

Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model

Tsung-Yuan Hsu* Chi-Liang Liu* Hung-Yi Lee

ZERO-SHOT READING COMPREHENSION

→ Training Reading Comprehension (RC) model without using data from target domain (ex. Chinese)

- **Motivation:** difficulty of collecting RC training data for every language
- **Method:** transfer knowledge from English RC to other language RC with multilingual BERT.
- **Dataset:** SQuAD v1.1, DRCD (Chinese), KorQuAD (Korean)

EM/F1 SCORE

Model	Train-set	EM	F1
Shao et al. 2018	Chinese	-	53.78
QANet	Chinese	66.10	78.10
English-BERT	Chinese	65.00	76.96
Chinese-BERT	Chinese	82.00	89.10
multi-BERT	Chinese	81.24	88.68
multi-BERT	English	63.31	78.82
multi-BERT	English + Chinese	82.63	90.10

EM/F1 scores over Chinese testing set.
English/Chinese-BERT is BERT pretrained on English/Chinese monolingual dataset.
(English dataset: SQuAD, Chinese dataset: DRCD)

CROSS-LINGUAL TRANSFER

→ Explore cross-lingual transferring ability of the method between different language-pairs

- **Dataset:**
 - Original: SQuAD, DRCD, KorQuAD
 - Translated: SQuAD and DRCD translated into other 4 languages with Google Translate

Training set	Testing set		
	English	Chinese	Korean
En	81.2/88.6	63.3/78.8	49.2/69.3
Zh	34.1/53.8	81.2/88.7	56.4/78.2
Kr	58.5/68.4	73.4/82.7	69.4/89.3
En-Er	67.5/76.4	56.5/72.5	37.2/56.3
En-Zh	59.7/71.4	61.4/78.8	49.0/72.7
En-Jp	53.3/64.9	62.4/76.7	50.4/72.0
En-Kr	41.7/50.1	56.7/71.6	47.1/70.8
Zh-En	26.6/44.1	57.7/71.7	40.5/59.5
Zh-Fr	23.4/39.8	44.9/62.0	39.6/59.9
Zh-Jp	25.5/42.6	60.9/72.4	44.9/65.7
Zh-Kr	26.5/42.2	58.2/69.5	47.4/67.7

EM/F1 score of multi-BERTs fine-tuned on different training sets and tested on different languages (En: English, Fr: French, Zh: Chinese, Jp: Japanese, Kr: Korean, xx-yy: translated from xx to yy). The text in bold means training data language is the same as testing data language.

TPOLOGY MANIPULATED

→ If the model only learns the semantic mapping between different languages, changing English typology order from SVO to SOV should improve the transfer ability from English to Korean significantly.

- **Dataset:**
 - Artificially created typology-manipulated dataset.

Typology	Example	Train	English	Chinese	Korean
SVO	En: I like you Ch: 我喜歡你	En	81.2/88.6	63.3/78.8	49.2/69.3
SOV	Jp: (僕は) 君が好きです Kr: (나는) 너를 좋아해	En-SOV	78.4/86.5	62.8/78.3	47.6/70.4
		En-VOS	79.4/87.1	59.1/74.6	46.2/67.0
FAKE SOV	En: I like you → I you like	En-VSO	79.4/87.1	60.9/76.8	44.2/65.4
		En-OSV	78.9/86.9	63.5/78.0	49.0/70.7
		En-OVS	73.6/82.5	57.6/72.1	45.8/67.0

Example of the difference of typology between languages and how artificially created typology-manipulated dataset is created.

EM/F1 scores on artificially created typology-manipulated dataset.

CODE SWITCHING

→ If tokens are represented in a language-agnostic way, the model may be able to handle code-switching data.

- **Dataset:**
 - Artificial code-switching datasets by word-by-word translation with given dictionaries and we substitute the words if the words are in the bilingual dictionaries.

Example	Train	Mix Lang.	EM	F1	Sub.
pred: second 法律 of 熱力學 (Zh) gt: second law of thermodynamics		None	81.17	88.63	0%
		Chinese	68.79	78.18	31%
pred: エレクトリック motors (Jp) gt: electric motors	English	French	65.7	77.43	61%
		Japanese	63.32	74.06	30%
		Korean	39.93	63.46	32%
pred: the 차이점 in 잠재력 에너지 (Kr) gt: the difference in potential energy					

Answers inferred on code-switching dataset. The predicted answers would be the same as the ground truths (gt) if we translate every word into English.

EM/F1 scores on artificial code-switching datasets generated by replacing some of the words in English dataset with synonyms in another language. (Sub. is the substitution ratio of the dataset)

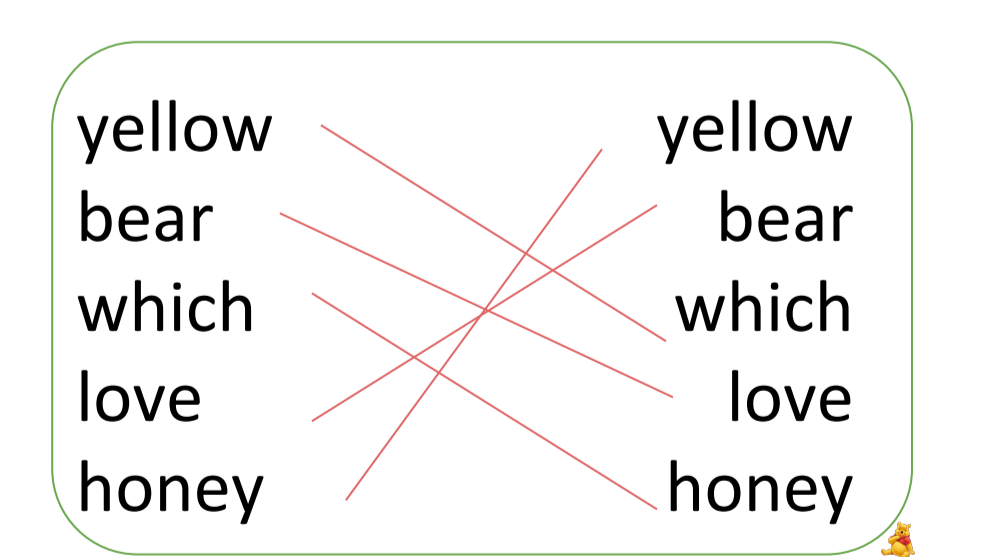
UNSEEN LANGUAGE DATASET

→ Examine if the model could tackle problems via simple literal pattern matching between question and answers.

- **Dataset:**
 - Modified datasets where every word is replaced with another word in the same vocabulary following a randomly generated 1-to-1 word mapping.

Train	Test	EM	F1
English	English-permuted	1.25	11.54
English	Chinese-permuted	5.02	17.49
Chinese	Chinese-permuted	8.91	25.67

EM/F1 scores over artificially created unseen languages (English-permuted and Chinese-permuted).



Concept of word mapping generated by random embeddings permutation

VISUALIZATION OF EMBEDDINGS

PCA visualization of the last BERT pretrained layer.

LEFT: before fine-tuning on SQuAD.
RIGHT: after fine-tuning on SQuAD

