ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E2.1

DeepText

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

Entregable E2.1: Nuevas representaciones vectoriales estáticas para el euskara utilizando los modelos del estado del arte (word2vec, Fasttext)

Responsable	E2.1
Tipo	Recurso
Ejercicio	2020

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E2.1

Índice

ELKARTEK 2020

Proyectos de investigación fundamental colaborativa

DeepText – Entregable E2.1

Introducción

Este entregable describe las representaciones vectoriales estáticas entrenadas utilizando los modelos del estado del arte word2vec y FastText dentro de la tarea T2.1 (Entrenamiento de representaciones estáticas) del proyecto DeepText. Estos sistemas entrenados con los corpus construidos en la tarea T1.1 nos servirán como punto de partida y de comparación respecto a los sistemas posteriores que se generarán a lo largo del proyecto.

Descripción

Se han generado embeddings estáticos utilizando los algoritmos Word2vec y FastText. Word2vec (Mikolov et al., 2013) es el primer algoritmo neuronal propuesto para obtener representaciones textuales estáticas de palabras. FastText (Bojanowski et al., 2017) propone mejoras sobre word2vec, ya que permite obtener representación no solo de palabras completas sino de unidades “sub-palabra”. Esta característica es especialmente adecuada para idiomas aglutinantes como el euskera.

Los únicos embeddings estáticos disponibles en euskera hasta ahora eran los distribuidos por Facebook utilizando el algoritmo FastText, y generados utilizando un enfoque basado en el modelo skip-gram, donde cada palabra se representa como un vector de n-gramas de caracteres (Bojanowski et al., 2017). Hasta ahora había dos representaciones estáticas disponibles: (i) **Wiki word vectors** (FastText-official-wikipedia) entrenados sobre Wikipedia utilizando el modelo skip-gram descrito en (Bojanowski et al., 2017) con los parámetros por defecto (window size=5, 3-6 length character n-grams and 5 negativos); y (ii) **Common Crawl word vectors** (FastText-official-common-crawl) entrenados sobre Common Crawl y Wikipedia utilizando CBOW con pesos posicionales, n-grams de caracteres de longitud 5, window size=5 y 10 negativos(Grave et al., 2018).

En DeepText se han generado y puesto a disposición pública, los siguientes embeddings estáticos:

Euskera

- Word2vec-BMC: Entrenados utilizando Word2vec sobre el corpus BMC (euskara) durante 10 epochs (frecuencia mínima 5, 300 dimensiones, window-size=5), tanto con el modelo skip-gram como con el modelo CBOW. Estos son los primeros embeddings Word2vec disponibles públicamente en euskera.

- Word2vec-Deeptext: Embeddings entrenados de la misma manera que Word2vec-BMC, pero utilizando el corpus final compilado en el marco de Deeptext.
- FastText-BMC: Entrenados utilizando FastText sobre el corpus BMC con los parametros por defecto (Bojanowski et al., 2017).
- FastText-Deeptext: Embeddings entrenados de la misma manera que Word2vec-BMC, pero utilizando el corpus final compilado en el marco de Deeptext.

Castellano

- Word2vec-Deeptext: Entrenados utilizando Word2vec sobre el corpus Deeptext_es (castellano, 1.000M tokens) durante 10 epochs (frecuencia mínima 5, 300 dimensiones, window-size=5), tanto con el modelo skip-gram como con el modelo CBOW.
- FastText-Deeptext: Entrenados utilizando FastText sobre el corpus Deeptext_es (castellano, 1.000M tokens) con los parámetros por defecto (Bojanowski et al., 2017).

Descarga de los recursos

Las representaciones estáticas entrenadas están accesibles en el siguiente link:.

http://www.deeptext.eus/es/node/3

Los embeddings aquí recogidos han sido entrenados con las herramientas originales Word2vec (https://code.google.com/archive/p/word2vec/) y Fasttext (https://fasttext.cc/), y exportados a formato textual.

Bibliografia

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135–146.