ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E1.2

DeepText

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

Entregable E1.2: Corpus multilingües (comparables y paralelos) de gran tamaño para euskera, castellano e inglés.

Responsable	Elhuyar
Tipo	Recurso
Ejercicio	2020

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E1.2

Índice

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E1.2

Introducción

Este entregable describe el entregable “Corpus multilingües paralelos de gran tamaño para euskera, y castellano” generado dentro de la tareas T1.1 y T2.2 del proyecto DeepText. Estos corpus podrán ser usados para entrenar distintos modelos de lenguaje neuronales multilingües de última generación creados en el proyecto.

Descripción

Para la creación del corpus final multilingüe paralelo de Castellano-Euskera, se han recolectado distintos corpus públicos y libres desde la web OPUS[1] (Tiedemann, 2012), añadiendo los corpus multilingües paralelos de los que se disponía anteriormente.

Todos estos textos paralelos, vienen por defecto alineados a nivel de frases, lo que dificulta la creación de corpus paralelo.

Los textos paralelos suelen ser utilizados para el entrenamiento de modelos neuronales para la traducción automática, pero también pueden utilizarse para entrenar modelos de lenguaje multilingües, como por ejemplo XLM (Lample et al., 2019), donde utilizan como objetivo el *Translation Language Modeling* (TLM) junto al habitual *Masked Language Modeling* (MLM).

La tabla 1 muestra la composición del corpus paralelo multilingüe compilado (eu-es). El tamaño del corpus paralelo es de un total de 141M de palabras.

Fuente	**Tamaño**
Bittor	47M
EITB-ParCC	24M
EJ (OpenData)	10M
EhuHac	10M
GipuzkoaNET	9M
Elhuyar	9M
MultiParaCrawl	9M
ParaCrowl	7M

1 https://opus.nlpl.eu/

OpenSub	5M
Consumer	1M
Otros: QED, aldaketa16, eitb24, EHU liburuak.	1M

Tabla 1: Composición de la versión final del corpus paralelo multilingüe (Es-Eu).

Descarga de los recursos

Los corpus recopilados no están disponibles en repositorios públicos debido a cuestiones legales. Los recursos están disponibles para usos específicos. Para obtener los recursos descritos en este documento los interesados deben ponerse en contacto con Xabier Saralegi (x.saralegi@elhuyar.eus) o con Iñaki San Vicente (i.sanvicente@elhuyar.eus).

Bibliografia

Tiedemann, J. (2012, May). Parallel Data, Tools and Interfaces in OPUS. In *Lrec* (Vol. 2012, pp. 2214-2218).

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291.