IxaMed at PharmacoNER Challenge 2019

Xabier Lahuerta and **Iakes Goenaga** and **Koldo Gojenola** and
Aitziber Atutxa and **Maite Oronoz**
IXA NLP Group
UPV/EHU University of the Basque Country
Donostia, Basque Country
xlahuerta001@ikasle.ehu.eus, iakes.goenaga,
koldo.gojenola, aitziber.atucha, maite.oronoz{@*ehu.eus*}

Abstract

The aim of this paper is to present our approach (IxaMed) on the PharmacoNER 2019 task. The task consists of identifying chemical, drug, and gene/protein mentions from clinical case studies written in Spanish. The evaluation of the task is divided in two scenarios: one corresponding to the detection of named entities and one corresponding to the indexation of named entities that have been previously identified. In order to identify named entities we have made use of a Bi-LSTM with a CRF on top in combination with different types of word embeddings. We have achieved our best result (86.81 F-Score) combining pretrained word embeddings of Wikipedia and Electronic Health Records (50M words) with contextual string embeddings of Wikipedia and Electronic Health Records. On the other hand, for the indexation of the named entities we have used the Levenshtein distance obtaining a 85.34 F-Score as our best result.

1 Introduction

The aim of this paper is to present our approach in the PharmacoNER 2019 task (Gonzalez et al., 2019), on Medical Entity Recognition and Concept Indexing. The task consists of identifying different types of entities in the clinical domain in Spanish. The evaluation of the task is divided in two scenarios: the detection of medical entities, and the linking of each entity with its corresponding Concept Unique Identifier, a task called Concept Indexing.

The training corpus contains a manually classified collection of clinical cases derived from Open access Spanish medical publications (SPACCC) (Intxaurrondo, 2018). It contains a total of 1,000 clinical cases (396,988 words). This kind of narrative shows properties of both the biomedical and medical literature as well as clinical records.

In order to carry out the tasks, for Named Entity Recognition we have made use of a Recurrent Neural Network (RNN) to identify named entities feeding it with different types of embeddings, combining pretrained word embeddings and contextualized character-level word embeddings or contextual string embeddings. Furthermore, for Concept Indexing task we have opted to use a simple but effective Levenshtein distance method. We have achieved a F-score of 86.81 identifying named entities and 85.34 in Concept Indexing.

2 Related work

The SemEval 2014 Task 7 (Pradhan et al., 2014) was similar to the present competition, except for the number and types of entities to be identified (diseases and others) and the fact that discontinuous entities were also included. Task 7 in SemEval 2014 also comprised two subtasks, medical entity recognition and concept indexation. To tackle the first subtask, different teams used approaches as MaxEnt, SVM or CRF in combination with the extraction of syntactic and semantic attributes. The authors in (Tang et al., 2014) obtained the best results in strict F-Score with 78.5 on the development set and 81.3 on the test set. Their results were 4.7 points higher than those of the second ranked team (Kaewphan et al., 2014).

For the second subtask, namely Concept Indexation, the solutions proposed were very similar among the different teams. As in the NER task, the winner was (Tang et al., 2014) with an accuracy of 74.1 on the test set. Their solution was based on the cosine similarity using Vector Space Model (VSM). The team in (Ghiasvand and Kate, 2014) assigned the Concept Unique Identifier (CUI) code by comparing candidate strings with the terms obtained from the training set and the contents in the Unified Medical Language Sys-

tem (UMLS). They also proposed a method based on edit distance, more precisely Levenshtein distance (Levenshtein, 1966). The second best team (Kaewphan et al., 2014) employed word embeddings, word2vec (Mikolov et al., 2013), for word representation and the cosine similarity to find the closest standard term in UMLS. As a novelty, they implemented a binary classification based on Support Vector Machines (SVMs).

In SemEval 2015 (task 14), the evaluation was the only difference compared to SemEval 2014 (task 7). Besides strict evaluation (correct CUI and complete entity identification), relaxed evaluation was also pursued (successful CUI assignment and partly successful entity identification). In this case, the winning team was (Pathak et al., 2015), which obtained in the strict evaluation an F-score of 75.7, and in the relaxed one an F-score of 78. The methods used were similar to those used in SemEval-2014. In this case, a CRF was used to detect entities and a SVM classifier to determine if these were joined or not (and thus catch discontinuous entities). Regarding Concept Indexing, they used basically customized look-ups, like Dictionary look-up (exact match of entity word permutations, LVG), Customized Dictionary look-up (split UMLS entities by function words), and Customized Dictionary look-up (list of possible UMLS spans and application of Levenshtein distance). The second highest ranked team (Leal et al., 2015) obtained, for strict evaluation, an F-score of 74 and in the relaxed one 76.5. They employed a CRF to identify entities (also discontinuous entities), and for Concept Indexing they applied exact match on the terminology content of the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) enriching it with an abbreviation dictionary built on the training set. They also implemented a comparison method exploiting SNOMED-CT tree structure, Lucene index and Levenshtein average after splitting each recognized entity and each SNOMED-CT candidate.

Besides these competitions in recent years, improvements have been made mostly in the entity recognition subtask using neural networks such as Bi-LSTM + CRFs (Lample et al., 2016). (Casillas et al., 2019) used the tool for the detection of entities in clinical texts in Spanish, obtaining improvements with respect to previous works (Perez et al., 2017), from an F1-Score of 70.30 to 72.01. Employing a similar system (Goenaga et al., 2018) obtained the first position at the last IberEval shared task (Hermenegildo Fabregat and Araujo, 2018).

3 Resources

Apart from the tools we will present in the following sections, we made use of external data with the intention of completing the information the system extracts from the corpus provided by the organization. For this purpose we employed word-embeddings (Mikolov et al., 2013) that we have calculated (window length = 1, dimensions = 300, algorithm = SkipNgram) from Electronic Health Records (50M words), together with pretrained word-embeddings (window=5, dimensions=300, algorithm= Skip-gram) that have been calculated with Wikipedia2Vec (Yamada et al., 2018).

On the other hand, we have also used contextual string embeddings (Akbik et al., 2018) we have calculated from Electronic Health Records (number of layers=1, hidden size=2,048, sequence length=250, mini batch size=32) and Wikipedia (number of layers=1, hidden size=1,024, sequence length=250, mini batch size=100).

4 Methods

In this section we will explore the different methods we have used to perform the two sub-tasks of the shared task.

4.1 Track 1: NER Offset and Entity Classification

In this section we present our approach in order to extract named entities in track 1 of the shared task. For this purpose we employed a neural network based architecture, more precisely an specific Bi-LSTM (a RNN subclass, (Hochreiter and Schmidhuber, 1997)) with a CRF on top of it (Lample et al., 2016; Ma and Hovy, 2016) using as input raw text and the word-embeddings we have mentioned in section 3. This kind of neural network is widely used to pursue sequence to sequence tagging (Ma and Hovy, 2016; Jagannatha and Yu, 2016). One of the advantages of using Bi-LSTM in contrast to other machine learning techniques such as SVM, Perceptron or CRFs is that the size of the context is automatically learned by the LSTM and there is no need to perform any complicated text preprocessing to obtain features to feed the tool.

One of the strengths of our approach is that it combines different types of embeddings based on different types of corpus. On one hand, we use embeddings that have been calculated on a general domain corpus (Wikipedia) and embeddings that have been calculated on a medical domain corpus (EHRs). On the other hand, we stack pretrained word embeddings, character-level embeddings and contextual string embeddings and we feed the neural network with them. While the pretrained word embeddings and character-level embeddings are well known by the scientific community, the contextual string embeddings have been introduced recently (Akbik et al., 2018). This type of embeddings is based on recent advances in neural Language Modeling (LM) that have allowed a language to be modeled as distributions over sequences of characters instead of words (Sutskever et al., 2014), (Graves, 2013), (Kim et al., 2016).

Recent work has shown that by learning to predict the next character on the basis of previous characters, such models learn internal representations that capture syntactic and semantic properties: even though trained without an explicit notion of word and sentence boundaries, they have been shown to generate grammatically correct text, including words, subclauses, quotes and sentences (Sutskever et al., 2014), (Graves, 2013), (Karpathy et al., 2015).

The main features of these contextual string embeddings or contextualized character-level word embeddings are the following:

- They can be pre-trained on large unlabeled corpora.
- They are able to capture the meaning of the words in context and are able to produce different embeddings for polysemous words depending on their usage.
- They model words and contexts as sequences of characters, to both better handle rare and misspelled words as well as model subword structures such as prefixes and endings.

Lastly, we have sent two runs for named entity recognition (track 1): one run with the setup mentioned above, Bi-LSTM + CRF stacking pretrained and contextual embeddings, and one run with the same setup and using the development corpus for training for a few epochs (fine-tuning) as a last step.

4.2 Track 2: Concept Indexing

The normalization of given named entities consists in linking named entities to concepts in standardized medical terminologies, allowing generalization across contexts. The task consists in assigning, to each term, its corresponding Concept Unique Index. For example, “corticoide”, “corticoides” and ”cortecostiroides” are all normalized to the same Concept (B-255877006). In our work, we made use of a Text Similarity based mapping from the given terms to different sets:

- The terms present in the training set. This set is limited but gives an account of standard and non-standard terms present in spontaneously written health records.
- SNOMED-CT terms that can be considered a standard terminology.

We tried approximate searching to guarantee a matching, by a string-based similarity measure, as the well-known Levenshtein distance, a standard soft-matching approach in text normalization. We computed the Levenshtein distance between the input string and the set of terms that served as reference. Edit distance is used to quantify similarity between two strings, counting the minimum number of operations required to transform one string into another. The most common metric is the Levenshtein Distance (Levenshtein, 1966) in where the basic edit operations are removal, insertion and substitution of a single character. This metric finds the minimum distance for each spontaneous diagnostic term (SpoDT) with respect to all standard Diagnostic Terms (DictDT), obtaining the best candidate match (see equation 1).

$$minLev(SpoDT, DictDTs) \quad (1)$$

Hence, strings were searched in the reference-set and ranked according to this distance.

Exact matching of spontaneous expressions in standard dictionaries is not a good option, because it obtains a low accuracy. By contrast, matching with respect to previously classified non-standard expressions is well-worthy. However, the results show a considerable boost when using as reference the set of spontaneous terms and the standard reference (SNOMED CT).

We also tried a different approach using a sequence-to-sequence approach that, although it

NER	
Basic	**Fine-tuned**
86.60	86.81

Table 1: The results we have obtained for NER task. Basic = Combination of word embeddings and contextual string embeddings as an input of a Bi-LSTM with a CRF on top. Fine-tuned = Basic setup + fine-tuning on development set.

Concept Indexing	
Levenshtein Dist. 1	**Levenshtein Dist. 2**
85.14	85.34

Table 2: Results for Concept Indexing task. Levenshtein Distance 1 = Levenshtein distance applied to the entities extracted by the basic setup. Levenshtein Distance 2 = Levenshtein distance applied to the entities extracted by the fine-tuned setup.

gave promising results (an F-Score around 65% for concept Indexing), it was around 20 absolute points below the simplest option of using the Levenshtein distance. We think that this could be interesting to examine the strengths and weaknesses of each approach, and try to combine their positive aspects in a single combined or ensemble system, but we leave it as future work.

5 Results

In this section we present the results we have achieved for both tracks, NER and Concept indexing respectively. For this purpose we have compiled all the results in tables 1 and 2. If we observe the results obtained for both tracks we see a logical correlation between F-Score obtained for NER and the F-Score obtained for Concept Indexing. In other words, the better is the result for NER the better is the result for Concept Indexing. This is due to the fact that we use the output of the NER system as input of the Concept Indexing system.

Furthermore, if we analyze the results for each track we can observe we surpass the F-Score of 85.00 in all cases, thus confirming the robustness of our approaches. For NER, applying a Bi-LSTM with a CRF on top and feeding this neural network with stacked pretrained and contextual embeddings we have achieved a F-Score of 86.60. In contrast, fine-tuning on development set the previously mentioned neural network we outperform this result by 0.21. Although the improvement is not significant we have met our goal, that is to say, we have outperformed the basic setup avoiding overfitting.

Moreover, we have applied Levenshtein distance in order to assign a concept index to named entities that have been identified by NER system. We have achieved a 85.14 of F-Score when the input for the Concept Indexing system are named entities extracted by the basic NER system and a 85.34 of F-Score when the input are the named entities extracted by the fine-tuned NER system.

6 Conclusions

The purpose of this work was to evaluate the feasibility of different approaches to medical entity detection and concept indexing. Entity detection was dealt with a sequential tagger that uses word embeddings and contextual string embeddings acquired from electronic health records and Wikipedia. Concept normalization was approached by Text Similarity techniques. Surprisingly, the Levenshtein-based system obtained relatively good results, and this aspect deserves a further study of the strengths and weaknesses of each approach.

Acknowledgements

This work has been partially funded by:

- The Spanish ministry (projects PROSA-MED: TIN2016-77820-C3-1-R, DOMINO: PGC2018-102041-B-I00, both from MCIU/AEI/FEDER, UE).
- The Basque Government (projects DETEAMI: 2014111003).

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Arantza Casillas, Nerea Ezeiza, Iakes Goenaga, Alicia Perez, and Xabier Soto. 2019. Measuring the Effect of Different Types of Unsupervised Word Representations on Medical Named Entity Recognition. *International Journal of Medical Informatics (https://doi.org/10.1016/j.ijmedinf.2019.05.022)*.

Omid Ghiasvand and Rohit J Kate. 2014. UWM: Disorder Mention Extraction from Clinical Text Using CRFs and Normalization Using Learned Edit Distance Patterns. In *SemEval@ COLING*, pages 828–832.

Iakes Goenaga, Aitziber Atutxa, Koldo Gojenola, Arantza Casillas, Arantza Daz de Ilarraza, Nerea Ezeiza, Maite Oronoz, Alicia Perez, and Olatz Perez de Viñaspre. 2018. A Hybrid Approach For Automatic Disability Annotation. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018). ISSN 1613-073. Vol-2150. Pages 31-36.*

Aitor Gonzalez, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. In *Proceedings of the BioNLP Open Shared Tasks (BioNLP-OST)*, pages 1–X, Hong Kong, China. Association for Computational Linguistics.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Juan Martinez-Romo Hermenegildo Fabregat and Lourdes Araujo. 2018. Overview of the DIANN Task: Disability Annotation Task. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018).*

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Ander Intxaurrondo. 2018. SPACCC (Version 2019-02-01) [Data set]. Zenodo.

Abhyuday N Jagannatha and Hong Yu. 2016. Structured prediction models for RNN based sequence labeling in clinical text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 856.

Suwisa Kaewphan, Kai Hakala, and Filip Ginter. 2014. UTU: Disease Mention Recognition and Normalization with CRFs and Vector Space Representations. In *SemEval@COLING*, pages 807–811.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

André Leal, Bruno Martins, and Francisco Couto. 2015. ULisboa: Recognition and normalization of medical concepts. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 406–411. Association for Computational Linguistics.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *ACL (1)*. The Association for Computer Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Parth Pathak, Pinal Patel, Vishal Panchal, Sagar Soni, Kinjal Dani, Amrish Patel, and Narayan Choudhary. 2015. ezDI: A Supervised NLP System for Clinical Narrative Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 412–416. Association for Computational Linguistics.

Alicia Perez, Rebecka Weegar, Arantza Casillas, Koldo Gojenola, Maite Oronoz, and Hercules Dalianis. 2017. Semi-supervised medical entity recognition: A study on Spanish and Swedish clinical corpora. *Journal of biomedical informatics*, 71:16–30.

Sameer Pradhan, Noémie Elhadad, Wendy W Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. In *SemEval@ COLING*, pages 54–62.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Yaoyun Zhang1 Jingqi Wang1 Buzhou Tang, Yonghui Wu1 Min Jiang, and Yukun Chen3 Hua Xu. 2014. UTH_CCB: a report for semeval 2014–task 7 analysis of clinical text. *SemEval 2014*, page 802.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2018. Wikipedia2Vec: An Optimized Tool for Learning Embeddings of Words and Entities from Wikipedia. *arXiv preprint 1812.06280*.