Give your text representation models some love: the case for Basque

Jon Zorrilla Gamboa

EPS

Madrid, 14 de marzo de 2023

Índice

1. Introducción y objetivos
2. Trabajo relacionado
3. Creando modelos en euskera
4. Evaluación y resultados
5. Discusión y conclusiones

Índice

1. Introducción y objetivos
   - Introducción
   - Objetivos
2. Trabajo relacionado
3. Creando modelos en euskera
4. Evaluación y resultados
5. Discusión y conclusiones

Introducción

- Los word-embeddings y los modelos entrenados nos han ayudado en la mayoría de problemas de procesamiento de lenguaje natural.
- Problemas: son difíciles de entrenar, y las compañías nos permiten hacer uso de modelos preentrenados.
- Surgen los modelos multilingüísticos. Modelos de lenguaje como BERT y Flair y word embeddings de FastText.

Objetivos

- Recrear modelos de lenguaje haciendo uso de un corpus más limpio y más completo y cambiando la tokenización de palabras debido a la naturaleza del euskera.
- Estudiar los resultados obtenidos mediante nuevos modelos y compararlos con los resultados obtenidos por modelos oficiales en las siguientes tareas: clasificación de temas, clasificación de sentimientos, POS tagging y Named Entity Recognition (NER).

Índice

1. Introducción y objetivos
2. Trabajo relacionado
3. Creando modelos en euskera
4. Evaluación y resultados
5. Discusión y conclusiones

Trabajo relacionado

- Word embeddings: Glove o Word2Vec. FastText propone una mejora al trabajar de manera más granular.
- Word embeddings contextuales, como Flair y BERT (mBERT).
- CAMEMBERT Y BERTeus modelos preentrenados para francés y euskera.
- Nuevos modelos: FastText BMC, Flair-BMC y BERTeus.

Índice

1. Introducción y objetivos
2. Trabajo relacionado
3. Creando modelos en euskera
4. Evaluación y resultados
5. Discusión y conclusiones

Creando modelos en euskera

- Hacer uso de un texto limpio y bien estructurado, además de lo más amplio posible. Se hace uso de un corpus en euskera basado en noticias y wikipedia, consistente de 224.6 millones de tokens, llamado BMC (Basque Media Corpus).
- Embeddings estáticos: FastText, entrenados con texto de wikipedia o texto de Common Crawl. En este caso, se usarán vectores de palabras entrenados en BMC.
- Word Embeddings contextuales: Flair distribuye sus propios embeddings preentrenados en euskera. En este caso se entrenarán los embeddings the Flair en BMC.
- Modelos de Lenguaje BERT entrenado en BMC.

Índice

1. Introducción y objetivos
2. Trabajo relacionado
3. Creando modelos en euskera
4. Evaluación y resultados
   - Clasificación de temas
   - Clasificación de sentimientos
   - POS Tagging
   - Named Entity Recognition (NER)
5. Discusión y conclusiones

Evaluación

- Modelos oficiales de FastText (Wikipedia o Common-Crawl) vs FastText-BMC.
- Modelos de embeddings oficiales de Flair vs Flair-BMC.
- Modelo multilingual de BERTm vs BERTeus.
- Tareas: clasificación de temas, clasificación de sentimientos, POS Tagging y NER.

Clasificación de temas

- Se hace uso de un dataset con 12.000 titulares del periódico Argia. Las noticias están clasificadas en 12 únicos temas. Agerri et al. (2020).

	Micro F1	Macro F1
Static Embeddings		
FastText-Wikipedia	65.00	54.98
FastText-Common-Crawl	28.82	3.73
FastText-BMC	69.45	60.14
Flair Embeddings		
Flair-official	65.25	55.83
Flair-BMC	68.61	59.38
BERT Language Models		
mBERT-official	68.42	48.38
BERTeus	**76.77**	**63.46**
Baseline		
TF-IDF Logistic Regression	63.00	49.00

Clasificación de sentimientos

- Se hace uso de un corpus de tweets conteniendo mensajes relacionados con temas culturales. El corpus contiene 3 clases (positivo, neutro y negativo) a partir de 2936 ejemplos. Agerri et al. (2020).

	micro F1	Macro F1
Static Embeddings		
FastText-Wikipedia	71.10	66.72
FastText-Common-Crawl	66.16	58.89
FastText-BMC	72.19	68.14
FlairEmbeddings		
Flair-official	72.74	67.73
Flair-BMC	72.95	69.05
BERT Language Models		
mBERT-official	71.02	66.02
BERTeus	**78.10**	**76.14**
Baseline		
SVM (San Vicente, 2019)	74.02	69.87

POS Tagging

- A partir del Basque UD Treebank, formado de 5274 frases de temas literarios y periodísticos. La tarea consiste en asignar a cada una de las palabras del texto su categoría gramatical. Agerri et al. (2020).

	Word Accuracy
Static Embeddings	
FastText-Wikipedia	94.09
FastText-Common-Crawl	91.95
FastText-BMC	96.14
FlairEmbeddings	
Flair-official	97.50
Flair-BMC	97.58
BERT Language Models	
mBERT-official	96.37
BERTeus	**97.76**
Baseline	
(Heinzerling and Strube, 2019)	96.10

Named Entity Recognition (NER)

- Se hace uso del corpus EIEC, consistente de 44.000 tokens para entrenamiento y 15.000 para test, formado por 4 tipos de entidades. La tarea consiste en identificar y categorizar información clave en el texto, tales como ubicación, persona, organización y miscelánea. Agerri et al. (2020).

	Precision	Recall	F1
Static Embeddings			
FastText-Wikipedia	72.42	50.28	59.23
FastText-Common-Crawl	72.09	45.31	55.53
FastText-BMC	74.12	67.33	70.56
Flair embeddings			
Flair-official	81.86	79.89	80.82
Flair-BMC	84.32	82.66	83.48
BERT Language Models			
mBERT-official	81.24	81.80	81.52
BERTeus	87.95	86.11	**87.06**
Baseline			
(Agerri and Rigau, 2016)	80.66	73.14	76.72

Índice

1. Introducción y objetivos
2. Trabajo relacionado
3. Creando modelos en euskera
4. Evaluación y resultados
5. Discusión y conclusiones

Discusión y conclusiones

- Los modelos BMC han mejorado todos los modelos previos en las 4 tareas. Agerri et al. (2020).

	Task			
	Topic Classification	Sentiment	POS	NER
Static Embeddings				
FastText-Wikipedia	65.00	71.10	94.09	59.23
FastText-Common-Crawl	28.82	66.16	91.95	55.53
FastText-BMC	69.45	72.19	96.14	70.56
Flair Embeddings				
Flair-official	65.25	72.74	97.50	80.82
Flair-BMC	68.61	72.95	97.58	83.48
BERT Language Models				
mBERT-official	68.42	71.02	96.37	81.52
BERTeus	**76.77**	**78.10**	**97.76**	**87.06**
Baselines	63.00	74.02	96.10	76.72

Discusión y conclusiones

- En general, los modelos formados por word embeddings monolinguales usando FastText y modelos de lenguaje como BERT y Flair entrenados en BMC obtienen unos resultados mucho mejores a los obtenidos por modelos públicos de PLN.
- A la hora de crear BERTeus, se han usado la misma estructura e hiperparámetros que para BERT oficial.
- Todo esto muestra la importancia de elegir un corpus bueno cuidadosamente para preentrenar los modelos en el mismo y hacer uso de una tokenización de palabras específica debido a la naturaleza del idioma.
- Como los modelos han usado los mismos hiperparámetros que los oficiales, aún tienen margen de mejora.
- Aplicable a todos los lenguajes minoritarios.

Bibliografía I

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. Give your text representation models some love: the case for basque. *CoRR*, abs/2004.00033, 2020. URL `https://arxiv.org/abs/2004.00033`.