ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E4.1

DeepText

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

Entregable E4.1: Marco de evaluación unificado BasqueGLUE

Responsable	**Elhuyar**
Tipo	**Recurso**
Ejercicio	**2021**

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E4.1

Índice

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E4.1

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 euskera, BasqueGLUE, que 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.

3. BasqueGLUE

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. A continuación

Named Entity Recognition

La primera tarea que hemos elegido para incluir en BasqueGLUE es la conocida tarea de Named entity recognition (NERC), una tarea de etiquetado de secuencias.
El conjunto de datos NERC se divide en dos subtareas: NERC dentro del dominio y NERC fuera del dominio. En cuanto a la métrica de evaluación se utilizará la media de los F1 de ambas subtareas.

Para la subtarea NERC dentro del dominio (NERCid), EIEC fue el conjunto de datos estándar anterior para NERC en euskera, utilizado también para evaluar BERTeus. EIEC se compone de textos procedentes de fuentes de noticias y está anotado siguiendo el esquema de anotación BIO en cuatro categorías: persona, organización, ubicación y varios. Con el objetivo de obtener un conjunto de datos más amplio, hemos fusionado EIEC con un nuevo conjunto de datos de NERC que contiene textos del periódico Naiz, que se anotó con las mismas pautas utilizadas en EIEC. Para este benchmark, fusionamos ambos conjuntos de datos y los dividimos en 3 partes, entrenamiento, desarrollo y evaluación, de forma que ambas fuentes están presentes en todas las divisiones.

En la subtarea NERC fuera del dominio (NERCood), el conjunto de entrenamiento está formado por datos del dominio de las noticias, mientras que el conjunto de evaluación contiene datos de artículos de Wikipedia. En concreto, el conjunto de entrenamiento se ha obtenido uniendo los datos de entrenamiento y de desarrollo para el NERC dentro del dominio. Para la evaluación fuera del dominio, se ha creado un nuevo conjunto de datos a partir de Wikipedia, que se ha anotado junto con los datos de Naiz, siguiendo

las mismas directrices. El conjunto de datos de Wikipedia se ha dividido en conjuntos de desarrollo y de evaluación para la tarea de NERCood.

Intent Classification (FMTODeu_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 Basque o FMTODeu.

Los ejemplos están anotados con una de las 12 clases de intención diferentes correspondientes a acciones relacionadas con la alarma, recordatorios o el tiempo. Denominaremos al conjunto de datos FMTODeu_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 (FMTODeu_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. FMTODeu 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 FMTODeu_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 (BHTCv2)

La clasificación de temas es otra tarea de clasificación de secuencias multiclase. El conjunto de datos que ofrecemos aquí se basa en el conjunto de datos BHTC. Contiene titulares de noticias (descripciones breves de artículos) del semanario vasco Argia. Las noticias se clasifican según doce categorías temáticas.

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E4.1

Aunque los ejemplos del BHTC original estaban en minúsculas y se les había eliminado la puntuación, decidimos mantener los textos originales para BasqueGLUE. También eliminamos varios ejemplos duplicados del set de entrenamiento y ejemplos que no estaban en euskera. A partir de ahora llamaremos a este conjunto de datos BHTCv2. Al igual que para la clasificación de intenciones, el rendimiento se mide utilizando la puntuación micro F1.

Sentiment Analysis (BEC):

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 Basque Election Campaign 2016 Opinion Dataset (BEC2016eu) es un nuevo conjunto de datos para la tarea de análisis de sentimientos, una tarea de clasificación de secuencias, que contiene tuits sobre la campaña de las elecciones vascas de 2016. El rastreo se realizó durante el periodo de campaña electoral (2016/09/09-2016/09/23), haciendo un seguimiento de los principales partidos y sus respectivos candidatos. Los tuits fueron anotados manualmente como positivos, negativos o neutros. La métrica que proponemos para esta tarea y conjunto de datos es la puntuación micro F1.

Stance Detection (VaxxStance)

La detección de posturas o stance (SD) es una de las tareas del campo de la detección de noticias falsas, también abordada como una tarea de clasificación de secuencias. Su objetivo es detectar la postura en los redes sociales sobre un tema muy controvertido y de moda. La tarea consiste en determinar si un tuit dado expresa una postura contraria (AGAINST), favorable (FAVOR) o neutra (NEUTRAL) hacia el tema.

El conjunto de datos VaxxStance ha sido incluido en BasqueGLUE. Se trata de tuits relacionados con el movimiento antivacunas. El conjunto de datos no incluía un conjunto de desarrollo, por lo que dividimos los datos de entrenamiento originales creando un nuevo conjunto de entrenamiento y desarrollo como resultado. La división se hizo de forma aleatoria, en un esfuerzo por proporcionar datos de desarrollo y hacer más consistentes las puntuaciones obtenidas en el nuevo benchmark, para una comparación más justa entre los modelos evaluados. Siguiendo la pista original de VaxxStance y la literatura de tareas compartidas de SD, medimos el rendimiento de los

sistemas mediante la puntuación F1 macro-average (MF1) de dos clases: FAVOR y AGAINST.

QNLI

Para incluir una tarea de QA en el benchmark, adaptamos el conjunto de datos de QA ElkarHizketak, un conjunto de datos de preguntas y respuestas conversacionales (QA) de bajos recursos para el euskera creado por voluntarios nativos. El conjunto de datos está construido sobre secciones de Wikipedia sobre personas y organizaciones conocidas y contiene alrededor de 400 diálogos y 1600 pares de preguntas y respuestas.

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.

WiC

Word in Context o WiC es una tarea de desambiguación del sentido de las palabras (WSD) incluida en el Benchmark SuperGLUE para el inglés. Está diseñada como una forma particular de clasificación binaria de pares de frases. Dados dos fragmentos de texto y una palabra polisémica que aparece en ambos (la posición de la palabra está marcado en ambos fragmentos), la tarea consiste en determinar si la palabra tiene el mismo sentido en ambas frases. El rendimiento se evalúa mediante la precisión.

Hemos generado un conjunto de datos tomando como punto de partida el corpus EPEC-EuSemcor. EPEC-EuSemcor es un corpus con etiquetas de sentido para el euskera. Este corpus contiene un conjunto de ocurrencias de sustantivos que han sido anotados con los sentidos de WordNet v1.6. Contiene 42.615 ocurrencias de sustantivos anotados manualmente, que corresponden a los 407 sustantivos vascos más frecuentes. Sólo hemos utilizado las ocurrencias con frases contextuales de 10 a 50 palabras de longitud.

El conjunto de datos del WiC en euskera sigue el diseño del conjunto de datos del WiC del inglés, emparejando frases de contexto para cada sustantivo para crear las instancias de la tarea de clasificación.

Adoptamos la misma estrategia que en el conjunto de datos en inglés para aumentar la claridad del conjunto de datos y eliminamos todos los pares cuyos sentidos eran conexiones de primer grado en en el grafo semántico de WordNet, incluidos los sentidos hermanos y los que pertenecen al mismo supersentido (supersense). Para esta poda se utilizó un mapeo de Wordnet v1.6 a Wordnet v3.0, como es habitual en otros conjuntos de datos relacionados con WSD.

Se separaron 1.400 y 600 instancias para los conjuntos de test y de desarrollo, respectivamente, con la restricción de no tener frases contextuales repetidos. Los pares restantes cuyas frases contextuales no formarán parte de los sets de evaluación y de desarrollo formaron nuestros datos de datos de entrenamiento (donde permitimos la repetición de las de las frases de contexto, para aumentar el tamaño de los datos de entrenamiento). Por último, nos aseguramos de que todas las divisiones estuvieran equilibradas para ejemplos positivos y negativos.

Coreference Resolution (EpecKorrefBin)

La última tarea que hemos seleccionado para nuestro benchmark BasqueGLUE es la resolución de correferencias, para la que ya existe un conjunto de datos para el euskera, EPEC-KORREF. Sin embargo, la resolución de correferencias implica la agrupación de menciones en entidades, sin que exista un formato fácil para abordar la tarea en su totalidad. Por eso, decidimos convertirla en una tarea de clasificación binaria, adoptando el mismo formato utilizado para la tarea Winograd Schema Challenge (WSC). En esta nueva tarea, el modelo tiene que predecir si dos menciones de un texto, que pueden ser pronombres, nombres o frases sustantivas, se refieren a la misma entidad. Adaptamos EPEC-KORREF a esta tarea, y nombramos a este conjunto de datos como EpecKorrefBin. Durante la creación de ejemplos, 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.

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E4.1

El resultado de este paquete de trabajo son los benchmarks para la evaluación de modelos de lenguaje en NLU para euskera y castellano, que estaran disponibles públicamente, y que servirá para comparar los modelos de lenguaje existentes con los nuevos modelos de lenguaje que se vayan creando, ayudando a progresar en el área del procesamiento del lenguaje para estos idiomas.

La tabla 1 muestra un resumen de las 9 tareas que forman parte del benchmark BasqueGLUE:

Corpus	**|Train|**	**|Dev|**	**|Test|**	**Tarea**	**Metrica**	**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

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

Tabla 1: Estadísticas de los datasets que conforman BasqueGLUE.

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
NERCid	CC BY-NC-SA 4.0
NERCood	CC BY-NC-SA 4.0
FMTODeu_intent	CC BY-NC-SA 4.0
FMTODeu_slot	CC BY-NC-SA 4.0
BHTCv2	CC BY-NC-SA 4.0
BEC2016eu	Twitter's license + CC BY-NC-SA 4.0
VaxxStance	Twitter's license + CC BY 4.0
QNLIeu	CC BY-SA 4.0
WiCeu	CC BY-NC-SA 4.0
EpecKorrefBin	CC BY-NC-SA 4.0

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

Descarga de los recursos

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

Referencias

G. Urbizu, I. San Vicente, X. Saralegi, R. Agerri, A. Soroa. 2022. BasqueGLUE: A Natural Language Understanding Benchmark for Basque. In proceedings of the 13th Edition of the Language Resources and Evaluation Conference (LREC2022).