Using Embeddings And Bi-LSTM+CRF Model To Detect Tumor Morphology Entities In Spanish Clinical Cases

Sergio Santamaria Carrasco
Paloma Martínez Fernández

Table of Contents

HULAT-UC3M Team

Sergio Santamaria Carrasco
Student of the master's degree in computer science and technology at Universidad Carlos III de Madrid

Paloma Martínez Fernández
Member of HULAT research group (hulat.inf.uc3m.es) and faculty member of Computer Science Department at Universidad Carlos III de Madrid

Methods & System Description

Step I: Preprocces

- ❑ Split into tokens and sentences
- ❑ Remove stop words
- ❑ Annotations transformed to BIOES schema

Table 1
Vocabulary statistics

	Training set	Development set	Training and Development set
Words	22118	21031	31329
Characters	126	123	137

Step II: Input Features

- ❑ **Words:** Embedding representation 300 dimensional vector
- ❑ **Part-of-Speech:** Embedding representation 40 dimensional vector
- ❑ **Characters:** Embedding representation 30 dimensional vector
- ❑ **Syllables:** Embedding representation 75 dimensional vector
- ❑ **Meaning Cloud Named Entities:** 15 dimensional one-hot vector

spaCy
fastText
meaning cloud

Methods & System Description

Step III: Deep Learning Architecture

Maximum character sequence length	30
Maximum syllable sequence length	10
Character convolutional filters	50 filters of size 3
Syllables convolutional filters	50 filters of size 3
First BiLSTM hidden state dimension	300 for the forward and backward layers
Second BiLSTM hidden state dimension	300 for the forward and backward layers
Dense layer units	200
Dropout	0.4
Optimizer	ADAM optimizer, learning rate: 0.001
Number of epochs	5

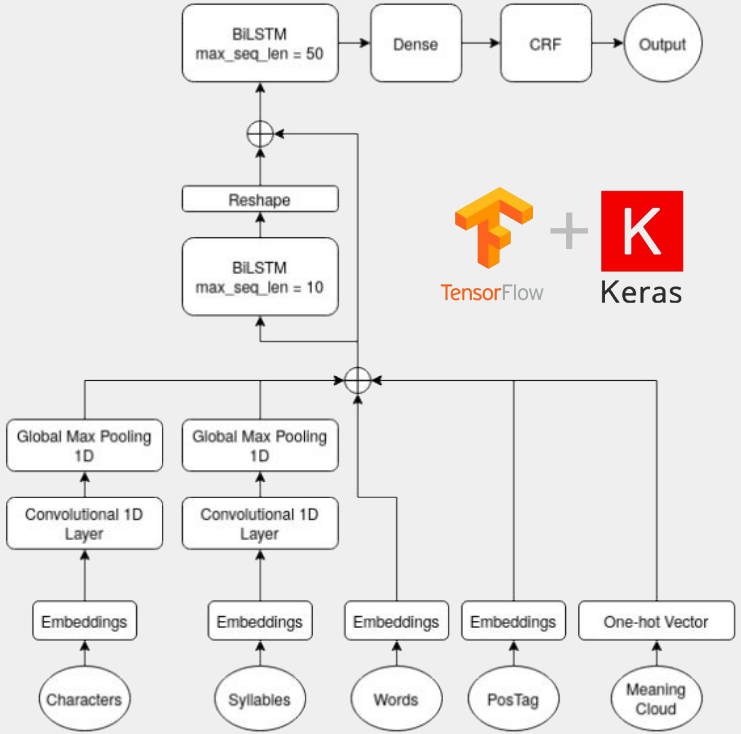

Figure 1: Architecture of the proposed model for named entity recognition.

https://github.com/ssantamaria94/CANTEMIST-Participation

Results

Hyperparameter Tuning

Table 2
Detailed hyper-parameter settings in the CANTEMIST task.

Parameters	Tuned range	Opt.
Sequence length	[(100,10), (50,10), (100,20), (50,20)]	(50,10)
Train batch size	[8, 32, 64]	32
Dev batch size	32	32
Test batch size	32	32
Learning rate	[0.01, 0.001, 0.0001]	0.001
Epoch number	[5, 7, 10 ,15]	5
Dropout	[0.4, 0.5]	0.4

Results CANTEMIST-NER

Table 3
Results of CANTEMIST-NER on the test set

Precision	Recall	F1-score
0.826	0.843	0.834

Error Analysis

Table 4
Proportion of errors made in CANTEMIST-NER task.

Incorrect boundaries	Missed	Incorrectly distinguished
39.68%	34.52%	25.80%

Conclusions and Future Word

- Our contribution is a relatively straightforward architecture to explore CANTEMIST task that gets competent results (0.834 F1-score).
- The vast majority of mistakes made by the model are made when it confuses terminology due to context:
  a. Incorrect boundaries
  b. Missing the tumor morphology mention (context required)
  c. Incorrectly distinguishing the tumor morphology mention
- In future work, our goal is to explore different ways to include document-level information, as well as examine to other deep learning architectures and other types of embeddings such as contextual embeddings and knowledge-based resources

Bibliography

[1] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.

[2] A. Miranda-Escalada, E. Farré, M. Krallinger, Cantemist corpus: gold standard of oncologyclinical cases annotated with CIE-O 3 terminology, 2020. URL: https://doi.org/10.5281/zenodo.3978041. doi: 10.5281/zenodo.3978041 , Funded by the Plan de Impulso de lasTecnologías del Lenguaje (Plan TL).

[3] L. Ratinov, D. Roth, Design challenges and misconceptions in named entity recognition, in:Proceedings of the Thirteenth Conference on Computational Natural Language Learning(CoNLL-2009), 2009, pp. 147–155.

[4] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subwordinformation, arXiv preprint arXiv:1607.04606 (2016).

[5] F. Soares, M. Villegas, A. Gonzalez-Agirre, J. Armengol-Estapé, S. Barzegar, M. Krallinger,Fasttext spanish medical embeddings, 2020. URL: https://doi.org/10.5281/zenodo.3744326.doi: 10.5281/zenodo.3744326 , Funded by the Plan de Impulso de las Tecnologías delLenguaje (Plan TL).

[6] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprintarXiv:1412.6980 (2014).

[7] F. Chollet, et al., Keras, 2015. URL: https://github.com/fchollet/keras.