On the cross-lingual transferability of multilingual prototypical models across NLU tasks

Oralie Cattan, Christophe Servan, Sophie Rosset

On the cross-lingual transferability of multilingual prototypical models across NLU tasks

Oralie Cattan[1,2], Christophe Servan[1], and Sophie Rosset[2]

[1]QWANT, Neuilly-sur-Seine, France

[2]Université Paris-Saclay, CNRS, LISN, Orsay, France

`inital.lastname@qwant.com | lastname@lisn.fr`

Qwant

LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

université PARIS-SACLAY

Overview

- Multilingual Natural Language Understanding (NLU) over 2 classification tasks:
  Intent Recognition and **Slot Filling**
- Classification: a supervised learning task
  - ☞ limited availability of high-quality annotated datasets
  - ☞ complex adaptation to new languages, especially for poorly endowed languages

Approaches explored:

- Transformer-based models: contextualized representations, pre-training and transfer learning
- Limitations:
  - ☞ underperformances [Pires et al., 2019, Conneau et al., 2020]: on poorly endowed languages, when facing data scarcity
  - ☞ instabilities [Zhang et al., 2021, Mosbach et al., 2021]: overfitting, label noise memorization or catastrophic forgetting

Contributions

1. A comparison of Transfer and Meta Learning approaches on a multilingual few-shot NLU benchmark for evaluating cross-lingual generalisation
2. A Transformers-based Prototypical Network that enhance cross-lingual representations based on a multilingual Transformer-based model (mBERT) [Devlin et al., 2019]
3. A new zero-shot scenario proposal

MultiATIS++ corpus

- MultiATIS++ [Xu et al., 2020]: 8 different other languages: Spanish (es), German (de), French (fr), Portuguese (pt), Hindi (hi), Chinese (zh), Japanese (ja), and Turkish (tr)
- 37,084 training examples and 7,859 test examples (Figure 1)
- Hindi and Turkish subcorpora cover only a subset of intents and slots with few labeled examples

EN	show		departures	from	atlanta	for	american
	O		O	O	B-fromloc.city_name	O	B-airline_name
ES	Muestra		salidas	desde	Atlanta	de	American
	O		O	O	B-fromloc.city_name	O	B-airline_name
PT	Mostre		partidas	de	Atlanta	da	American
	O		O	O	B-fromloc.city_name	O	B-airline_name
DE	Zeige		Abflüge	von	Atlanta	für	American
	O		O	O	B-fromloc.city_name	O	B-airline_name
FR	Montrer	des	départs	d’	Atlanta	pour	American
	O	O	O	O	B-fromloc.city_name	O	B-airline_name
ZH	显示	美国航空	从	亚特兰大	出发的航班		
	O	B-airline_name	O	B-fromloc.city_name	O		
JA	アトランタ	発	アメリカ	便を表示する			
	B-fromloc.city_name	O	B-airline_name	O			
HI	अमेरिकन	के	लिए	अटलांटा	से	प्रस्थान	दिखाएं
	B-airline_name	O	O	B-fromloc.city_name	O	O	O
TR	atlanta	‘	dan	american	kalkislarini	goster	
	B-fromloc.city_name	O	O	B-airline_name	O	O	

Figure 1: English training example and its translated versions in MultiATIS++.

- Three configuration
  1. **target only**: training with only the target language data
  2. **multilingual**: training on the concatenation of all of the 9 languages and testing the model for each target language
  3. **zero-shot**: training on the concatenation of all training datasets from all languages except the one we want to test

Approach proposed

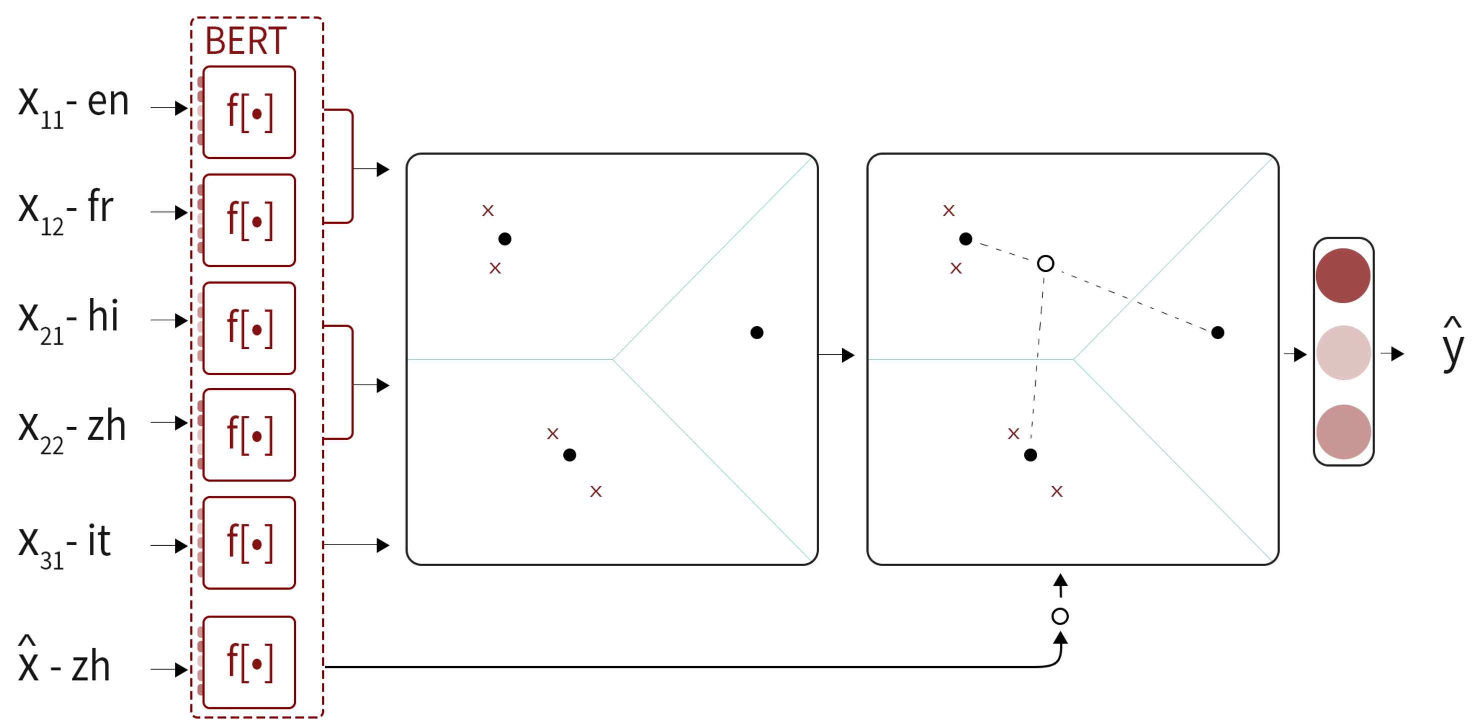

Figure 2: Transformer-based Prototypical Neural Network

- Episodic learning: N-way, k-shot classification tasks, support and query sets
- Prototypical Neural Network [Snell et al., 2017] (PNN): *learn the metric space to perform classification*
- From convolution and recurrent features extractors to a mBERT-based PNN encoder (Figure 2)

Results

config.	encoder	en	es	de	zh	ja	pt	fr	hi	tr
target only	**mBERT**	98.54	**97.31**	98.43	**97.09**	**97.20**	**97.54**	**98.88**	90.93	83.36
	mBERT + PNN 5w1s	97.46	95.14	97.18	96.35	95.53	96.80	97.11	84.95	85.17
	mBERT + PNN 5w10s	**98.77**	96.97	**98.54**	97.0	96.64	97.42	97.98	**91.33**	**89.33**
multilingual	**mBERT**	98.42	97.98	98.59	97.65	97.45	**98.3**	98.46	95.33	**93.93**
	mBERT + PNN 5w1s	95.33	93.71	95.93	95.89	94.42	94.00	94.78	91.4	90.91
	mBERT + PNN 5w10s	**99.87**	**98.54**	**98.60**	**98.67**	**98.54**	**98.32**	**98.66**	**95.49**	92.61
zero-shot	**mBERT**	96.42	**97.98**	**97.54**	96.71	**97.45**	97.42	**97.87**	**94.37**	**91.61**
	mBERT + PNN 5w1s	93.73	92.02	93.27	95.62	91.73	93.51	93.28	90.51	89.92
	mBERT + PNN 5w10s	**96.47**	97.87	96.86	**97.65**	96.64	**98.10**	97.45	93.17	90.67

Table 1: Averaged intent accuracies obtained with PNNs on 5-way k-shot classification $k \in [1, 10]$ (best scores are marked in bold) and baseline results.

config.	encoder	en	es	de	zh	ja	pt	fr	hi	tr
target only	**mBERT**	95.64	85.52	94.88	92.93	93.13	91.71	92.78	85.12	78.22
	mBERT + PNN	**95.76**	**87.40**	**95.63**	**93.45**	**93.93**	**92.22**	**93.13**	**85.70**	**82.67**
multilingual	**mBERT**	96.02	88.03	95.03	93.63	93.01	92.31	91.18	87.39	86.83
	mBERT + PNN	**98.40**	**92.09**	**97.12**	**95.50**	**97.24**	**95.81**	**96.80**	**89.59**	**88.39**
zero-shot	**mBERT**	**94.10**	**87.14**	**94.23**	**92.17**	**92.61**	**91.59**	**90.79**	86.14	85.86
	mBERT + PNN	93.25	86.99	93.57	91.82	92.38	91.19	90.39	**87.49**	**86.83**

Table 2: Averaged slot F1s obtained with PNNs on 5-way 10-shot and baseline results (highest scores are marked in bold).

Conclusions

- Performance gains achieved by exploiting language interrelationships learnt with transfer learning
- Competitive NLU systems for under-resources languages with only a fraction of training examples using PNNs

References

[Conneau et al., 2020] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzm'an, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Mosbach et al., 2021] Mosbach, M., Andriushchenko, M., and Klakow, D. (2021). On the stability of fine-tuning {bert}: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

[Pires et al., 2019] Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

[Snell et al., 2017] Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4077–4087. Curran Associates, Inc.

[Xu et al., 2020] Xu, W., Haider, B., and Mansour, S. (2020). End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. Association for Computational Linguistics.

[Zhang et al., 2021] Zhang, T., Wu, F., Katiyar, A., Weinberger, K. Q., and Artzi, Y. (2021). Revisiting few-sample {bert} fine-tuning. In *International Conference on Learning Representations*.