# FRE @ BC8 SympTEMIST track: Named Entity Recognition

Ander Martinez[1*] and Nuria García-Santa[1],

[1]AI & Computing Research Group, Fujitsu Research of Europe Ltd., Spain

*Corresponding author: Tel: +44 (0) 204 512 3064, E-mail: ander.martinez@fujitsu.com

## Abstract
This paper describes our submission on the *SympTEMIST* Named Entity Recognition (NER) shared subtask at *BioCreative 2023*. We submitted two systems based on a RoBERTa architecture LLM trained on Spanish-language clinical data available at *HuggingFace* model repository. The techniques that we used for both systems are Conditional Random Fields (CRF) and Byte-Pair Encoding dropout (BPE dropout). In the second system we also included Sub-Subword feature based embeddings (SSW). Our systems obtained strict F1-score 0.727 and 0.728 with and without SSW, respectively.

## Introduction
Named Entity Recognition (NER) is one of the cornerstones of text mining. It is particularly useful when applied to the clinical context where Electronic Health Record (EHR) often consists of many unstructured clinical notes containing entities such as diseases, procedures, drugs, and symptoms. The case for symptoms is particularly challenging since a single symptom can be written in many ways with varying degrees of detail. NER is necessary to go from unstructured information to structured information to perform downstream tasks. The performance of the downstream task directly depends on the performance of the NER task. For this reason, we find the symptom NER subtask proposed by SympTEMIST of particular interest.

NER, as a classical Natural Language Processing (*NLP*) task, has a long history. Besides simple n-gram matching, a popular approach to NER used to be *Hidden Markov Models* (1, 2). An improvement over HMM was applying CRFs (3, 4). With the popularization of *Deep Learning* (DL), *Recurrent Neural Networks* (RNNs) became popular for NER (5).

NER models are trained with hand-labelled (*gold standard*) data. This kind of data is costly to produce and therefore usually exists in limited amount. However, DL networks usually need substantial amounts of data to start producing satisfactory results. Because of this, *large language models* (LLM), such as *BERT* (6), RoBERTa (7), became popular for NER. These models are trained on unlabelled data and serve as a basis for other downstream tasks. Nowadays, the LLM-based approach remains the most popular, according to the number of publications.

Our submitted systems are based on a RoBERTa architecture LLM trained on Spanish-language clinical data available at *HuggingFace* model repository[1]. The techniques that we used are Conditional Random Fields (CRF), BPE-dropout, and Sub-Subword feature based embeddings

[1] Model name: `PlanTL-GOB-ES/roberta-base-biomedical-clinical-es`

(SSW) for one of the systems. All these techniques will be briefly introduced in the Techniques section.

## Techniques

This section describes the strategies used for our NER models. The first two, *CRF* and *BPE dropout*, were used for both submissions. One of the systems used the *sub-subword features* technique, while the other one did not.

### Conditional Random Fields

The original BERT paper (6) demonstrated the usability of BERT for NER, but it did not use CRF. Later authors, such as (8), showed that CRF improved the results in some cases. CRF can model the probability of transitioning from one output label to the next one. In NER tasks following a schema such as `BIO`, Beginning-Inside-Outside (9), CRF can help avoid impossible transitions. CRF is usually used in conjunction with the *Viterbi* algorithm to consider different output sequences.

### BPE Dropout

BPE dropout (10) was introduced as an alternative to (11), where they found that the main drawback to the subword regularization method is its complexity since it requires training a unigram language model and it uses *Expectation–Maximization* (*EM*) and *Viterbi* (12) algorithms to sample segmentations.

One of the benefits of *BPE dropout* is that it works on *BPE* vocabulary models (13), same as (usually) used by *RoBERTa*, and as such, we did not need to rebuild the vocabularies. In comparison, the *unigram language model subword regularization* method uses a statistical model and dynamic programming to be able to sample different segmentations from the same sequence. BPE dropout uses random noise to discard certain merge-operations, randomly generating a different sequence of subwords each time. This is so because BPE does not store the frequencies of each subword, only the order of the merge-operations. Merge-operations are discarded with a probability *p*, which is usually 0.1. Provilkov et al. (10) concluded through several experiments that BPE dropout achieves better results. Our systems used *BPE dropout* during training, with a dropout probability *p* of 0.1.

### Sub-subword Features

We used the Sub-subword feature method (14) in one of the systems to expose the character-level information to the network. According to (15), the sub-subword features method helps regularize the systems with little training data. The method consists in building the embedding matrices from the n-gram features of the subwords in the vocabulary. The features used to produce the embeddings are selected by an algorithm before training, and the neural network that produces the embeddings is trained with the rest of the model.

Since we used a RoBERTa LLM to build the NER models, we did not want to discard its (sub-)word embeddings. Before training the NER model using the sub-subword features embeddings, we fit the feature-to-embedding (FTE) network to produce embeddings similar to those included with the RoBERTa model. We used *Mean Squared Error* (MSE) training for this purpose. After this step, the NER model was used normally (using CRF and BPE dropout).

## Experiments

In order to choose the best approach for our submissions, we performed some experiments using the provided training-data (16). The data provided contained 750 documents. The documents were segmented into sentences using Spanish-language NLTK *punkt*. We avoided splitting sentences when that would split a labelled entity. After sentence segmentation the dataset contained 12009 sentences. Of these sentences we made a training, validation and test datasets that contained 11009, 500 and 500 sentences, respectively.

| System | F1 score (entity level) |
|---|---|
| `Softmax` | 65.78% |
| `Softmax+BPEd` | 72.12% |
| `CRF+BPEd` | 75.77% |
| `CRF+BPEd+bias` | 78.03% |
| `CRF+BPEd+bias+SSWF` | 77.92% |

**Table 1** *Results of experiments.*

We used BIO encoding for the entities. In preliminary experiments we did not find any benefit in using `S-` or `E-` tags. We first tried using a SoftMax layer on top of an LLM model. We tried different Spanish-language models available at *HuggingFace* and finally the model by (17) gave best results for us, with 65.78% F1 score. We used BPE dropout to improve the F1 score to 72%.

We observed that our models were producing invalid transitions, such as outputting `` `I-SINTOMA` `` labels without a preceding `` `B-SINTOMA` ``. For this reason, we decided to try using CRF on top of the LLM-based NER model, which improved the F1 score. Since our predictions were still producing invalid transitions, we initialized the CRF transition matrix to disallow `O-` to `I-` transitions. This gave us the best results.

We also tried using the Sub-subword features approach described in the techniques section. This did not improve the F1 score for us. The results are summarized in Table 1.

We trained all the models for 25 epochs with batches of 15 sentences and learning rate of 2e-5. *AdamW* optimizer was used.

## Conclusions

Since multiple submissions were allowed for each team, we submitted two systems corresponding to the *CRF+BPEd+bias* and *CRF+BPEd+bias+SSWF* from Table 1 but trained on the whole training data. At the moment of writing this report we do not know how our models were positioned in the final ranking, but we do know that they got 72.67% and 72.77% F1 scores respectively and that the mean and median of all submissions were 0.61 and 0.7, respectively. We also know that our submissions got the best recall values of all submissions.

We reproduce the results as reported by the organizers in Table 2. The scores P and R stand for precision and recall. The scores prefixed by "o_" show their overlapping counterpart. We only considered strict F1 score to optimize our models.

| Team's Name | Run name | P | R | F1 | o_P | o_R | o_F1 |
|---|---|---|---|---|---|---|---|
| FRE | 1-roberta | 0.7231 | 0.7303 | 0.7267 | 0.8616 | 0.8702 | 0.8658 |
| FRE | 2-roberta_ssw | 0.7154 | **0.7403** | 0.7277 | 0.8487 | **0.8782** | 0.8632 |

**Table 2** Results reported by the organizers. Scores prefixed by "o_" report overlapping results.

## References


1. J. Mayfield, P. McNamee, and C. Piatko. (2003) Named Entity Recognition using Hundreds of Thousands of Features. In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* [Internet]. p. 184–7. Available from: https://aclanthology.org/W03-0429

2. S. Morwal, N. Jahan, and D. Chopra. (2012) Named Entity Recognition using Hidden Markov Model (HMM). *Int J Nat Lang Comput*. **1**, 15–23.

3. J. Lafferty, A. McCallum, and F. Pereira. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In. Available from: https://www.semanticscholar.org/paper/Conditional-Random-Fields%3A-Probabilistic-Models-for-Lafferty-McCallum/f4ba954b0412773d047dc41231c733de0c1f4926

4. J. R. Finkel, T. Grenager, and C. Manning. (2005) Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* [Internet]. Ann Arbor, Michigan: Association for Computational Linguistics; p. 363–70. Available from: https://aclanthology.org/P05-1045

5. S. Chowdhury, X. Dong, L. Qian, et al. (2018) A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. *BMC Bioinformatics*. **19**(Suppl 17), 499.

6. J. Devlin, M. W. Chang, K. Lee, et al. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. arXiv; Available from: http://arxiv.org/abs/1810.04805

7. Y. Liu, M. Ott, N. Goyal, et al. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach [Internet]. arXiv; Available from: http://arxiv.org/abs/1907.11692

8. F. Souza, R. Nogueira, and R. Lotufo. (2020) Portuguese Named Entity Recognition using BERT-CRF [Internet]. arXiv; Available from: http://arxiv.org/abs/1909.10649

9. L. A. Ramshaw and M. P. Marcus. (1995) Text Chunking using Transformation-Based Learning [Internet]. arXiv; Available from: http://arxiv.org/abs/cmp-lg/9505040

10. I. Provilkov, D. Emelianenko, and E. Voita. (2020) BPE-Dropout: Simple and Effective Subword Regularization [Internet]. arXiv; Available from: http://arxiv.org/abs/1910.13267

11. T. Kudo. (2018) Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* [Internet]. Melbourne, Australia: Association for Computational Linguistics; p. 66–75. Available from: https://www.aclweb.org/anthology/P18-1007

12. A. Viterbi. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Inf Theory*. **13**(2), 260–9.

13. R. Sennrich, B. Haddow, and A. Birch. (2015) Neural Machine Translation of Rare Words with Subword Units. *ArXiv150807909 Cs* [Internet]. Available from: http://arxiv.org/abs/1508.07909

14. A. Martinez, K. Sudoh, and Y. Matsumoto. (2021) Sub-Subword N-Gram Features for Subword-Level Neural Machine Translation. 自然言語処理. **28**(1), 82–103.

15. A. Martinez. (2021) The Fujitsu DMATH Submissions for WMT21 News Translation and Biomedical Translation Tasks. In: *Proceedings of the Sixth Conference on Machine Translation* [Internet]. Online: Association for Computational Linguistics; p. 162–6. Available from: https://aclanthology.org/2021.wmt-1.13

16. S. Lima-López, E. Farré-Maduell, L. Gasco-Sánchez, et al. (2023) Overview of SympTEMIST at BioCreative VIII: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In: *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*.

17. C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, et al. (2021) Biomedical and Clinical Language Models for Spanish: On the Benefits of Domain-Specific Pretraining in a Mid-Resource Scenario.