ELKARTEK 2020
Proyectos de investigación fundamental colaborativa

DeepText – Entregable E2.3

DeepText

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

E2.3: Nuevos modelos de lenguaje multilingües para el euskara, castellano e inglés utilizando corpus monolingües y paralelos

Responsable	**IXA**
Tipo	**Recurso**
Ejercicio	**2021**

ELKARTEK 2020
Proyectos de investigación fundamental colaborativa
DeepText – Entregable E2.3

Índice

ELKARTEK 2020
Proyectos de investigación fundamental colaborativa

DeepText – Entregable E2.3

Introducción

Este entregable describe el modelo de lenguaje multilingüe llamado IXAmBERT el cuál ha sido pre-entrenado para inglés, español y euskera dentro de la tarea T2.3 (Entrenamiento de modelos de lenguaje contextuales multilingües) del proyecto DeepText. Este modelo entrenado con los corpus recogidos en el paquete de trabajo 1 junto con las Wikipedias en inglés, español y euskera, ha supuesto un avance en el estado del arte y supone un hito importante para el desarrollo de aplicaciones basadas en redes neuronales para el euskera y castellano. Avanzando así además, en el estado del arte de las aplicaciones de Procesamiento de Lenguaje Natural (PLN) para estas lenguas. El modelo ha sido utilizado con éxito al transferir conocimientos adquiridos del inglés al euskera en un sistema de control de calidad conversacional. Además, IXAmBERT ha sido utilizado en diferentes tareas del paquete de trabajo 3 de DeepText.

Descripción

Pre-entrenamiento del modelo

Se ha entrenado un modelo multilingüe para euskera, español y castellano con parte del corpus de PT1 y las Wikipedia en inglés, español y euskera. Mas concretamente, hemos pre-entrenado un modelo BERT multilingüe con la intención de realizar experimentos de transferencia entre idiomas con muchos recursos (el inglés) a idiomas con menos recursos (el español y el euskera). Este tipo de experimentos pueden ser llevados a cabo con el modelo oficial mBERT, pero al cubrir tantos idiomas el euskera no está del todo bien representado. Para crear este nuevo modelo multilingüe que contiene únicamente inglés, castellano y euskera, hemos seguido la misma configuración que en el modelo BERTeus. Reutilizamos el mismo corpus del modelo del euskera monolingüe y añadimos la Wikipedia en inglés y español con 2,5M y 650M tokens respectivamente. Debido al desequilibrio de los tamaños de los corpus de entrada, hemos utilizado las mismas estrategias de sobremuestreo y creación de vocabulario de subpalabras propuestas en Lample y Conneau (2019). Como resultado, tenemos un vocabulario de subpalabras multilingüe del tamaño de 112K tokens. Para el entrenamiento del modelo se ha utilizado la técnica de "*whole word masking*", así como las funciones objetivo típicas en arquirecturas BERT como son el MLM (M*asked-Language*

Modeling) y NSP (N*ext Sentence Prediction*), con un ritmo de aprendizaje (*learning rate*) de 1e-4 y un tamaño de lotes de 256 sub-palabras.

Evaluación en búsqueda de respuestas conversacionales

Hemos evaluado el modelo IXAmBERT en la tarea de la búsqueda de respuestas conversacionales en euskara. Para ello, hemos entrenado IXAmBERT en un conjunto de datos de búsqueda de respuestas conversacionales y hemos comparado los resultados con otros modelos multilingües que incluyen euskara, como mBERT. Dada una pregunta y un pasaje como entrada, los sistemas conversacionales de búsqueda de respuestas están diseñados para encontrar un extracto relevante dentro del pasaje que responde a la pregunta. Para modelar la componente conversacional, es importante conocer la historia de la conversación (pares pregunta-respuesta) que se ha producido previamente a la pregunta en curso, ya que esta historia es probablemente que haya provocado dicha pregunta. En otras palabras, para entender la pregunta actual, es posible que sea necesario tener en cuenta el historial de la conversación, ya que la pregunta a tratar podría tener referencias a preguntas o respuestas anteriores. Por lo tanto, este tipo de sistemas reciben como entrada el historial de diálogo con los pares de preguntas/respuestas previas.
En los experimentos se han utilizado dos conjuntos de datos: el dataset ElkarHizketak (Otegi el a., 2020) para el euskara, y el dataset Quack para el Inglés (Choi et al., 2018). Para medir la transferencia entre lenguas que ofrecen los modelos de lenguaje multilingües, hemos diseñado dos tipos de experimentos:

- *Zero-shot*: en este experimento, el modelo es entrenado utilizando Quack (Inglés) y evaluado en el conjunto de datos en euskara (test de Elkarrizketak).
- *Transfer-learning*: El modelo *zero-shot* es entrenado de nuevo utilizando en un conjunto de datos en euskara (train de Elkarrizketak), que pertenece además al mismo dominio que el conjunto de datos de evaluación.

La métrica utilizada para comparar los modelos es la llamada F1, la media harmónica sobre la precisión y el *recall* calculados sobre el solapamiento de la secuencia de palabras seleccionadas por el sistema con la secuencia palabras que forman la respuesta real.

En el escenario del sistema conversacional, IXAmBERT obtiene mejores resultados que modelos de lenguajes multilingües existentes (mBERT) como se muestra en la siguiente tabla:

Modelo	Zero-shot	Transfer learning
Baseline	28,7	28,7
mBERT	31,5	37,4
IXAmBERT	38,9	**41,2**
mBERT+historico	33,3	28,7
IXAmBERT+historico	**40,7**	40,0

La tabla muestra de forma clara que el modelo IXAmBERT es preferible al modelo multiBERT, un modelo de lenguaje multilingüe pre-entrenado simultáneamente con artículos de Wikipedia de 104 idiomas diferentes y publicado en Devlin et al. (2019).

Descarga de los sistemas

El modelo IXAmBERT está accesible en la siguiente página web.
http://www.deeptext.eus/resources

Para poder utilizar el modelo, al ser un modelo de hugginface, es suficiente con instalar Transformers de Huggingface (https://huggingface.co/transformers/). Transformers está testado en Python 3.6+, PyTorch 1.3.1+ y TensorFlow 2.3+.

Del mismo modo, si se quiere utilizar el modelo IXAmBERT, que se encuentra en Huggingface, es suficiente con incluir las siguientes líneas de código:

```
tokenizer=BertTokenizer.from_pretrained('ixa-ehu/ixambert-base-cased')
model=BertModel.from_pretrained('ixa-ehu/ixambert-base-cased')
```

Publicaciones relacionadas

Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa, Eneko Agirre. (2020). Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque. Proceedings of The 12th Language Resources and Evaluation Conference, pp. 429–435. European Language Resources Association. ISBN: 979-10-95546-34-4

Bibliografía

Aduriz, I., Aranzabe, M. J., Arriola, J. M., Atutxa, A.,Diaz de Ilarraza, A., Garmendia, A., and Oronoz, M. (2003). Construction of a basque dependency treebank.InTreebanks and Linguistic Theories, Vaxjo, Sweden

Agerri R., San Vicente I., Campos J.A., Barrena A., Saralegi X., Soroa A., Agirre E. (2020) Give your Text Representation Models some Love: the Case for Basque. Proceedings of LREC.

Alegria, I., Arregi, O., Ezeiza, N., and Fern ´andez, I.(2006). Lessons from the development of a named en-tity recognizer for Basque.Procesamiento del lenguajenatural, 36:25–37.

Aranzabe, M. J., Atutxa, A., Bengoetxea, K., de Ilarraza,A. D., Goenaga, I., Gojenola, K., and Uria, L. (2015).Automatic conversion of the basque dependency treebank to universal dependencies. In Proceedings of thefourteenth international workshop on treebanks an linguistic theories (TLT14), pages 233–241.

Cañete J., Chaperon G., Fuentes R., Ho J., Kang H., Pérez J. (2020) Spanish Pre-Trained BERT Model and Evaluation Data. PML4DC at ICLR 2020

Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W. t., Choi, Y., Liang, P., and Zettlemoyer, L. (2018). QuAC: Question answering in context. arXiv preprint arXiv:1808.07036.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.(2019). BERT: Pre-training of Deep Bidirectional Trans-formers for Language Understanding. InProceedings ofthe 2019 Conference of the North American Chapter ofthe Association for Computational Linguistics: HumanLanguage Technologies, Volume 1 (Long and Short Pa-pers), pages 4171–4186.

Lample, G. and Conneau, A. (2019). Crosslingual language model pretraining. arXiv preprint arXiv:1901.07291.

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692

Otegi A., Agirre A., Campos J.A., Soroa A., Agirre E. (2020). Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque. Proceedings of The 12th Language Resources and Evaluation Conference, pp. 429–435.

Qu, C., Yang, L., Qiu, M., Croft, W. B., Zhang, Y., and Iyyer, M. (2019). BERT with History Answer Embedding for Conversational Question Answering. In SIGIR ’19.