Neural Networks approaches focused on French Spoken Language Understanding: application to the MEDIA Evaluation Task

Sahar Ghannay[†] **Christophe Servan**[‡] **Sophie Rosset**[†]

†Université Paris-Saclay,
CNRS, LIMSI,
91405 Orsay, France
`lastname@limsi.fr`

‡ QWANT
61 rue de Villier
92200 Neuilly-sur-Seine, France
`inital.lastname@qwant.com`

Abstract

In this paper, we present a study on a French Spoken Language Understanding (SLU) task: the MEDIA task. Many works and studies have been proposed for many tasks, but most of them are focused on English language and tasks. The exploration of a richer language like French within the framework of a SLU task implies to recent approaches to handle this difficulty. Since the MEDIA task seems to be one of the most difficult, according to several previous studies, we propose to explore Neural Networks approaches focusing of three aspects: firstly, the Neural Network inputs and more specifically the word embeddings; secondly, we compared French version of BERT against the best setup through different ways; Finally, the comparison against State-of-the-Art approaches. Results show that the word embeddings trained on a small corpus need to be updated during SLU model training. Furthermore, the French BERT fine-tuned approaches outperform the classical Neural Network Architectures and achieves state of the art results. However, the contextual embeddings extracted from one of the French BERT approaches achieve comparable results in comparison to word embedding, when integrated into the proposed neural architecture.

1 Introduction

Spoken language understanding (SLU) module is a key component for a spoken language dialogue system. It consists on semantically analyse user queries and identifies text spans that mention semantic concepts. SLU task can fall into three sub-tasks: domain classification, intent classification, and slot-filling (Tur and Mori, 2011). The latter is the task that interests us in this study.

Over the past five years, the studies developed for SLU task are based on neural network architectures (Yao et al., 2014; Mesnil et al., 2015; Guo et al., 2014; Zhang and Wang, 2016; Dinarelli et al., 2017; Simonnet et al., 2017; Korpusik et al., 2019; Ghannay et al., 2020). Recent approaches take benefit from contextual or language model embeddings such as BERT (Devlin et al., 2019). Korpusik et al. (2019) investigated the transfer ability of a pre-trained BERT representation for English SLU tasks. But, as far as we know, there are no such studies on a French SLU task.

Following Ghannay et al. (2020)'s study, many avenues can be explored. In their study, the word embeddings have been frozen during training (are not updated), since Lebret et al. (2013) show that fine-tuned word embeddings show very similar performance and provide comparable results. However, the evaluation of whether updating the embeddings during SLU model training improves or not the results, for SLU task, is less studied. In addition, their SLU model is fed only with word embeddings, and they did not use any additional features, thus there are some rows for improvements. Finally, Béchet and Raymond (2019) benchmarked several SLU tasks and proposed a difficulty hierarchy in which the MEDIA evaluation (Bonneau-Maynard et al., 2006) seems to be the most difficult SLU task.

Contributions: This study focuses on a French SLU task: the MEDIA evaluation, in which we firstly propose the evaluation of whether updating the word embeddings during training can improve the results, according to several scenarios. Secondly, we propose to use a BiLSTM-CNN architecture (Ma, 2016) that integrates character embeddings as additional features, using a convolution layer. Finally, we propose to evaluate the performance of BERT approaches against the BiLSTM-CNN architecture and State-of-the-Art on the MEDIA task (Simonnet, 2019) through different ways: i) We propose to fine-tune BERT on SLU task using two french models: CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020). ii) based on the results of i) we propose to integrate the extracted BERT contextual embeddings to the BiLSTM-CNN architecture and compare it to word embeddings.

2 SLU Model descriptions

This section describe the SLU models used in this study. The first two models are based on BiLSTM and its update: the BiLSTM-CNN. The NeuroNLP2 implementation[1] was used for both BiLSTM implementations. The third model is based on the BERT models.

BiLSTM (Bidirectional long short-term memory) architecture has been proven to be relevant to model output dependencies on SLU tasks (Yao et al., 2014; Mesnil et al., 2015).

To further improve the performance of our SLU model, we propose to use a BiLSTM-CNN (convolutional neural network) architecture (Ma, 2016) that integrates character embeddings using a convolution layer, in addition to the word embeddings.

Finally, we propose to fine-tune BERT (Devlin et al., 2019) on SLU task using two french models: CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020). The CamemBERT model is trained on the French part of the OSCAR corpus (Suárez et al., 2019) composed of 138GB of raw text, and FlauBERT (Le et al., 2020) on 71GB of heterogeneous French corpora.

3 Experiments

In this section, we present the experiments we performed using our approaches and their setup[2]. For both BiLSTM and BiLSTM-CNN, we made some hyper-parameters tuning by varying the number of layers $l \in \{1, 2, 3\}$, the size of the BiLSTM hidden layers $n \in \{128, 256, 512\}$ and the batch size $b \in \{16, 32, 64\}$. For BiLSTM-CNN, in addition to the other parameters, the window size is set to 3 and the number of filters (dimension of character embeddings) is set to $s \in \{30, 50, 100\}$.

3.1 Data

Experiments are conducted on the French MEDIA[3] corpus, composed of 1258 transcribed dialogues, which is about hotel reservation and information (Bonneau-Maynard et al., 2006). The corpus was manually annotated, following a BIO model, with semantic concepts characterized by a label and its value. The corpus is split into three parts: a training corpus composed of 13k sentences, a development corpus composed of 1.3k sentences, and a test corpus composed of 3.5k sentences.

3.2 Word embeddings training

One of the aim of our experiments is to see whether the update of the word embeddings during training of the SLU model (*update*) improves or not the results by mainly varying the data used to train the word embeddings and the hyper-parameters of the SLU model.

Following Ghannay et al. (2020)'s results, we propose to use CBOW word embeddings approach from word2vec (Mikolov et al., 2013), which is trained using the default parameters using three different corpora setup. The first one is a small and task-dependent corpus: training set of **MEDIA** corpus is used, by keeping all the words due to the small data size. A huge and out-of-domain corpus was used as second setup: the French Wikipedia dump (**WIKI**), which is composed of 573 million words using a vocabulary size of 923k words. Finally, both corpora (noted **WIKI+MEDIA**).

The common parameters used to train our word embeddings are: window size=5, negative sampling=5, dimension=300. They have been selected based on previous studies (Pennington et al., 2014; Bojanowski et al., 2017). Note that the out of vocabulary (OOV) words are represented by null vectors.

[1] https://github.com/XuezheMax/NeuroNLP2

[2] the code an data needed to run the experiments are available here: https://github.com/saharghannay/MEDIA_Eval

[3] MEDIA is publicly available for academic use: https://catalogue.elra.info/en-us/repository/browse/ELRA-S0272/

3.3 Results

3.3.1 Embeddings update

Those experiments aim to observe the impact of the update (noted *update*) of CBOW word embeddings or their freeze (noted *freeze*) while training of SLU module. We proposed different training setups for the word embeddings (MEDIA, WIKI and WIKI+MEDIA), presented in section 3.2. Also, the BiLSTM number of layers is set from 1 to 3.

Results summarized in Table 1 show that when the word embeddings are trained on MEDIA data, the update of the word embeddings is helpful and improves the results in terms of F1 score, whatever the size of the architecture. However, when the embeddings are trained on WIKI or WIKI+MEDIA data the update of the embeddings while training is not helpful and degrades the results. Thus, the best results are obtained using the BiLSTM architecture composed of 3 hidden layers, using one of the CBOW embeddings trained on WIKI or WIKI+MEDIA corpora, that obtain comparable results in terms of F1 score (86.40 vs. 86.69).

Config.	*Update*			*Freeze*		
Train Emb.	**#nb. BiLSTM layers**			**#nb. BiLSTM layers**		
	1	**2**	**3**	**1**	**2**	**3**
MEDIA	84.18	84.18	85.35	72.36	79.57	80.69
WIKI	84.73	85.82	86.47	84.11	86.06	86.40
WIKI +MEDIA	84.84	85.35	86.00	84.08	85.74	**86.69**

Table 1: Performance on Test MEDIA in terms of F1 score of CBOW word embeddings approach trained on three corpora (MEDIA, WIKI and WIKI+MEDIA), using the BiLSTM architecture.

#Layer	**WIKI**			**WIKI+MEDIA**		
	Character embeddings dimensions					
	30	**50**	**100**	**30**	**50**	**100**
1	84.38	84.59	84.85	84.13	84.47	84.73
2	85.88	86.43	86.18	86.20	86.75	86.34
3	87.02	87.05	**87.40**	87.29	87.01	87.30

Table 2: Results on Test MEDIA in terms of F1 score using embeddings trained on both wiki and wiki+MEDIA corpora, using the BiLSTM-CNN architecture. (The word embeddings are frozen)

3.3.2 Character embeddings evaluation

In this section, we propose to use a BiLSTM-CNN architecture (Ma, 2016) that integrates character embeddings as additional features, using a convolution layer. We experiment the use of different character embeddings dimensions, and different numbers of BiLSTM layers. Based on the results in section 3.3.1, for those experiments, we used the embeddings trained on both WIKI and WIKI+MEDIA corpora, which are frozen during the BiLSTM-CNN training.

Results summarized in table 2, show that the use of character embeddings as additional features was helpful and improves the performance in comparison to the results in table 1. We observe that, both embeddings trained on WIKI and WIKI+MEDIA achieve comparable results. This shows that, we don't need to use both a task-dependent corpus and another out-of-domain corpus to train the word embeddings. Note that, we observed the same thing, when the embeddings trained on WIKI data are fine-Tuned on MEDIA data. The best results (F1=87.40) achieved using the embeddings trained on WIKI data, using the appropriate parameters: 3 BiLSTM layers and character embeddings dimension of 100. Note that, beyond these values the performance of the system drops slightly.

3.3.3 Comparison to French BERT

In this section, we propose to evaluate the performance of BERT approaches on the MEDIA task through two different ways. The experimental results are summarized in table 3 using F1 score and the Concept Error Rate (CER), which is estimated like the Word Error Rate (WER). The CER is used to compare our approach with the State-of-the-Art proposed by Simonnet (2019) and noted *biRNN-EDA*. We also reported the results of the best system presented in table 2.

i) We propose to fine-tune BERT on SLU task using two French models: CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020) base models. Results in table 3 show the performance of BERT's models on SLU task. The best results achieved using CamemBERT base model trained on ccnet data. It yields 29.35% of relative improvement in terms of CER reduction in comparison to the baseline (7.56 vs 10.7). In addition, it outperforms the proposed BiLSTM-CNN system and improves the prediction of

Architecture	Embed. Training data	Embed.'s approach	F1	CER
biRNN-EDA	–	–	–	10.7
BiLSTM-CNN	WIKI	CBOW (dim=300)†	87.40	9.88
		CBOW (dim=768)	86.80	10.11
FineTune BERT (i)	oscar 138 GB	CamemBERT-base	89.18	7.93
	ccnet 135 GB	CamemBERT-base	**89.37**	**7.56**
	heterogeneous corpus 71 GB	FlauBERT-base	89.04	8.13
BiLSTM	ccnet 135GB (ii)	CamemBERT-base (dim=768)	86.59	10.45
BiLSTM-CNN		CamemBERT-base (dim=768)	87.15	10.11

Table 3: Performance on Test MEDIA in terms of F1 and CER scores of the proposed systems († is the best system presented in table 2).

some tags: "nom, chambre-fumeur, objet, ...". We observe that the models CamemBERT base (trained on OSCAR data) and FlauBERT obtain competitive results in terms of F1 and CER scores. Note that CamemBERT and FlauBERT base models achieve better results than the large models.

ii) Last, we propose to integrate the extracted BERT's contextual embeddings to the BiLSTM and BiLSTM-CNN architectures, instead of CBOW word embeddings. Based on the results of i) we used the embeddings extracted from CamemBERT base model trained on ccnet data. After tokenizing the MEDIA corpus, the CamemBERT model was applied on the resulting data to extract the embeddings of 768 dimensions, for each sub-word from the last transformer layer. The token embeddings corresponds to the sum of its sub-word embeddings. A new CBOW embeddings is trained on WIKI data with dimension 768 to have comparable results. Results (last 2 lines) show that the use of CamemBERT contextual embeddings achieves competitive results in comparison to CBOW embeddings whatever the architecture used (BiLSTM or BiLSTM-CNN). Those results corroborate the results reported by Ghannay et al. (2020), in which, CBOW and ELMo (Peters et al., 2018) embeddings obtained comparable results in terms of F1 score (86.06 vs. 86.42). Last, the results with BiLSTM and BiLSTM-CNN architectures reveals the importance of character embeddings, even when they are combined with contextual embeddings.

4 Conclusions and future work

The paper presented a study focuses on French Spoken Language Understanding (SLU) task using the MEDIA corpus. First we proposed the evaluation of whether updating the word embeddings during training improves or not the results, according to several scenarios.Second, we proposed to use a BiLSTM-CNN architecture that integrates character embeddings as additional features. Last, we proposed to evaluate the performance of BERT approaches on the MEDIA task through different ways.

Experimental results show, that the word embeddings needed to be updated during SLU model training are the ones trained on small corpus like MEDIA. However, It is better for word embeddings trained on huge and out-of-domain to be frozen, since those word embeddings have captured enough general semantic and syntactic characteristics relevant to SLU task. More, The word embeddings trained on WIKI and WIKI+MEDIA achieve comparable results. This shows that, we don't need to use both a task-dependent corpus and another out-of-domain corpus to train the word embeddings. In addition, we observed the usefulness of character embeddings when added as additional features. Regarding the evaluation of the performance of BERT approaches, the fune-tuning of CamemBERT and Flaubert base models show that the best results are achieved using CamemBERT base model trained on ccnet data. It yields to 29.35% of relative improvement in terms of CER reduction in comparison to the baseline (7.56 vs 10.7). Finally, the integration of the extracted CamemBERT's contextual embeddings to the BiLSTM and BiLSTM-CNN architectures reveal that contextual embeddings achieves competitive results in comparison to CBOW word embeddings whatever the architecture, and confirm the importance of character embeddings.

For future work, we propose to evaluate the performance of Bert's contextual embeddings extracted from different encoder's layers, and to make in-depth error analysis for the different systems.

Acknowledgements

This work has been partially funded by the LIHLITH project (ANR-17-CHR2-0001-03), and supported by ERA-Net CHIST-ERA, and the “Agence Nationale pour la Recherche” (ANR, France).

References

Frédéric Béchet and Christian Raymond. 2019. Benchmarking benchmarks: introducing new automatic indicators for benchmarking spoken language understanding corpora. In *Interspeech*, Graz, Austria.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5.

Hélène Bonneau-Maynard, Christelle Ayache, Frédéric Bechet, Alexandre Denis, Anne Kuhn, Fabrice Lefevre, Djamel Mostefa, Matthieu Quignard, Sophie Rosset, Christophe Servan, and Jeanne Villaneau. 2006. Results of the French Evalda-Media evaluation campaign for literal understanding. In *lrec*, Genoa.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, Minneapolis, Minnesota. Association for Computational Linguistics.

Marco Dinarelli, Vedran Vukotic, and Christian Raymond. 2017. Label-dependency coding in Simple Recurrent Networks for Spoken Language Understanding. In *Interspeech*, Stockholm, Sweden.

S. Ghannay, A. Neuraz, and S. Rosset. 2020. What is best for spoken language understanding: small but task-dependant embeddings or huge but out-of-domain embeddings? In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8114–8118.

Daniel Guo, Gokhan Tur, Wen-tau Yih, and Geoffrey Zweig. 2014. Joint semantic utterance classification and slot filling with recursive neural networks. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 554–559. IEEE.

Mandy Korpusik, Zoe Liu, and James Glass. 2019. A comparison of deep learning methods for language understanding. In *Interspeech, September 15–19, 2019, Graz, Austria*, Graz, Austria.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Rémi Lebret, Joël Legrand, and Ronan Collobert. 2013. Is deep learning really necessary for word embeddings? Technical report, Idiap.

Eduard Ma, Xuezheand Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *ACL*. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Grégoire Mesnil, Yann Dauphin, Kaisheng Yao, Yoshua Bengio, Li Deng, Dilek Hakkani-Tur, Xiaodong He, Larry Heck, Gokhan Tur, Dong Yu, and Geoffrey Zweig. 2015. Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(3):530–539.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *arXiv preprint arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Edwin Simonnet. 2019. *Deep learning applied to spoken langage understanding*. Theses, Université du Maine.

Edwin Simonnet, Sahar Ghannay, Nathalie Camelin, Yannick Estève, and Renato De Mori. 2017. ASR error management for improving spoken language understanding. In *Interspeech 2017*, Stockholm, Sweden.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. *Challenges in the Management of Large Corpora (CMLC-7) 2019*, page 9.

Gokhan Tur and Renato De Mori. 2011. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley & Sons.

Kaisheng Yao, Baolin Peng, Yu Zhang, Dong Yu, Geoffrey Zweig, and Yangyang Shi. 2014. Spoken language understanding using long short-term memory neural networks. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 189–194. IEEE.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *IJCAI*, pages 2993–2999.