elhuyar
ezagutuz aldatzea

Universidad del País Vasco · Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

LREC 2022

BasqueGLUE:
A Natural Language Understanding Benchmark for Basque

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

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

g.urbizu@elhuyar.eus

INDEX

BasqueGLUE:

Introduction

Motivation

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

Motivation

- Transformer → Big Language Models (LM)
- Remarkable results on NLU tasks, via transfer learning:
  - Pretrain on huge unannotated text corpus
  - Finetune on dowstream 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

Design

DESIGN

Metodology

- 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

DESIGN

Tasks

Corpus	Task	Domain
NERC_{id}	NERC	News
NERC_{ood}		News, Wikipedia
FMTODeu_{intent}	Intent classification	Dialog system
FMTODeu_{slot}	Slot filling	Dialog system
BHTCv2	Topic classification	News
BEC2016eu	Sentiment analysis	Twitter
VaxxStance	Stance detection	Twitter
QNLI_{eu}	QA/NLI	Wikipedia
WiC_{eu}	WSD	Wordnet
EpecKorrefBin	Coreference resolution	News

DESIGN

NERC ($NERC_{id}$, $NERC_{ood}$)

- Sequence labeling task
- 2 subtaks:
  - 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											
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.

DESIGN

Intent Classification (FMTODeu$_{intent}$)

- 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

Intent Classification (FMTODeu$_{intent}$)
Text: Alarma ezarri gaurko 6:00etan
Set the alarm today at 6:00am
Intent: `alarm/set_alarm`

Slot filling (FMTODeu$_{slot}$)

- 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

Slot Filling (FMTODeu$_{slot}$)					
Tokens:	Euria	egingo	du	gaur	?
Labels:	`B-weather/attribute`	`O`	`O`	`B-datetime`	`O`
Is it going to rain today?					

DESIGN

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, 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

DESIGN

Sentiment analisis (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`

DESIGN

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

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

DESIGN

Question Answering ($QNLI_{eu}$)

- 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: wether the sentence contains the aswer 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 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`

DESIGN

Word in Context (WiC_{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

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`

DESIGN

Coreference Resolution (EpecKorrefBin)

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

- Metric: Accuracy

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`

Summary

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.

Evaluation

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

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

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)

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

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

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

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.

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.

López de Lacalle, M., Saralegi, X., and L´opez, 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.

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

References

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.

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.

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.

Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

https://github.com/Elhuyar/BasqueGLUE

g.urbizu@elhuyar.eus