ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E1.1

DeepText

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

Entregable E1.1: Corpus monolingües de gran tamaño para euskera y castellano para los cinco dominios especificados.

Responsable	Elhuyar
Tipo	Recurso
Ejercicio	2020

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E1.1

Índice

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E1.1

Introducción

Este entregable describe el recurso “Corpus monolingües de gran tamaño para euskera y castellano” elaborado dentro de la tareas T1.1 y T2.2 del proyecto DeepText. Estos corpus servirán para entrenar distintos modelos de lenguaje neuronales de última generación para euskera y castellano generados en el proyecto.

Descripción

Para entrenar tanto las representaciones de palabras estáticas como los modelos de lenguaje neuronales, son necesarios grandes corpus. Con ese objetivo, construimos un corpus en euskera y otro en castellano compuesto por textos de distintas fuentes y de varios dominios.

El Corpus de euskera construido se compone de textos recolectados de artículos de noticias, Wikipedia, literatura, ciencia, twitter y subtítulos. En el caso de las noticias, literatura y ciencia, se han obtenido textos de más de una fuente. El texto se ha extraído con scripts creados a medida para cada fuente, utilizando herramientas[1] de conversión de formato en los caso de los archivos en formatos no textuales (epub o pdf) y filtros de idiomas[2]. Tras la extracción de los textos, se han limpiado y normalizado.

La tabla 1 muestra la composición del corpus final compilado en euskera. El corpus contiene contenido recogido hasta diciembre de 2020, con un total de 395M de palabras.

Fuente	**Tipo**	**Tamaño**
Wikipedia	enciclopedia	40M
Berria	news	97M
EiTB	news	28M
Argia	news	8M
Medios Locales	news	87M
Gara	news	3M

1 https://github.com/kevinboone/epub2txt, https://github.com/jalan/pdftotext, http://manpages.ubuntu.com/manpages/bionic/man1/unrtf.1.html
2https://github.com/saffsd/langid.py (Heafield et al., 2020)

Armiarma	reviews	2M
Opensubtitles	subtitles	4M
Twitter	twitter	43M
Literatura (booktegi, susa, erein, armiarma, teatrotestuak)	Literature (Poetry, Drama & Prose)	24M
Ciencia (ikergazte, ekaia, ehu-argitalpenak, ADDI, zientzia.eus, zth, ztc, zientzia-kaiera)	Science (articles, book & news)	58M

Tabla 1: Composición de la versión final del corpus en euskera.

El Corpus de castellano construido se compone en textos recolectados de artículos de noticias, Wikipedia, twitter y varios corpus de gran tamaño de contenido variado.

La tabla 2 muestra la composición del corpus final construido para castellano. El corpus contiene contenido recogido hasta diciembre de 2020, con una cantidad total de palabras superior a los 30M.

Fuente	**Tipo**	**Tamaño**
Wikipedia	enciclopedia	599M
Efe	news	175M
El Pais	news	327M
Diario Vasco	news	77M
Twitter	twitter	3.399M
Spanish unanotated corpora[3]	mix	3.000M
Oscar[4]	web	25.928M

Tabla 2: Composición de la versión final del corpus en castellano.

3 https://github.com/josecannete/spanish-corpora
4 https://oscar-corpus.com/ (Suárez et al., 2020)

DeepText – Entregable E1.1

La versión final del corpus del castellano, supera con creces las necesidades de tamaño fijadas en este proyecto, llegando a los 30.000M de palabras. Debido a eso, se ha utilizado un subconjunto de corpus variado de gran calidad de más de 1.000M de palabras para el entrenamiento de representaciones textuales estáticas SOTA.

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

Heafield, K., Kshirsagar, R., & Barona, S. (2015, July). Language identification and modeling in specialized hardware. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 384-389).

Suárez, P. J. O., Romary, L., & Sagot, B. (2020, July). A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. In ACL 2020-58th Annual Meeting of the Association for Computational Linguistics.