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

Sahar Ghannay[1], Christophe Servan[2], Sophie Rosset[1]

[1]Université Paris-Saclay, CNRS, LIMSI — lastname@limsi.fr

[2]QWANT — initial.lastname@qwant.com

Introduction

SLU task:

- Semantically analyse sentences and identifies text spans that mention **semantic concepts**.
- BIO tagging : format that allows to delimits mentions in utterances
- Evaluation in Concept Error Rate (CER) and F1-score

Hyp	I	want	to	book	A	room
Concept	Command				number	Object
Tag	Command-B	command-I	command-I	command-I	number-B	object-B

Goal:

- Focus on French SLU MEDIA task: one of the most difficult, according to several previous studies
- Explore Neural Networks (NN) approaches focusing of three aspects:
  - ➡The NN inputs: more specifically the word embeddings
  - ➡Compare French version of BERT against the best setup through different ways
  - ➡The comparison against State-of-the-Art approaches.

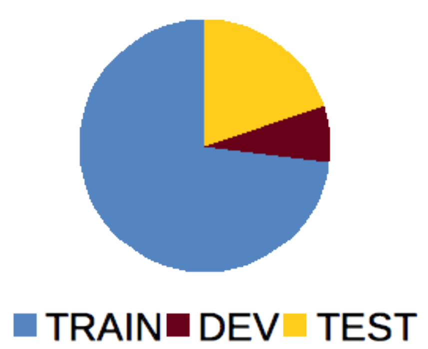
Experiments

SLU model descriptions:

1. BiLSTM (Bidirectional long short-term memory) architecture
2. BiLSTM-CNN (convolutional neural network) architecture that integrates character embeddings using a convolution layer, in addition to the word embeddings
3. Fine-tune BERT on SLU task using two french models: CamemBERT and FlauBERT.
   - The CamemBERT model is trained on the French part of the OSCAR corpus composed of 138GB of raw text, and FlauBERT on 71GB of heterogeneous French corpora.

MEDIA corpus:

- Touristic information
- French corpus
- 74 concepts
- 17.8k utterances



Word embeddings training

One of the aim of our experiments is to see whether the update of the 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.

➡Use CBOW word embeddings approach (dim=300) using three different corpora setup:

- Small and task-dependent corpus: training set of **MEDIA**
- A huge and out-of-domain corpus: the French Wikipedia dump (**WIKI**) (573 million words, vocabulary size of 923k words)
- Both corpora: noted **WIKI+MEDIA**

Results

Embeddings update

- Evaluate the impact of the update (noted ***update***) of CBOW embeddings or their freeze (noted ***freeze***) while training of SLU module.
- Experiment different training setups for CBOW (**MEDIA**, **WIKI** and **WIKI+MEDIA**), and the BiLSTM number of layers (1, 2 or 3).

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**

➡Embeddings trained on MEDIA: update is helpful
➡Embeddings trained on WIKI or WIKI+MEDIA: update of the is not helpful
➡Best results obtained using the BiLSTM (3 hidden layers), using one of the CBOW embeddings trained on **WIKI** or **WIKI+MEDIA** corpora

Character embeddings evaluation

- Integrates character embeddings as additional features
- Experiment the use of different character embeddings dimensions (30, 50 or 100), and different numbers of BiLSTM layers

#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

➡The use of character embeddings as additional features is helpful
➡Both embeddings trained on WIKI and WIKI+MEDIA achieve comparable results

- task-dependent corpus seems useless compared to huge and out-of-domain corpora (wiki)

Embeddings update

- Evaluate the performance of BERT approaches on the MEDIA task through two different ways:
  - I)Fine-tune BERT on SLU task using two French models: CamemBERT and FlauBERT
  - II)Integrate the extracted BERT's contextual embeddings to the BiLSTM and BiLSTM-CNN architectures, instead of CBOW

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

➡I)CamemBERT base model trained on ccnet data achieves the best results

- 29.35% of relative improvement in terms of CER reduction in comparison to the baseline (biRNN-EDA)
- outperforms BiLSTM-CNN system and improves the prediction of some tags: "nom, chambre-fumeur, objet, …'
- Achieves comparable results to FlauBERT base model

➡II) the use of CamemBERT contextual embeddings achieves competitive results in comparison to CBOW embeddings whatever the architecture used (BiLSTM or BiLSTM-CNN)

- the results with BiLSTM and BiLSTM-CNN architectures reveals the importance of character embeddings, even when they are combined with contextual embeddings.

Conclusions

➡Word embeddings need to be updated when trained on task-dependent corpora
➡Word embeddings trained on huge and out-of-domain data can captured enough general characteristics relevant to SLU task
➡Word embeddings trained on both task-dependent corpora and out-of-domain corpora are useless.
➡French BERT model (CamemERT) achieved best results on French MEDIA SLU task (F1=89.37 and CER=7.56)