ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E4.2

DeepText

Nueva generación de modelos neuronales de inteligencia artificial para transformar las tecnologías de la lengua en la industria del País Vasco

Entregable E4.3: Informe de evaluación de modelos sobre BasqueGLUE y SpanishGLUE.

Responsable	**Elhuyar**
Tipo	**Informe**
Ejercicio	**2021**

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E4.2

Índice

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E4.2

1. Introduccion

Este Informe presenta los trabajos realizados dentro de la tarea T4.3 del proyecto DeepText. En concreto se detallan los resultados de la evaluación realizada sobre distintos modelos de lenguaje neuronales para euskera y castellano en diversas tareas de NLU, utilizando los benchmark BasqueGLUE y SpanishGLUE desarrollados en este mismo proyecto (Entregables 4.1 y 4.2, respectivamente).

Se han seleccionado modelos existentes con el objetivo de comparar el estado del arte con los modelos desarrollados en Deeptext. El trabajo realizado tiene también como objetivo evaluar si BasqueGLUE y SpanishGLUE son adecuados como benchmark de referencia para los presentes y futuros modelos neuronales.

Este documento se organiza de la siguiente manera a partir de aquí: la sección dos presenta los experimentos realizados y resultados obtenidos para el Euskera, mientras que la sección 3 hace lo propio con el Español. En los dos casos, se presentan los datos de evaluación, los modelos utilizados y por último la evaluación realizada.

2. Evaluación de modelos en euskera (BasqueGLUE)

Datasets

BasqueGLUE está formada por 9 tareas de NLU, que abarcan un amplio abanico de tamaños de datasets, de dificultades de las tareas, así como el dominio de estas. Cada tarea será evaluada por una única métrica, y la división de los datasets para entrenamiento, validación y evaluación están definidas.

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E4.2

Corpus	|Train|	|Dev|	|Test|	Tarea	Métrica	Dominio
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	9.652	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[1]	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

Tabla 1: Estadisticas de los conjuntos de datos que conforman BasqueGLUE

Modelos

Hemos comparado varios modelos utilizando BasqueGLUE como plataforma de evaluación. dos modelos para implementar el baseline. Los modelos incluídos en la comparación son los siguientes:

- **BERTeus** (Agerri et al., 2020): BERTeus es un modelo BERT base (12-layer, 768-hidden, 12-heads, 125M parameters) para el euskera entrenado anteriormente dentro del proyecto Deeptext sobre 224M de textos de tokens de fuentes de noticias y Wikipedia.
- **ElhBERTeu** (Urbizu et al., 2022): Es también un modelo BERT base que hemos entrenado en Deeptext con un corpus monolingüe de mayor tamaño

[1] F1 macro-average (MF1) de dos clases: FAVOR y AGAINST.

recogido recientemente, al que llamaremos ElhBERTeu. Para entrenar ElhBERTeu, aumentamos el corpus BMC (Agerri et al., 2020) utilizado para BERTeus. En concreto, se añadieron datos de 2020 y 2021 de las mismas fuentes de BMC, así como nuevas fuentes de noticias y textos de otros dominios.

- **Euscrawl** (Artetxe et al. 2022): Modelo basado en la arquitectura roberta (Liu et al., 2019) base (12-layer, 768-hidden, 12-heads, 125M parameters), preentrenado sobre un corpus similar a BMC.
- **Euscrawl_L** (Artetxe et al., 2022): Modelo similar al anterior, pero basado en la arquitectura roberta large (24-layers, 1024-hidden, 16-heads, 355M parameters).

No se incluye multilingual BERT modelos multilingües en esta evaluación porque ya en (Agerri et al., 2020) se demostró que BERTeus supera ampliamente en rendimiento a mBERT.

Evaluación

Se ha realizado la evaluación de cada uno de los modelos haciendo fine tuning para cada tarea, con los siguientes parámetros: Learning rate = 3e-5, Batch size = 32.
El número de epochs se optimiza sobre el conjunto de validación para cada tarea y modelo sobre un máximo de 10 epochs, utilizando 5 inicializaciones aleatorias. Los resultados mostrados en la tabla 2 son de una única ejecución con el epoch e inicialización con mejor rendimiento sobre el conjunto de validación. El resultado correspondiente a NERC es la media de las dos tareas NERC_id and NERC_ood. Se toma esta decisión para evitar que una tarea tenga más importancia que el resto en el promedio final.

Modelo	**AVG**	**NERC**	**Fintent**	**Fslot**	**BHTC**	**BEC**	**Vaxx**	**QNLI**	**WiC**	**coref**
		F1	**F1**	**F1**	**F1**	**F1**	**MF1**[2]	**Acc**	**Acc**	**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
Euscrawl	73.43	82.95	79.94	**77.14**	78.26	**71.58**	53.55	73.42	**73.29**	**70.70**
Euscrawl_L	**74.65**	**83.57**	**82.34**	75.03	**79.23**	70.12	61.91	**77.22**	72.43	70.02

Tabla 2: Resultados de la evaluacón sobre BasqueGLUE

[2] F1 macro-average (MF1) de dos clases: FAVOR y AGAINST.

La tabla 2 presenta los resultados de la evaluación para los modelos en euskera. Si observamos el promedio, Euscrawl_L obtiene el mejor resultado, mostrando que obtiene un rendimiento global más robusto. Si miramos las tareas individualmente, vemos que Euscrawl y Euscrawl_L salen vencedores en un el mismo número de tareas (4/9), aunque Euscrawl_L lo hace con major margen. ElhBERTeu supera a ambos en la tarea de stance detection, donde Euscrawl_L queda segundo, con un amplio margen sobre Euscrawl.

A la pregunta de si BasqueGLUE conforma un benchmark de evaluación adecuado, los resultados muestran que hay un claro margen de mejora en todas las tareas, lo que demuestra que los conjuntos de datos de BasqueGLUE presentan desafíos reales para los actuales modelos de lenguaje de estado del arte. GLUE se publicó un promedio de 70 para el mejor modelo base, mientras que SuperGLUE reportó un promedio de 74,6 para el mejor modelo, similar a la puntuación media de 74,65 obtenida por Euscrawl_L en BasqueGLUE.

3. Evaluación de modelos en Castellano (SpanishGLUE)

Datasets

SpanishGLUE está formada por 6 tareas de NLU, que abarcan un amplio abanico de tamaños de datasets, de dificultades de las tareas, así como el dominio de estas. Cada tarea será evaluada por una única métrica, y la división de los datasets para entrenamiento, validación y evaluación están definidas.

Dataset	Train	Dev	Test	Tarea	Métrica	Dominio
NERCconll2002	8.324	1.916	1.518	NERC	F1	News
FMTODes_intent	3.417	1.900	1.348	Intent classification	F1	Dialog system
FMTODes_slot	29.471	16.252	11.695	Slot filling	F1	Dialog system
MLDoc	9.458	1.000	4.000	Topic classification	F1	News
InterTass2020	4.802	2.465	1.500	Sentiment analysis	MF1	Twitter

SQAC	15.036	1.864	1.910	QA (QNLI)	Acc	Wikipedia

Tabla 3: Estadisticas de los conjuntos de datos que conforman SpanishGLUE

Modelos

Hemos comparado los modelos entrenados en Deeptext con otros modelos existentes para castellano. También incluímos en la comparación modelos multilingües como MBERT (Devlin et al., 2019) e IXAmBERT (Otegi et al., 2020). Los modelos incluídos en la comparación son los siguientes:

- **MBERT** (Devlin et al.,2019): Multilingual BERT es un modelo que utiliza la arquitectura BERT base (12-layer, 768-hidden, 12-heads, 125M parameters), entrenado utilizando textos de Wikipedia en 104 idiomas.
- **MBERTeus** (Otegi et al., 2020): Modelo basado en BERT base, utilizando datos exclusivamente de euskera, castellano e inglés.
- **Beto** (Cañete et al., 2020): Modelo basado en la arquitectura BERT base, sobre un corpus de 3K millones de palabras compilado de diversas fuentes digitales[3].
- **Roberta-b-bne** (Gutierrez Fandiño et al., 2022): Modelo basado en la arquitectura roberta (Liu et al., 2019) base (12-layer, 768-hidden, 12-heads, 125M parameters) preentrenado sobre un corpus de 135K millones de palabras extraídas del Archivo Web del Español construido por la Biblioteca Nacional de España entre los años 2009 y 2019.
- **Roberta-l-bne** (Gutierrez Fandiño et al., 2022): Modelo similar al anterior preentrenado utilizando la arquitectura roberta large, con una configuración superior (24-layers, 1024-hidden, 16-heads, 355M parameters).
- **IxaBERTesV1**: Modelo basado en la arquitectura BERT base entrenado en el proyecto Deeptext (Vease entregable E2.2).
- **IxaBERTesV2**: Modelo basado en la arquitectura ROBERTA base entrenado en el proyecto Deeptext (Vease entregable E2.2).

Evaluación

Se ha realizado la evaluación de cada uno de los modelos haciendo fine tuning para cada tarea, con los siguientes parámetros: Learning rate = 3e-5, Batch size = 32.
El número de epochs se optimiza sobre el conjunto de validación para cada tarea y modelo sobre un máximo de 5 epochs.

[3] https://github.com/josecannete/spanish-corpora

Modelo	AVG	NERC	Intent classification	Slot filling	Topic classification	Sentiment analysis	SQAC
		F1	F1	F1	F1	MF1	ACC
Mbert	81.77	86.38	98.44	89.79	95.45	36.99	83.56
mBERTeus	82.16	86.63	98.52	89.94	96.50	40.45	80.94
IxaBERTesV1 (Bert)	80.61	85.06	97.92	86.80	96.28	39.37	78.21
IxaBERTesV2 (RoBerta)	83.10	86.72	98.89	89.61	96.25	**48.90**	78.24
Beto	**83.93**	86.95	**98.96**	**90.59**	96.15	46.48	**84.47**
Roberta-b-bne	83.54	87.12	98.52	88.50	96.23	46.59	84.32
Roberta-l-bne	83.79	**87.48**	98.74	90.13	**97.02**	47.34	82.01

Tabla 4: Resultados para los modelos de castellano sobre SpanishGLUE.

La tabla 4 muestra los resultados de la evaluación realizada sobre SpanishGLUE para los modelos presentados en la sección anterior. Los resultados muestran resultados mixtos. En primer lugar, observamos que los modelos monolingües superan notablemente a los multilingües. En segundo lugar, Beto obtiene el mejor resultado global, aunque las diferencias con respecto a los modelos con arquitectura Roberta no son significantes. Si observamos las tareas individuales, Beto muestra el mejor rendimiento en 3 de las 6 tareas, pero se ve superado tanto los modelos entrenados sobre el corpus BNE en las otras tres[4]. Entre los modelos desarrollados en el proyecto Deeptext, podemos concluir que IxaBERTesV2 es un modelo competitivo respecto al estado del arte, que incluso obtiene el mejor resultado en la tarea de Análisis de Sentimiento.

A la pregunta de si SpanishGLUE conforma un benchmark de evaluación adecuado, los resultados muestran que hay un claro margen de mejora en todas las tareas, excepto en ‘intent classification’ y en ’topic classification’, donde todos los sistemas obtienen resultados notablemente altos. La primera se ha incluido por ir en conjunto con la tarea de slot filling, ambas relacionadas con el desarrollo de chatbots. La segunda se ha incluído en SpanishGLUE por ser un dataset de uso extendido presente en varias evaluaciones de la literatura (Cañete et al., 2020; Gutierrez Fandiño et al., 2022). El resto de los conjuntos de datos de SpanishGLUE presentan desafíos reales para los actuales modelos de lenguaje de estado del arte, con especial mención

[4] Los resultados son más bajos que los reportados por (Gutierrez Fandiño et al., 2022) para algunas tareas comunes en ambas evaluaciones. Sin embargo hay que señalar que en el trabajo mencionado se explora un mayor número de hiperparametros.

al dataset de análisis de sentimiento, que tiene la peculiaridad de que incluye tweets en 6 variantes del español, lo que representa un reto aún mayor.

Referencias

García-Vega, M., Díaz-Galiano, M. C., García-Cumbreras, M. A., Del Arco, F. M. P., Montejo-Ráez, A., Jiménez-Zafra, S. M., ... & Chiruzzo, L. (2020, September). Overview of tass 2020: Introducing emotion detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) Co-Located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain* (pp. 163-170).

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., ... & Villegas, M. (2022). Spanish language models. *Procesamiento del Lenguaje Natural*, *v. 68*, p. 39-60, mar. 2022. ISSN 1989-7553

Canete, J., Chaperon, G., Fuentes, R., Ho, J. H., Kang, H., & Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, *2020*, 2020.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa, and Eneko Agirre. 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, Marseille, France. European Language Resources Association.

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 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, Marseille, France. European Language Resources Association.