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 E4.2: Marco de evaluación unificado SpanishGLUE

Responsable	**Elhuyar**
Tipo	**Recurso**
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 entregable describe el recurso “Marco de evaluación unificado BasqueGLUE”, elaborado dentro de la tarea T4.1 del proyecto DeepText. Se trata de un conjunto de datasets para distintas tareas de NLU, que servirá como benchmark para evaluar el rendimiento de los distintos modelos de lenguaje neuronales para el euskera, tanto los ya existentes, como los desarrollados en Deeptext, pero también los futuros.

2. Tarea

Se ha creado un nuevo benchmark para la evaluación de los modelos de lenguaje para castellano: SpanishGLUE. Este benchmark sigue el diseño de los benchmark GLUE y SuperGLUE del inglés. Para la selección de tareas, se ha seguido los siguientes criterios, definidos en SuperGLUE:

- Entidad de la tarea: Las tareas deben poner a prueba la capacidad de un sistema para entender y razonar sobre los textos en euskera y castellano.
- Dificultad de la tarea: Las tareas deben estar fuera del alcance de los actuales sistemas del estado del arte, pero pueden ser resueltas por la mayoría de los hablantes nativos con formación universitaria.
- Evaluabilidad: Las tareas deben tener una métrica automática de rendimiento que se correlacione adecuadamente con los juicios humanos de calidad de los sistemas.
- Datos públicos: En SuperGLUE, se exige que las tareas tengan datos de entrenamiento públicos, para minimizar los riesgos que implica utilizar conjuntos de datos de nueva creación. Sin embargo, en el caso del euskera, debido al número relativamente bajo de recursos anotados disponibles, hemos decidido incluir algunos datasets nuevos, que se han creado recientemente para otros fines, o se han construido a partir de conjuntos de datos previamente anotados y bien conocidos, Tomamos esta decisión con el objetivo de incluir tareas más diversas para poder evaluar mejor las capacidades NLU de los modelos.
- Formato de las tareas: Se ha optado por tareas que tenían formatos de entrada y salida relativamente simples, para evitar obligar a los investigadores a crear arquitecturas complejas para tareas específicas.
- Licencias: Los datasets de tareas deben estar disponibles bajo licencias que permitan su uso y redistribución con fines de investigación.

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E4.2

3. SpanishGLUE

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.

Named Entity Recognition:

La primera tarea que hemos elegido para incluir en SpanishGLUE es la conocida tarea de Named entity recognition (NERC), una tarea de etiquetado de secuencias. Ya existen numeroso datasets para castellano para esta tarea, por lo que se ha tomado la decisión de utilizar un datasets existente. Entre los datasets disponibles, hemos elegido el dataset de la tarea compartida CoNLL-2002 (Tjong Kim Sang, 2002). Anotado con 4 tipos de entidades (personas, localizaciones, organizaciones, y otros) en formato BIO, este dataset es el más usado para la tarea, aunque el mejor resultado hasta la fecha no supera el 89% de F1-score (Gutierrez Fandiño et al., 2022).

Intent Classification (FMTODes_intent):

La siguiente tarea incluida en el benchmark es la clasificación de intenciones, una tarea de NLU en el campo de los sistemas de diálogo que tiene como objetivo identificar la intención que el usuario denota en una frase, entre un conjunto de clases predefinidas. Por lo tanto, se aborda como una tarea de clasificación de secuencias multiclase. El conjunto de datos que seleccionamos para la tarea es el Facebook Multilingual Task Oriented Dataset for Spanish o FMTODes (López de Lacalle et al., 2020).

Los ejemplos están anotados con una de las 12 clases de intención diferentes correspondientes a acciones relacionadas con la alarma, recordatorios o el tiempo. FMOTDes
es una versión corregida del conjunto de datos original (Schuster et al.,2019). Una revisión inicial del conjunto de datos español, reveló que un número significativo de

traducciones incluyó notables errores de ortografía, morfológicos, sintácticos o léxicos. Para su corrección se revisaron los conjuntos de desarrollo, entrenamiento y test y se procedió a hacer las correcciones correspondientes tanto de los textos como de las anotaciones. Las expresiones con un elevado número de fallos fueron descartadas directamente para aligerar el esfuerzo manual. Por la misma razón, el test fue solo parcialmente revisado. Denominaremos al conjunto de datos FMTODes_intent en este benchmark, para diferenciarlo de la tarea siguiente tarea también incluida en el conjunto de datos. Mantenemos la división original de train/dev/test. Se utilizará la métrica micro F1 para la evaluación.

Slot Filling (FMTODes_slot):

La tercera tarea es el slot_filling, una tarea que también proviene del campo de los sistemas de diálogo y que suele realizarse junto con la tarea de clasificación de intenciones. El objetivo es identificar las entidades asociadas a las intenciones expresadas en las frases formuladas por el usuario. FMTODes incluye estas anotaciones de entidades.

La tarea es una tarea de etiquetado de secuencias similar a NERC, siguiendo el esquema de anotación BIO sobre 11 categorías. Llamaremos al conjunto de datos FMTODes_slot, y al igual que para la clasificación de intenciones, mantenemos la división original de train/dev/test. Se utilizará la métrica micro F1 para la evaluación.

Topic Classification (ML-doc):

La clasificación de temas es otra tarea de clasificación de secuencias multiclase. El conjunto de datos que incluido en SpanishGLUE es la parte en castellano de ML-doc (Schwenk & Li, 2018). Este es un dataset de uso extendido presente en varias evaluaciones de la literatura (Cañete et al., 2020; Gutierrez Fandiño et al., 2022). El corpus contiene noticias de la agencia Reuters (Lewis et al., 2004). Las noticias se clasifican según cuatro categorías temáticas: CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social) and MCAT (Markets). Al igual que para la clasificación de intenciones, el rendimiento se mide utilizando la puntuación micro F1.

Sentiment Analysis (InterTASS-2020):

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E4.2

El análisis de sentimiento es una tarea bien conocida presente en la mayoría de los benchmarks de NLU. El objetivo es clasificar correctamente la polaridad de los textos dados, entre las clases positiva, neutra o negativa en nuestro caso.

El Taller de Analisis de Sentimiento (TASS) (Villena-Román et al. 2013), organizado desde 2012 a publicado varios corpus para distintas tareas de análisis de sentimiento en español. Para SpanishGLUE, hemos adoptado el corpus InterTASS-2020 (García-Vega et al. 2020), anotado para la tarea de detección de polaridad, tal y como se plantea en la tarea 1.2 de TASS 2020. Este corpus contiene tweets en 6 variantes del español (México, Uruguay, Chile, Perú, Costa Rica y España), lo que representa un reto aún mayor para los sistemas. Los tweets están anotados en base a tres categorías (positivo, negativo, neutro). La métrica que proponemos para esta tarea y conjunto de datos es la puntuación macro F1, ya que la distribución de las clases no es equitativa.

QNLI (sqac):

Para incluir una tarea de QA en el benchmark, adaptamos el conjunto de datos de QA SQAC (Gutiérrez-Fandiño et al., 2022). SQAC es un conjunto de datos para realizar QA extractivo, sin preguntas incontestables. Se crea a partir de textos extraídos de la Wikipedia española, artículos enciclopédicos, artículos de boletines de Wikinews y textos procedentes del corpus Ancora (Taulé, Martí y Recasens, 2008), que incluye noticias y literatura. Anotado por hablantes nativos, contiene un 18.817 preguntas y respuestas, divididas en tres partes: train, development y test.

Adaptamos este conjunto de datos a una tarea de clasificación binaria de pares de frases, siguiendo el diseño de QNLI para el inglés, la cual está incluida en GLUE. Formamos un ejemplo con cada par de preguntas y cada frase de contexto y, a continuación, filtramos los pares con la menor coincidencia léxica entre la pregunta y la frase en el caso de los ejemplos negativos, hasta que nos quedamos con un conjunto de datos equilibrado. Como métrica de evaluación, seguimos el diseño del QNLI y utilizamos la precisión.

limitamos los pares de menciones a los que se encuentran en la misma frase o en frases consecutivas.

Para crear ejemplos negativos, seleccionamos pares de menciones que incluyeran un pronombre o tuvieran el mismo tipo de mención (por ejemplo, que ambos fueran nombres propios de lugares). A continuación, en un intento de hacer la tarea más difícil, filtramos los pares de menciones más similares entre los ejemplos positivos y también los que eran más diferentes de los negativos. Como medidas de similitud se

utilizaron la distancia Levenshtein para los ejemplos positivos y el ratio de conjunto de tokens (token set ratio) para los negativos. Por último, nos aseguramos que todas las divisiones están equilibradas para los ejemplos positivos y negativos.

En resumen, estos son las 6 tareas que forman parte del benchmark SpanishGLUE:

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	F1	Twitter
SQAC	15.036	1.864	1.910	QA (QNLI)	Acc	Wikipedia

Tabla 1: Estadísticas de los datasets que conforman SpanishGLUE.

Cada dataset tiene su propia licencia, debido a que la mayoría de ellos derivan de datasets preexistentes. La tabla 2 muestra un resumen de dichas licencias. El resto de ficheros que componen el benchmark, incluyendo el script de evaluación de los sistemas se distribuyen bajo licencia Creative Commons Attribution 4.0 (CC BY 4.0).

Dataset	License
NERCconll2002	Unknown
FMTODes_intent	CC BY-NC-SA 4.0
FMTODes_slot	CC BY-NC-SA 4.0
MLDoc	CC BY-NC 4.0
InterTass2020	Twitter's license + CC BY-NC-SA 4.0
SQAC	CC BY-SA 4.0

Tabla 2: Resumen de licencias de los conjuntos de datos que conforman Spanish.

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E4.2

Descarga de los recursos

El benchmark SpanishGLUE está accesible en la siguiente página web: https://github.com/Elhuyar/SpanishGLUE

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

López de Lacalle, M., Saralegi, X., & San Vicente, I. (2020, May). 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* (pp. 2796-2802).

Schuster, S., Gupta, S., Shah, R., and Lewis, M. (2019). Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805, June.

Holger Schwenk and Xian Li. 2018. A Corpus for Multilingual Document Classification in Eight Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Villena-Román, Julio, Lana-Serrano, Sara, Martínez-Cámara, Eugenio, González-Cristobal, José Carlos. 2013. Revista de Procesamiento del Lenguaje Natural, 50, pp 37-44.