Attention-Informed Mixed-Language Training for Zero-shot Cross-lingual Task-oriented Dialogue Systems

Zihan Liu, Genta Indra Winata, Zhaojiang Lin,
Peng Xu, Pascale Fung

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

CAiRE
Center for Artificial Intelligence Research

Background

- Supervised neural-based approaches have shown the effectiveness for natural language processing.

Background

- Supervised neural-based approaches have shown the effectiveness for natural language processing.
- However, they heavily rely on large amounts of training data, which makes them not scalable to low-resource languages.

Background

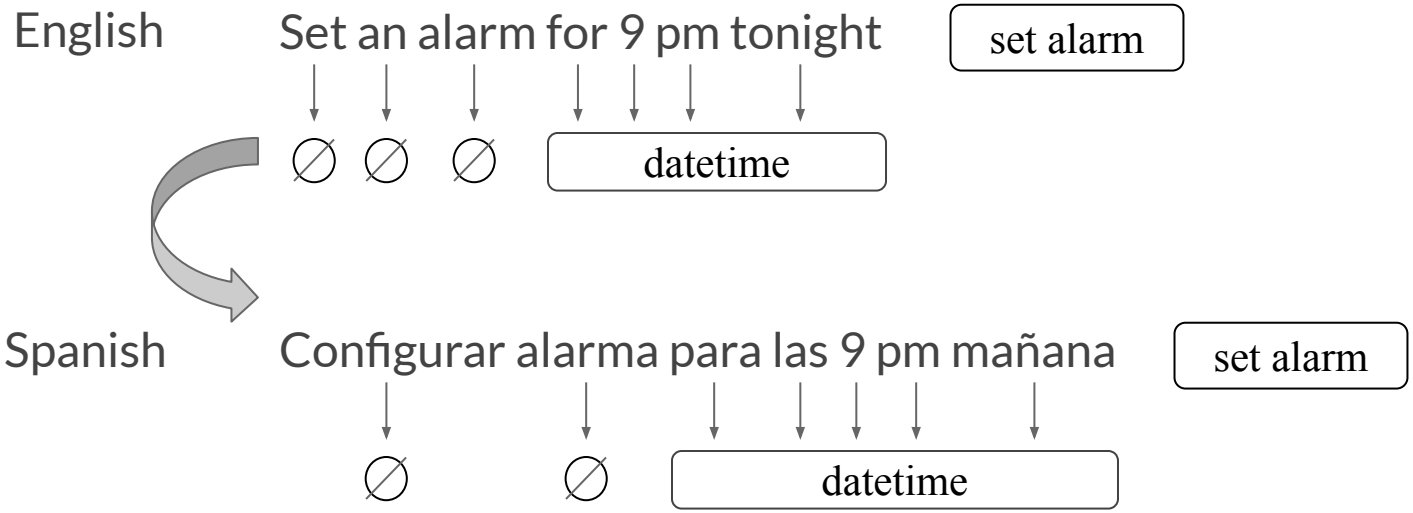
- Supervised neural-based approaches have shown the effectiveness for natural language processing.
- However, they heavily rely on large amounts of training data, which makes them not scalable to low-resource languages.
- A straightforward idea is to adapt the model from the high-resource language into the low-resource languages.

Cross-lingual Task-oriented Dialogue Systems

- Dialogue State Tracking (DST)

Cross-lingual Task-oriented Dialogue Systems

- Natural Language Understanding (NLU)



Straightforward solutions

1. Translate training set from source language to target langauge

2. Translate test samples

Straightforward solutions

1. Translate training set from source language to target langauge

2. Translate test samples

Problems

1. We need large amounts of resources to build machine translation systems.
2. Machine translation systems perform badly if the source language and target languages are unrelated languages (e.g., English and Chinese).

Cross-lingual Adaptation

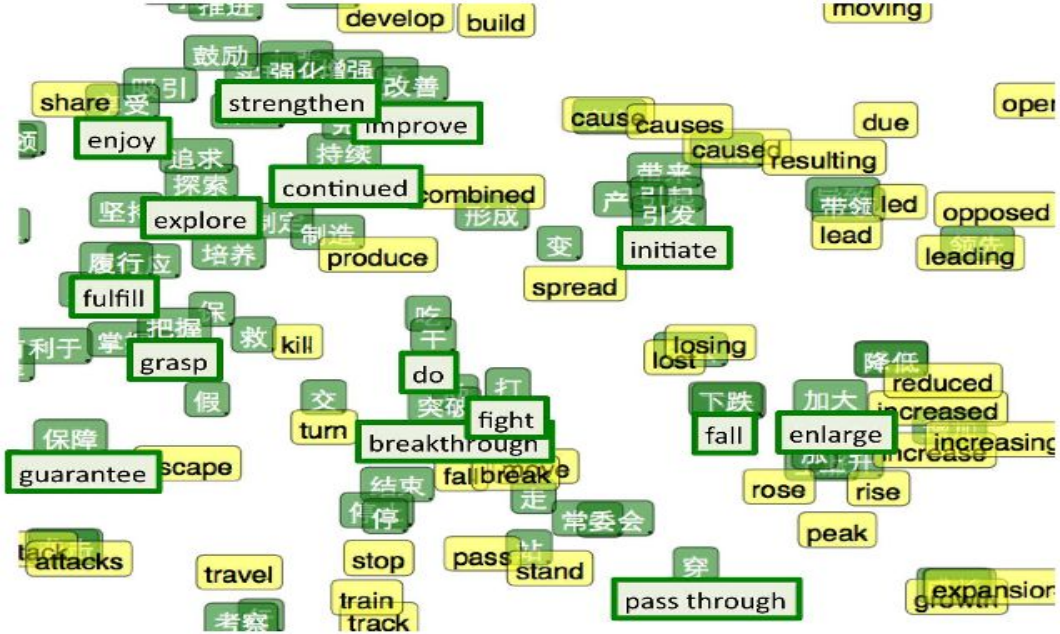
Leverage the inter-connections among languages

English Systems

Cross-lingual Adaptation

Leverage the inter-connections among languages

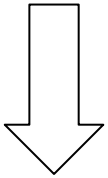
English Systems



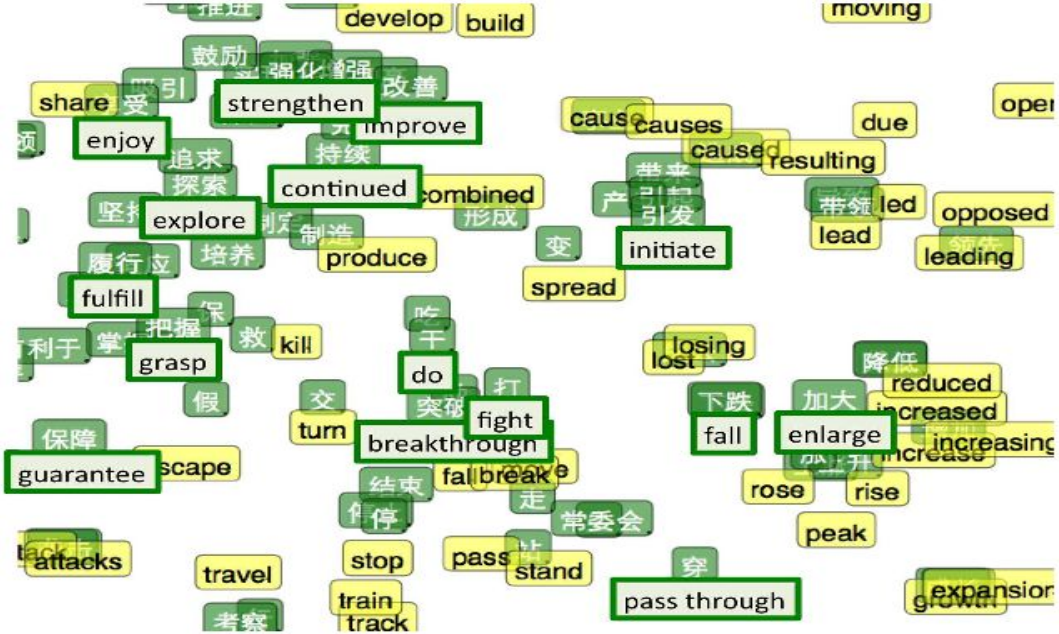
Cross-lingual Adaptation

Leverage the inter-connections among languages

English Systems



Chinese Systems

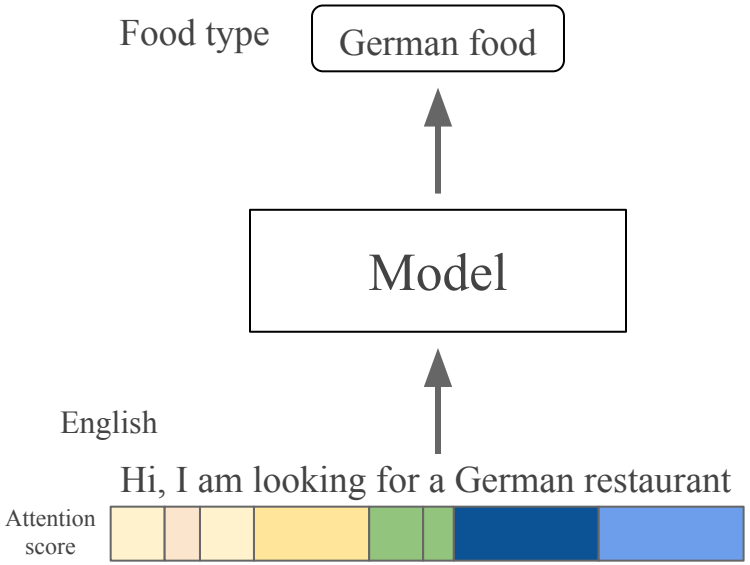


Related work

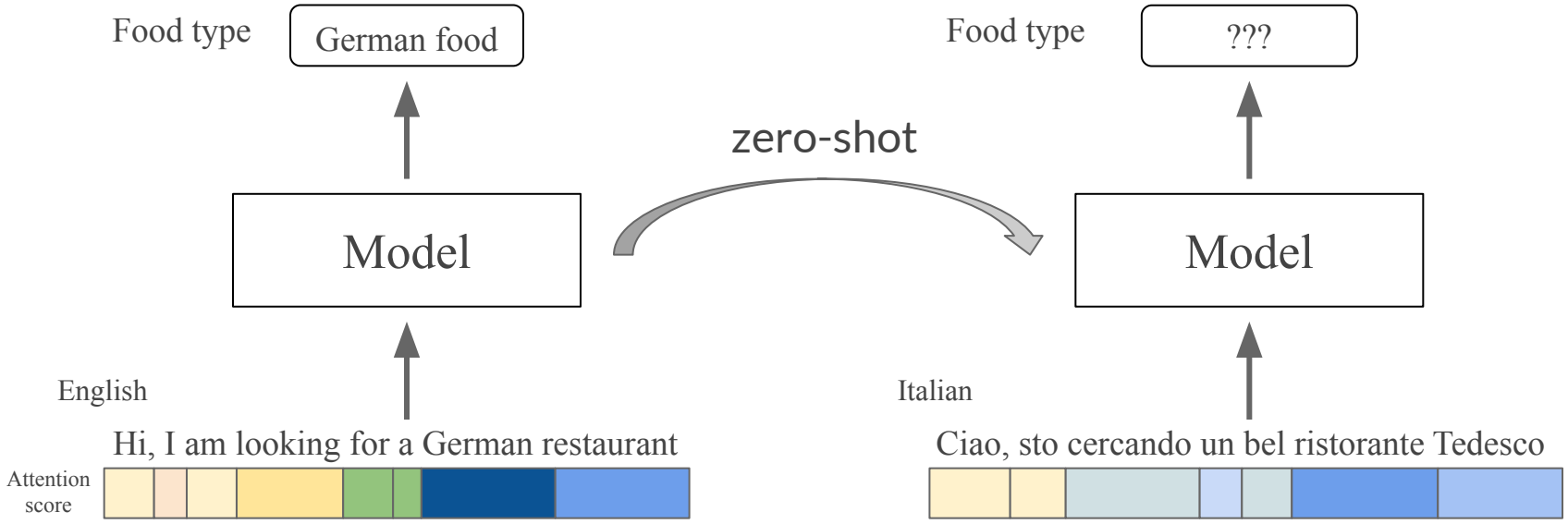
- Chen et al (2018)[1] utilized large amounts of parallel data or bilingual dictionary to build zero-shot cross-lingual DST systems.
- Schuster et al (2019)[2] also leveraged extensive parallel data to build zero-shot cross-lingual NLU systems.
- Collecting bilingual resources is expensive and time-consuming, our work only utilizes very few word pairs as bilingual resources.

[1] XL-NBT: A Cross-lingual Neural Belief Tracking Framework
[2] Cross-Lingual Transfer Learning for Multilingual Task Oriented Dialog

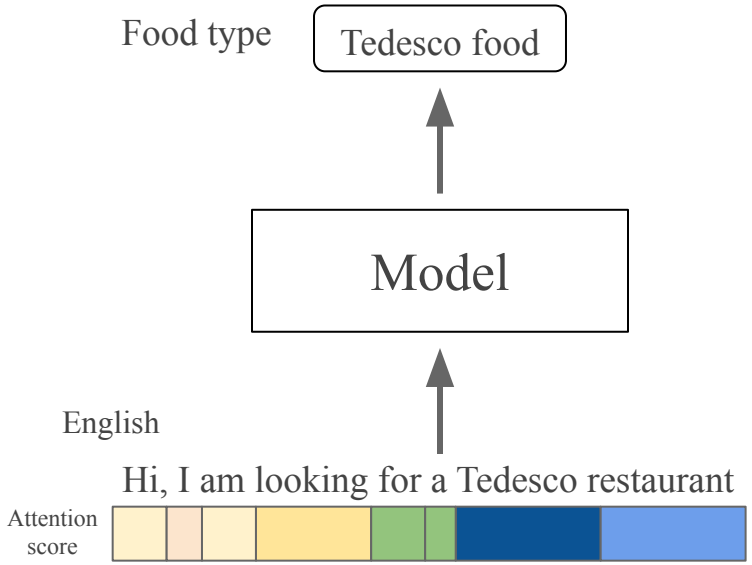
Intuition of Mixed-Language Training



Intuition of Mixed-Language Training

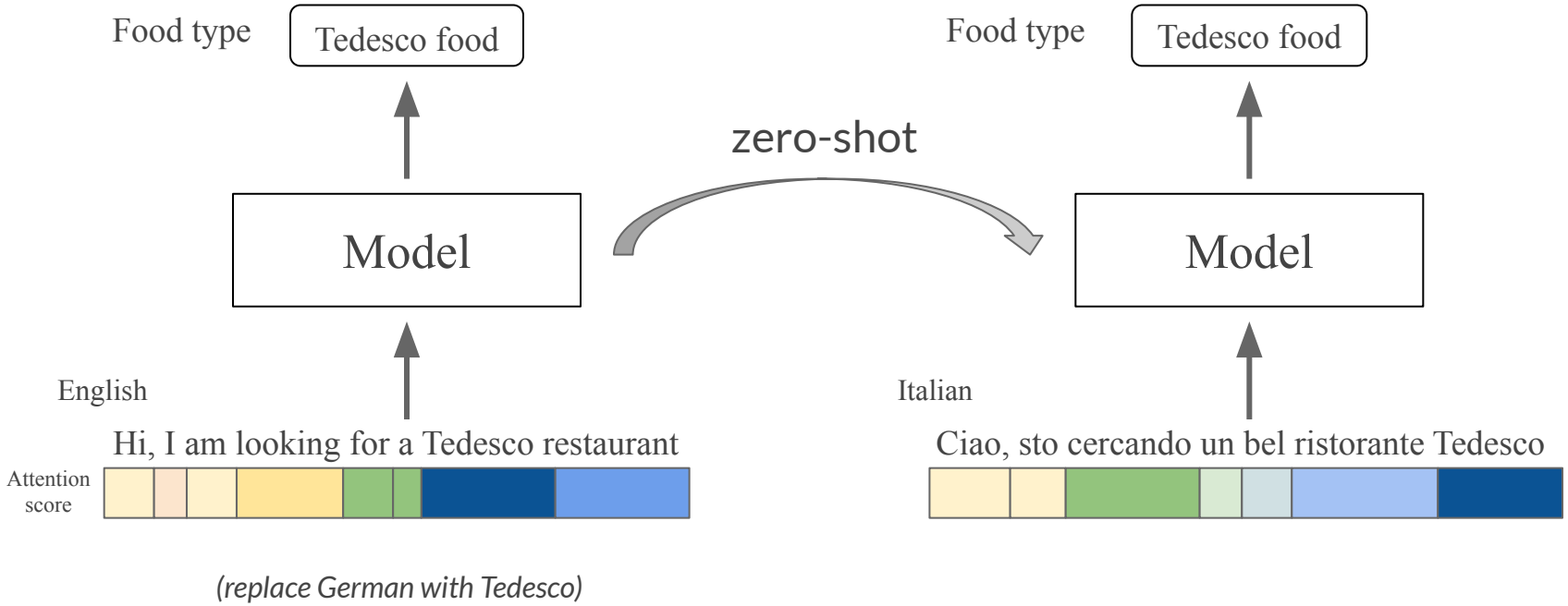


Intuition of Mixed-Language Training



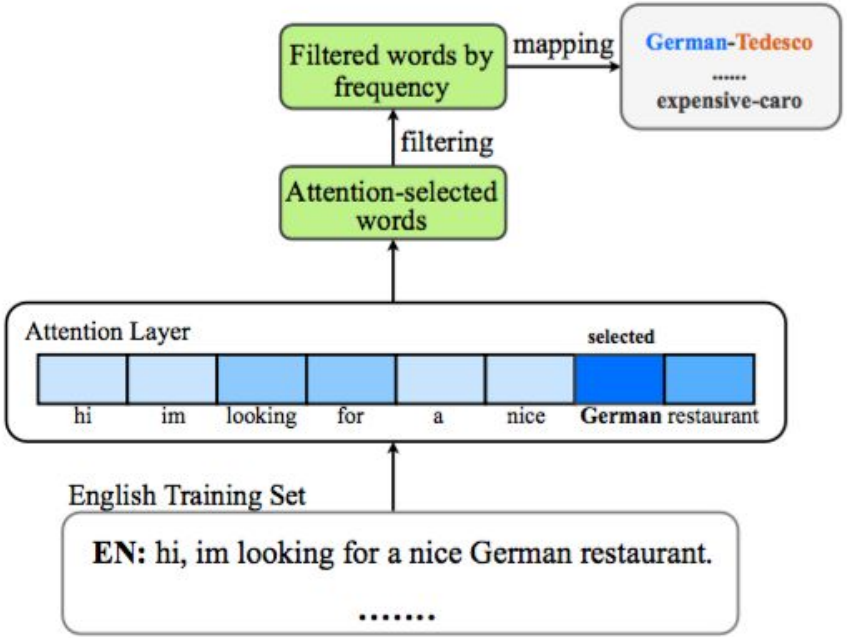
(replace German with Tedesco)

Intuition of Mixed-Language Training

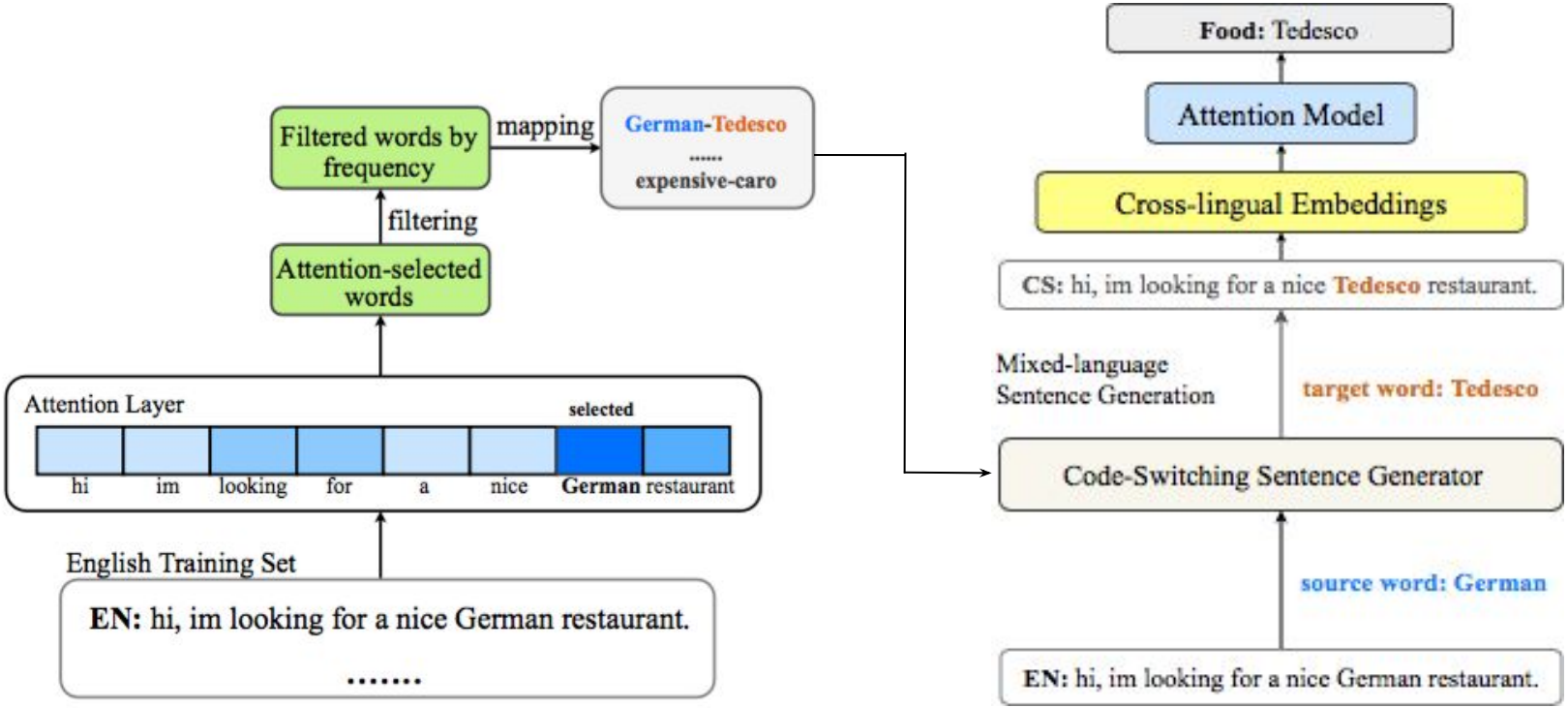


(replace German with Tedesco)

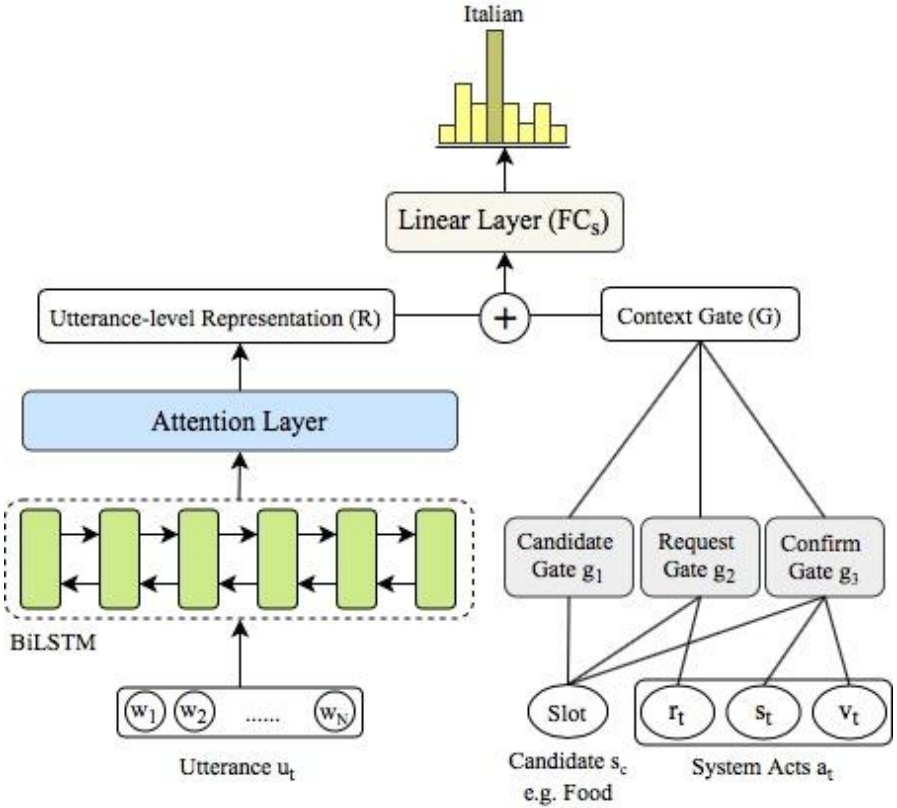
Attention Layer to select keywords



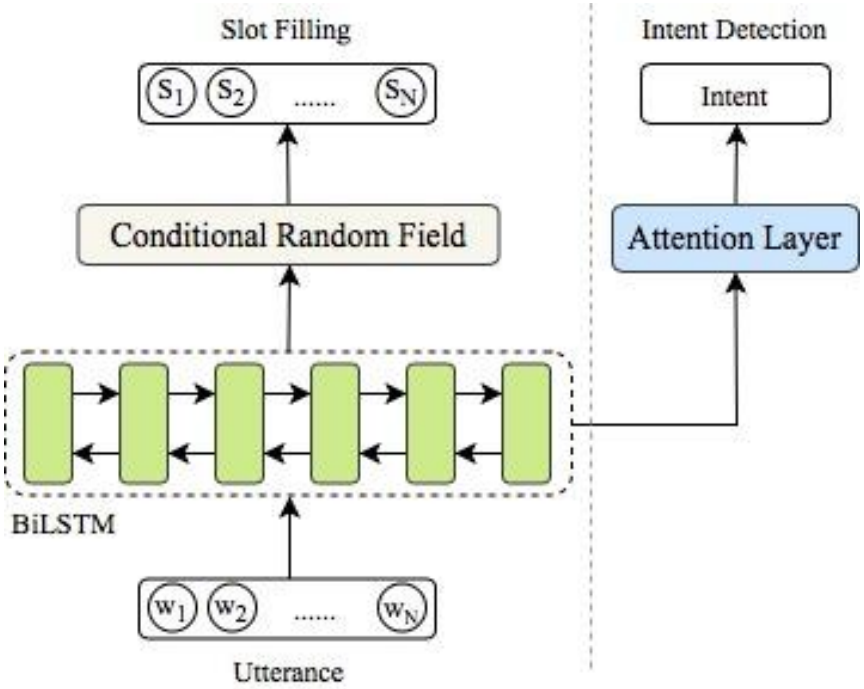
Mixed-Language Training



Dialogue State Tracking (DST)



Natural Language Understanding (NLU)



Zero-shot Results in DST Task

Model	German								
	slot acc.			joint goal acc.			request acc.		
	BASE	MLT_O	MLT_A	BASE	MLT_O	MLT_A	BASE	MLT_O	MLT_A
MUSE	60.69	68.58	**71.38**	21.57	30.61	**36.51**	74.22	80.11	**82.99**
XLM (MLM)*	52.21	66.26	**68.25**	14.09	29.45	**31.29**	75.15	78.48	**80.22**
+ Transformer	53.81	65.81	**68.55**	13.97	30.87	**32.98**	76.83	78.95	**81.34**
XLM (MLM+TLM)*	58.04	65.39	**66.25**	16.34	29.22	**29.83**	75.73	78.86	**79.12**
+ Transformer	56.52	66.81	**68.88**	16.59	31.76	**33.12**	78.56	81.59	**82.96**
Multi. BERT*	57.61	67.49	**69.48**	14.95	30.69	**32.23**	75.31	83.66	**86.27**
+ Transformer	57.43	68.33	**70.77**	15.67	31.28	**34.36**	78.59	84.37	**86.97**
Ontology Matching†		*24*			-			*21*	
Translate Train†		*41*			-			*42*	
Bilingual Dictionary‡		*51.74*			*28.07*			*72.54*	
Bilingual Corpus‡		*55*			*30.84*			*68.32*	
Supervised Training		*85.78*			*78.89*			*84.02*	

Model	Italian								
	slot acc.			joint goal acc.			request acc.		
	BASE	MLT_O	MLT_A	BASE	MLT_O	MLT_A	BASE	MLT_O	MLT_A
MUSE	60.59	73.55	**76.88**	20.66	36.88	**39.35**	79.09	82.24	**84.23**
Multi. BERT*	53.34	65.49	**69.48**	12.88	26.45	**31.41**	76.12	84.58	**85.18**
+ Transformer	54.56	66.87	**71.45**	12.63	28.59	**33.35**	77.34	82.93	**84.96**
Ontology Matching†		*23*			-			*21*	
Translate Train†		*48*			-			*51*	
Bilingual Dictionary‡		*73*			*39.01*			*77.09*	
Bilingual Corpus‡		*72*			*41.23*			*81.23*	
Supervised Training		*88.92*			*80.22*			*91.05*	

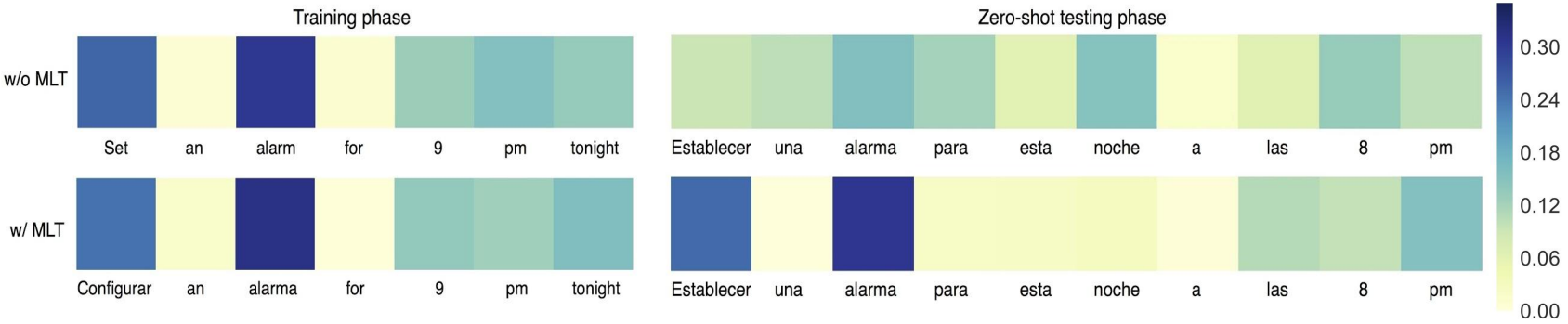
Zero-shot results for the target languages on Multilingual WOZ 2.0. MLT_A denotes our approach (attention-informed MLT), which utilizes the same number of word pairs (90 word pairs) as MLT_O (MLT based on ontology).

Zero-shot Results in NLU Task

Model	Spanish						Thai					
	Intent acc.			Slot F1			Intent acc.			Slot F1		
	BASE	MLT_H	MLT_A	BASE	MLT_H	MLT_A	BASE	MLT_H	MLT_A	BASE	MLT_H	MLT_A
RCSLS	37.67	77.59	**87.05**	22.23	**59.12**	57.75	35.12	68.63	**81.44**	8.72	29.44	**30.42**
XLM (MLM)	60.8	75.11	**83.95**	38.55	63.29	**66.11**	37.59	46.34	**65.31**	8.12	19.03	**20.43**
+ Transformer	62.33	82.83	**85.63**	41.67	66.53	**67.95**	40.31	57.27	**68.55**	11.45	26.02	**27.45**
XLM (TLM+MLM)	62.48	81.34	**84.91**	42.27	65.71	**66.48**	31.62	50.34	**65.25**	7.91	19.22	**19.88**
+ Transformer	65.32	83.79	**87.48**	44.39	66.03	**68.55**	37.53	68.62	**72.59**	12.84	26.56	**27.98**
Multi. BERT	73.73	77.51	**86.54**	51.73	**74.51**	74.43	28.15	52.25	**70.57**	10.62	24.41	**28.47**
+ Transformer	74.15	82.9	**87.88**	54.28	**74.88**	73.89	26.54	53.84	**73.46**	11.34	26.05	**27.12**
Zero-shot SLU†		*46.64*			*15.41*			*35.64*			*12.11*	
Multi. CoVe		*53.34*			*22.50*			*66.35*			*32.52*	
Multi. CoVe w/ auto		*53.89*			*19.25*			*70.70*			*35.62*	
Translate Train		*85.39*			*72.87*			*95.85*			*55.43*	

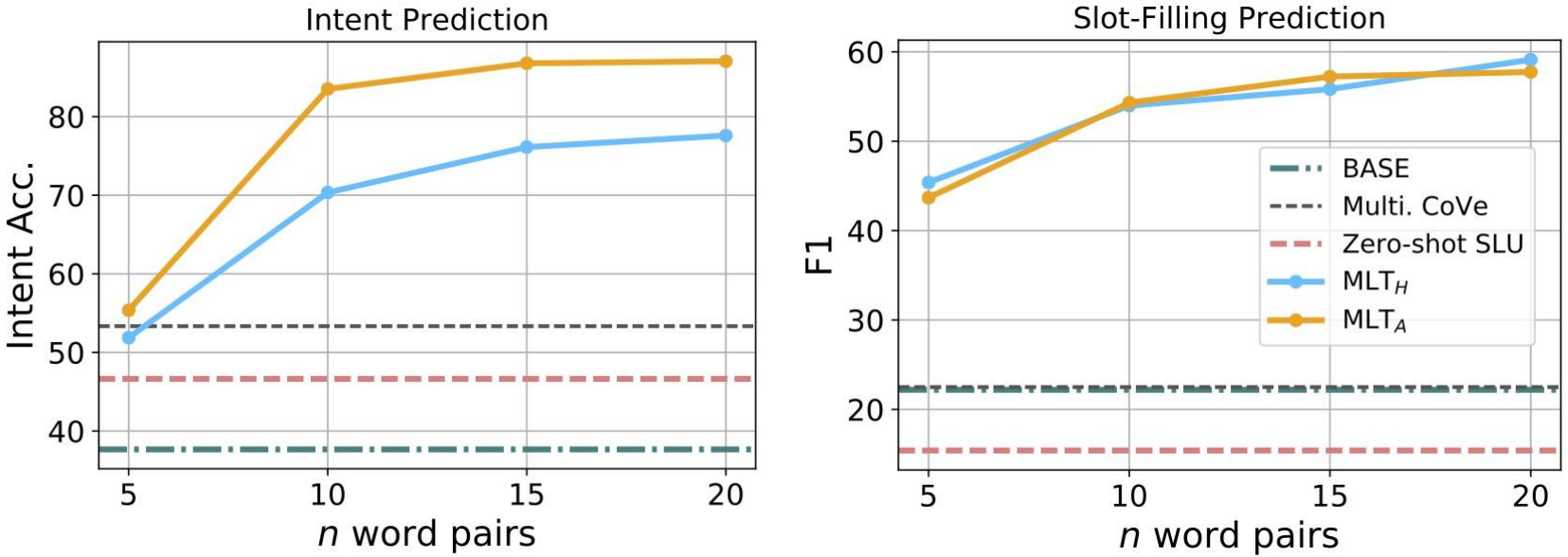
Zero-shot results on multilingual NLU dataset (Schuster et al. 2019), and the number of word pairs on both MLT_H and MLT_A is 20.

Visualization



Attentions on words in both training and testing phases.

Zero-shot Results in NLU Task



The dynamics of the NLU task: intent and slot-filling results with different numbers of word pairs on Spanish test data.

Conclusion

- We propose attention-informed mixed-language training for cross-lingual task-oriented dialogue systems.
- Our approach utilizes very few task-related parallel word pairs base on the attention scores.
- The task-related words have a generalization ability to other words that have similar semantics in target languages.

Thanks!

Check our code

https://github.com/zliucr/mixed-language-training