DeepText

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

Entregable E2.4: Nuevas meta-representaciones vectoriales estáticas multilingües (Recurso).

Responsable	IXA
Tipo	Recurso
Ejercicio	2022

ELKARTEK 2020
Proyectos de investigación fundamental colaborativa
DeepText – Entregable E2.2

Índice

ELKARTEK 2020
Proyectos de investigación fundamental colaborativa
DeepText – Entregable E2.2

Introducción

Este entregable describe representaciones textuales estáticas bilingües/multilingües desarrolladas dentro de la tarea T2.4 (Nuevas representaciones textuales estáticas bilingües/multilingües) dentro del proyecto DeepText. Las representaciones han sido generadas utilizando un nuevo método para construir representaciones bilingües que aminora los problemas de los métodos tradicionales (Ormazabal et al., 2021). Las representaciones están centradas en las lenguas del proyecto, Euskara, Castellano e Inglés. Para todos estos pares se han generado representaciones vectoriales estáticas (embeddings) y se han evaluado su efectividad en la tarea de generación de diccionarios bilingües. Las conclusiones de los experimentos indican que la calidad de los embeddings es muy buena. Estos nuevos embeddings se distribuyen bajo una licencia libre[1]. Como consecuencia, la comunidad PLN del País Vasco dispone de un nuevo recurso que permite desarrollar aplicaciones multilingües en los idiomas co-oficiales del País Vasco, así como el Inglés.

Avanzando así además, en el estado del arte de las aplicaciones de Procesamiento de Lenguaje Natural (PLN) para estas lenguas.

Descripción

Las representaciones de palabras en forma de vectores continuos (embeddings) son necesarias para acometer una gran variedad de aplicaciones dentro del Procesamiento de Lenguaje Natural (PLN). Los embeddings de palabras son vectores de dimensión fija (normalmente entre 100 y 300 dimensiones) que representan a la propia palabra, y cuyo rasgo más característico es que vectores de palabras similares están cerca en el espacio N-dimensional formado por el conjunto de todos los embeddings (Mikolov et al, 2013; Pennington et al. 2014; Bojanowski et al., 2017).

Las representaciones vectoriales de palabras pueden serlo en más de un idioma, de tal forma que palabras de idiomas diferentes que comparten un mismo significado obtienen representaciones vectoriales que estás próximas unas de otras en el espacio multilingüe común. Así, estas representaciones son extremadamente útiles en tareas PLN croslíngües, incluidas la generación automática de diccionarios bilingües.

[1] http://www.deeptext.eus/eu/node/3

ELKARTEK 2020
Proyectos de investigación fundamental colaborativa
DeepText – Entregable E2.2

Nuevo Método para construir representaciones de palabras bilingües

La investigación reciente sobre representaciones de palabras en varios idiomas ha estado dominada por enfoques llamados "*mapping* no supervisados" (Artetxe et al, 2017). Estos métodos requieren de dos pasos. Primeramente, se construyen embeddings monolingües por separado utilizando documentos en cada idioma. Seguidamente, se aprende una función de *mapping*, que que alinean los embeddings monolingües entre sí, dando como resultado un conjunto de embeddings en un espacio bilingüe. Dichos métodos se basan en el principio básico de que la estructura de los espacios monolingües son similares, pero hay varios trabajos que ponen en duda esta suposición (Søgaard et al., 2018;Nakashole and Flauger, 2018; Patra et al., 2019). El nuevo método generado es un enfoque alternativo que no tiene esta limitación. En lugar de alinear dos espacios monolingües pre-entrenados, el método funciona manteniendo los *embedidngs* del idioma de destino fijos, y aprendiendo un nuevo conjunto de embeddings para el idioma de origen, que están alineados con ellos. Es un método no supervisado, que requiere un pequeño diccionario con palabras "semilla" alineadas (por ejemplo, una lista de palabras homógrafas que se escriban de idéntica forma en los dos idiomas, una lísta con números, etc) como la única forma de supervisión.

El método se describe con más detalle en el artículo (Ormazabal et al, 2021) que se presentó en la conferencia más prestigiosa sel área del PLN, el congreso ACL, que según el ranking de la SCIE, tiene una valoración de A++[2].

Nuevas meta-representaciones vectoriales estáticas multilingües

Utilizando el método descrito en el apartado anterior, se han construido nuevos embeddings bilingües para los pares euskara-castellano y euskara-inglés (además de los embeddings que se describen en el artículo). Los embeddings han sido entrenados utilizando un corpus compuesto de textos provenientes de Wikipedia, tanto en castellano (600M de palabras), inglés (1200M de palabras) y euskara (60M de palabras). Siguiendo las indicaciones del artículo, utilizamos el corpus de la lengua con mayores recursos (correspondiente al Castellano e Inglés) como base, es decir, estimamos los embeddings del euskara utilizando los vectores alineándolos a los correspondientes a sus traducciones en la lengua base. Como diccionario "semilla", utilizamos una lista de palabras homógrafas entre cada par de idiomas, así como una lista de 20 traducciones entre ellos. En total, se han calculado embeddings para 200 mil palabras en los tres idiomas. La dimensión de los vectores es 300.

[2]Según los criterios del CENAI en el área de computación, los congresos A++ de SCIE son equivalentes a artículos en revistas del cuartil Q1 según el JCR.

Hemos evaluado las representaciones vectoriales generadas en la llamada tarea BLI (*Bilingual Lexical Induction,* Inducción lexical bilingüe), que consiste en encontrar la correspondiente traducción de una palabra de la lengua fuente a la lengua destino. Data una palabra en la lengua fuente, el método de evaluación consiste en encontrar las palabras en la lengua destino que estén más cercanas a él, utilizando algún criterio basado en las representaciones vectoriales de las palabras en el espacio compartido. Una forma de encontrar estas palabras es utilizar el ángulo formado por los embeddings de las palabras. Sin embargo, ésta técnica, llamada de "vecinos más cercanos", es conocida por dar unos malos resultados en tareas BLI, principalmente debido al fenómeno llamado *hubness:* la tendencia de ciertas palabras de estar cerca de otras muchas en el espacio N-dimensional. Por ello, en esta evaluación utilizamos el método llamado CLSL y presentado en (Conneau et al., 2018), que atenúa el efecto del *hubness* en los resultados.

El método lo evaluamos utilizando dos diccionarios bilingües ES-EU y EN-EU, y comprobando si entre las alternativas propuestas por nuestro algoritmo está la palabra correcta. La métrica utilizada es la llamada P@k, que dice si la traducción correcta se encuentra dentro de los K primeras alternativas propuestas por nuestro método.

Resultados

Los resultados obtenidos son los siguientes:

	P@1	P@10	P@20
ES-EU	**27.03%**	**61.05%**	**65.99%**
EN-EU	**16.12%**	**37.4%**	**44.21%**

Tabla 1: Resultados

Se puede observar que los resultados en P@1 no son especialmente buenos, pero que la efectividad del sistema aumenta drásticamente si se consideran las primeras 10 alternativas sugeridas por el sistema. Los resultados para el par ES-EU son sensiblemente mejores que el equivalente EN-EU, pero estos resultados pueden ser explicados por dos motivos principales. Por un lado, la proximidad mayor del castellano al euskara, en comparación con el Inglés. Esto hace que haya muchas más palabras homógrafas compartidas entre los dos idiomas, lo que es un requisito importante de nuestro método. Por otro lado, es de esperar que el diccionario usado en la evaluación BLI ES-EU sea de una calidad sensiblemente mayor que el EN-EU, ya

que éste primero cuenta con una mayor tradición lexicográfica. En todo caso, los resultados presentados son los mejores obtenidos hasta la fecha para embeddings bilingües para estos pares de idiomas.

Descarga de los sistemas

Los representaciones estáticas bilingües ES-EN y ES-EU están accesibles en la siguiente página web.

http://www.deeptext.eus/resources

Publicaciones asociadas al trabajo

Aitor Ormazabal, Mikel Artetxe, Aitor Soroa, Gorka Labaka, and Eneko Agirre. 2021. Beyond Offline Mapping: Learning Cross-lingual Word Embeddings through Context Anchoring. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6479–6489, Online. Association for Computational Linguistics.

Bibliografía

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In Proceedings of the 55th Annual Meeting of the Association for Computational Lin- guistics (Volume 1: Long Papers), pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

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.

Alexis Conneau, Guillaume Lample, Marc-Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In Proceedings of the 6th International Conference on Learning Representations (ICLR 2018).

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

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.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. 2019. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3336–3341, Hong Kong, China. Association for Computational Linguistics.

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

Ndapa Nakashole and Raphael Flauger. 2018. Characterizing departures from linearity in word translation. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 221–227, Melbourne, Australia. Association for Computational Linguistics.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, and Graham Neubig. 2019. BLISS in non-isometric embedding spaces.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics.