Experiments from LIMSI at the French Named Entity Recognition Coarse-grained task

Sahar Ghannay[1][0000−0002−7531−2522], Cyril Grouin[1][0000−0001−5809−188X], and Thomas Lavergne[1][]

Université Paris-Saclay, CNRS, LIMSI, 91405 Orsay, France
first.last@limsi.fr

Abstract. This paper presents the participation of the LIMSI team in the HIPE 2020 Challenge on the Coarse-grained named entity recognition task for French. Our approach jointly predicts the literal and metonymy entities. For this, a CamemBERT base model and a CRF model were used. We submitted three systems: a joint model using only CamemBERT, a joint model extended with a CRF layer, and a CamemBERT model without joint option. Experimental results show that the second system achieved best results on the literal tags (F1=.814) while the third system performed best (F1=.667) on the metonymy tags. The second system allowed us to obtain our best results on both the dev and test datasets for the literal tags. Nevertheless, we observed a difference on the metonymy tags where our first system obtained best results on the dev dataset (F1=.663) while our third system performed best on the test dataset (F1=.667).

Keywords: Named Entity Recognition, historical texts, contextual word embeddings

1 Introduction

In 2011 and 2012, two corpora have been produced and annotated into extended named entities. The 2011 Quaero corpus focused on broadcast news [7] while the 2012 Quaero corpus is composed of press archives in French from December 1880 [6]. Those two corpora were used for NLP Challenges that included both coarse-grained and fine-grained named entities, and several named entity imbrications such as the metonymy phenomena. The current HIPE 2020 Challenge builds on the annotation guidelines produced during the 2011 and 2012 Quaero NLP Challenges [7, 6].

Specifically for the HIPE 2020 challenge [4], one main issue concerns the digitization of texts from distinct times (from 1798 to 2018 on the French data) with digitization errors such as insertion and deletion of characters (e.g., "*oppositipjn*" instead of "opposition"), including insertion and deletetion of spaces which produces tokenization issues ("*limitrop he*" vs. "limitrophe" or "*rég iment*" vs. "régiment", producing two

tokens instead of only one). Digitization errors mainly occur on grammatical words.[1] Nevertheless, one may find such errors in named entity, which makes a NER task more difficult (e.g., in the person name "*Picqu_art*" instead of "Picquart" or in the town name "*Glascow*" instead of "Glasgow").

The CLEF HIPE 2020 challenge proposed several tasks (coarse-grained and fine-grained named entity recognition (NER), and entity linking) in three languages (English, French, German). We are interested in the sub-task 1.1 called NERC Coarse-grained for French, that concerns the recognition and classification of entity mentions according to coarse-grained types (Person, Location, Organisation and Product). For this task we distinguish two coarse types of the entity mention token: according to the literal sense and to the metonymic sense, named respectively **literal** and **metonymy** tags. These coarse types correspond to NE-COARSE-LIT and NE-COARSE-METO columns in the data.

For this challenge we proposed three neural systems that take benefit from contextual word embeddings using Camembert [22] model. This model was extended to jointly predict literal and metonymy NE tags, in addition to the use of a CRF layer on the top of this model to further improve the predictions by taking advantage from neighborhoods labels.

The paper is organized along the following lines: Section 2 presents related work on NER task. Section 3 describes the proposed NER system. The experimental setup and results are described in Section 4, just before the conclusion (Section 5).

2 Related work

The Named Entity Recognition (NER) task consists in identifying text spans that mention named entities (people names, companies, location) and classifying them into predefined categories (Person, Location, Organisation and Product). NER task is a key component of several Natural Language Processing (NLP) applications such as information retrieval [8], text understanding [29], question answering [23].

For decades, the NER task has been widely studied and different approaches have been proposed. Traditional approaches can fall into three categories [28, 15]: rule-based [10], unsupervised learning [5], and feature-based supervised learning approaches [30, 17]. Recent approaches are based on neural network architectures in which hidden features are discovered automatically. Generally the NER architecture can be regarded as a composition of an encoder (CNN, BiLSTM (bi-directional long-short term memory), RNN, transformer, *etc*.) and a decoder (BiLSTM, CRF, *etc*.) [18]. The first NER neural model was proposed by [9] which is based on unidirectional LSTM architecture. Collobert *et al.* [2] proposed a CNN-CRF architecture enriched by character-level embeddings. Lample *et al.* [14] proposed a BiLSTM-CRF architecture that takes benefit from both word and character-level embeddings. State of the art NER systems leverage recent advances in deep learning and recent approaches that take benefit from contextual or language model embeddings such as BERT [20, 16, 18].

[1] The most common errors are found in short grammatical words (error/correct form): Cn/Un, co/ce, cotte/cette, do/de, k/à, lai/lui, lo/le, on/en, quo/que, uno/une.

3 Proposed NER system

The proposed NER system is based on CamemBERT [22] model which we extended to jointly predict both NE tags: literal and metonymy. In the following subsections we briefly define the CamemBERT model and then present the proposed joint NER model.

3.1 CamemBERT

The CamemBERT model is based on RoBERTa (Robustly Optimized BERT Pretraining Approach) [19] which is based on BERT (Bidirectional Encoder Representations from Transformers) [3].

BERT's model architecture is a multi-layer bidirectional Transformer [27] encoder, trained with a masked language modeling and Next Sentence Prediction objectives. RoBERTa was proposed to improve BERT pre-training procedure by dynamically changing the masking pattern applied to the training data, removing the next sentence prediction task, and training with larger batches and longer sequences, on more data, and for longer.

Similar to BERT and RoBERTa, CamemBERT is a multi-layer bidirectional Transformer. It uses the original architectures of BERTBASE (12 layers, 768 hidden dimensions, 12 attention heads, 110M parameters) and BERTLARGE (24 layers, 1024 hidden dimensions, 12 attention heads, 110M parameters). CamemBERT is similar to RoBERTa, using the improved pre-training procedure. However it uses the whole-word masking and the SentencePiece tokenization [12] instead of WordPiece [25]. For more details about CamemBert we refer the reader to Martin et al. [22].

3.2 Joint NER

As we mentioned before, the NER system has to train jointly both literal and metonymy NE tags. Inspired by previous work on Intent Classification and Slot Filling [1], we propose to extend CamemBERT for this purpose. Hence, the final hidden states of the tokens h_2 , ..., h_T (excluding the first special token (<s>) fed into two softmax layers to classify over literal and metonymy tags. Specifically, each tokenized input word fed into a SentencePiece tokenizer and the hidden state of the first sub-token is used as input to the softmax classifiers. The literal and metonymy tags are predicted respectively as :

$$y_i^{LIT} = softmax(W h_i + b), i \in 1...N \tag{1}$$

$$y_i^{METO} = softmax(W h_i + b), i \in 1...N \tag{2}$$

where h_i is the hidden state of the first sub-token for the word w_i.

To jointly learn literal and metonymy tags, the learning objective is to maximize the conditional probability defined as follows:

$$p(y^{LIT}, y^{METO}|w) = \prod_{i=1}^{N} p(y_i^{LIT}|w) p(y_i^{METO}|w) \tag{3}$$

The model is fine-tuned end-to-end via minimizing the cross-entropy loss.

3.3 CRF

For the NER task, label predictions are dependent on surrounding words' predictions. Thus, for a given input sentence, it is helpful to consider the correlations between neighborhood labels and jointly decode the best label chain. It has been shown that the use of conditional random fields (CRF) [13] layer on top of BiLSTM (bi-directional long-short term memory) encoder improves many sequence labeling task including NER [21]. For that reason, we propose to add a CRF layer for modeling NE label dependencies, on top of the joint CamemBERT model.

4 Experiments and results

4.1 Data description

For the Coarse-grained NER task, we used the French datasets provided by the organizers. The corpus is divided into $train$, dev and $test$ sets, which are composed respectively of 158, 43 and 43 documents from distinct periods of time. As the document size is very long to be processed with our model, we decided to split the document to several sentences of length $\geq l$.

Two rules are considered during splitting: i) take into account the NE tags *i.e.* we have to reach the end of the tag before splitting, ii) take into account the end of the document, if the remaining part of the document is less than or equal to $2 * l$, than we consider the remaining document as a sentence. Both rules are applied to train and dev datasets, while only the second rule was applied to the test dataset since annotations were masked.

After hyper-parameter fine-tuning, we defined the minimum length size to 50, thus the sentence length varies from 50 to 149. Table 1 reports sentence numbers for each data set.

Table 1. Sentence numbers for each data set for French corpus

Data sets	**#sentences**
train	3080
dev	693
test	755

4.2 Training details

We used the CamemBERT base model as provided by its authors,[2] which is composed of 12 layers, 768 hidden dimensions and 12 attention heads. CamemBERT is pre-trained on the French part of the OSCAR [26] corpus: a pre-filtered and pre-classified version

[2] https://camembert-model.fr/

of Common Crawl, composed of 138GB of raw text and 32.7B tokens after subword tokenization.

For fine-tuning, all hyper-parameters are tuned on the development (dev) set. The minimum sentence length is selected from [10,20,50]. The max sequence length is 256. The batch size is selected from [32, 64, 128]. The maximum number of epochs is 100. For optimization we used Adam [11] with an initial learning rate of 5e-5. The dropout probability is 0.1.

4.3 Results

This section reports the results of the best three submitted systems, namely:

- sys1: fine-tuning the joint NER model without CRF layer;
- sys2: fine-tuning the joint NER model with CRF layer;
- sys3: fine-tuning the CamemBERT base model without joint option, hence for this model the literal and metonymy tags are concatenated and considered as only one tag. This system has more tags to predict than $sys1$ and $sys2$;

The results are evaluated at entity and document levels in terms of micro and macro Precision, Recall and F1-measure considering two scenarios : exact (strict) and fuzzy (relaxed) boundary matching, for both literal and metonymy tags [24].

The best systems are selected based on the results on the dev set, whose results in terms of micro Precision, Recall and F1-measure for both tags are summarized in Table 2. Note that our system is often the second best on French.

Table 2. Micro Precision, Recall and F1-score results for literal and metonymy tags on dev data set in both strict and fuzzy scenarios.

Tag	Systems	Strict			Fuzzy		
		F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
LIT	sys1	0.873	0.874	0.872	0.92	0.921	0.919
	sys2	**0.882**	0.879	0.886	**0.929**	0.925	0.933
	sys3	0.873	0.87	0.875	0.925	0.922	0.927
METO	sys1	**0.663**	0.725	0.611	**0.663**	0.725	0.611
	sys2	0.646	0.711	0.593	0.646	0.711	0.593
	sys3	0.612	0.573	0.657	0.612	0.573	0.657

Similarly as on the dev set, for literal tag, the three systems achieve comparable results. For metonymy, $sys1$ achieves better results than $sys3$ and $sys2$, yielding respectively to 8.82% and 2.63% of improvements in terms of micro F1 in strict and fuzzy scenarios, while $sys2$, that includes a CRF layer on top of the joint NER model, improves the results at the document level in terms of macro F1 (0.68) by 2.25% and 1.34% respectively for $sys3$ and $sys1$ in both scenarios.

Results on test data in terms of micro Precision, Recall and F1-measure for both tags are summarized in Table 3.

Table 3. Micro Precision, Recall and F1-score results for literal and metonymy tags on test data set in both strict and fuzzy scenarios.

Tag	systems	Strict			Fuzzy		
		F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
LIT	sys1	0.801	0.791	0.811	0.897	0.886	0.908
	sys2	**0.814**	0.799	0.829	0.896	0.88	0.913
	sys3	0.807	0.798	0.818	**0.898**	0.887	0.909
METO	sys1	0.603	0.69	0.536	0.603	0.69	0.536
	sys2	0.627	0.696	0.571	0.637	0.707	0.58
	sys3	**0.667**	0.647	0.688	**0.675**	0.655	0.696

We observe that for literal tag the proposed systems obtain comparable results.

The prediction of metonymic tags is not an easy task, since it represents only 0.27% B-org and 0.05% of I-org (in each of train, dev and test sets) and a few rare tags B-Loc in train and dev and B-time in test. Considering, the prediction of metonymy tags as a separate task (the case for $sys1$ and $sys2$), this may cause generalization problems. The results we obtained on the test data confirm this hypothesis. Indeed, the $sys3$ trained on the concatenation of both tags achieved the best results in terms of micro F1 w.r.t. $sys1$ and $sys2$, which is not the case on dev. Thus, the concatenation of both tags is helpful to predict metonymy tag and to avoid their frequency problem. *i.e.* considering those labels BI-loc_BI-org vs. BI-loc_O, it easier for the system to distinguish the LOC category with or without metonymy.

Last, $sys1$ achieves the best results in terms of macro F1 (0.747) in comparison to $sys2$ (.738) and $sys3$ (.733).

5 Conclusions

In this paper, we presented our participation in the CLEF HIPE 2020 Challenge on the Coarse-grained named entity recognition task for French. The proposed approach jointly predicts the literal and metonymic entities. For this, a CamemBERT base model and a CRF model were used. We submitted three systems: a joint model using only CamemBERT, a joint model extended with a CRF layer, and a CamemBERT model without joint option.

On the test dataset, we achieved our best results on the literal tags using our second system (F1=.814) while on the metonymy tags, our third system performed best (F1=.667). Our second system allowed us to obtain the best results on both the dev and test datasets for the literal tags. Nevertheless, we observed a difference on the metonymy tags where our first system obtained best results on the dev dataset (F1=.663) while our third system performed best on the test dataset (F1=.667); surprisingly, differences between first and third systems are about 5 points in strict F1-score.

References

1. Chen, Q., Zhuo, Z., Wang, W.: Bert for joint intent classification and slot filling. arXiv preprint arXiv:1902.10909 (2019)

2. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of machine learning research **12**(Aug), 2493–2537 (2011)
3. 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. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
4. Ehrmann, M., Romanello, M., Flückiger, A., Clematide, S.: Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020). Lecture Notes in Computer Science (LNCS), vol. 12260. Springer (2020)
5. Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Unsupervised named-entity extraction from the web: An experimental study. Artificial intelligence **165**(1), 91–134 (2005)
6. Galibert, O., Rosset, S., Grouin, C., Zweigenbaum, P., Quintard, L.: Extended named entities annotation on OCRed documents: From corpus constitution to evaluation campaign. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (May 2012)
7. Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., Quintard, L.: Proposal for an extension of traditional named entities: from guidelines to evaluation, an overview. In: Proc of LAW. Jeju-do, South Korea (2011)
8. Guo, J., Xu, G., Cheng, X., Li, H.: Named entity recognition in query. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 267–274 (2009)
9. Hammerton, J.: Named entity recognition with long short-term memory. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. pp. 172–175. Association for Computational Linguistics (2003)
10. Kim, J.H., Woodland, P.C.: A rule-based named entity recognition system for speech input. In: Sixth International Conference on Spoken Language Processing (2000)
11. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
12. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018). https://doi.org/10.18653/v1/D18-2012, https://www.aclweb.org/anthology/D18-2012
13. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
14. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural Architectures for Named Entity. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 260–270. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/N16-1030
15. Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering (2020)

16. Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., Li, J.: A Unified MRC Framework for Named Entity Recognition. arXiv preprint arXiv:1910.11476 (2019)
17. Liao, W., Veeramachaneni, S.: A simple semi-supervised algorithm for named entity recognition. In: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing. pp. 58–65 (2009)
18. Liu, M., Tu, Z., Wang, Z., Xu, X.: LTP: A New Active Learning Strategy for Bert-CRF Based Named Entity Recognition. arXiv preprint arXiv:2001.02524 (2020)
19. 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)
20. Luoma, J., Pyysalo, S.: Exploring Cross-sentence Contexts for Named Entity Recognition with BERT. arXiv preprint arXiv:2006.01563 (2020)
21. Ma, X., Hovy, E.: End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1064–1074. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1101, https://www.aclweb.org/anthology/P16-1101
22. Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: Camembert: a tasty french language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)
23. Mollá, D., Van Zaanen, M., Smith, D.: Named entity recognition for question answering (2006)
24. Moosavi, N.S., Strube, M.: Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 632–642. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1060, https://www.aclweb.org/anthology/P16-1060
25. Schuster, M., Nakajima, K.: Japanese and korean voice search. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5149–5152. IEEE (2012)
26. Suárez, P.J.O., Sagot, B., Romary, L.: Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. Challenges in the Management of Large Corpora (CMLC-7) 2019 p. 9 (2019)
27. 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)
28. Yadav, V., Bethard, S.: A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In: COLING (2018)
29. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: ERNIE: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129 (2019)
30. Zhou, G., Su, J.: Named entity recognition using an HMM-based chunk tagger. In: proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 473–480. Association for Computational Linguistics (2002)