TALP at eHealth-KD Challenge 2020: Multi-Level Recurrent and Convolutional Neural Networks for Joint Classification of Key-Phrases and Relations

Salvador Medina[a], Jordi Turmo[a]

[a]Universitat Politècnica de Catalunya, Campus Nord, Carrer de Jordi Girona, 1, 3, 08034 Barcelona, Spain

Abstract

This article describes the model presented by the TALP Team to IberLEF's eHealth Knowledge Discovery 2020 shared task[1]. The model iterates over the idea of using a single model for simultaneously identify key-phrases and their relationships. Taking into account the new transfer-learning sub-task presented for 2020's edition of eHealthKD, our model does not rely on any domain-specific knowledge nor handcrafted features. Our model was competitive in all four sub-tasks, ranking in 2nd, 3rd, 4th and 1st position respectively.

Keywords

NERC, Relation Extraction, eHealth NLP, Contextual Embeddings

1. Introduction

This article describes the design choices and training strategy behind the model presented by the TALP team for IberLEF's eHealth Knowledge Discovery 2020 shared task[1]. This shared task consists of identifying relevant key-phrases and relationships among them in Electronic Health documents from Spanish Medline. eHealthKD 2020's edition includes two significant additions respect to previous editions: an ensemble dataset created by combining predictions from previous 2019 edition's model outputs when applied to an unlabelled dataset, and a new transfer-learning sub-task.

Our model iterates over our team's 2019 model[2] by leveraging several pre-trained word-level text representation models, as well as taking advantage of the automatically labelled corpus in a pre-training step. In particular, we make use of the pre-trained Word2Vec and FastText Medical Word Embedding for Spanish models[3] from Barcelona Super-computing Center, which were trained using the SciELO database and a health-related subset of the Wikipedia. We use these two models to add context-specific knowledge to our model, which we believe was one of the shortcomings of the superseded model. However, the results suggest using these word representations does not represent an appreciable improvement over the general-purpose ones for Scenarios 1 to 3 and may even be detrimental for Scenario 4.

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)
EMAIL: smedina@cs.upc.edu (S. Medina); turmo@cs.upc.edu (J. Turmo)
ORCID: 0000-0003-2473-8571 (S. Medina); 0000-0002-7521-1115 (J. Turmo)

2. System Description

Our model expects a document and a source token index as input and generates a sequence of labels for each key-phrase and relation class. Input documents are parsed using FreeLing's dependency parser and each one of their tokens are encoded using either a BERT, a Word2Vec or a FastText pre-trained word-embedding model. The model then applies convolution filters to the encoded tokens of the input documents, combines the word-level filter's outputs of each input token and the specified source token with sentence-level embeddings of the documents, and outputs the boundaries of each key-phrase containing the source token as well as the likelihoods that every other token is the target of a relation having the specified source token's key-phrases as a source.

In order to generate all possible relations, the model should be run for every input token and have the all raw likelihoods combined across every one of them. This approach of looking at a single input token at a time is inspired by attention-based translation models such as the Transformer, in which the model comes up with the most likely output token one at a time, conditioned to the previously generated tokens and the whole untranslated document.

2.1. Internal structure of the model

A visual representation of the model's structure is shown in Figure 1. The network is composed of a set of shared intermediate layers and two independent output layers. The intermediate layers include a Bidirectional Gated Recurrent Unit layer followed by a set of convolution filters. The recurrent units' and convolution outputs are finally concatenated and fed to a fully connected layer. The output layers consist of a fully connected layer followed by a Conditional Random Field layer.

This structure lets the model look at both the local and global contexts of each of the input tokens. Particularly, the local context is captured by the recurrent units' output and the non-pooled convolution layer's output, while the global context is captured by the max-pooled convolution layer's output. Additional global context information is added when the BERT-based model is used by concatenating the encoding of the auxiliary *CLS* token.

The global context information and the target token's local context information are added to all time-steps before being fed to the fully connected shared layer. The final outputs are then generated by a Conditional Random Field (CRF) layer. Output CRF layers have proven to improve the capabilities of GRU and LSTM networks in low-resource sequence tagging tasks[4].

2.2. Output generation and decoding

As described in Section 2, our system receives the sequence of tokens of a document and a token's index and outputs the bounds of the innermost key-phrase to which the token belongs. These bounds are encoded and decoded by assigning a Begin, Inside, Unitary and End tag to each token included in that key-phrase and Out to every other token (*BIOUE-tag*). One limitation of this approach is the fact that just one key-phrase is decoded for each token index, but this is not an issue in our case, as key-phrases may subsume but not overlap other key-phrases.

For each input token, our model outputs the list of relations' probabilities between the innermost entity to which the token belongs and each one of the tokens in the document is

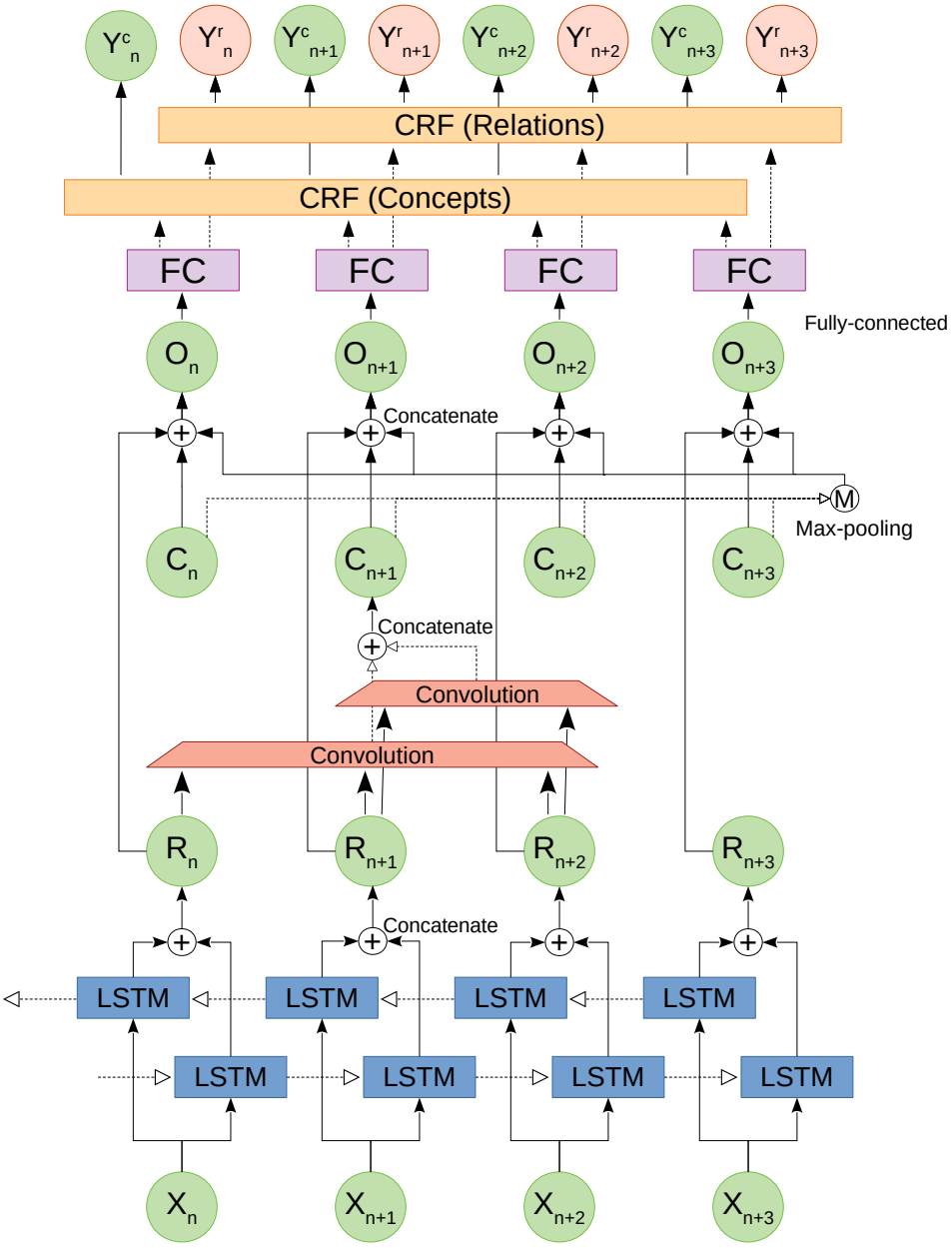

Figure 1: Schematic architecture of the identification artificial neural network

predicted. Note that for the source token, we only consider the innermost entity whereas for the target tokens we consider all parent entities. Consequently, our method does not allow for overlapping relations from the same source token. This restriction is imposed so that the encoded sequence is not ambiguous. A visual representation of relations' probability predictions is shown in Figure 2. Relations are predicted from the target key-phrase if the aggregated score

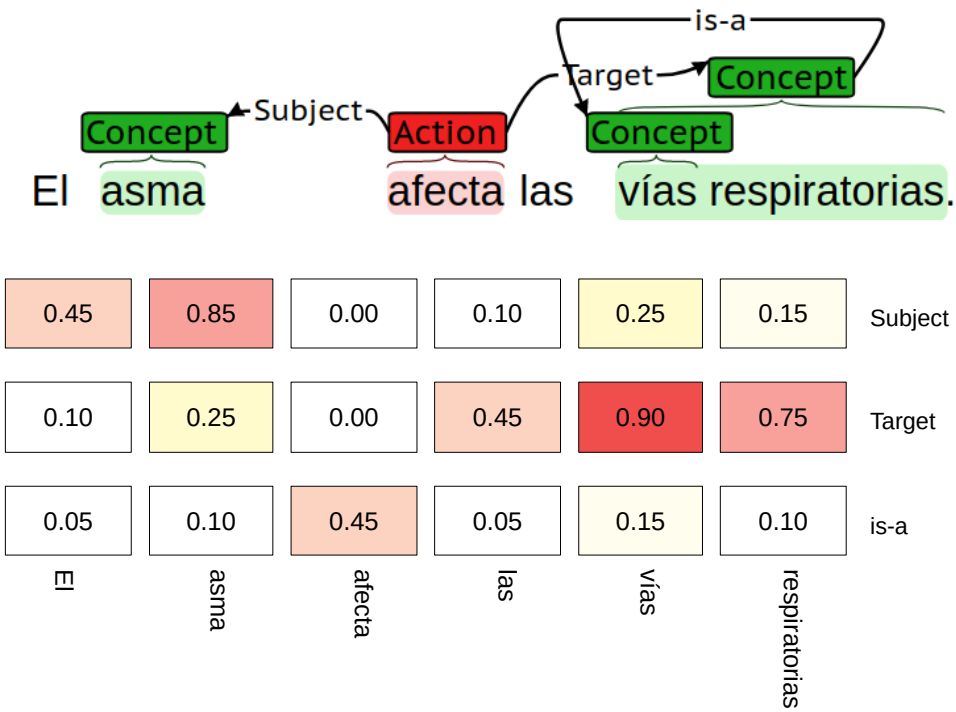

Figure 2: Visual representation of how relations and key-phrases are encoded by the network.

inside a key-phrase span surpasses a threshold. Only the key-phrase with the highest score is selected if multiple key-phrases overlap.

2.3. Input features

As previously mentioned, our model process the documents at the token level. We represent each token by a vector, which results from the concatenation of the features listed below:

- One-Hot encoding of the *category* and *type* fields of the token's Part-of-Speech Tag from FreeLing's tag-set.
- Normalized vector encoding the dependencies found in the path between the token and the target token (the one that is being decoded). It is computed by adding the one-hot encoding representation of the dependency class for each hop in the dependency path and normalizing the resulting vector, not considering its direction. For instance, the representation of the token "I" in "I eat fish" when the target token is "fish" would be a vector with $\sqrt{2}$ in the positions corresponding to "subj" (subject) and "cd" (direct complement); whereas for "eat" it would be a vector with a single 1 in the "cd" position.
- One-Hot encoding of the distance between the token and the target token.

- Word-embedding of the token. We consider 4 alternative pre-trained word embedding models:
  - Concatenation of the last output layers of a multi-language general-purpose BERT[5] model[1] with no fine tuning.
  - Word2Vec and FastText Medical Word Embedding for Spanish models from Barcelona Super-computing Center[2][3].
  - FastText Spanish Unannotated Corpora from SUC[3][6]

2.4. Pre-training with the ensemble corpus

Due to the comparatively large number of parameters in our model respect to the size of the training dataset, overfitting can be an issue. We prevent this by using the relatively larger but inaccurate ensemble in a pre-training phase. In order not to let our model's variables fall into local minima that would make our model mimic previous' years models, we randomly add documents from IberLEF 2020's training corpus. Furthermore, we increase dropout and gradually decrease the learning rate for the training and fine-tuning training steps.

2.5. Single-scenario training and fine-tuning

In the general evaluation scenario, the loss function has to balance accuracy for both the key-phrase recognition and relation extraction tasks. This may be problematic, as the parameter updates made by the optimizer to improve one task might be detrimental for the other task. However, in evaluation scenarios 2 and 3, that is, independent key-phrase recognition and relation extraction tasks, the model does not have to generate both outputs. Consequently, on the one hand, we can use an uncompromising loss function. On the other hand, this means not being able to exploit the correlation between tasks, so it might as well lead to worse performance.

To study this effect, we suggest thee different single-scenario training strategies: using the general model with no alteration whatsoever, fine-tuning the general model's outputs with independent loss function for a few epochs, or training the specific model from scratch. Note that in the case of scenario 3, we decode the key-phrases using the gold truth rather than the model's output for all three strategies; and concatenate a one-hot-encoding of the key-phrase labels to the input for the *from-scratch* strategy. Table 2 shows the results for all three single-scenario training strategies.

2.6. Trainable parameters and computational resources

All models were trained using the TensorFlow® 1.15 framework for Python® 3.6 in an 8 core Intel® Xeon® E5-2620 v4 CPU at 2.10GHz, 16GB of DDR4 RAM, a GeForce® GTX 1070 GPU

[1]We used the *BERT-Base, Multilingual Cased* model (104 languages, 12-layer, 768-hidden, 12-heads, 110M parameters) from the authors' repository (https://github.com/google-research/bert)

[2]We used April 15, 2020's SciELO + Wikipedia, 300 dimensions version of Medical Word Embedding for Spanish, which can be downloaded from https://zenodo.org/record/3744326

[3]We used the 300 dimensions sub-word binary model from https://github.com/dccuchile/spanish-word-embeddings/blob/master/emb-from-suc.md

Model	Scenario 1			Scenario 4		
	P	R	F_1	P	R	F_1
Vicomtech	0.679364	0.652315	**0.665564**	0.594009	0.535521	0.563251
UH-MAJA-KD	0.634542	0.615741	0.625	0.608321	0.49813	0.547739
Talp-UPC (submission)	0.626969	0.626389	0.626679	0.604724	0.563772	**0.58353**
Talp-UPC (BERT)	0.629630	0.627306	**0.628466**	0.464271	0.555970	0.506000
Talp-UPC (BERT FT)	0.602778	0.600000	0.601386	0.551309	0.618937	**0.583169**
Talp-UPC (W2V Health)	0.573148	0.606268	**0.589243**	0.382219	0.485488	0.427708
Tapl-UPC (FastText Health)	0.569444	0.598832	0.583768	0.400291	0.496906	0.443396
Tapl-UPC (FastText General)	0.574537	0.592646	0.583451	0.418363	0.496059	**0.453910**

Table 1
Final evaluation results of *IberLEF2020's eHealth-KD* scenarios 1 and 4. We also include the evaluation for the context-specific corpora. BERT FT is fine-tuned using the the transfer-learning development corpus.

and a 7200rpm 1TB Seagate® HDD.

BERT-based and Word2Vec/FastText-based models were trained for a total of 128 and 96 epochs respectively, divided among the pre-training, training and fine-tuning steps. Training epochs were evenly distributed between pre-training and training steps for models with no fine-tuning. When fine-tuning was applied (transfer-learning or single-task scenarios), pre-training was shortened by 16 epochs.

For each word representation model, independent models were trained with 8, 32 and 64 convolution filters of sizes 3 and 5; and 8, 32 and 64 single-layer recurrent units.

3. Results

Tables 1 and 2 show the results for all four scenarios of IberLEF's 2020 eHealth Knowledge Discovery shared task for the best scoring submissions. Our model is competitive in all the scenarios, landing in second, third, fourth and first position respectively.

The scores obtained when using the context-specific Word2Vec (W2V Health), FastText (FastText Health) and general-purpose FastText (FastText General) models are shown in the bottom half of Table 1. With these models, see a drop of 0.04 in F_1 score in scenario 1 respect to the BERT-based model and a 0.13 in scenario 4. The difference in score for scenario 1 between the context-specific and general-purpose models is negligible, whilst we see a drop of 0.01 in F_1 for scenario 4.

Table 1 also shows the independent results of our BERT-based model when fine-tuned with the transfer-learning development corpus. The fine-tuned model sees an increase of 0.08 in F_1 score for Scenario 4 while only seeing a 0.03 drop in Scenario 1.

As it can be seen in Table 2, our model is least competitive in scenario 3, in which it is outperformed by IXA-NER-RE's model by a margin of almost 0.06 in F_1 score. However, this score is matched the unsubmitted *from-scratch* model described in Section 2.5. Similarly, the *general* model strategy is comparable to SINAI's model in scenario 2.

Model	Scenario 2			Scenario 3		
	P	R	F_1	P	R	F_1
SINAI	0.844633	0.806655	**0.825207**	0.627063	0.365385	0.461725
Vicomtech	0.821622	0.820144	0.820882	0.671679	0.515385	0.583243
IXA-NER-RE	0.726733	0.660072	0.6918	0.647887	0.619231	**0.633235**
UH-MAJA-KD	0.820255	0.808453	0.814312	0.629237	0.571154	0.59879
Talp-UPC (fine-tuned)	0.807218	0.82464	0.815836	0.646635	0.517308	0.574786
Talp-UPC (general)	0.841727	0.808290	**0.824670**	0.501923	0.617021	0.553552
Talp-UPC (from-scratch)	0.821942	0.810284	0.816071	0.592308	0.678414	**0.632444**

Table 2
Final evaluation results of *IberLEF2020's eHealth-KD* Scenarios 2 and 3. We also include the evaluation for the two additional single-scenario training strategies, which were not submitted.

4. Discussion

The joint key-phrase classification and relation extraction model presented by our team for the previous edition of IberLEF's eHealth Knowledge Discovery shared task outperformed every other participant model by a wide margin. This confirmed our belief that a joint model has the potential to exploit the mutual information between the two tasks and provide better evaluation results than traditional step-by-step architecture. The improvement was, however, less appreciable for the key-phrase classification task.

After comparing our model to the rest of the participant's submissions, we hypothesised that one of the main shortcomings of ours was the absolute lack of context-specific knowledge. For this year's edition, we decided to explore different alternatives to tackle this. But since a new transfer learning scenario was added, whose evaluation score would probably be compromised if the source model relied too heavily upon context-specific features, we opted for adding this context-specific information in a way that would not significantly alter the model's structure nor make it less general with handcrafted rules. Particularly, we opted for swapping the general-purpose word representation model by a health-specific one.

Unfortunately, the results show that the use of context-specific word embeddings does not substantially improve upon general-purpose embeddings and even leads to worse results in the transfer-learning scenario. Not only that, but we have also shown that contextual word embeddings such as BERT and XLNet significantly outperform predictive word embedding models such as Word2Vec and FastText. Moreover, the concatenation of this second word representation does not seem to provide any additional information over the original, whilst it makes the model more complex in terms of the number of trainable parameters.

Several hypotheses may explain these unsatisfactory results. First, we argue that although the documents' language register is formal, the use of technical terms is limited. Similarly, relation classes and specially key-phrase categories are arguably general, as pointed out by the results obtained in Scenario 4. Secondly, predictive word embedding models may not be able to capture the medical terms' semantic information to a degree that can be used by our model, but rather more explicit features may be preferable.

5. Conclusions

In this article, we have described the main characteristics of the model that we have developed for TALP team's submission to IberLEF's 2020 eHealth Knowledge Discovery shared task. Our model follows the trend started by our team's 2018's model, which consists of using a single network with shared weights that jointly performs the key-phrase recognition and relation extraction tasks to leverage the mutual information between the two. It has proven to be competitive against the other participants model's, especially in the general and transfer-learning scenarios, ranking in second and first position respectively. The transfer-learning scenario particularly highlights the adaptability and context-independence of our model.

Three main improvements were made over the previous year's model: adaptive learning-rate for pre-training, single scenario fine-tuning and context-specific word vector representations. The last of which has been rather underwhelming though, and we conclude that adding context-specific information to our model is still an unresolved issue.

Besides the aforementioned limitation, we see other shortcomings to our model that still need to be tackled to more accurately capture the mutual information between the two knowledge discovery tasks. Among these improvements, we would like to point out two that we believe are more promising:

- Use a trainable combination function for the outputs generated by the model for different source tokens in a document. Our current model, on the other hand, uses a simple union operation to join the predictions for the different tokens of single key-phrase.
- Use of fine-tuned context-specific contextual word embedding model. The use of context-specific predictive word embeddings have proven not successful for our model, but general-purpose contextual word embeddings can be fine-tuned with context-specific unlabelled corpora.

Acknowledgments

This contribution has been partially funded by the Spanish Ministry of Economy (MINECO) and the European Union (TIN2016-77820-C3-3-R and AEI/ FEDER,UE).

References

[1] A. Piad-Morffis, Y. Gutiérrez, H. Cañizares-Diaz, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, A. Montoyo, Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2020, in: Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2020, Spain, September, 2020., 2020.

[2] S. Medina Herrera, J. Turmo Borras, 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.

[3] F. Soares, M. Villegas, A. Gonzalez-Agirre, M. Krallinger, J. Armengol-Estapé, Medical word embeddings for Spanish: Development and evaluation, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 124–133. URL: https://www.aclweb.org/anthology/W19-1916. doi:10.18653/v1/W19-1916.

[4] Z. Huang, W. Xu, K. Yu, Bidirectional lstm-crf models for sequence tagging, arXiv preprint arXiv:1508.01991 (2015).

[5] 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).

[6] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.