

LSI_UNED at CLEF eHealth2021: Exploring the effects of transfer learning in negation detection and entity recognition in clinical texts

Hermenegildo Fabregat¹, Andres Duque^{1,2}, Lourdes Araujo^{1,2} and Juan Martinez-Romo^{1,2}

¹NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos. Universidad Nacional de Educación a Distancia (UNED)

²Instituto Mixto de Investigación - Escuela Nacional de Sanidad (IMIENS)

Abstract

This paper describes the approach presented by the LSI_UNED team in the *Multilingual Information Extraction* task (*SpRadIE*) of CLEF eHealth 2021. The proposed system is a deep learning stack designed for separately detecting negation hedge cues and other biomedical entities in the task. Transfer learning techniques are applied for studying whether pre-trained weights from a different negation detection task can be effectively incorporated into the model for improving a baseline system trained only with the provided data. The system obtains promising results in the task, obtaining the second best F1 score, and the best precision score among all participant systems.

Keywords

Biomedical information extraction, Transfer learning, Negation detection

1. Introduction

Named Entity Recognition (NER) is the task that aims to detect a particular set of entities within a text. It represents one of the key steps in the process of information extraction in any specific domain. In the field of biomedicine, entity detection is of paramount importance for successfully performing subsequent tasks in the information extraction pipeline, such as relation extraction or document classification. Considering the huge amount of information currently available in the biomedical domain, including research papers, clinical notes or medical reports, the development of automatic systems able to perform accurate NER in those types of documents will definitely lead to better health support systems.

In this context, the *eHealth Evaluation Lab* conducted at the *Conference and Labs of the Evaluation Forum* (CLEF) 2021 [1] is a great opportunity for testing systems designed for solving these kind of tasks related to the biomedical domain. In particular, Task 1 of the eHealth 2021 challenge, named *Multilingual Information Extraction (SpRadIE)* [2] focuses on the detection of biomedical entities in clinical texts (radiology reports) written in the Spanish language.

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ gildo.fabregat@lsi.uned.es (H. Fabregat); aduque@lsi.uned.es (A. Duque); lurdes@lsi.uned.es (L. Araujo); juaner@lsi.uned.es (J. Martinez-Romo)

📞 0000-0001-9820-2150 (H. Fabregat); 0000-0002-0619-8615 (A. Duque); 0000-0002-7657-4794 (L. Araujo); 0000-0002-6905-7051 (J. Martinez-Romo)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

In this paper, we present a deep learning architecture designed for taking advantage of the use of transfer learning techniques in the detection of a particular type of entity, in this case negation hedge cues. For this purpose, the proposed architecture is a pipeline with two different branches receiving the same input, each branch being a particular neural network. One of the networks will perform the detection of negation hedge cues, while the other network will be used for recognizing the rest of the entities proposed in the task. In order to analyze the effect and possible improvements offered by the use of transfer learning techniques in detecting negation hedge cues, the network performing this subtask will be initialized either randomly, as well as the other network, in a usual setup, or with information from a different task oriented to negation detection, in a transfer learning setup.

The rest of the paper is organized as follows: Section 2 briefly presents some systems that faced similar tasks in past competitions. Details of the task addressed in this paper are given in Section 3, and the proposed system is presented in Section 4. Results obtained in the competition are shown and discussed in Section 5, while Section 6 is devoted to analyze some systematic errors detected during the development of the system. Finally, some conclusions and future lines of work are depicted in Section 7.

2. Background

The identification and classification of named entities is a deeply studied field in Natural Language Processing (NLP), and more particularly in the biomedical domain. The use of classical NLP approaches has led to the development of well-established systems in the literature such as Metamap [3]. These classical approaches include look-up dictionaries [4] and rule-based systems like PROPER [5] or TextDetective [6]. Machine learning systems [7] and especially deep learning techniques [8, 9], however, represent the current state of the art in biomedical NER. The development of specific biomedical word embeddings [10] and language models [11] have been key to the huge success of these systems.

Many different tasks related to biomedical NER have been proposed in evaluation campaigns such as CLEF eHealth 2015 [12], TASS eHealth-KD 2018 [13] or IberLEF eHealth-KD 2019 [14] and 2020 [15]. As mentioned before, the use of deep learning approaches for addressing these tasks has grown exponentially in the past few years, to the point of representing the vast majority of participating systems. Many of those systems propose deep learning stacks mainly based on Bidirectional Long Short Term Memory (Bi-LSTM) layers followed by Conditional Random Field (CRF) layers for performing entity detection and classification [16, 17, 18]. The use of techniques based on the Transformer architecture [19] such as BERT [20] has also gained high popularity in these tasks since their publication [21, 22].

In addition to the aforementioned challenges and evaluation campaigns, other works addressing biomedical NER tasks in the Spanish language have been recently developed. Deep learning methods are applied in [23] for the identification and subsequent anonymization of named entities within radiology reports. Transfer learning techniques based on contextualized word embeddings are employed in [24] for detecting pharmacological entities (substances, compounds and proteins) in Spanish clinical cases, improving previous results obtained with standard and general domain word embeddings.

3. Task: Multilingual Information Extraction

Task 1 of the eHealth Evaluation Lab at CLEF 2021 (*SpRadIE*) aims at the detection and classification of biomedical entities and hedge cues in radiology reports written in the Spanish language. The participating systems are asked to recognize ten different classes, separated into seven entities (anatomical entity, finding, location, measure, type of measure, degree and abbreviation) and three hedge cues (negation, uncertainty and conditional temporal). In order to achieve a good performance, systems must adequately deal with some casuistries inherent to NER tasks in the biomedical domain: long entities, discontinuous entities, overlaps or polysemy.

The dataset provided by the organizers consists of anonymized ultrasonography reports from the radiology department of a pediatric hospital in Argentina. Further information regarding the original annotation criteria, which was slightly modified for this task, can be found in [25]. The dataset contains 169 documents for training purposes and 92 documents for development purposes, all of them annotated using the BRAT format [26]. System testing is performed with an additional test set of 207 unseen documents. The development dataset is divided into two types of documents: *same-sample* documents, whose vocabulary is similar to the one in the training corpus, and *held-out* documents containing words that do not usually occur in the training corpus.

Finally, evaluation of the participating systems is carried out using Precision, Recall and F1 metrics over the Jaccard index between the predicted and the reference entities. Two different F1 measures are computed: exact F1 only considers exact matches of the predicted entities, while lenient F1 is a more relaxed metric that computes a score regarding the overlapping between the predicted entity and the reference.

4. System Description

The proposed system is a deep learning architecture that is mainly focused on two particular types of layers: Bidirectional Long Short-Term Memory layers (Bi-LSTM) and Conditional Random Field layers (CRF). Input documents are processed forwards and backwards thanks to the Bi-LSTMs, and each token from the documents is finally classified through the CRF layer.

4.1. Pre-processing

Since the reports from the dataset are initially annotated using the BRAT format, it is important to transform this annotation into a format that can be used for representing the final classes to which each token can belong. For this purpose, we use the BILOU annotation scheme, widely used in different NER tasks. This scheme discriminates between the beginning (B), inside (I) and last (L) tokens of a particular entity, as well as entities composed of a unique (U) token, and tokens in the document that are out (O) of any entity.

The final output of our system is designed for taking into account discontinuities and overlapped entities, both of them being two of the most repeated linguistic challenges within the provided corpus. After inspecting the training and development dataset, the entities that present a greater number of discontinuities and overlappings are *Location*, *Findings* and *Abbreviations*. Moreover, we are particularly interested in treating *Negation* hedge cues separately, in order

to analyze whether additional information coming from a different negation detection tasks is able to provide useful knowledge to our network. Due to these considerations, we model those four classes (Location, Finding, Abbreviation and Negation) in a separate way, and gather the remaining six entities in the same output structure. Hence, the four separate entities can be modelled by only using BILOU labels (since they will be modelled in separate output vectors). On the other hand, since the output for the six remaining entities is being represented in the same vector, the entity type has to be combined with the BILOU labels when considering these entities: for instance, using label B-Measure for distinguishing it from B-Degree or B-Anatomical_Entity.

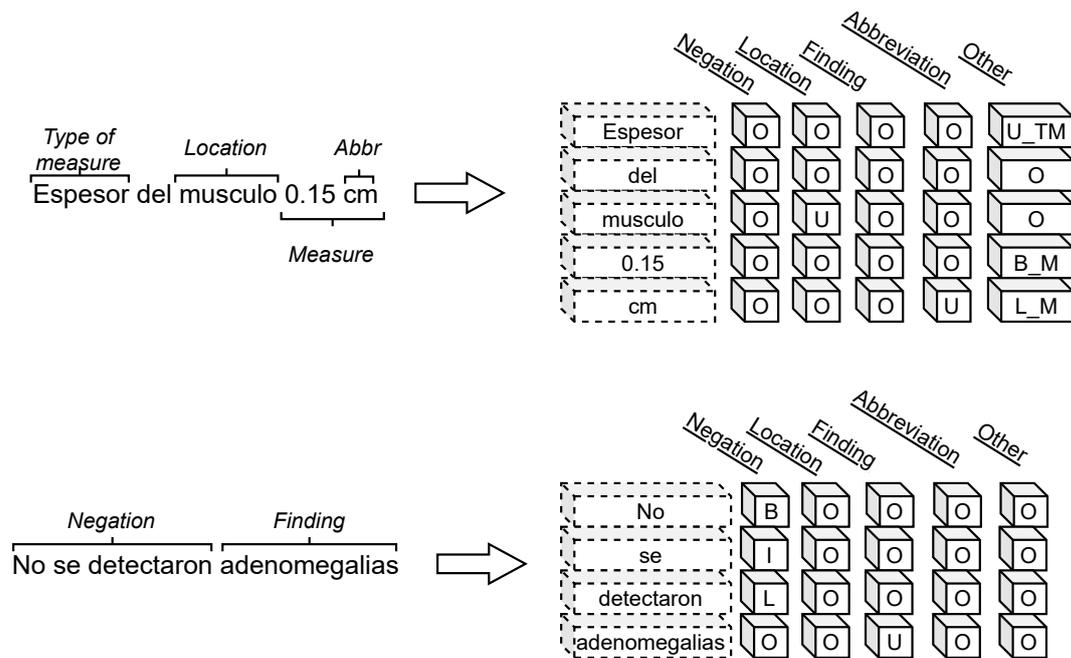


Figure 1: Example of transformation from BRAT format to BILOU annotation scheme.

Figure 1 shows an example of transforming sentences annotated with the BRAT format to the described BILOU annotation scheme. The top part of the figure shows a sentence in which we find a unique token labelled as *Type_of_measure*, another one as *Location*, and a unique *Abbreviation* embedded within a *Measure*. These elements are represented as unique tokens (U) in the *Location* and *Abbreviation* vectors. As neither *Type_of_Measure* nor *Measure* have independent output vectors, particular labels must be used within the *Other* vector for specifying that a unique *Type_of_Measure* (U_TM) and a begin and last *Measure* token (B_M and L_M) are used for representing "Espesor" and "0.15 cm" are used, respectively. The bottom part of the figure represents the use of begin, inside and last tokens (B, I and L) within the *Negation* vector for representing "No se detectaron", and a unique (U) token in the *Finding* vector for representing "adenomegalias".

4.2. Input Features

The features used for representing the input of the proposed deep learning stack are the following:

- **Word embeddings:** Two different pre-trained word embeddings from different sources are used for text representation. On the one hand, we use general domain Spanish 100-dimensional word embeddings in FastText, trained on Common Crawl and Wikipedia [27]. On the other hand, also 100-dimensional embeddings in FastText, generated from Spanish clinical texts [28], are also tested in order to analyze the differences and potential improvements.
- **Character embeddings:** The use of character embeddings may help in decreasing loss information caused by the reduction of dimensionality in word embeddings. We train character embeddings from scratch using a convolutional layer for generating a 16-dimensional character vector for each token in the document.
- **Casing, punctuation and formatting information:** An additional 8-position one-hot vector is used for modeling different casing scenarios, as well as information about punctuation marks and other formatting issues: uppercased first letter, term ending in comma, term ending in dot, term being a number, term being mostly numeric (over 50% of the characters being digits), term containing any digit, term containing any other punctuation marks, and other cases. Through this feature, we encode information that is usually omitted by word embeddings.

4.3. Main Architecture

The main design of the proposed deep learning stack is shown in Figure 2. Vectors representing word embeddings, character embeddings and casing information are concatenated and fed into two different pipelines, both of them consisting of a Bi-LSTM layer followed by a dense layer and a Conditional Random Field that performs the final classification. As mentioned in Section 2, this combination of Bi-LSTM and CRF layers has shown high performance in different NER tasks in the past few years. Although modern BERT-based architectures might offer better results, they have been avoided in this case due to the small size of the training dataset provided by the organizers. In the proposed system, the first pipeline is used for detecting all the possible entities in the dataset except for the negation hedge cues, while the second pipeline performs independent detection of negation hedge cues. Four parallel CRF layers are used in the first pipeline for classifying the aforementioned most frequently overlapping entities (*Location*, *Finding* and *Abbreviation*), and the set of remaining entities. A single CRF layer is used in the second pipeline for classifying negation hedge cues.

A final post-processing step based on rules is applied to the output of the deep learning architecture for solving systematic errors. The proposed rules are as follows:

- Use of the regular expression “ $([0-9]+)([a-zA-Z]+)$ ” for finding terms such as “128cm”. In those cases, the expression “cm” is added to the list of entities as an *Abbreviation*.
- Use of a more complex regular expression based on the previous one for trying to ensure the annotation of three-dimensional measures such as “2.5 x 2.5 x 128cm”.

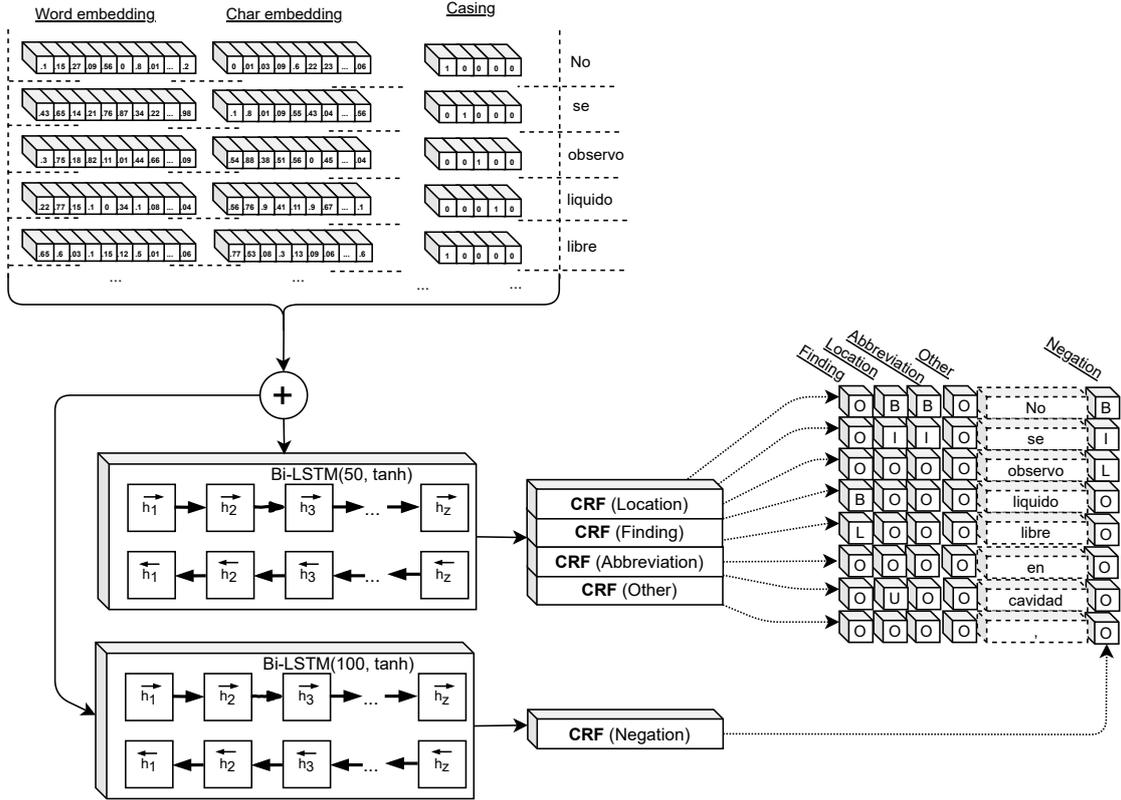


Figure 2: Proposed deep learning stack.

- Some scenarios have been identified in which no annotation is generated for some abbreviations. This last rule tries to cover the full annotation of the following measure abbreviations: “cc”, “cm”, “mm”, “ml”, “l”, “kg”, “g”, “mg”.

4.4. Transfer Learning

As it has been depicted in previous sections, the main objective of this system is to allow the analysis of potential improvements that can be obtained by applying transfer learning techniques to the independent identification of negation hedge cues. For this purpose, we compare the performance of the system when the initial weights of both Bi-LSTM networks are randomly initialized with that achieved when incorporating pre-trained weights to the Bi-LSTM network that performs negation detection. These pre-trained weights are extracted from a different negation detection task. In particular, the weights are generated in the process of training a different deep learning stack for detecting negation scopes and triggers, using for this purpose the SFU Review SP-NEG corpus [29]. The deep learning stack used for that separate task is described in detail in [30], and is based on the combination of a Bi-LSTM layer followed by a dense neural network performing the final classification. The transfer learning

process is achieved by using the weights from this Bi-LSTM trained on the negation task for initialising the weights of the Bi-LSTM devoted to the detection of the negation hedge cues in the eHealth task.

According to the widely recognised and adopted categorisation of transfer learning techniques presented in [31], the proposed approach would be a case of inductive transfer learning, in which the two tasks involved in the process are different, although their domains are strongly related, and also labelled data are available in both the source and target tasks (negation detection and eHealth NER, respectively). This is a similar setting to multi-task learning, however, in this case only one task is optimized for achieving high performance, instead of trying to learn both tasks simultaneously. Regarding deep transfer learning categorizations such as the proposed in [32], this would be a case of network-based deep transfer learning.

5. Results

Four different runs were allowed in the test phase of the *CLEF eHealth* challenge. In consequence, we have prepared four different settings of our system for submitting them to the organizers in this phase. In those settings, we combine the two different word embeddings mentioned in Section 4.2, this is, general domain and clinical embeddings, and also we explore the two main settings of the deep learning architecture: classic weight initialization (random, no transfer learning), and transfer learning initialization, both of them applied to the Bi-LSTM layer related to negation detection. These four settings will be denoted as follows in the experiments: **Classic+General** for classic initialization and general domain embeddings, **Transfer+General** for transfer learning approach and general domain embeddings, **Classic+Clinical** for classic initialization and clinical embeddings, and **Transfer+Clinical** for transfer learning approach and clinical embeddings.

As detailed in Section 3, two different development sets (*same-sample* and *held-out*) were provided by the organizers, and a simple test set was used for the final scores of the participating systems. Tables 1, 2 and 3 show the scores achieved using the different settings of our system, for the *same-sample* and *held-out* development dataset and for the test dataset respectively, using the official metrics of the task.

Table 1

Results achieved by the LSI_UNED team in the CLEF eHealth *same-sample* development dataset, for the lenient matching and exact matching metrics. F1, precision (P) and recall (R) values for each metric are expressed as percentages. Bold indicates the best setting for each metric.

Setting	Same-sample development					
	Lenient			Exact		
	F1	P	R	F1	P	R
Classic+General	83.49	87.63	80.35	80.45	84.38	77.47
Transfer+General	84.02	88.38	80.67	80.82	84.98	77.60
Classic+Clinical	84.43	88.85	80.95	82.04	86.34	78.66
Transfer+Clinical	84.18	89.06	80.50	81.37	86.11	77.77

Table 2

Results achieved by the LSI_UNED team in the CLEF eHealth *held-out* development dataset, for the lenient matching and exact matching metrics. F1, precision (P) and recall (R) values for each metric are expressed as percentages. Bold indicates the best setting for each metric.

Setting	Held-out dev.					
	Lenient			Exact		
	F1	P	R	F1	P	R
Classic+General	77.80	89.36	69.56	74.56	85.56	66.72
Transfer+General	78.89	89.27	71.24	75.36	85.17	68.11
Classic+Clinical	75.89	85.73	68.58	72.82	82.23	65.82
Transfer+Clinical	76.70	86.67	69.41	73.72	83.26	66.74

Table 3

Results achieved by the LSI_UNED team in the CLEF eHealth test dataset, for the lenient matching and exact matching metrics. F1, precision (P) and recall (R) values for each metric are expressed as percentages. Bold indicates the best setting for each metric.

System setting	Test dataset					
	Lenient			Exact		
	F1	P	R	F1	P	R
Classic+General (Run 3)	83.66	90.88	77.51	80.14	87.06	74.25
Transfer+General (Run 1)	83.88	90.28	78.33	80.07	86.17	74.76
Classic+Clinical (Run 4)	83.71	89.75	78.43	79.57	85.30	74.55
Transfer+Clinical (Run 2)	83.77	89.73	78.55	79.82	85.50	74.84

Some insights can be drawn from the results obtained by the proposed settings of our system regarding the different development datasets and the test dataset. It can be seen that, in general, transfer learning techniques applied to negation detection provide some improvements to the classic approach, particularly in the case of the *held-out* development dataset, while the best results for the *same-sample* dataset are achieved using the classic approach. However, in this *same-sample* dataset the differences are quite small. Regarding the use of general domain or clinical embeddings, again the most noticeable differences occur in results for the *held-out* dataset. Considering the test dataset, we can observe that the setting that obtains the best F1 measure in the lenient matching metric uses general domain embeddings and transfer learning techniques. However, the small differences between the four different runs submitted for the task indicate that the good performance shown by our system is more attributable to the proposed deep learning architecture (Bi-LSTM + CRF) than to the use of transfer learning techniques or different embeddings. In addition, we can observe that the results achieved in the test dataset are quite close to those obtained in the *same-sample* development dataset, and higher than those obtained in the *held-out* development dataset. This might indicate that the test dataset developed by the organizers is possibly more similar to the *same-sample* development dataset, and hence to the training dataset.

Table 4 illustrates the behaviour of the four different configurations of the system for each of the entities and hedge cues in the task: *Abbreviation (Abb)*, *Anatomical_Entity (AE)*, *Conditional_Temporal (CT)*, *Degree (Deg.)*, *Finding (Find.)*, *Location (Loc.)*, *Measure (Meas.)*, *Negation (Neg.)*, *Type_of_Measure (TM)* and *Uncertainty (Unc.)*. System configurations are the same as shown in Tables 1, 2 and 3: Only F1 score for the lenient evaluation is shown in order to simplify the table.

Table 4

Results achieved by the different runs of the LSI_UNED team in the CLEF eHealth test dataset, for each of the proposed entities. Metric is lenient F1, expressed as a percentage. Bold indicates the best setting for each entity.

Setting	Entities									
	Abb.	AE	CT	Deg.	Find.	Loc.	Meas.	Neg.	TM	Unc
C+G	90.70	82.08	57.14	44.44	75.34	65.89	88.89	92.09	86.28	70.06
T+G	91.22	82.83	36.36	46.93	73.01	66.87	88.40	92.26	88.25	72.33
C+C	91.04	81.67	50.00	65.71	71.93	66.87	88.32	94.50	89.26	66.62
T+C	92.20	82.50	50.00	63.77	73.03	65.54	87.52	90.05	89.28	66.71

As mentioned before, no major differences are found when comparing neither the “Classic” against the “Transfer” initialization schemes, nor the “General” against the “Clinical” word embedding models employed. The main differences can be seen regarding entity *Degree*, for which the use of clinical embeddings clearly improves the results compared to those obtained when using general domain embeddings. On the other hand, general domain embeddings offer quite better results than clinical embeddings for hedge cue *Uncertainty*. The use of the proposed transfer learning setting brings slight improvements for entities *Abbreviation*, *Anatomical_Entity*, *Type_of_Measure* and *Uncertainty*, while negation hedge cues only benefit from this transfer learning technique when using general domain embeddings. All these results reinforce the idea that the good results offered by the system are a consequence of the proposed deep learning architecture, over and above the use of transfer learning technique or specific embeddings.

Finally, Table 5 shows the comparison of results obtained by the best run of each system participating in the *CLEF eHealth 2021* task (*SpRadIE*), according and ordered by the F1 measure for the lenient matching metric, as provided by the organizers.

Comparing our system with other participating systems, our best run is ranked second in the task, out of seven participants. Moreover, we are able to obtain the highest precision scores, both in the lenient and in the exact matching metrics. Regarding F1, our team is just 1.63% behind the best system in the lenient metric, and 0.19% in the exact metric, while the differences between our results and those obtained by the third best system are much higher (5.41% and 6.94%, respectively). In addition, thanks to the information provided by the organizers upon completion of the evaluation, we know that our runs are able to obtain the best F1 lenient values for detecting entities *Finding* and *Measure*, and the second best for entities such as *Abbreviation*, *Degree* or *Negation*. Since *Finding*, *Abbreviation* and *Negation* are considered separately for

Table 5

Results achieved by the participating systems in the CLEF eHealth test dataset, for the lenient matching and exact matching metrics. F1, precision (P) and recall (R) values for each metric are expressed as percentages. Results are ordered by the F1 lenient metric, and bold indicates the results of our best setting.

System	Test dataset					
	Lenient			Exact		
	F1	P	R	F1	P	R
EdIE-KnowLab	85.51	87.24	83.85	80.26	81.88	78.70
LSI_UNED	83.88	90.28	78.33	80.07	86.17	74.76
ctb madrid	78.47	78.62	78.32	73.13	73.27	72.99
HULAT_MA	75.64	78.38	73.08	64.92	67.28	62.73
SINAI	73.70	86.07	64.43	67.96	79.37	59.42
SWAP	59.17	70.18	51.14	47.84	56.75	41.35
ims_unipd	16.00	9.29	57.62	9.38	5.45	33.77

classification (see Section 4.1), this could indicate that using separate classifiers for each entity might bring important improvements.

6. Error Analysis

In this section we present some systematic errors affecting the performance of the proposed system that were detected during the development phase:

- The system is not able to process some of the discontinuous entities included in the dataset. These entities represent 4.64% of the *held-out* dataset and 4.12% of the *same-sample* dataset. The annotation of some of those entities has been avoided in order to prevent the system from mislearning particular entities. For instance, the original text “VIA BILIAR intra y extrahepatica” should result in the detection of entities “VIA BILIAR extrahepatica” and “VIA BILIAR intra hepatica”. However, it is particularly difficult to design an annotation scheme for representing both entities in the training step, hence our system is not taking this particular case into account.
- Although the system includes different CRF layers for addressing overlapping entities, the total number of CRF layers is less than the total number of entities in the task. As mentioned in Section 4.1, only those entities most frequently involved in overlapping issues were selected for being classified in a separate CRF layer, from the preliminary study of the provided corpus. The main reason for this decision was to reduce the complexity of the final deep learning stack. However, although the remaining cases of overlapping entities may represent a small proportion of the total number of cases, some errors might come from this design choice.
- Documents within the training corpus were not tokenized and contained misspellings due to the specific nature of medical texts. Our system does not consider special solutions for misspelling errors, and the tokenization step only considers whitespace as a token

delimiter. This fact usually leads to recall issues that, in our case, are addressed by considering subword information through the use of FastText embeddings.

7. Conclusions and Future Work

In this paper we described the deep learning architecture proposed for the *Multilingual Information Extraction* task (*SpRadIE*) of *CLEF eHealth 2021*. We explored the use of transfer learning techniques taking advantage of information from negation detection tasks, and we also analyzed the differences in results when using general domain embeddings and clinical embeddings. The obtained results are quite promising, especially with regard to the proposed deep learning stack, composed of Bi-LSTM layers and CRF classifiers, and dividing the classification of those entities more likely to appear embedded or in a discontinuous form within the dataset. Improvements provided by the use of transfer learning were only found in specific settings. We obtained the second best F1 score among the participants in the task, not far behind the first place, and the best precision score in the task.

One of the first future lines of work should be exploring further decomposition of the annotation scheme used in the documents, for analyzing the effects of classifying, for instance, each entity separately. We consider that transfer learning techniques are a promising line of research within the task, however, a different secondary task more related to the main NER task should probably be found for the effects of this transfer learning to be noticed. We consider that, in this task, the influence of negation hedge cues within the addressed task is not strong enough for transfer learning from the considered secondary task to make a real difference. Finally, further exploration of the different word embedding models considered for this work might be an interesting research line. For instance, the combination of both general and clinical word embeddings, either by averaging or concatenating them could offer some additional insights on the behaviour of the different models.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under Grant PID2019-106942RB-C32, as well as project EXTRAE II (IMIENS 2019) and the research network AEI RED2018-102312-T (IA-Biomed).

References

- [1] H. Suominen, L. Goeuriot, L. Kelly, L. A. Alemany, E. Bassani, N. Brew-Sam, V. Cotik, D. Filippo, G. González-Sáez, F. Luque, P. Mulhem, G. Pasi, R. Roller, S. Seneviratne, R. Upadhyay, J. Vivaldi, M. Viviani, C. Xu, Overview of the CLEF eHealth Evaluation Lab 2021, in: *CLEF 2021 - 12th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS)*, Springer, 2021.
- [2] V. Cotik, L. A. Alemany, D. Filippo, F. Luque, R. Roller, J. Vivaldi, A. Ayach, F. Carranza, L. D. Francesca, A. Dellanzo, M. F. Urquiza, Overview of CLEF eHealth Task 1 - SpRadIE:

- A challenge on information extraction from Spanish Radiology Reports, in: CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS, 2021.
- [3] A. R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program., in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, p. 17.
 - [4] Z. Yang, H. Lin, Y. Li, Exploiting the performance of dictionary-based bio-entity name recognition in biomedical literature, *Computational Biology and Chemistry* 32 (2008) 287–291.
 - [5] K.-i. Fukuda, T. Tsunoda, A. Tamura, T. Takagi, et al., Toward information extraction: identifying protein names from biological papers, *Pacific Symposium on Biocomputing*. *Pacific Symposium on Biocomputing* 707 (1998) 707–718.
 - [6] J. Tamames, Text detective: a rule-based system for gene annotation in biomedical texts, *BMC bioinformatics* 6 (2005) 1–8.
 - [7] P.-T. Lai, M.-S. Huang, T.-H. Yang, W.-L. Hsu, R. T.-H. Tsai, Statistical principle-based approach for gene and protein related object recognition, *Journal of cheminformatics* 10 (2018) 1–9.
 - [8] Q. Wei, T. Chen, R. Xu, Y. He, L. Gui, Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks, *Database* 2016 (2016).
 - [9] Y. Wu, M. Jiang, J. Xu, D. Zhi, H. Xu, Clinical named entity recognition using deep learning models, in: *AMIA Annual Symposium Proceedings*, volume 2017, American Medical Informatics Association, 2017, p. 1812.
 - [10] Y. Zhang, Q. Chen, Z. Yang, H. Lin, Z. Lu, BioWordVec, improving biomedical word embeddings with subword information and MeSH, *Scientific data* 6 (2019) 1–9.
 - [11] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* 36 (2020) 1234–1240.
 - [12] L. Goeuriot, L. Kelly, H. Suominen, L. Hanlen, A. Névéol, C. Grouin, J. Palotti, G. Zucon, Overview of the CLEF eHealth Evaluation Lab 2015, in: J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. Jones, E. San Juan, L. Capellato, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Springer International Publishing, Cham, 2015, pp. 429–443.
 - [13] E. M. Cámara, Y. Almeida-Cruz, M. C. Díaz-Galiano, S. Estévez-Velarde, M. Á. G. Cumbreiras, M. G. Vega, Y. Gutiérrez, A. Montejo-Ráez, A. Montoyo, R. Muñoz, A. Piad-Morffis, J. Villena-Román, Overview of TASS 2018: Opinions, Health and Emotions, in: *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, co-located with 34th SEPLN Conference (SEPLN 2018)*, Sevilla, Spain, September 18th, 2018, volume 2172 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018, pp. 13–27. URL: http://ceur-ws.org/Vol-2172/p0_overview_tass2018.pdf.
 - [14] A. Piad-Morffis, Y. Gutiérrez, J. P. Consuegra-Ayala, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, A. Montoyo, Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2019, in: *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., 2019, pp. 1–16. URL:

http://ceur-ws.org/Vol-2421/eHealth-KD_overview.pdf.

- [15] A. Piad-Morffis, Y. Gutiérrez, H. Cañizares-Díaz, S. Estevez-Velarde, R. Muñoz, A. Montoyo, Y. Almeida-Cruz, Overview eHealth-KD 2020, in: 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, September 23th, 2020, volume 2664 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 71–84. URL: http://ceur-ws.org/Vol-2664/eHealth-KD_overview.pdf.
- [16] A. Bravo, P. Accuosto, H. Saggion, LaSTUS-TALN at IberLEF 2019 eHealth-KD Challenge: Deep Learning Approaches to Information Extraction in Biomedical Texts, in: IberLEF@SEPLN, 2019, pp. 51–59.
- [17] H. Fabregat, A. D. Fernandez, J. Martinez-Romo, L. Araujo, NLP_UNED at eHealth-KD Challenge 2019: Deep Learning for Named Entity Recognition and Attentive Relation Extraction, in: IberLEF@SEPLN, 2019, pp. 67–77.
- [18] A. R. Pérez, E. Q. Caballero, J. M. Alvarado, R. C. Linares, J. P. Consuegra-Ayala, UH-MAJA-KD at eHealth-KD Challenge 2020, in: 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, September 23th, 2020, volume 2664 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 125–135. URL: http://ceur-ws.org/Vol-2664/eHealth-KD_paper5.pdf.
- [19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv preprint arXiv:1706.03762 (2017).
- [20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [21] S. Medina Herrera, J. Turmo Borrás, Talp-upc at eHealth-KD challenge 2019: A joint model with contextual embeddings for clinical information extraction, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019): co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019): Bilbao, Spain, September 24th, 2019, CEUR-WS.org, 2019, pp. 78–84.
- [22] A. G. Pablos, N. Pérez, M. Cuadros, E. Zotova, Vicomtech at eHealth-KD Challenge 2020, in: 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, September 23th, 2020, volume 2664 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 102–111. URL: http://ceur-ws.org/Vol-2664/eHealth-KD_paper3.pdf.
- [23] I. Perez-Diez, R. Perez-Moraga, A. Lopez-Cerdan, J.-M. Salinas-Serrano, M. de la Iglesia-Vaya, De-identifying Spanish medical texts-Named Entity Recognition applied to radiology reports, *Journal of Biomedical Semantics* 12 (2021) 1–13.
- [24] L. Akhtyamova, P. Martínez, K. Verspoor, J. Cardiff, Testing Contextualized Word Embeddings to Improve NER in Spanish Clinical Case Narratives, *IEEE Access* 8 (2020) 164717–164726.
- [25] V. Cotik, D. Filippo, R. Roller, H. Uszkoreit, F. Xu, Annotation of Entities and Relations in Spanish Radiology Reports, in: RANLP, 2017, pp. 177–184.
- [26] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, BRAT: a web-based tool for NLP-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics,

2012, pp. 102–107.

- [27] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, arXiv preprint arXiv:1802.06893 (2018).
- [28] A. Gutiérrez-Fandiño, J. Armengol-Estapé, C. P. Carrino, O. De Gibert, A. Gonzalez-Agirre, M. Villegas, Spanish Biomedical and Clinical Language Embeddings, arXiv preprint arXiv:2102.12843 (2021).
- [29] S. M. J. Zafra, M. Taulé, M. T. Martín-Valdivia, L. A. U. López, M. A. Martí, SFU ReviewSP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns, *Lang. Resour. Evaluation* 52 (2018) 533–569. URL: <https://doi.org/10.1007/s10579-017-9391-x>. doi:10.1007/s10579-017-9391-x.
- [30] H. Fabregat, L. Araujo, J. Martínez-Romo, Deep learning approach for negation trigger and scope recognition, *Proces. del Leng. Natural* 62 (2019) 37–44. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5950>.
- [31] S. J. Pan, Q. Yang, A survey on transfer learning, *IEEE Transactions on knowledge and data engineering* 22 (2009) 1345–1359.
- [32] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, C. Liu, A survey on deep transfer learning, in: *International conference on artificial neural networks*, Springer, 2018, pp. 270–279.