IXA at eHealth-KD Challenge 2021: Generic Sequence Labelling as Relation Extraction Approach*

Edgar Andrés[1] [0000−0002−8190−8963]

UPV/EHU IXA research group, Bilbao, Spain
dgar.andres@ehu.eus

Abstract. The eHealth-KD 2021 is the automatic extraction of knowledge challenge from health documents written in Spanish with a small selection of sentences from different domains and languages to encourage cross-lingual and transfer learning approaches, we use the pre-trained Language Model (LM), namely XML-RoBERTa-base, to provide a Cross-lingual representation of tokens and the ability to transfer learning from general domains. Our group participated in all the proposed scenarios; the main one (F1 0.499), Entity Recognition (ER) (F1 0.653) and Relation Extraction (RE) (F1 0.430). The present system was designed as a pipeline of generic sequence labellers, each of them independently fine-tuned for each subtask. The generic sequence labeller consists of a feed-forward network that learns how to align a sequence of tokens into a sequence of labels regardless of the language and domain. This simple straightforward system ranked in the third position in the main and Entity recognition scenario and widely outperformed the other systems in the relation extraction scenario.

Keywords: eHealthKD 2021 , Knowledge Discovery , Natural Language Processing , Deep Learning

1 Introduction

eHealthKD series provide nice scenarios to build and evaluate Natural Language Processing systems on the medical domain. This year eHealthKD2021 [1] includes a selection of sentences not exclusively from medical texts but from other domains and different languages. In the last years, the amount of medical texts, regardless of format, has grown exponentially and accordingly, the interest in its processing for several clinical purposes. In this paper, we propose a system to extract entity mentions and their semantic relation type occurring in Spanish texts in the

* Supported by organization UPV/EHU IXA research group.
IberLEF 2021, September 2021, Málaga, Spain.

context of the eHealthKD2021 evaluation task. The system was built in two steps. We first identify and classify entity mentions in the sentence, and afterwards, we classify the relation type of identified entity pairs. Sentences are encoded using fine-tuned XLM-RoBERTa [2], which is a neural language model trained in multiple languages including Spanish. We transfer the general knowledge using a pre-trained model of the XLM-RoBERTa-base language model and fine-tuning it for the tasks of identifying entities and relations. We propose a simple yet robust model, where each component is trained separately. This strategy, contrary to joint models, makes learning easier and faster (focusing on one task at a time) and gives flexibility for domain and language adaptation. With this system, we hypothesize the idea that generic sequence labellers could competitively handle the relation extraction task while defining suitable formats to represent the problems and accurate ways to conditionate LM.

2 State Of The Art

Actual systems focus on retrain LM for span detection and entity detection [3, 4], LM we use for the task is highly related with the result we achieve for specific domains [5, 6], since LMs have appeared we see that performance of NLP tasks are directly related to the LM we use to represent textual data. Relation Extraction (RE) approaches faced the first revolution on [7] reaching high-performance systems [8–10] those are mainly based on BERT technologies and derivatives. In the clinical domain RE [11, 12] we can encounter a high-performance system based on specific techniques such as novel architectures of Bi-LSTM cells. SOTA Domain-agnostic approaches [13, 14] follow the idea of improving the LM representation using adaptive techniques for required task Sequence Labelling (SL) or RE. Cross-Lingual performance [15] is mainly derived from the appearance of the BERT model and the cross-lingual features it provides. SL [16, 17] even is an extensively researched field, is nowadays widely used in new application fields such as clinical data mining.

3 System Description

Generic sequence labelling The 2-stage system to extract entity mentions and their semantic relation type occurring in Spanish texts is based on a pipeline of fine-tuned generic sequence labellers as described in 1. We use a feed-forward network (FFN) to compute the probability $\tilde{y_i} = FFN(x_i)$ for each token, where each value in $\tilde{y_i}$ represents the score for a tag in a target tag set. Equation 1 shows how we formalized the the feedforward network.

$$FFN(x_i) = softmax(W_e x_i + b_e) \tag{1}$$

We decided to apply a pipeline of sequence labellers to 1) keep the model as simple as possible and 2) avoiding over-fitting of the model, as it could learn specific dependencies in training. For the final prediction, we apply an *argmax* function over the label probability distribution obtained for each token. The sequence labelling is learned minimizing the cross-entropy loss shown in Equation 2.

$$\mathcal{L}^t = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{y_i} \log \tilde{\mathbf{y}}_\mathbf{i} \tag{2}$$

Where $\mathbf{y_i}$ is the true label vector for the input token x_i, and N is the number of instances in the training set for the task. As you deduce, the input file format is composed of two columns, containing x_i and y_i pairs per line, the different examples are separated by empty lines. Finally, we use the special token "jump_line" to define the end of a text.

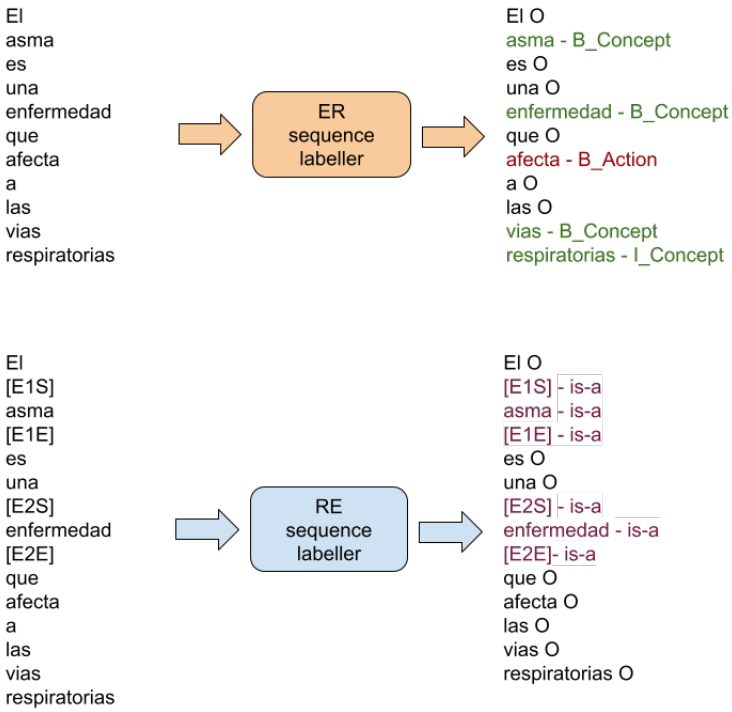

Fig. 1. 2-step relation extraction.

Subtask A: Entity recognition The input provided by the organizer as BRAT standoff format (.ann), was split into texts keeping line jumps, then texts were

divided into tokens keeping white spaces. Those tokens were aligned with the labels following Inside Outside Beginning format (IOB). This format does not capture overlapped and disjoint entities. The output of the system was converted again into (.ann) files.

Subtask B: Relation extraction We applied once again the same tokenization strategy exposed for entity recognition. In this case, as we already have the entities identified in the previous step, we pairwise each possible combination generating a repeated example per pair, entity markers [18] are added surrounding entities to avoid overfitting. In this case, we align the entities with the relation type. The output of the system was transformed into final (.ann) files.

Training setup We used huggingface transformers [19] for default training parameters setup, both systems were trained over respective train set and fine-tunned with respective dev set, both performed 40 epochs with a batch size of 40 examples, each fine-tune maximized the f1 score described in Conll2005 shared task [20]. The best model of 11 / 12 checkpoints out of 1100 / 12000 total steps were respectively used for entity recognition / relation extraction. Both models were calculated in 30 minutes each using a single NVIDIA Titan V.

4 Results

In the following Table 1 we summarize the results in the three different scenarios over the official Test set, the best results for each metric are highlighted with bold characters. The system gets competitive remarks in whole scenarios winning the third one (Relation extraction) with outstanding results. Although the good results we encounter low precision stats, this is due to the generic Language model we used (XLM-RoBERTa), we encountered similar issues in the previous series [21]. .

	Scenario 1			**Scenario 2**			**Scenario 3**		
Model	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Vicomtech	0.541	0.535	**0.531**	0.700	**0.747**	0.684	0.542	0.283	0.372
PUCRJ-PUCPR-UFMG	**0.568**	0.503	0.528	**0.715**	0.697	**0.706**	0.367	0.205	0.263
uhKD4	0.485	0.374	0.423	0.518	0.537	0.527	**0.556**	0.222	0.318
Baseline	0.337	0.177	0.232	0.350	0.272	0.306	0.438	0.017	0.033
IXA	0.465	**0.539**	0.499	0.614	0.698	0.653	0.454	**0.409**	**0.430**

Table 1. Results of eHealth-KD 2021 task. We summarize the top fourth systems: Vicomtech [22], PUCRJ-PUCPR-UFMG [23], uhKD4 [24] and ours

5 Conclusions

Simple compositions of fine-tunned FFN and LM can accurately describe the target language, this is sufficient to perform competitively in prediction tasks via sequence labelling regardless of domain and language, in this way we define generic sequence labelling. We conclude that the sequence labelling task is extensible to many tasks like seq2seq or classification with competitive performance at low cost as we have seen in several approaches [25, 21]. This time we expand the idea enforcing the necessity of new simple mathematical modelling techniques to handle huge amount of complex data as we have seen in RE task.

References

1. Alejandro Piad-Morffis, Yoan Gutiérrez, Suilan Estevez-Velarde, Yudivián Almeida-Cruz, Rafael Muñoz, and Andrés Montoyo. Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2021. *Procesamiento del Lenguaje Natural*, 67(0), 2021.
2. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.
3. Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.
4. Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*, 2020.
5. Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.
6. Sapan Shah, Sreedhar Reddy, and Pushpak Bhattacharyya. A retrofitting model for incorporating semantic relations into word embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1292–1298, 2020.
7. Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. A general framework for information extraction using dynamic span graphs. *arXiv preprint arXiv:1904.03296*, 2019.
8. Jue Wang and Wei Lu. Two are better than one: Joint entity and relation extraction with table-sequence encoders. *arXiv preprint arXiv:2010.03851*, 2020.
9. Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. *arXiv preprint arXiv:2010.13415*, 2020.
10. Angrosh Mandya, Danushka Bollegala, and Frans Coenen. Graph convolution over multiple dependency sub-graphs for relation extraction. In *COLING*, pages 6424–6435. International Committee on Computational Linguistics, 2020.
11. Di Zhao, Jian Wang, Yijia Zhang, Xin Wang, Hongfei Lin, and Zhihao Yang. Incorporating representation learning and multihead attention to improve biomedical cross-sentence n-ary relation extraction. *BMC bioinformatics*, 21(1):1–17, 2020.

12. Zhiheng Li, Zhihao Yang, Yang Xiang, Ling Luo, Yuanyuan Sun, and Hongfei Lin. Exploiting sequence labeling framework to extract document-level relations from biomedical texts. *BMC bioinformatics*, 21:1–14, 2020.
13. Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, 2020.
14. Hao Huang, Guodong Long, Tao Shen, Jing Jiang, and Chengqi Zhang. Rate: Relation-adaptive translating embedding for knowledge graph completion. *arXiv preprint arXiv:2010.04863*, 2020.
15. Jouni Luoma and Sampo Pyysalo. Exploring cross-sentence contexts for named entity recognition with bert. *arXiv preprint arXiv:2006.01563*, 2020.
16. Hai-Long Trieu, Thy Thy Tran, Khoa NA Duong, Anh Nguyen, Makoto Miwa, and Sophia Ananiadou. Deepeventmine: end-to-end neural nested event extraction from biomedical texts. *Bioinformatics*, 36(19):4910–4917, 2020.
17. Bin Ji, Jie Yu, Shasha Li, Jun Ma, Qingbo Wu, Yusong Tan, and Huijun Liu. Span-based joint entity and relation extraction with attention-based span-specific and contextual semantic representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 88–99, 2020.
18. Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics.
19. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.
20. Xavier Carreras and Lluís Màrquez. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)*, pages 152–164, 2005.
21. E Andrés, O Sainz, A Atutxa, and O Lopez de Lacalle. Ixa-ner-re at ehealth-kd challenge 2020: Cross-lingual transfer learning for medical relation extraction. In *Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@ SEPLN*, volume 2020, 2020.
22. Aitor García-Pablos, Naiara Pérez, and Montse Cuadros. Vicomtech at ehealth-kd challenge 2021: Deep learning approaches to model health-related text in spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, 2021.
23. Lucas Pavanelli, Elisa Terumi Rubel Schneider, Yohan Bonescki Gumiel, Thiago Castro Ferreira, Lucas Ferro Antunes de Oliveira, João Vitor Andrioli de Souza, Giovanni Pazini Meneghel Paiva, Claudia Maria Silva e Oliveira, Lucas Emanuel Cabral Moro, Emerson Cabrera Paraiso, Eduardo Labera, and Adriana Pagano. Pucrj-pucpr-ufmg at ehealth-kd challenge 2021: A multilingual bert-based system for joint entity recognition and relation extraction. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, 2021.
24. Dayany Alfaro-González, Dalianys Pérez-Perera, Gilberto González-Rodríguez, and Antonio Jesús Otaño-Barrera. uhkd4 at ehealth-kd challenge 2021: Deep

learning approaches for knowledge discovery from spanish biomedical documents. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, 2021.
25. Edgar Andrés Santamaría. End to end approach for i2b2 2012 challenge based on cross-lingual models. 2020.