CEUR-WS.org/Vol-2664/cantemist_paper8.pdf

Deep Neural Model with Contextualized-word Embeddings for Named Entity Recognition in Spanish Clinical Text

Renzo Rivera-Zavala[a], Paloma Martinez[a]

[a]Carlos III University of Madrid, Avda. Universidad, 30, 28911 Leganés, Madrid

Abstract

In this work, we introduce a Deep Learning architecture for concepts related to cancer Named Entity Recognition in Spanish clinical cases texts. We propose a deep neural approach based on two Bidirectional Long Short-Term Memory (Bi-LSTM) network and Conditional Random Field (CRF) network using character and contextualized-word embeddings to deal with the extraction of semantic, syntactic and morphological features. The approach was evaluated on the CANTEMIST Corpus obtaining an F-measure of 82.3% for NER.

Keywords

Natural Language Processing, Clinical Texts, Deep Learning, Contextual Information

1. Introduction

Currently, the number of biomedical literature is growing at an exponential rate. Therefore, the efficient access to information on biological, chemical, and biomedical data described in scientific articles, patents, or e-health reports is a growing interest in the biomedical industry, research, and so forth. In this context, improved access to biomedical concept mentions in biomedical texts is a crucial step prior to downstream tasks such as drug and protein interactions, chemical compounds, adverse drug reactions, among others.

Named Entity Recognition (NER) is a crucial task in biomedical Information Extraction (IE), intending to automatically extract and identify mentions of concepts of interest in running text, typically through their mention offsets or by classifying individual tokens whether they belong to entity mentions or not. There are different approaches to address the NER task. Dictionary-based methods, which are limited by the size of the dictionary, spelling errors, the use of synonyms, and the constant growth of vocabulary. Rule-based methods and Machine Learning methods usually require both syntactic and semantic features as well as specific language and domain features. One of the most effective methods is Conditional Random Fields (CRF) [1] since CRF is one of the most reliable sequence labeling methods. Recently, deep learning-based methods have also demonstrated state-of-the-art performance for English [2, 3, 4] texts by automatically learning relevant patterns from corpora, which allows language and

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: renzomauricio.rivera@alumnos.uc3m.es (R. Rivera-Zavala); pmf@inf.uc3m.es (P. Martinez)
ORCID: 0000-0002-0877-7063 (R. Rivera-Zavala)

© 2020 Copyright 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

IberLEF 2020, September 2020, Málaga, Spain.
CEUR Workshop Proceedings (CEUR-WS.org)

domain independence. However, pre-trained models in Spanish and even more in the biomedical domain are limited, for to the best of our knowledge there is only one public available work that addresses the generation of Spanish biomedical word embeddings [5, 6].

In this paper, we propose a deep neural model with two Bi-LSTM layers and a CRF layer. To do this, we adapt the NeuroNER model proposed in [7] for track 1 (NER offset and entity classification) of the CANTEMIST task [8]. Specifically, we have extended NeuroNER by adding contextualized-word information and information about overlapping or nested entities. Moreover, in this work, we use an existing pre-trained contextualized-word model as well as our train from scratch contextualized-word model: i) a word2vec/FastText Spanish Billion Word Embeddings models [9], which were trained on the 2014 dump of Wikipedia ii) our medical word embeddings for Spanish trained using the FastText model and iii) a sense-disambiguation embedding model [10].

Experiment results on CANTEMIST tasks showed that our features representation improved each of separate representations, implying that LSTM-based compositions play different roles in capturing token-level features for NER tasks, thus making improvements in their combination. Moreover, the use of specific domain contextualized-word vector representations outperforms general domain word vector representations.

2. Materials and Methods

In this section, we first describe the corpora used to generate our train from the scratch contextualized-word representation, the training procedure and the pre-trained contextualized-word models used in our study. Then, we describe our system architecture for offset and entity classification. Finally, the datasets used for training, validating, and evaluating our deep learning model performance.

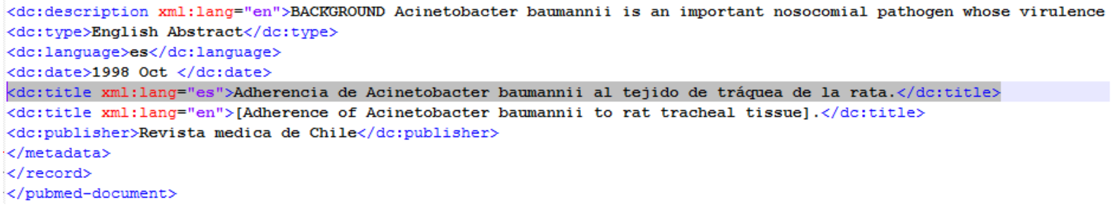
2.1. Corpora

The corpora were gathered from Spanish biomedical texts from different multilingual biomedical sources:

1. The Spanish Bibliographical Index in Health Sciences (IBECS - http://ibecs.isciii.es) corpus that collects scientific journals covering multiple fields in health sciences,
2. Scientific Electronic Library Online (SciELO - https://scielo.org/es/) corpus gathers electronic publications of complete full-text articles from scientific journals of Latin America, South Africa, and Spain,
3. MedlineNLM corpus obtained from the PubMed free search engine (https://www.ncbi.nlm.nih.gov/pubmed/),
4. The MedlinePlus corpus (an online information service provided by the U.S. National Library of Medicine - https://medlineplus.gov/), consists of Health topics, Drugs and supplements, Medical Encyclopedia and Laboratory test information, and
5. The UFAL corpus (https://ufal.mff.cuni.cz/ufal_medical_corpus) is a collection of parallel corpora of medical and general domain texts.

Table 1
Biomedical Spanish corpus details.

Collection\Corpus	IBECS	SciELO	MedlineNLM	MedlinePlus	UFAL
Documents	168,198	161,710	330,928	1,063	265,410
Words	23,648,768	26,169,655	4,710,191	217,515	41,604,517
Unique Words	184,936	159,997	20,942	5,099	198,424



```xml
<dc:description xml:lang="en">BACKGROUND Acinetobacter baumannii is an important nosocomial pathogen whose virulence
<dc:type>English Abstract</dc:type>
<dc:language>es</dc:language>
<dc:date>1998 Oct </dc:date>
<dc:title xml:lang="es">Adherencia de Acinetobacter baumannii al tejido de tráquea de la rata.</dc:title>
<dc:title xml:lang="en">[Adherence of Acinetobacter baumannii to rat tracheal tissue].</dc:title>
<dc:publisher>Revista medica de Chile</dc:publisher>
</metadata>
</record>
</pubmed-document>
```

Figure 1: Dublin core format for biomedical corpus.

Source corpus details are described in Table 2.1.

All the corpora are in XML (Dublin core format) and TXT format files. XML files were processed for extract only raw text from specific XML tags such as "title" and "description" from Spanish labels, based on the Dublin Core format as shown in Figure 1. TXT files were not processed. Raw texts from all files were compiled in a single TXT file. Texts were processed, setting all to lower, removing punctuations, trailing spaces and stop words and used as input to generate our word embeddings. Sentences preprocessing (split and tokenized) were made using Spacy [1], an open-source python library for advanced multi-language natural language processing.

2.2. Contextualized-word models

The use of word representations from pre-trained unsupervised methods is a common practice and a crucial step in NER pipelines. Previous word embedding models such as Word2Vec [11], Glove [12], and FastText [13] focused on context-independent word representations. However, in the last few years models are focused on learning context-dependent word representations, such as ELMo [14], CoVe [15], and the state-of-the-art BERT model [16].

BERT is a context-dependent word representation model based on a masked language model and trained using the transformer architecture [16]. Previous models such as RNN (LSTM GRU) combines two unidirectional layers (i.e., Bi-LSTM) to address the sequential nature of natural language, as a replacement for the sequential approach the BERT model employs a much faster attention-based approach. BERT is pre-trained in two unsupervised "artificial" tasks: i) masked language modeling that predicts randomly masked words in a sequence, and hence can be used for learning bidirectional representations by jointly conditioning on both left and right contexts in all layers and ii) next sentence prediction in order to train a model that understands sentence

[1] https://spacy.io/

Table 2
Contextualized-word models details.

Detail	**BSC-BERT**	**bert-base-multilingual-cased**	**BETO cased**
Language	Spanish	104 languages	Spanish
Domain	Biomedical	General	General
Type	Contextual Word	Contextual Word	Contextual Word
Corpus size	6 billion	3,300M	3 billion
Vocab size	200k	120k	31k
Hidden size	768	768	1024
Algorithm	BERT train	BERT train	BERT train

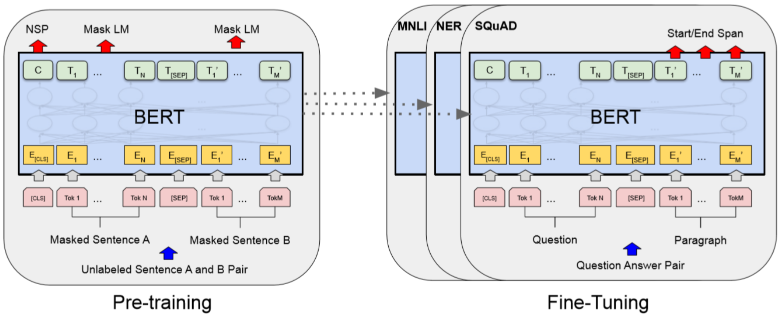


Figure 2: BERT pre-training and fine-tuning architecture overview (source [16]).

relationships. The transformer layer has two sub-layers: a multi-head self-attention mechanism, and a position-wise fully connected feed-forward network, followed by a normalization layer. Even though BERT learns a lot about language through pre-training, it is possible to adapt the model by adding a customized layer on top of BERT outputs and then new training is done with specific data (this phase is called fine-tuning). We refer readers [16] for a more detailed description of BERT. An overview of the BERT architecture can be seen in Figure 2.

Due to the benefits of the BERT model, we adopted the multilingual cased [16] and the BETO [17] pre-trained BERT models. Moreover, we trained from scratch a Biomedical Spanish model (BSC-BERT) with 12 transformer layers (12-layer, 768-hidden, 12-heads, 110Mparameters) and a SoftMax output layer to perform the NER task. First, we replace the WordPiece tokenizer with the SentencePiece implementation [18] and the Spacy [19] Spanish tokenizer for sentence and subword segmentation. We train with a batch size of 128 sequences for 1,000,000 steps, which is approximately 40 epochs over the 4 million word corpus. We use Adam with a learning rate of 1e-4. We use a dropout probability of 0.15 on all layers and a gelu activation function. Training of BSC-BERT was performed on 1 Cloud TPU, 8vCPUs Intel(R) Xeon(R) CPU @ 2.30GHz and 16GB memory. Details of the train and pre-trained models can be seen in Table 2.2.

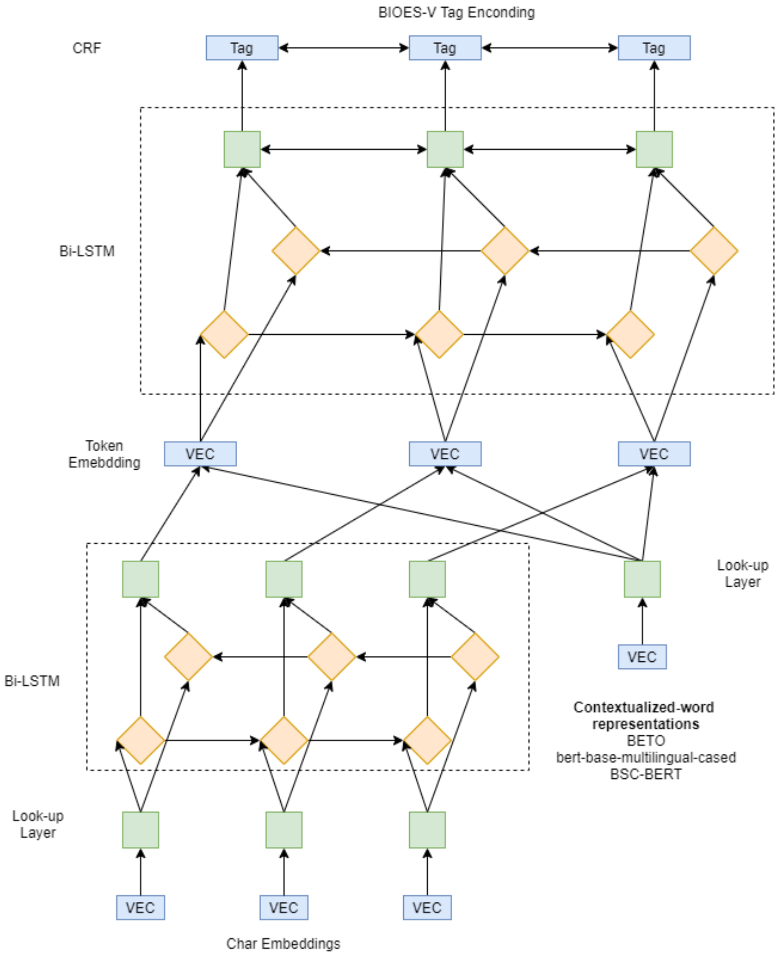


Figure 3: The architecture of the Bi-LSTM CRF model for tumor morphology identification.

2.3. System Description

Our approach involves the adaption of a state-of-art NER model named NeuroNER as proposed in [7], based on a deep learning network with a preprocess step, learning transfer from pre-trained models, two recurrent neural network layers and a last layer for CRF (see Figure 3). The input for the first Bi-LSTM layer is character embeddings. In the second layer, we concatenate character embeddings from the first layer with contextualized-word representations for the second Bi-LSTM layer. Finally, the last CRF layer obtains the most suitable labels for each token using a tag encoding format. For more details about NeuroNER, please refer to [7].

Our contribution consists of extending the NeuroNER system with additional features. In particular, adding contextualized-word representations and the extended BMEWO-V encoding format has been added to the network.

The BMEWO-V encoding format distinguishes the B tag for entity start, the M tag for entity continuity, the E tag for entity end, the W tag for a single entity, and the O tag for other tokens

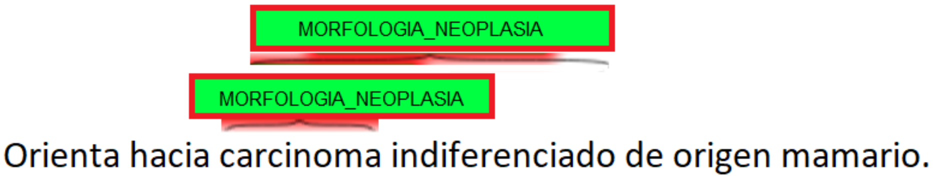


Figure 4: BRAT annotation example from CANTEMIST corpus sentence.

Table 3
Tokens annotated with BMEWO-V encoding in the ConLL-2003 format.

token	offset start	offset end	nested-tag	tag
Orienta	0	6	O	O
hacia	8	13	O	O
carcinoma	15	25	V-MORFOLOGIA_NEOPLASIA	B-MORFOLOGIA_NEOPLASIA
indiferenciado	27	32	O	E-MORFOLOGIA_NEOPLASIA
de	33	35	O	O
origen	36	42	O	O
mamario	43	50	O	O
.	51	52	O	O

that do not belong to any entity. The V tag allows us to represent nested entities. BMEWO-V is similar to other previous encoding formats [20]; however, it allows the representation of nested and discontinuous entities. As a result, we obtain our sentences annotated in the CoNLL-2003 format [21]. An example of the BMEWO-V encoding format applied to the sentence "Orienta hacia carcinoma indiferenciado de origen mamario." ("Orients towards undifferentiated carcinoma of mammary origin.") can be seen in Figure 4 and Table 3.

2.3.1. First Bi-LSTM layer using character embeddings

Word embedding models are able to capture syntactic and semantic information. However, other linguistic information, such as morphological information and orthographic transcription are not exploited. According to [22], the use of character embeddings improves learning for specific domains and is useful for morphologically rich languages such as Spanish. For this reason, we decided to include the character-level representations to obtain morphological and orthographic information from words. Each word is decomposed into its character n-grams and initialized with a random dense vector which is then learned. We used a 25-feature vector to represent each character. In this way, tokens in sentences are represented by their corresponding character embeddings, which are the input for our Bi-LSTM network.

Table 4
CANTEMIST subsets details.

Dataset	Subset	Documents	Sentences	Entities
CANTEMIST	Train	501	19359	6529
	Valid	250	9489	3246
	Test	5231	88406	

2.3.2. Second Bi-LSTM layer using word and Sense embeddings

The input for the second Bi-LSTM layer is the concatenation of character embeddings from the first layer with the pre-trained contextualized-word representations of the tokens in a given input sentence. The second layer goal is to obtain a sequence of probabilities for each tag in the BMEWO-V encoding format. In this way, for each input token, this layer returns six probabilities (one for each tag in BMEWO-V). The final tag should be with the highest probability for each token.

2.3.3. Last layer based on Conditional Random Fields (CRF)

To improve the accuracy of predictions, a Conditional Random Field (CRF) [1] model is trained, which takes as input the label probability for each independent token from the previous layer and obtains the most probable sequence of predicted labels based on the correlations between labels and their context. Handling independent labels for each word shows sequence limitations. For example, considering the drug sequence labeling problem, an "I-MORFOLOGIA-NEOPLASIA" tag cannot be found before a "B-MORFOLOGIA-NEOPLASIA" tag or a "B-MORFOLOGIA-NEOPLASIA" tag cannot be found after an "I-MORFOLOGIA-NEOPLASIA" tag. Finally, once tokens have been annotated with their corresponding labels in the BMEWO-V encoding format, the entity mentions must be transformed into the BRAT format. V tags, which identify nested or overlapping entities, are generated as new annotations within the scope of other mentions.

3. Evaluation

As it was described above, our system is based on a deep network with two Bi-LSTM layers and the last layer for CRF. We evaluate our NER system using the train, validation, and test subsets from the CANTEMIST dataset provided by the CANTEMIST task organizers [8]. Detailed information for each subset can be seen in Table 4. The CANTEMIST dataset is a manually annotated corpus of 5,982 clinical cases written in Spanish and annotated with mentions of concepts related to cancer, namely tumor morphology. CANTEMIST NER task can be addressed as a binary classification task only considering one entity type "MORFOLOGIA-NEOPLASIA" ("tumor morphology").

The CANTEMIST task considers three subtasks. In this work, we address the first track considering offset recognition and entity classification of tumor morphology. The F-measure is used as the main metric where true positives are entities which match with the gold standard entity boundaries and type. A detailed description of evaluation can be found in the CANTEMIST

Table 5
Contextualized-word models details.

Detail	BSC-BERT	bert-base-multilingual-cased	BETO cased
Language	Spanish	104 languages	Spanish
Domain	Biomedical	General	General
Type	Contextual Word	Contextual Word	Contextual Word
Corpus size	6 billion	3,300M	3 billion
Vocab size	200k	120k	31k
Hidden size	768	768	1024
Algorithm	BERT train	BERT train	BERT train

Table 6
Contextualized-word models out of the task results for entity classification on CANTEMIST test subset.

Model	Precision (%)	Recall (%)	F-Score (%)
bert-base-multilingual-cased	80.41	76.78	78.55
BETO cased	84.68	79.02	81.66
BSC-BERT	**83.59**	**81.63**	**82.60**

web [2].

The NER task is addressed as a sequence labeling task. For the NER track, we tested different configurations with various pre-trained contextualized-word models. The pre-trained models and their parameters are summarized in Table 5.

In Table 6, we compare the different pre-trained models on the validation subset. As shown in Table 6 specific domain contextualized-word models outperform general domain models by almost 1 points. For the test subset, we applied our best system configuration NeuroNER + BSC-BERT obtaining an f-score of 82.30% for offset detection and entity classification. Despite we predict results for more than 5K documents, the evaluation is only performed in 300 of them.

4. Conclusions

In this work, we propose a system for the detection of concepts related to cancer in clinical narrative written in Spanish. We address the named entity recognition task as a sequence labeling task. We proposed a deep learning approach only using dense vector representations features instead of hand-crafted word-based features. We proved that as in other tasks such as NER, the use of dense representation of words such as word-level and character-level are helpful for named entity recognition. The proposed system achieves satisfactory performance with F-score over 82.3% for the test subset. The extension of NeuroNER network is domain-independent and could be used in other fields, although generic pre-trained contextualized-word

[2] https://temu.bsc.es/cantemist/?p=3975

representations are used, new pre-trained biomedical Spanish contextualized-word model has been generated for this work.

We found that the text preprocessing step had a significant impact on the entity offset recognition and classification. Separating words by the hyphen '-' caused some errors. Abbreviation recognition is a difficult task due to ambiguity and length, even more for very short abbreviations (1-2 letters) due to their high level of ambiguity. Long entities consisting of more than 5 tokens are hard to identify correctly. Moreover, words not present in the pre-trained model's vocabulary are not recognized in entity offset recognition and classification; therefore, we decided to initialize words not present in the vocabulary with random vectors.

As future work, we plan to generate contextualized-word representations integrating biomedical knowledge into our systems such as SNOMED-CT or UMLS. The motivation would be to see whether contextualized-word representations generated with biomedical knowledge can help to improve the results and provide a deep learning model for biomedical NER and concept indexing.

Acknowledgments

This work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain (DeepEMR project TIN2017-87548-C2-1-R).

References

[1] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, pp. 282–289. URL: http://dl.acm.org/citation.cfm?id=645530.655813.

[2] W. Hemati, A. Mehler, Lstmvoter: chemical named entity recognition using a conglomerate of sequence labeling tools, Journal of Cheminformatics 11 (2019) 3. URL: https://doi.org/10.1186/s13321-018-0327-2. doi:10.1186/s13321-018-0327-2.

[3] M. Pérez-Pérez, O. Rabal, G. Pérez-Rodríguez, M. Vazquez, F. Fdez-Riverola, J. Oyarzábal, A. Valencia, A. Lourenço, M. Krallinger, Evaluation of chemical and gene/protein entity recognition systems at biocreative v.5: the cemp and gpro patents tracks, 2017.

[4] V. Suárez-Paniagua, R. M. R. Zavala, I. Segura-Bedmar, P. Martínez, A two-stage deep learning approach for extracting entities and relationships from medical texts, Journal of Biomedical Informatics 99 (2019) 103285. URL: http://www.sciencedirect.com/science/article/pii/S1532046419302047. doi:https://doi.org/10.1016/j.jbi.2019.103285.

[5] M. M. K. M. Armengol-Estapé Jordi, Soares Felipe, Pharmaconer tagger: a deep learning-based tool for automatically finding chemicals and drugs in spanish medical texts, Genomics Inform 17 (2019) e15–. URL: http://genominfo.org/journal/view.php?number=557. doi:10.5808/GI.2019.17.2.e15. arXiv:http://genominfo.org/journal/view.php?number=557.

[6] F. Soares, M. Villegas, A. Gonzalez-Agirre, M. Krallinger, J. Armengol-Estapé, Medical word embeddings for Spanish: Development and evaluation, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 124–133. URL: https://www.aclweb.org/anthology/W19-1916.

[7] F. Dernoncourt, J. Y. Lee, P. Szolovits, NeuroNER: an easy-to-use program for named-entity recognition based on neural networks, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 97–102. URL: https://www.aclweb.org/anthology/D17-2017. doi:10.18653/v1/D17-2017.

[8] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.

[9] C. Cardellino, Spanish Billion Words Corpus and Embeddings, 2016. URL: http://crscardellino.me/SBWCE/.

[10] A. Trask, P. Michalak, J. Liu, sense2vec - A fast and accurate method for word sense disambiguation in neural word embeddings, CoRR abs/1511.06388 (2015). URL: http://arxiv.org/abs/1511.06388. arXiv:1511.06388.

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, CoRR abs/1310.4546 (2013). URL: http://arxiv.org/abs/1310.4546. arXiv:1310.4546.

[12] J. Pennington, R. Socher, C. Manning, Glove: Global Vectors for Word Representation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014) 1532–1543. URL: http://aclweb.org/anthology/D14-1162. doi:10.3115/v1/D14-1162. arXiv:1504.06654.

[13] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, CoRR abs/1607.04606 (2016). URL: http://arxiv.org/abs/1607.04606. arXiv:1607.04606.

[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, volume 1, Association for Computational Linguistics (ACL), 2018, pp. 2227–2237. doi:10.18653/v1/n18-1202. arXiv:1802.05365.

[15] B. McCann, J. Bradbury, C. Xiong, R. Socher, Learned in translation: Contextualized word vectors, Technical Report, 2017. arXiv:1708.00107.

[16] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, Technical Report, 2019. URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[17] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained bert model and evaluation data, in: to appear in PML4DC at ICLR 2020, 2020.

[18] T. Kudo, J. Richardson, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, 2018. arXiv:1808.06226.

[19] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.

[20] A. Borthwick, J. Sterling, E. Agichtein, R. Grishman, Exploiting diverse knowledge sources via maximum entropy in named entity recognition, in: Sixth Workshop on Very Large Corpora, 1998. URL: https://www.aclweb.org/anthology/W98-1118.

[21] E. F. Tjong Kim Sang, F. De Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 142–147. URL: https://www.aclweb.org/anthology/W03-0419.

[22] W. Ling, C. Dyer, A. W. Black, I. Trancoso, R. Fermandez, S. Amir, L. Marujo, T. Luis, Finding function in form: Compositional character models for open vocabulary word representation, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 1520–1530. URL: https://www.aclweb.org/anthology/D15-1176. doi:10.18653/v1/D15-1176.