Multilingual Transformers and Cross-Lingual Information Retrieval

Project by Will Taft

Our task

Build a model that ranks documents by relevance to a query, even when the document and query are in a different language.

The old approach

- Term interaction models which take as input the similarity of the embeddings for every query term with every document term.
- Cross-lingual embeddings make the model cross-lingual.

Reason to think we can do better

Results applying transformers to document ranking for both long and short documents (Yang 2019):

Model	2011 AP	2011 P30	2012 AP	2012 P30	2013 AP	2013 P30	2014 AP	2014 P30
QL	0.3576	0.4000	0.2091	0.3311	0.2532	0.4450	0.3924	0.6182
RM3	0.3824	0.4211	0.2342	0.3452	0.2766	0.4733	0.4480	0.6339
DRMM (Guo et al., 2016)	0.3477	0.4034	0.2213	0.3537	0.2639	0.4772	0.4042	0.6139
DUET (Mitra et al., 2017)	0.3576	0.4000	0.2243	0.3644	0.2779	0.4878	0.4219	0.6467
K-NRM (Xiong et al., 2017)	0.3576	0.4000	0.2277	0.3520	0.2721	0.4756	0.4137	0.6358
PACRR (Hui et al., 2017)	0.3810	0.4286	0.2311	0.3576	0.2803	0.4944	0.4140	0.6358
MP-HCNN (Rao et al., 2019)	0.4043	0.4293	0.2460	0.3791	0.2896	0.5294	0.4420	0.6394
BiCNN (Shi et al., 2018)	0.4293	0.4728	0.2621	0.4147	0.2990	0.5367	0.4563	0.6806
BERT	**0.4697**	**0.5040**	**0.3073**	**0.4356**	**0.3357**	**0.5656**	**0.5176**	**0.7006**

Document ranking with a transformer

- Present document ranking as a sequence classification task.
- Is [SEP]Q[SEP]DOC[SEP] relevant or non-relevant?
- Take the pretrained transformer and finetune on a sequence classification task like this.

Making it multilingual

- Facebook's XLM
- Pretrains transformers on many monolingual corpora with a shared BPE code vocabulary.
- TRANSLATE-TRAIN is with fine tuning on all languages. The bottom is zero-shot. This is for a sentence pair classification task.

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Δ
Machine translation baselines (TRANSLATE-TRAIN)																
Devlin et al. (2018)	81.9	-	77.8	75.9	-	-	-	-	70.7	-	-	76.6	-	-	61.6	-
XLM (MLM+TLM)	85.0	80.2	80.8	80.3	78.1	79.3	78.1	74.7	76.5	76.6	75.5	78.6	72.3	70.9	63.2	76.7
Machine translation baselines (TRANSLATE-TEST)																
Devlin et al. (2018)	81.4	-	74.9	74.4	-	-	-	-	70.4	-	-	70.1	-	-	62.1	-
XLM (MLM+TLM)	85.0	79.0	79.5	78.1	77.8	77.6	75.5	73.7	73.7	70.8	70.4	73.6	69.0	64.7	65.1	74.2
Evaluation of cross-lingual sentence encoders																
Conneau et al. (2018b)	73.7	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4	65.6
Devlin et al. (2018)	81.4	-	74.3	70.5	-	-	-	-	62.1	-	-	63.8	-	-	58.3	-
Artetxe and Schwenk (2018)	73.9	71.9	72.9	72.6	73.1	74.2	71.5	69.7	71.4	72.0	69.2	71.4	65.5	62.2	61.0	70.2
XLM (MLM)	83.2	76.5	76.3	74.2	73.1	74.0	73.1	67.8	68.5	71.2	69.2	71.9	65.7	64.6	63.4	71.5
XLM (MLM+TLM)	**85.0**	**78.7**	**78.9**	**77.8**	**76.6**	**77.4**	**75.3**	**72.5**	**73.1**	**76.1**	**73.2**	**76.5**	**69.6**	**68.4**	**67.3**	**75.1**

Our attempt

- Took the an XLM transformer pretrained on 100 languages, including English and Swahili.
- Because documents were long, created training data with a sliding window. The window inherited the relevance judgement from the entire document.

  [SEP]Q[SEP]DOC[SEP]

- This creates a model to give relevance judgements on passages. We consider the highest scoring passage to rank the document.
- Produced a model that gives the same prediction on every sequence.

What we should try next

- Fine tune on an English only dataset. It’s more important to have short documents with accurate relevance judgements for training.

A second approach we tried:

- Better word embeddings.
- This approach failed; the supervised word embeddings are much better.

	XLM (mlm-100)	**XLM (mlm-tlm-15)**	**MUSE supervised**
Average cosine similarity between direct translation words pairs	0.324 on 506 in vocabulary word pairs	0.18 on 1228 in vocabulary words pairs	0.430 on 4279 in vocabulary word pairs