Literature</td>
<td>24M</td>
</tr>
<tr>
<td>Others</td>
<td>7M</td>
</tr>
<tr>
<td>Total</td>
<td>575M</td>
</tr>
</tbody>
</table>

Corpora used to pre-train ElhBERTeu.
EVALUATION

Results

- We fine-tune both models separately for each task
- lr = 3e-5 and maximum of 10 epochs, and 5 runs
- Results on test split using best performing model over the development set
EVALUATION

Results

- We fine-tune both models separately for each task
- \( lr = 3e-5 \) and maximum of 10 epochs, and 5 runs
- Results on test for the best performing model over the development set

- Transformers\(^1\) library for sequence labelling and text classification tasks
- Jiants\(^2\) toolkit for span classification tasks (WiC and Coreference)

\(^1\)https://github.com/huggingface/transformers
\(^2\)https://github.com/nyu-mll/jiant
EVALUATION

Results

- We fine-tune both models separately for each task
- lr = 3e-5 and maximum of 10 epochs
- Results on test for the best performing model over the development set

- Transformers library for sequence labelling and text classification tasks
- Jiants toolkit for span classification tasks (WiC and Coreference)

<table>
<thead>
<tr>
<th>Model</th>
<th>AVG</th>
<th>NERC F1</th>
<th>F_{intent} F1</th>
<th>F_{slot} F1</th>
<th>BHTC F1</th>
<th>BEC F1</th>
<th>Vaxx MF1</th>
<th>QNLI acc</th>
<th>WiC acc</th>
<th>coref acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERTeus</td>
<td>73.23</td>
<td>81.92</td>
<td>82.52</td>
<td>74.34</td>
<td>78.26</td>
<td>69.43</td>
<td>59.30</td>
<td>74.26</td>
<td>70.71</td>
<td>68.31</td>
</tr>
<tr>
<td>ElhBERTeus</td>
<td>73.71</td>
<td>82.30</td>
<td>82.24</td>
<td>75.64</td>
<td>78.05</td>
<td>69.89</td>
<td>63.81</td>
<td>73.84</td>
<td>71.71</td>
<td>65.93</td>
</tr>
</tbody>
</table>
Conclusions
Conclusions

We introduced BasqueGLUE:

- First NLU benchmark for Basque
- Useful to evaluate large LMs in a robust and general way
- 9 diverse NLU tasks that require language understanding

BasqueGLUE is freely available at https://github.com/Elhuyar/BasqueGLUE
CONCLUSIONS

Conclusions

We introduced BasqueGLUE:

- First NLU benchmark for Basque
- Useful to evaluate large LMs in a robust and general way
- 9 diverse NLU tasks that require language understanding

Evaluation:

- Finally able to compare models exhaustively at NLU in Basque
- We compared 2 models: Berteus & ElhBERTeu
- Conclude that the best model is ElhBERTeu

BasqueGLUE is freely available at https://github.com/Elhuyar/BasqueGLUE
References


References


elhuyar

ezagutuz aldatzea

https://github.com/Elhuyar/BasqueGLUE

g.urbizu@elhuyar.eus