Projecting Heterogeneous Annotations for Named Entity Recognition

Rodrigo Agerri[0000−0002−7303−7598] and German Rigau[0000−0003−1119−0930]

HiTZ Center - Ixa
University of the Basque Country UPV/EHU
rodrigo.agerri@ehu.eus

Abstract. In this paper we describe our participation in the CAPITEL at IberLEF 2020 shared task on Named Entity Recognition (NER). Our objectives to participate in the shared task were two-fold: (i) to benchmark current rich multilingual representations of text with respect to monolingual models trained specifically for Spanish; (ii) to study various methods of projecting annotations from several sources into a final target prediction. Our results show that monolingual models, even for a large language such as Spanish, perform better in this particular NER benchmark. Furthermore, our projection method indicates that substantial gains in performance can be obtained by projecting annotations from various heterogeneous sources to obtain the final prediction. Our submission obtained the best score, substantially outperforming other participants of the CAPITEL 2020 NER task.

Keywords: Named Entity Recognition · Information Extraction · Natural Language Processing.

1 Introduction

Named Entity Recognition (NER) is a widely studied Natural Language Processing (NLP) task. Briefly, the task involves annotating any mentions of entities (usually proper names) occurring in running text. The most common annotated corpora for NER focuses on four types of named entities: Locations, Organizations, Persons and Other (Miscellaneous) entities. Spanish NER has been well-studied, as it was one of the languages proposed in the CoNLL NER shared tasks [21, 20].

As for many other NLP tasks, current best performing models for NER are those based on large pre-trained language models which allow to build rich representations of text based on contextual word embeddings. These approaches are based on character-based models like Flair [4] or masked language models like BERT [7]. Furthermore, multilingual versions of these models have been

trained: the multilingual version of BERT [7] was trained for 104 languages. More recently, XLM-RoBERTa [6] was trained for 100 languages.

These publicly available deep learning multilingual models for text excel in tasks involving high-resourced languages such as English, but their performance drops when applied to low-resource languages [2]. This may occur for a number of reasons. First, each language has to share the quota of substrings and parameters with the rest of the languages represented in the pre-trained multilingual model. As the quota of substrings partially depends on corpus size, this means that larger languages such as English or Spanish are better represented than lower resourced languages such as Basque [2]. Moreover, multilingual models also seem to behave better for structurally similar languages [11].

In our submission for the CAPITEL 2020 NER task [17] we leverage both these multilingual models as well as other monolingual models trained specifically for Spanish. Furthermore, we project the annotations provided by each system into a target final prediction. The projection of several source annotations into a target is loosely inspired by a method originally designed for projection of annotations across languages [1]. Our projection method indicates that substantial gains in performance (around 1.3 points in F1 score) can be obtained by projecting annotations from various heterogeneous sources into a final target prediction. Our submission obtained the best score, substantially outperforming other participants of the CAPITEL 2020 NER task.

2 Related Work

Deep learning methods in NLP rely on the ability to represent words as continuous vectors on a low dimensional space, called word embeddings. The first approaches generated static word embeddings [14, 5], namely, they provided a unique vector-based representation for a given word independently of the context in which the word occurs. This means that polysemy cannot be represented. Thus, if we consider the word 'bank', static word embedding approaches will generate only one vector representation even though such word may have different senses, namely, 'financial institution','bench', etc.

In order to address this problem, contextual word embeddings were proposed. The idea is to be able to generate different word representations according to the context in which the word appears. Currently there are many approaches to generate such contextual word representations, but we will focus on those that have had a direct impact, in terms of performance, for the Named Entity Recognition task. First, Flair [4] representations are built following a LSTM-based architecture and trained as language models. Second, the models based on the transformer architecture [22] and of which BERT is perhaps the most popular example [7].

The multilingual counterpart of BERT, called mBERT, is a single language model pre-trained from corpora in more than 100 languages. Another standout model is XLM-RoBERTa [6] also based on the transformer architecture which provides a pre-trained language model for 100 languages trained on 2.5 TB

of Common Crawl text. Both mBERT and XLM-RoBERTa enable to perform transfer knowledge across languages [8, 16, 11], although in this paper we will use them in a monolingual setting for Spanish NER.

2.1 Flair

Flair refers to a system based on a BiLSTM architecture [9] and to a specific type of character-based contextual word embeddings. Flair (embeddings and system) have been successfully applied to sequence labeling tasks obtaining state-of-the-art results for a number of Named Entity Recognition (NER) and Part-of-Speech tagging benchmarks [4].

Flair embeddings consist of sequences of characters. More specifically, sentences are processed into sequences of characters and feed into a character-level Long short-term memory (LSTM) model. For each sentence, a forward LSTM language model processes the its sequence of characters from the beginning of the sentence to the last character of the word we are modeling. Furthermore, a backward LSTM performs the same operation going from the end of the sentence up to the first character of the word. The extracted hidden states contain information propagated from the end and the beginning of the sentence up to the first and the last character of the target word. Finally, the resulting two hidden states are concatenated to generate the final embedding.

Pooled embeddings are a type of Flair embeddings which consider global information in order to generate the final word embedding [3]. In this approach embeddings are kept into a *memory* which is later used in a pooling operation to obtain a global word representation. This representation will be the concatenation of all the local Flair contextualized embeddings obtained for a given word. It should be consider that pooling operation is involved in the process of fine-tuning the Flair pre-trained models, not in the process of training the language models themselves. We use the default pooling operation, *min*, which computes a vector of all element-wise minimum values [3].

2.2 Transformers

LSTM-based language models such as the one presented in the previous section cannot capture long-range sequence information. Furthermore, they are quite hard to train at a large scale (see [10], especially Figure 7). In order to address these issues, the Transformer architecture was proposed [22], based on multi-headed self-attention and positional encoding. The most popular Transformer is BERT [7], which pre-trains a Transformer encoder on the Masked Language Model (MLM) and Next Sentence Prediction (NSP) tasks. BERT is composed by stacked layers of Transformer encoders [22]. More specifically, in this paper we will use the $\text{BERT}_{\text{BASE}}$ configuration which contains 12 Transformer encoder layers, a hidden size of 768 and 12 self-attention heads for a total of 110M parameters.

The MLM task is designed as follows: For a input sequence of n tokens $x_1, x_2, ..., x_n$, 15% are selected as masking candidates. From those candidates,

80% of them are masked (they are replaced with the [MASK] token), 10% are replaced by a random word and the last 10% is left unchanged. For the NSP task, two segments are selected from the training corpus, A and B. In 50% of the cases B is the true next segment for A. For the rest, B is just a random segment. The model is trained to optimize the sum of the means of the MLM and NSP likelihoods.

It should be noted that the benefits of the NSP task during the pre-training process has been questioned [23, 13, 12]. Thus, other transformer proposals such as RoBERTa train without the NSP task, showing strong performance on the same downstream tasks.

XLM-RoBERTa relies exclusively on the MLM objective. The biggest update that XLM-Roberta offers is a significantly increased amount of training data, 2.5TB of Common Crawl clean data [6]. As for BERT, in this paper we use the base version of XLM-RoBERTa. The reason being that their base versions fit for fine-tuning into a standard GPU card with 12GB of RAM.

3 Experimental Setup

Named entities were originally annotated using the BIO encoding which identifies the Beginning, the Inside and the Outside of named entities. Later on the BILOU model[1] was proposed to mark tokens as the Beginning, the Inside and the Last tokens of multi-token entities as well as Unit-length entities [18]. Although the CAPITEL corpus is originally released using the BILOU model, we experiment with both type of encodings.

The CAPITEL (Corpus del Plan de Impulso a las Tecnologías del Lenguaje) has been developed by the PlanTL, the Royal Spanish Academy (RAE) and the Secretariat of State for Digital Advancement (SEAD) of the Ministry of Economy. These organizations signed an agreement for developing a linguistically annotated corpus of Spanish news articles, with the objective of extending the language resource infrastructure for the Spanish language. CAPITEL is composed of contemporary news articles and contains annotations for Universal Dependencies and Named Entities. The NER portion of the corpus contains around one million words.

For the experiments performed for this paper, we use a number of publicly available models:

1. Multilingual BERT (mBERT).
2. XLM-RoBERTa (base).
3. BETO, a monolingual Spanish BERT trained with Wikipedia and Spanish data from the OPUS corpus [19].
4. Flair official models for Spanish.

Additionally we trained the following monolingual language models for Spanish:

[1] Nowadays also known as the BIOES encoding: Beginning, Inside, Outside, End of entity and Single entity.

1. Flair-GW: Flair character-based language model trained on the Spanish Wikipedia and the Gigaword 3rd edition corpus, containing around 11GB of text.
2. Flair-Oscar: Flair language model trained on the OSCAR Spanish corpus [15], which contains 157GB of Common Crawl text cleaned and deduplicated.

The Flair embeddings for Flair-GW and Flair-Oscar were trained with the following parameters: Hidden size 2048, sequence length of 250, and a mini-batch size of 100. The rest of the parameters were left in their default setting. For Flair-GW, training was done for 5 epochs over the full training corpus. The training took around 5 days in a Nvidia Titan V GPU. With respect to Flair-Oscar, only one epoch was performed, requiring around a month to complete it.

4 Results

Table 1 reports only the best results obtained during the experimentation. Each of the S1-S8 results is the average of five randomly initialized runs. Flair models were trained using the default parameters, although we experimented adding FastText embeddings to the Flair and Pooled embeddings. We used 10 percent of the training data for development of the Flair models. In the case of the Transformer models described in the previous section, we used the full training set for hyperparameter fine-tuning. For XLM-RoBERTa we used a maximum sequence length of 128, mini-batch 16, 5e-5 learning rate, and 4 epochs. For mBERT and BETO best results were obtained using the same hyperparameters as for XLM-RoBERTa but increasing the sequence length to 256.

Table 1. Overview results on both development and test data.

		Development			Test		
	System	Precision	Recall	F1 score	Precision	Recall	F1 score
S1	Flair-Oscar + FT	89.65	89.36	89.51	88.86	88.63	88.74
S2	Flair-Oscar + FT (dev)	89.67	89.53	89.60	88.97	88.75	88.86
S3	Pool-Oscar + FT (dev)	**89.85**	89.63	**89.79**	89.07	**88.85**	88.96
S4	Pool-Oscar + FT e1	89.78	**89.72**	89.75	**89.29**	88.82	**89.07**
S5	Flair-Oscar + FT BIO	89.71	89.58	89.64	89.19	88.78	88.99
S6	BETO	89.64	89.34	88.99	87.19	88.36	87.77
S7	mBERT	87.90	88.90	88.40	87.03	87.75	87.39
S8	XLM-RoBERTa	88.29	89.54	88.91	87.37	88.48	87.92
P1	S2-S3-S6-S7-S8	**91.32**	**90.77**	**91.04**	90.70	88.11	89.38
P2	S2-S4-S6-S7-S8	91.10	90.59	90.84	**90.81**	88.06	89.42
P3	**S3-S4-S6-S7-S8**	91.19	90.72	90.96	90.50	**90.17**	**90.34**

Out of the many experiments performed with the three Flair language models (Official, GW and Oscar), the best performing language model in every possible

configuration was the Flair-Oscar model combined with the FastText embeddings trained on Wikipedia. In fact, Flair-Oscar was the best single system by a substantial margin. Apart from this, S2 and S3 show the small gains obtained by adding the 10 percent used for development for the final evaluation. Furthermore, S3 was trained when the progress of training the language model was at half epoch, whereas S4 was trained using the final Oscar language model based on one epoch. Finally, S5 is the same model as S1 but using BIO encoding instead of the original BILOU encoding from the CAPITEL corpus. The best overall invididual system was S4, significantly outperforming the multilingual and monolingual Transformer models.

With respect to the transformer models, it can be seen that in general their results are lower than those obtained by the Flair-Oscar models. During the development phase they all performed very closely although in the final, official results XLM-RoBERTa was slightly superior to the rest. Furthermore, results also show that mBERT performed worst and that XLM-RoBERTa obtains very similar results to the monolingual models.

The last three rows of Table 1 report the three best projections. Once we had the best 8 systems, we proceeded to project their predictions for any possible combination of the 8 systems. The best three systems were picked based on two criteria: the F1 score obtained on the development data and the number of No-agreements recorded by each projection.

The projections were performed using 5 predictions as source. We tested various strategies and the one we finally used to report the final results was, interestingly enough, the simplest of them all. It uses a very simple methodology based on the number of agreements between the predicted labels of the 5 source annotations: if agreement is >= 3 then project, otherwise, project "O".

As we could not compute F1 scores on the official test set released by the shared task, we simply picked the projection which recorded fewer No-agreements. This corresponds to the best overall system (P3), which uses S3, S4, S6, S7 and S8 as source to obtain the final prediction.

5 Concluding Remarks

In this paper we have described the experiments performed for our participation in the CAPITEL 2020 shared task on Named Entity Recognition. Even though the best results are obtained by the Flair-Oscar monolingual models, our results indicate that multilingual pre-trained models such as XLM-RoBERTa are performing increasingly close to monolingual models for a large-resourced language such as Spanish. Furthermore, we also show the benefits of projecting named entity annotations from various heterogeneous sources in order to substantially improve performance (around 1.3 points in F1 score over the best individual system).

Acknowledgments

This work has been partially funded by the Spanish Ministry of Science, Innovation and Universities (DeepReading RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE) and by *Ayudas Fundación BBVA a Equipos de Investigación Científica 2018* (BigKnowledge). Rodrigo Agerri is funded by the RYC-2017-23647 fellowship and acknowledges the donation of a Titan V GPU by the NVIDIA Corporation.

References

1. Agerri, R., Chung, Y., Aldabe, I., Aranberri, N., Labaka, G., Rigau, G.: Building named entity recognition taggers via parallel corpora. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (2018)
2. Agerri, R., San Vicente, I., Campos, J.A., Barrena, A., Saralegi, X., Soroa, A., Agirre, E.: Give your text representation models some love: the case for basque. In: Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020). pp. 4781–4788 (2020)
3. Akbik, A., Bergmann, T., Vollgraf, R.: Pooled contextualized embeddings for named entity recognition. In: NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics. p. 724–728 (2019)
4. Akbik, A., Blythe, D., Vollgraf, R.: Contextual string embeddings for sequence labeling. In: COLING 2018, 27th International Conference on Computational Linguistics. pp. 1638–1649 (2018)
5. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. arXiv:1911.02116 (2019)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
8. Heinzerling, B., Strube, M.: Sequence tagging with contextual and non-contextual subword representations: A multilingual evaluation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 273–291. Association for Computational Linguistics, Florence, Italy (Jul 2019)
9. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF Models for Sequence Tagging (2015)
10. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
11. Karthikeyan, K., Wang, Z., Mayhew, S., Roth, D.: Cross-lingual ability of multilingual bert: An empirical study. In: International Conference on Learning Representations (ICLR) (2020)

12. Lample, G., Conneau, A.: Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291 (2019)
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. pp. 3111–3119 (2013)
15. Ortiz Suárez, P.J., Sagot, B., Romary, L.: Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. pp. 9–16. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019 (2019)
16. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual bert? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 4996–5001 (2019)
17. Porta-Zamorano, J., Espinosa-Anke, L.: Overview of CAPITEL Shared Tasks at IberLEF 2020: NERC and Universal Dependencies Parsing. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) (2020)
18. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning. pp. 147–155 (2009)
19. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: LREC. vol. 2012, pp. 2214–2218 (2012)
20. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. pp. 142–147 (2003)
21. Tjong Kim Sang, E.F.: Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In: Proceedings of CoNLL-2002. pp. 155–158. Taipei, Taiwan (2002)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
23. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint arXiv:1906.08237 (2019)