







Extraction of Entities in Health Domain Documents Using Recurrent Neural Networks

Erick Barrios González¹ , Mireya Tovar Vidal¹ ,
Guillermo De Ita Luna¹ , and José A. Reyes-Ortiz² 

¹ Faculty of Computer Science, Benemerita Universidad Autonoma de Puebla,
14 sur y Av. C.U., San Claudio, Puebla, Mexico

erick.barrios@alumno.buap.mx, mireya.tovar@correo.buap.mx

² Universidad Autonoma Metropolitana, Av. San Pablo Xalpa 180, Azcapotzalco,
02200 Mexico City, Mexico

jaro@azc.uam.mx

Abstract. This paper reviews the subtask “A” of *eHealth-KD challenge 2021*, using a strategy oriented to the use of recurrent neural networks. In subtask “A” the entities are identified by document and their types (Concepts, actions, predicates, and references). This paper mainly compares various word embedding models, annotation styles in BIO, BILOU, and an own annotation style to distinguish entities composed of more than one word from entities of a single word. Is also proposed a solution based on POS tagging to improve the results of systems that use recurrent neural networks by changing the pre-processing. An evaluation process was performed with the following well-known metrics: precision, recall, and F₁.

Keywords: Entity extraction · Bidirectional long short-term memory · Natural language processing

1 Introduction

Natural language processing (NLP) methods are increasingly being used to extract knowledge from unstructured health texts. The organization of medical information can be helpful in clinical analyses, reduce the number of medical errors, or help make more appropriate decisions in some instances. Currently, it is easier to have medical information in electronic format and even to have that information structured. However, information in traditional media (such as publications, academic manuscripts, clinical reports) is of little use in selecting the most appropriate information in each case, whether in the clinical or research setting. Therefore, it exists the necessity for creating new ways of extracting knowledge from health texts and for the information to be comprehensive and reliable for any study or analysis to be carried out with that information.

Therefore, subtask “A” of *eHealth-KD challenge 2021* proposes identifying entities of a document and their respective types [13]. These entities are all relevant terms (of a single word or multiple words) representing semantically essential elements in a sentence. There are four types of entities:

- Concept: Identify a term, concept, relevant idea in the domain of knowledge of sentence.
- Action: Identify a process or modification of other entities. It can be indicated by a verb or verbal construction or by nouns.
- Predicate: Identifies a function or filter from a set of elements, which has a semantic label in the text, and is applied to an entity with some additional arguments.
- Reference: Identifies a textual element that refers to an entity (of the same sentence or a different one).

In Fig. 1 there are examples of how entities would be classified into concepts, actions, predicates, and references. For example, “asma” (asthma), “enfermedad” (disease), “vías respiratorias” (respiratory tract) are concepts, the word “afecta” (affects) is an action, the word “esta” (this) is a reference, and “mayores” (older) is a predicate.



Fig. 1. Entity recognition example [13].

This paper explores two strategies, a solution based in a own annotation style to distinguish entities composed of more than one word from entities of a single word, and a solution based on POS tagging to improve performance in learning models based on recurrent neural networks to identify entities in a corpus of the health domain.

The document is structured as follows: Sect. 2 presents the works related to this research. Section 3 shows the proposed solution for the identification of entities. Section 4 shows the results obtained. Finally, Sect. 5 contains the conclusion.

2 Related Work

The following describes work related to the identification of entities mainly using recurrent neuronal networks:

In [8] an entity identification model for any domain in general is presented. This model combines a bidirectional LSTM network and conditional random fields (BiLSTM-CRF) adding “Multi-Task Learning” (MTL) and a framework “called Mixture of Entity Experts” (MoEE). This work uses information “from CoNLL-2003 English NER (Named entity recognition)”, the entities found are classified into 4 types: people, locations, organizations and miscellaneous. Experiments were made with different word embeddings; the model used FastText (freeze settings) model and obtained a 69.53% with the measure F_1 .

In *eHealth-KD challenge 2019* in [5] a BILOU annotation scheme is used. The architecture inputs are represented by vectors, POS tag embedding and word embedding respectively, two types of rules have been applied to the result of deep learning: The first set of rules is aimed at correcting frequent errors expanding or reducing the scope of a detected keyword phrase, or modifying its type.

Also in *eHealth-KD challenge 2019* in [1] a system is proposed with a strategy that uses BiLSTM with a CRF (Conditional random fields) layer for the identification of entities. In addition, domain-specific word embeds are used. For each token that the sentence was divided into, the input for that token consists of a list of three feature vectors: Character encodings, POS tagging vectors, and word indexes.

In [2] an architecture based on a BiLSTM with a CRF classifier was proposed. The corresponding classes are encoded in BIO annotation style.

During *eHealth-KD challenge* in 2020, [12] uses the BMEWO-V tag system and BiLSTM layers as contextual encoders. The label “None” is included in the latter for cases where there is no entity present.

Also in *eHealth-KD challenge 2020* in [4] the proposed system uses the BILUOV tagger that uses a character encoding layer to transform the representation of the characters of each token into a single vector that captures the morphological dependencies of the token. The character encoding layer consists of an embedded layer, followed by a BiLSTM layer. Four independent instances are trained with this architecture, each one corresponding to a type of entity.

Finally, in *eHealth-KD challenge* in 2021, in the system proposed in [11] it obtains fourth place with a solution based on BiLSTM aimed at embedding words and characters; it also makes use of information POS of spacy and the BILUOV tagger.

3 Proposed Solution

The proposed solution for the detection of entities, of subtask “A” of *eHealth-KD challenge*, is divided into two stages described below.

3.1 Information Pre-processing

This section shows the basic structure for pre-processing, the description of the corpus, the description of the word embedding models, and the description of the tagging patterns used.

Pre-processing Structure: For the pre-processing of the information, the following steps have been considered:

1. Cleaning the text: In this step the corpus provided is divided into several sentences and irrelevant signs are eliminated.
2. Tagging: The sentences from the previous step are tokenized and their POS tagging is added to each word.
3. Annotation style: In this step, a vector is created for each annotation style, as well as vectors of the additional information attached.
4. Information vectors: The vectors of words and characters are converted into vectors of numbers, a dictionary of terms is created, where each term has its equivalent in number.
5. Word Embedding Model: This model receives all the words, tags, and expressions before they are vectorized to convert the information into an array of vectors.

Tagging: Grammar tagging (part-of-speech tagging) complements the information in the text; for this, it is essential to consider that grammatical tagging is usually different for Spanish and English and depends on the tool used to tag. That is why two spacy libraries will be used, one for English (`in_core_web_lg` of 714 mb) and another for Spanish (`is_core_news_lg` of 542 mb), but they have the same tagging format.

In order to improve the performance of the systems, it is proposed to attach additional information using POS tagging.

From the training corpus provided, content is made of the appearance of each POS tag as an entity. For example, there are 342 nouns as entities within all the sentences in the training body (nouns appear most often as entities within the text).

For each entity that consists of more than one word, its occurrences are counted according to the POS tagging pattern. For example, in the entity “asilos de ancianos” (nursing homes), the pattern would be “NOUN ADP NOUN.” An example of how the frequency of these patterns is captured is shown in Table 1.

Table 1. Frequency of appearance of entities (made up of more than one word) based on POS tagging.

POS tagging pattern	Frequency
<i>NOUN ADJ</i>	33
<i>PROPN PROP</i>	9
<i>NOUN ADP NOUN</i>	8
<i>NOUN ADP PROP</i>	6
<i>PROPN PROP PROP</i>	3

Finally, with the POS tagging and the patterns found, it is intended to build a proposal for a system where words are classified according to the frequency of their appearance.

Annotation and Proposal Styles: For the annotation styles, the following styles have been considered: BIO (Beginning, Inside, Outside), BILOU (Beginning, Inside, Last, Outside, Unique) [7] and an own style annotation aimed at differentiating entities that consist of more than one word from those that do not. To implement the annotation style, we consider the classification of the entities (concepts, actions, predicates, and references) and add the label “None” for the words in the text that do not belong to any group of entities mentioned above.

Table 2. Example of own annotation style.

Token	Tag
Algunos	Predicate
asilos	P_Concept
de	P_Concept
ancianos	P_Concept
cuentan	O_
con	O_
unidades	Concept
de	O_
cuidados	Action

The own annotation style proposed is shown in Table 2, which we will call the annotation style “P”, the labels that begin with the prefix “P_” refer to the entities that are made up of more than one word, “O_” for words that are not entities and no prefix for words that are single word entities. As mentioned, this annotation style is aimed at differentiating entities that are made up of more than one word that are not, and gives a simpler approach than the BIO and BILOU annotation style provide.

Information Vectors: The information on the frequency of the POS tagging was manually classified into 6 ranges: very high frequency (0.75–1), high frequency (0.5–0.75), medium frequency (0.35–0.5), low frequency (0.1–0.35), very low frequency (0–0.1) and no frequency (0). The classification is obtained from the probability of occurrence of the patterns found. It is calculated within a range of 0 to 1, the frequency of appearance of a tag is divided into the frequency of appearance of that tag recognized as an entity. For example, if 678 nouns appear and only 534 times those nouns are entities, the words that are tagged as a noun

have a probability of appearing of $(534/678) = 0.78$ and a probability of 0 if not never appears as an entity. For entities made up of more than one word, sequences that fulfill previously established patterns were searched and subsequently classified as previously shown using the BIO (prefixes B-, I-, O-) or BILOU (prefixes B-, I-, L-, O-, U-) annotation style. In Table 3, we can see how the sequence of words would look, with the annotation style added.

Table 3. Example of tag sequence created with annotation style BIO and BILOU.

Word	POS	BIO	BILOU
algunos	DET	B_LOW_FRECUENCY	U_LOW_FRECUENCY
asilos	NOUN	B_LOW_FRECUENCY	B_LOW_FRECUENCY
de	ADP	I_LOW_FRECUENCY	I_LOW_FRECUENCY
ancianos	NOUN	I_LOW_FRECUENCY	L_LOW_FRECUENCY
cuentan	VERB	B_MEDIUM_FRECUENCY	U_MEDIUM_FRECUENCY
con	ADP	B_VERYLOW_FRECUENCY	U_VERYLOW_FRECUENCY

3.2 Identification of Entities

For the identification of entities, it is proposed to use two systems, one BiLSTM with CNN and the other BiLSTM with CRF.

In Fig. 2, we can see how the pre-processing output files are the input for the entity identification system; each system will return a vector of numbers as a result, which must be converted using the term dictionaries to get the final file in BRAT format. For the training and prediction of the network, we have three different vectors; these are considered the “X” axis: Vector of words, vector POS and vector probabilities POS. In contrast, the vector with the annotation styles that contains the entity classification for training is the “Y” axis.

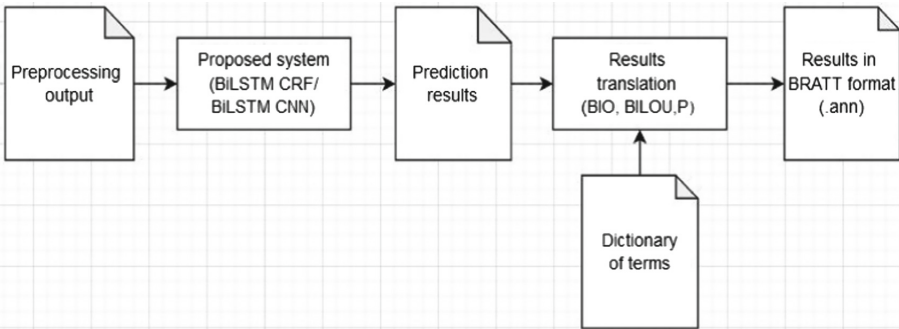


Fig. 2. Steps to identify entities.

Table 4. Format example *BRAT standoff*.

T1	Concept	3 10; 11 19	sistema vascular
T2	Predicate	26 29	red
T3	Concept	33 38; 39 49	vasos sanguíneos
T4	Concept	54 60	cuerpo

4 Results

4.1 Description of the Corpus

The corpus provided by *eHealth Challenge 2021* will be used. These corpora are classified mainly in 2 ways:

The corpus provided by *eHealth Challenge 2021* will be used for the pre-processing part. These corpora are classified mainly in 2 ways:

- Training (For system training): Made up of 100 sentences and approximately 822 distinct words in spanish (vocabulary).
- Testing (For system test): Made up of 100 sentences, 50 in english and 50 in spanish, with approximately a vocabulary with approximately 531 distinct words in spanish and approximately 607 distinct words in english.

Each corpus comprises two files, one “.ann” that contains the answers, and another “.txt” that contains the text to be treated. The file “.txt” is organized by sentences, for example, “El sistema vascular es la red de vasos sanguíneos del cuerpo.” As shown in Table 4, the entities and relationships are correctly classified in BRAT standoff format.

4.2 Pre-processing Results

This section shows the results obtained from the pre-processing stage:

The training corpus (training) provided has 728 entities (464 concept type, 73 predicate type, 18 reference type, and 173 action type); all entities are in Spanish. For the evaluation corpus (testing), we have a total of 934 entities, of which 451 are in Spanish (323 concept type, 37 predicate type, 6 reference type, and 85 action type) and 483 are in English (364 concept type, 63 predicate type, two reference type, and 54 action type).

4.3 Results: Word Embedding

Several embedding patterns as well as various text have been compiled in order to create a corpus and various word embedding patterns. A Wikipedia library for Python was used to create this corpus, allowing searching and saving the information in texts. In Table 5, different features are shown, such as the number of words used for their training, dimensions, the language, and the tool used for creating the model.

Table 5. Word embedding models.

Tool	Algorithm	Words	Dimension	Language	Source
<i>Wor2vec.</i>	<i>Skip-gram</i>	20 million	490	Spanish	Created
<i>Wor2vec.</i>	<i>Cbow</i>	20 million	490	Spanish	Created
<i>FastText.</i>	<i>Skip-gram</i>	20 million	490	Spanish	Created
<i>FastText.</i>	<i>Skip-gram</i>	40 million	360	Spanish/English	Created
<i>FastText.</i>	<i>Skip-gram</i>	40 million	700	Spanish/English	Created
<i>FastText.</i>	<i>Cbow</i>	600 billion	300	English	<i>FastText</i> [10]
<i>FastText.</i>	<i>Skip-gram</i>	600 billion	300	English	<i>FastText</i> [10]
<i>FastText.</i>	<i>Skip-gram</i>	16 billion	300	English	<i>FastText</i> [10]
<i>GloVe.</i>	–	42 billion	300	English	<i>stanford.edu</i> [6]

We can see in Table 5 that half of the models described were created from the compiled corpus that is made up of 40 million words with texts in English and Spanish, with an approximate 50% per language, approximately a vocabulary of 900,000 different words, and a total size of 360 MB. The other half of the embedding models were obtained from their respective official sources [10] and [6].

Table 6. Example of tag sequence created with annotation style BIO and BILOU.

Word	POS	BIO	BILOU
algunos	DET	B_LOW_FRECUENCY	U_LOW_FRECUENCY
asilos	NOUN	B_LOW_FRECUENCY	B_LOW_FRECUENCY
de	ADP	L_LOW_FRECUENCY	L_LOW_FRECUENCY
ancianos	NOUN	L_LOW_FRECUENCY	L_LOW_FRECUENCY
cuentan	VERB	B_MEDIUM_FRECUENCY	U_MEDIUM_FRECUENCY
con	ADP	B_VERYLOW_FRECUENCY	U_VERYLOW_FRECUENCY

4.4 System Results for Entity Identification

This section shows the results of the systems proposed to identify entities. First, the systems obtained are presented, and finally, the results obtained evaluating the systems.

System BiLSTM-CRF: The architecture of this system is mainly composed of a layer BiLSTM and a layer CRF that benefits the classification of entities, the architecture of this system is specifically composed of: 1 Embedding layer, 3 BiLSTM layers with 490 units, 1 LSTM layer with 980 units, one dense layer with nine units and 1 CRF layer with nine units.

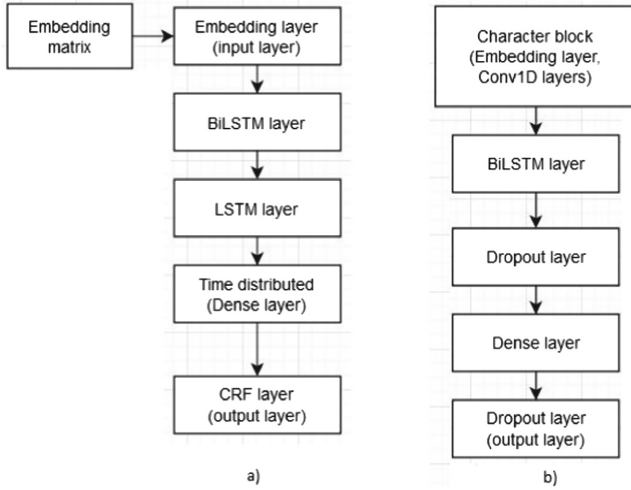


Fig. 3. a) Architecture BiLSTM-CRF, b) Architecture BiLSTM-CNN.

In Fig. 3, we can see the order of the layers used in that system. Finally, the network was trained with 21 epochs and a batch size of 110.

System BiLSTM-CNN: The architecture of this system is mainly composed of BiLSTM layers and CNN layers oriented to character embedding [9]. The character block is in charge of analyzing the words at the character level later to perform the classification, employing a BiLSTM network [3]. The architecture of this system is specifically composed of: 1 block of characters CNN, 1 BiLSTM layer with 128 units, one dropout layer with 128 units with a dropout rate of 0.5, 1 dense layer with 128 units, and one dropout layer with nine units with a dropout rate of 0.5.

In Fig. 3 we can see the order of the layers used in that system. Finally, the network trained with 21 epochs and a batch size of 110.

4.5 Evaluation of the Proposed Systems

The evaluation was performed with the evaluation corpus provided by [13]. For this, the precision, recall, and F_1 metrics proposed by [13] were used.

The results were obtained from the average of 20 executions of each model. The best results were obtained with FastText embedding models. The model that worked best was the self-created one with 20 million words and 490 dimensions (FastText).

In Table 7 A comparison of the annotation styles in the different entity identification systems is shown. It can be seen that the “P” annotation style got the best results at F_1 on both systems. In Table 7 is shown that “P” annotation style is better than BIO and BILOU annotation styles for this case.

Table 7. System results for entity identification, comparing annotation styles.

System	Recall	Precision	F ₁
BiLSTM-CNN - Annotation P	0.66852	0.48535	0.56169
BiLSTM-CNN - Annotation BIO	0.68438	0.47151	0.55817
BiLSTM-CNN - Annotation BILOU	0.67668	0.46965	0.55407
BiLSTM-CRF - Annotation P	0.67858	0.52135	0.58949
BiLSTM-CRF - Annotation BIO	0.68005	0.49503	0.57263
BiLSTM-CRF - Annotation BILOU	0.69358	0.47524	0.56396

In Table 8. A comparison of the systems with the “P” annotation style is shown, including the POS tagging proposal with different annotation styles. POS tagging proposal does not give an improvement.

Table 8. Results of systems for identification of entities, comparing the POS tagging proposal.

System	Recall	Precision	F ₁
BiLSTM-CNN - Annotation P (base)	0.66852	0.48535	0.56169
BiLSTM-CNN (P) - POS Annotation P	0.64834	0.48068	0.55165
BiLSTM-CNN (P) - POS Annotation BIO	0.63449	0.48971	0.55201
BiLSTM-CNN (P) - POS Annotation BILOU	0.65229	0.48264	0.55407
BiLSTM-CRF - Annotation P (base)	0.67858	0.52135	0.58949
BiLSTM-CRF (P) - POS Annotation P	0.67502	0.51250	0.58211
BiLSTM-CRF (P) - POS Annotation BIO	0.63101	0.51020	0.56201
BiLSTM-CRF (P) - POS Annotation BILOU	0.63633	0.51950	0.57103

4.6 Results of the Systems Proposed in *eHealth-KD Challenge 2021*

Table 9 shows a comparison of the results of subtask “A” with the participants of the *eHealth-KD challenge 2021*. It is important to note that these results of our proposal were after the *eHealth-KD challenge 2021*. Finally, Table 9 shows the neural network architecture on which the system of each participant is based.

Table 9. Comparison of results from subtask “A” (results obtained after the *eHealth-KD challenge 2021*.)

Team	Precision	Recall	F_1	Architecture
PUCRJ-PUCPR-UFMG	0.71491	0.69733	0.70601	BERT
Vicomtech	0.69987	0.74706	0.68413	BERT
IXA	0.61372	0.69840	0.65333	BERT
UH-MMM	0.54604	0.68503	0.60769	BiLSTM
Proposed system	0.52135	0.67858	0.58949	BiLSTM
uhKD4	0.51751	0.53743	0.52728	BiLSTM
Yunnan-Deep	0.52036	0.27166	0.33406	BiLSTM

5 Conclusion

The exposed BiLSTM architectures were implemented in this work. The novelities proposed in this work were the proposed solution directed to POS tagging and the “P” annotation style. Despite implementing the proposal directed to POS tagging with different annotation styles, in Table 8 can be seen that the implemented proposal decreases the score F_1 of the systems. However, the proposed “P” annotation style slightly improves compared to the other annotation styles reviewed in this article.

Finally, the best result obtained with a system based on a BiLSTM network is 0.60769 in F_1 , and the best-proposed system has a result of 0.5894 , as shown in Table 9, for which it can be considered as an acceptable result for a BiLSTM network. However, it is expected to explore other forms of pre-processing that can help to improve the F_1 of the systems.

References

1. Alvarado, J.M., Caballero, E.Q., Pérez, A.R., Linares, R.C.: UH-MAJA-KD at eHealth-KD challenge 2019. Deep learning models for knowledge discovery in Spanish eHealth documents. In: Cumbreiras, M., et al. (eds.) Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, CEUR Workshop Proceedings, Bilbao, Spain, 24 September 2019, vol. 2421, pp. 85–94. CEUR-WS.org (2019). http://ceur-ws.org/Vol-2421/eHealth-KD_paper_9.pdf
2. Bravo, À., Accuosto, P., Saggion, H.: LaSTUS-TALN at IberLEF 2019 eHealth-KD challenge. Deep learning approaches to information extraction in biomedical texts. In: Cumbreiras, M.Á.G., et al. (eds.) Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019. CEUR Workshop Proceedings, Bilbao, Spain, 24 September 2019, vol. 2421, pp. 51–59. CEUR-WS.org (2019). http://ceur-ws.org/Vol-2421/eHealth-KD_paper_5.pdf
3. Chiu, J.P.C., Nichols, E.: Named entity recognition with bidirectional LSTM-CNNs. CoRR abs/1511.08308 (2015). <http://arxiv.org/abs/1511.08308>

4. Consuegra-Ayala, J.P., Palomar, M.: UH-MatCom at eHealth-KD challenge 2020. In: Cumberras, M.Á.G., et al. (eds.) Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, 23 September 2020. CEUR Workshop Proceedings, vol. 2664, pp. 112–124. CEUR-WS.org (2020). http://ceur-ws.org/Vol-2664/eHealth-KD_paper4.pdf
5. Fabregat, H., Fernandez, A.D., Martínez-Romo, J., Araujo, L.: NLP_UNED at eHealth-KD challenge 2019: deep learning for named entity recognition and attentive relation extraction. In: Cumberras, M.Á.G., et al. (eds.) Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019. CEUR Workshop Proceedings, vol. 2421, pp. 67–77. CEUR-WS.org (2019). http://ceur-ws.org/Vol-2421/eHealth-KD_paper_7.pdf
6. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation (2015). <https://nlp.stanford.edu/projects/glove/>
7. Linet, C.Z.J.: Reconocimiento de entidades nombradas para el idioma español utilizando Conditional Random Fields con características no supervisadas. Universidad Católica San Pablo - Perú, Tesis de maestría (March 2017)
8. Liu, Z., Winata, G.I., Fung, P.: Zero-resource cross-domain named entity recognition. CoRR abs/2002.05923 (2020). <https://arxiv.org/abs/2002.05923>
9. Ma, E.: Besides word embedding, why you need to know character embedding? (2018). <https://towardsdatascience.com/besides-word-embedding-why-you-need-to-know-character-embedding-6096a34a3b10>
10. Mikolov, T., Grave, E., Bojanowski, P., Puhersch, C., Joulin, A.: Advances in pre-training distributed word representations. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC 2018 (2018)
11. Monteagudo-García, L., Marrero-Santos, A., Fernández-Arias, M.S., Cañizares-Díaz, H.: UH-MMM at eHealth-KD challenge 2021. In: Proceedings of the Iberian Languages Evaluation Forum, IberLEF 2021 (2021)
12. Pérez, A.R., Caballero, E.Q., Alvarado, J.M., Linares, R.C., Consuegra-Ayala, J.P.: UH-MAJA-KD at eHealth-KD challenge 2020. In: Cumberras, M.Á.G., et al. (eds.) Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, 23 September 2020. CEUR Workshop Proceedings, vol. 2664, pp. 125–135. CEUR-WS.org (2020). http://ceur-ws.org/Vol-2664/eHealth-KD_paper5.pdf
13. Piad-Morfis, A., Estevez-Velarde, S., Gutiérrez, Y., Almeida-Cruz, Y., Montoyo, A., Muñoz, R.: Overview of the eHealth knowledge discovery challenge at IberLEF 2021. *Proces. del Leng. Natural* **67**, 233–242 (2021). <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6392>