#### Evaluating Multilingual Text Encoders for Unsupervised Cross-Lingual Retrieval

Robert Litschko<sup>1</sup>, Ivan Vulić<sup>2</sup>, Simone Paolo Ponzetto<sup>1</sup>, Goran Glavaš<sup>1</sup>

<sup>1</sup>Data and Web Science Group, University of Mannheim, Germany <sup>2</sup>Language Technology Lab, University of Cambridge, UK



Contact: litschko@informatik.uni-mannheim.de

#### **Motivation**

- Pre-trained Transformers achieve strong performance in supervised NLP and have been adopted for multilingual NLP
  - BERT, GPT-3, RoBERTa, ...
  - De facto standard in NLU and NLG
- Multilingual Text Encoders render Cross-lingual Word Embeddings (CLWE) effectively obsolete

To which extend does this generalize to unsupervised Cross-lingual Information Retrieval (CLIR)?

### Contribution

- Systematic comparison of multilingual text encoders on:
  - A. Document-level CLIR (CLEF-2003), 9 language pairs
  - B. Sentence-level CLIR (Europarl), 6 language pairs
- Benchmark different types of models:
  - A. Baselines
  - B. Models based on multilingual transformers (mBERT, XLM)
  - C. Similarity-specialized sentence encoders

#### Language pairs

- $\mathsf{EN} \twoheadrightarrow \{\,\mathsf{FI},\,\mathsf{IT},\,\mathsf{RU},\,\mathsf{DE}\,\}$
- $\mathsf{DE} \longrightarrow \{ \mathsf{FI}, \mathsf{IT}, \mathsf{RU} \}$
- FI → { IT, RU }



#### A. Baselines

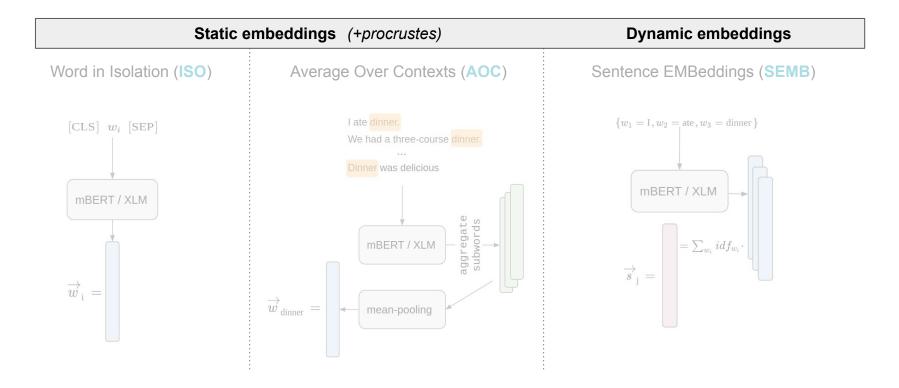
- Machine Translation baseline (MT-IR):
  - A. Translate query into target language with Google Translate
  - B. Monolingual retrieval: Unigram Language Modelling
- CLWE Baseline (Proc-B):
  - A. Map monolingual embedding spaces into cross-lingual space via procrustes
  - B. Represent documents / queries by their idf-weighted sum
  - C. Retrieval: Rank documents by cosine-similarity
- Adjusted (fair) CLWE Baseline (**Proc-B**<sub>LEN</sub>):
  - A. Use first 128 word-pieces for query- and document-embeddings

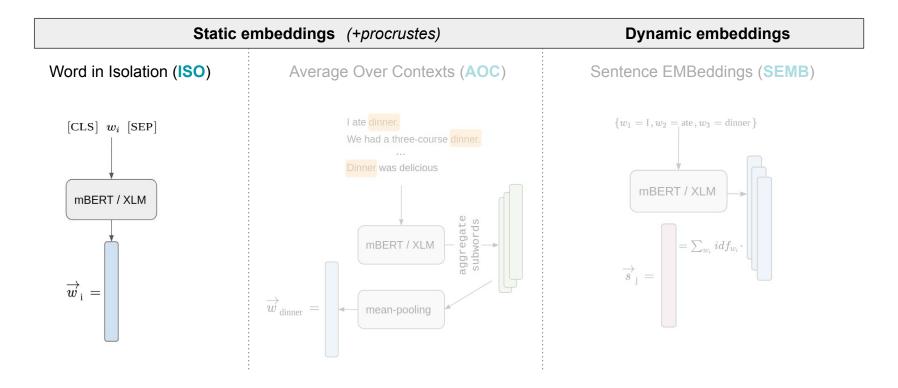


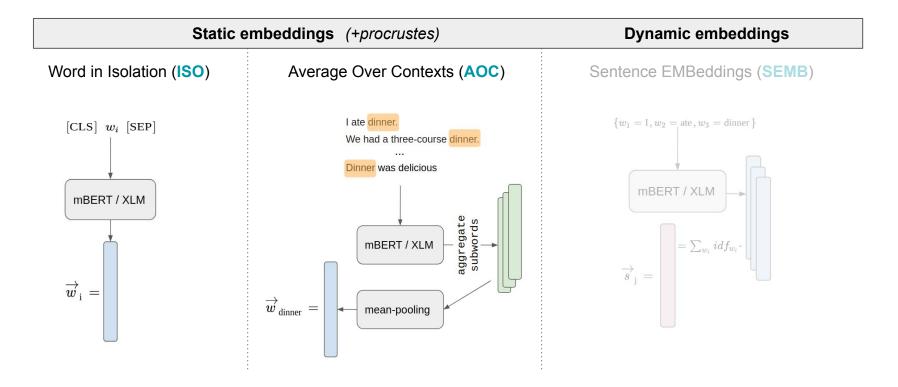


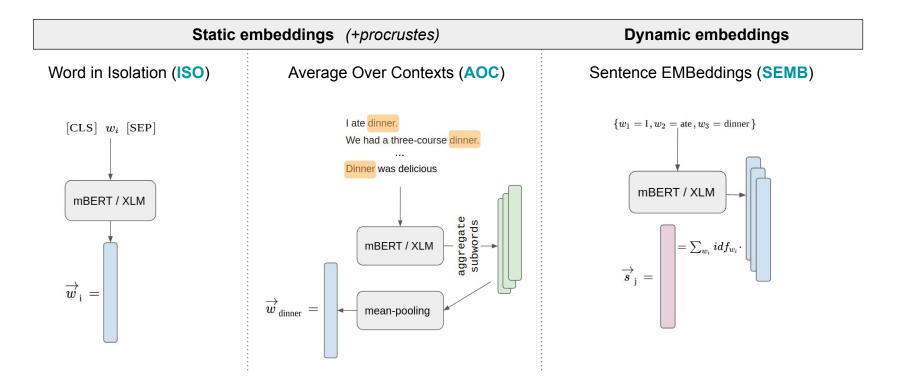
Procrustes (Mikolov et al. 2013)

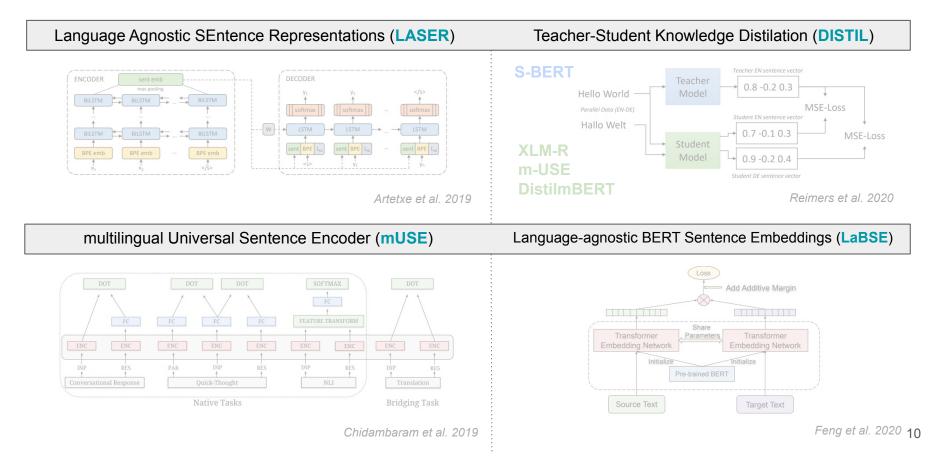
$$W_{L_1} = rgmin_W ||X_{L_1}W - X_{L_2}||$$

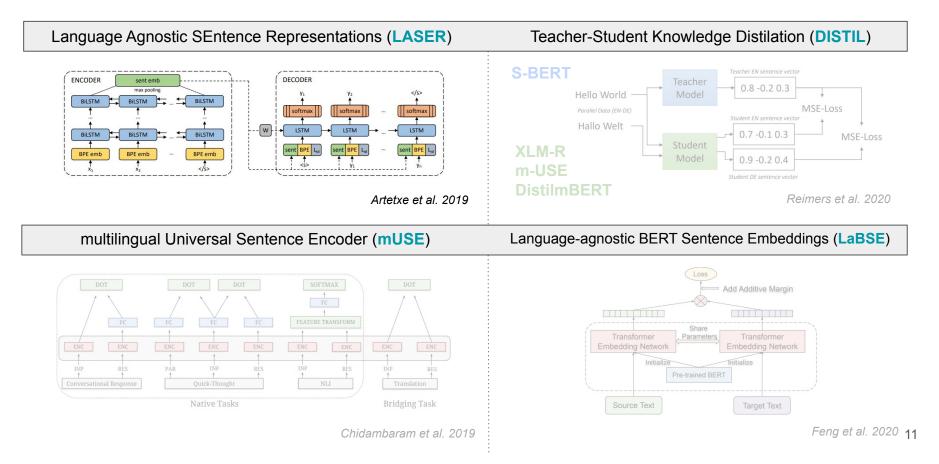


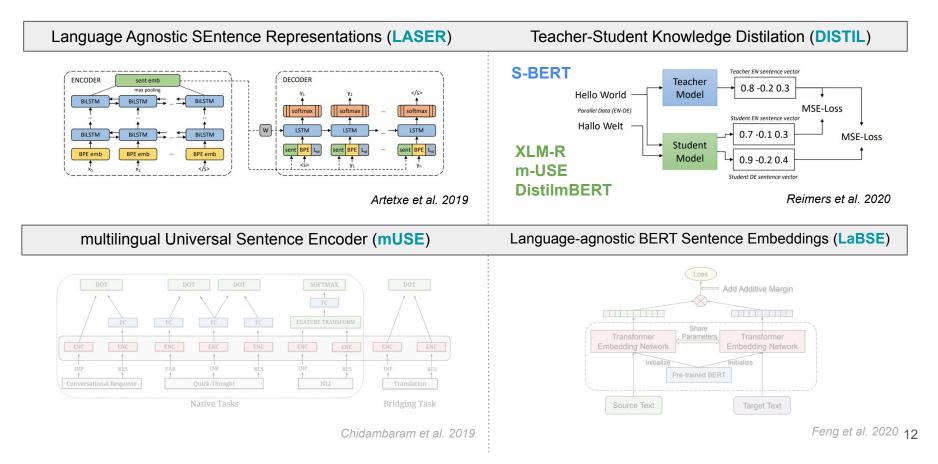


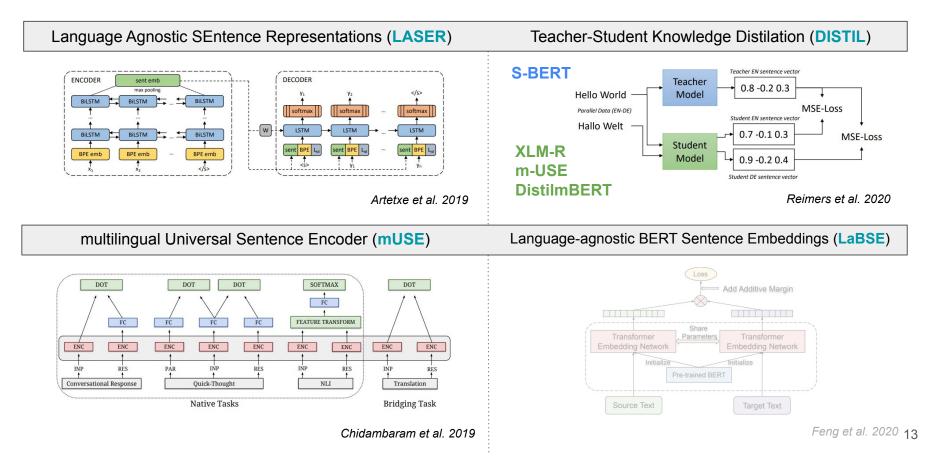


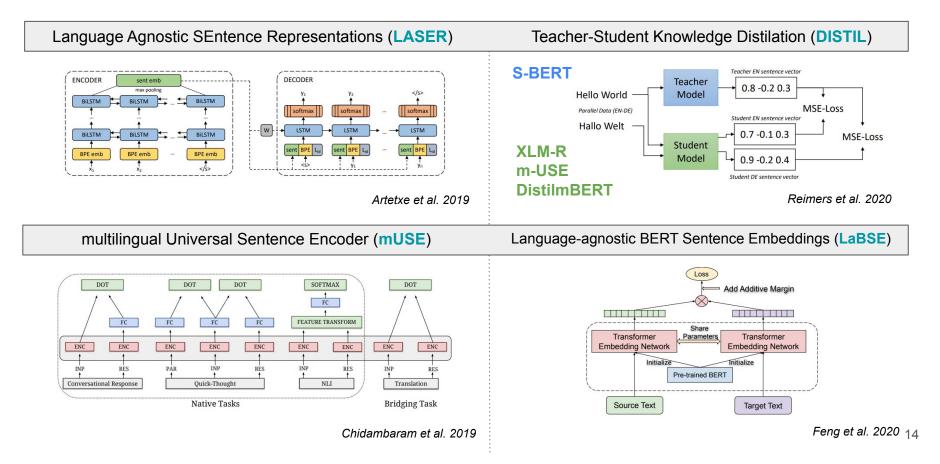














#### **Document-level CLIR Results**

	EN-FI	EN-IT	EN-RU	EN-DE	DE-FI	DE-IT	DE-RU	FI-IT	FI-RU	AVG	w/o FI
Baselines											
MT-IR	.276	.428	.383	.263	.332	.431	.238	.406	.261	.335	.349
Proc-B	.258	.265	.166	.288	.294	.230	.155	.151	.136	.216	.227
Proc-B <sub>LEN</sub>	.165	.232	.176	.194	.207	.186	.192	.126	.154	.181	.196
Models based on	multiling	ual Transj	formers								
SEMB <sub>XLM</sub>	.199*	.187*	.183	.126*	.156*	.166*	.228	.186*	.139	.174	.178
SEMB <sub>mBERT</sub>	.145*	.146*	.167	.107*	.151*	.116*	.149*	.117	.128*	.136	.137
AOC <sub>XLM</sub>	.168	.261	.208	.206*	.183	.190	.162	.123	.099	.178	.206
<b>AOC</b> <sub>mBERT</sub>	.172*	.209*	.167	.193*	.131*	.143*	.143	.104	.132	.155	.171
ISO <sub>XLM</sub>	.058*	.159*	.050*	.096*	.026*	.077*	.035*	.050*	.055*	.067	.083
<b>ISO</b> <sub>mBERT</sub>	.075*	.209	.096*	.157*	.061*	.107*	.025*	.051*	.014*	.088	.119
Similarity-specia	lized sent	ence enco	ders (with	parallel a	data supe	ervision)					
DISTILFILTER	.291	.261	.278	.255	.272	.217	.237	.221	.270	.256	.250
DISTIL <sub>XLM-R</sub>	.216	.190*	.179	.114*	.237	.181	.173	.166	.138	.177	.167
DISTILUSE	.141*	.346*	.182	.258	.139*	.324*	.179	.104	.111	.198	.258
DISTIL	т .294	.290*	.313	.247*	.300	.267*	.284	.221*	.302*	.280	.280
LaBSE	.180*	.175*	.128	.059*	.178*	.160*	.113*	.126	.149	.141	.127
LASER	.142	.134*	.076	.046*	.163*	.140*	.065*	.144	.107	.113	.094
m-USE	.109*	.328*	.214	.230*	.107*	.294*	.204	.073	.090	.183	.254

- None of the models outperform the CLWE baseline (Mean Average Precision; MAP)
- After adjusting for length AOC, SEMB come reasonably close (Proc-B<sub>LEN</sub>)

#### **Document-level CLIR Results**

	EN-FI	EN-IT	EN-RU	EN-DE	DE-FI	DE-IT	DE-RU	FI-IT	FI-RU	AVG	w/o FI
Baselines											
MT-IR Proc-B Proc-B <sub>LEN</sub>	.276 .258 .165	<b>.428</b> .265 .232	.383 .166 .176	<b>.263</b> .288 .194	<b>.332</b> .294 .207	<b>.431</b> .230 .186	.238 .155 .192	<b>.406</b> .151 .126	.261 .136 .154	<b>.335</b> .216 .181	<b>.349</b> .227 .196
Models based on	multiling	ual Transj	formers								
SEMB <sub>XLM</sub> SEMB <sub>mBERT</sub> AOC <sub>XLM</sub> AOC <sub>mBERT</sub> ISO <sub>XLM</sub> ISO <sub>mBERT</sub>	.199* .145* .168 .172* .058* .075*	.187* .146* .261 .209* .159* .209	.183 .167 .208 .167 .050* .096*	.126* .107* .206* .193* .096* .157*	.156* .151* .183 .131* .026* .061*	.166* .116* .190 .143* .077* .107*	.228 .149* .162 .143 .035* .025*	.186* .117 .123 .104 .050* .051*	.139 .128* .099 .132 .055* .014*	.174 .136 .178 .155 .067 .088	.178 .137 .206 .171 .083 .119
Similarity-special						· · · · · ·					
DISTIL <sub>FILTER</sub> DISTIL <sub>XLM-R</sub> DISTIL <sub>USE</sub> DISTIL <sub>DistilmBERT</sub> LaBSE LASER m-USE	.291 .216 .141* <b>.294</b> .180* .142 .109*	.261 .190* .346* .290* .175* .134* .328*	.278 .179 .182 .313 .128 .076 .214	.255 .114* .258 .247* .059* .046* .230*	.272 .237 .139* .300 .178* .163* .107*	.217 .181 .324* .267* .160* .140* .294*	.237 .173 .179 <b>.284</b> .113* .065* .204	.221 .166 .104 .221* .126 .144 .073	.270 .138 .111 <b>.302</b> * .149 .107 .090	.256 .177 .198 .280 .141 .113 .183	.250 .167 .258 .280 .127 .094 .254

- Mixed results: Three models generally outperform Proc-B
- LASER exhibits inferior results on CLIR (Bi-LSTM vs. Transformer)
- **DISTIL** FILTER: priori stopword filtering deteriorates performance

#### Sentence-level CLIR Results

	EN-FI	EN-IT	EN-DE	DE-FI	DE-IT	FI-IT	AVG	w/o F
Baselines								
MT-IR	.639	.783	.712	.520	.676	.686	.669	.723
Proc-B	.143	.523	.415	.162	.342	.137	.287	.427
Models based o	n multilingud	ul Transforme	ers					
SEMB <sub>XLM</sub>	.309*	.677*	.465	.391*	.495*	.346*	.447	.545
SEMB <sub>mBERT</sub>	.199*	.570	.355	.231*	.481*	.353*	.365	.469
AOC <sub>XLM</sub>	.099	.527	.274*	.102*	.282	.070*	.226	.361
AOC <sub>mBERT</sub>	.095*	.433*	.274*	.088*	.230*	.059*	.197	.312
ISO <sub>XLM</sub>	.016*	.178*	.053*	.006*	.017*	.002*	.045	.082
ISO <sub>mBERT</sub>	.010*	.141*	.087*	.005*	.017*	.000*	.043	.082
Similarity-speci	alized senter	ice encoders	(with parallel	data supervis	sion)			
DISTIL <sub>XLM-R</sub>	.924*	.944*	.942*	.911*	.919*	.915*	.849	.882
DISTILUSE	.084*	.960*	.952*	.137	.920*	.072*	.521	.944
DISTILDistilmBE	RT .817*	.902*	.902*	.810*	.842*	.793*	.844	.882
LaBSE	.971*	.972*	.964*	.948*	.954*	.951*	.960	.963
LASER	.974*	.976*	.969*	.967*	.965*	.961*	.969	.944
m-USE	.079*	.951*	.929*	.086*	.886*	.039*	.495	.922

- Still underperform compared to translation-based baseline MT-IR (Mean Reciprocal Rank; MRR)
- Models outperform **Proc-B**, improvement expected due to shorter sequence lengths

#### Sentence-level CLIR Results

			DE-FI	DE-IT	FI-IT	AVG	w/o FI
.639 .143	.783 .523	.712 .415	.520 .162	.676 .342	.686 .137	.669 .287	.723 .427
multilingu	al Transforme	ers.					
.309* .199*	.677* .570	.465 .355 274*	.391* .231*	.495* .481*	.346* .353*	.447 .365	.545 .469 .361
.095*	.433*	.274*	.088*	.230*	.059*	.197	.312
.010*	.178*	.087*	.005*	.017*	.000*	.043	.082
ized senter	nce encoders	(with parallel	data supervis	sion)			
.924* .084* .817* .971* .974*	.944* .960* .902* .972* <b>.976</b> *	.942* .952* .902* .964* .969*	.911* .137 .810* .948* .967*	.919* .920* .842* .954* .965*	.915* .072* .793* .951* .961*	.849 .521 .844 .960 <b>.969</b>	.882 .944 .882 .963 <b>.944</b> .922
	.143 multilingud .309* .199* .099 .095* .016* .010* ized senter .924* .084* .817* .971*	.143 .523 multilingual Transforme .309* .677* .199* .570 .099 .527 .095* .433* .016* .178* .010* .141* ized sentence encoders of .924* .944* .084* .960* .817* .902* .971* .972* .974* .976*	.143 .523 .415   multilingual Transformers .309* .677* .465   .199* .570 .355   .099 .527 .274*   .016* .178* .053*   .010* .141* .087*   ized sentence encoders (with parallel   .924* .944* .942*   .084* .960* .952*   .817* .902* .902*   .971* .972* .964*   .974* .976* .969*	.143 .523 .415 .162   multilingual Transformers   .309* .677* .465 .391*   .199* .570 .355 .231*   .099 .527 .274* .102*   .095* .433* .274* .088*   .016* .178* .053* .006*   .010* .141* .087* .005*   ized sentence encoders (with parallel data supervision) .137 .817* .902* .911*   .084* .960* .952* .137 .810* .971* .972* .964* .948*   .971* .972* .964* .948* .948* .949* .948*	.143 .523 .415 .162 .342   multilingual Transformers   .309* .677* .465 .391* .495*   .199* .570 .355 .231* .481*   .099 .527 .274* .102* .282   .095* .433* .274* .088* .230*   .016* .178* .053* .006* .017*   .010* .141* .087* .005* .017*   ized sentence encoders (with parallel data supervision) .924* .944* .942* .911* .919*   .084* .960* .952* .137 .920* .810* .842*   .971* .902* .902* .810* .842*   .971* .972* .964* .948* .954*   .974* .976* .969* .967* .965*	.143 .523 .415 .162 .342 .137   multilingual Transformers   .309* .677* .465 .391* .495* .346*   .199* .570 .355 .231* .481* .353*   .099 .527 .274* .102* .282 .070*   .095* .433* .274* .088* .230* .059*   .016* .178* .053* .006* .017* .002*   .010* .141* .087* .005* .017* .000*   ized sentence encoders (with parallel data supervision) .924* .944* .942* .911* .919* .915*   .084* .960* .952* .137 .920* .072*   .817* .902* .902* .810* .842* .793*   .971* .972* .964* .948* .954* .951*   .974* .976* .969* .967* .965* .961*	.143.523.415.162.342.137.287multilingual Transformers.309*.677*.465.391*.495*.346*.447.199*.570.355.231*.481*.353*.365.099.527.274*.102*.282.070*.226.095*.433*.274*.088*.230*.059*.197.016*.178*.053*.006*.017*.002*.045.010*.141*.087*.005*.017*.000*.043ized sentence encoders (with parallel data supervision).924*.944*.942*.911*.919*.915*.849.084*.960*.952*.137.920*.072*.521.817*.902*.902*.810*.842*.793*.844.971*.972*.964*.948*.954*.951*.960.969*.967*.965*.961*.969

- All models substatially outperform both baselines
- Caveat: Models trained on parallel data (effectively supervised retrieval)

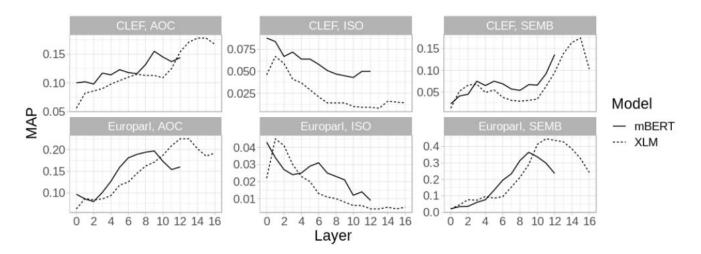


#### Input Sequence-length

Length	$\text{SEMB}_{\text{mBERT}}$	$SEMB_{XLM}$	DIST <sub>use</sub>	DIST <sub>XLM-R</sub>	DIST <sub>DmBERT</sub>	mUSE	LaB SE	LASER
64	.104	.128	.235	.167	.237	.254	.127	.094
128	.137	.178	.258	.162	.280	.247	.125	.035
256	.117	.158	.230	.146	.250	.197	.096	.024

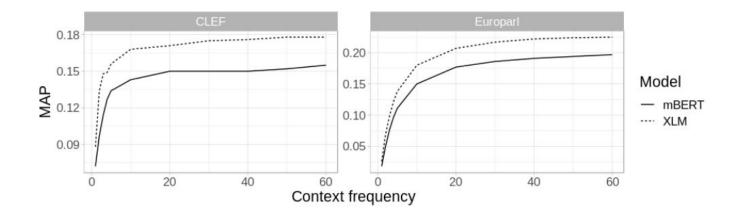
- Multilingual encoders effectively truncate long documents
- Encoding longer chunks of documents works slightly worse:
  - More difficult to encode longer portion of text semantically accurate
  - If relevance signal not within 128 tokens, it often does not appear beyond

#### Layer Selection



- There exist no universally optimal layer
- Lexical Sematics: ISO performs best on representations from lower layers
- Compositional Semantics: AOC / SEMB achieve best performance on higher layers

### Number of Contexts in AOC



- AOC embedding as average representation of the same term in different sentences
- Number of contexts is capped (hyperparameter)
- Performance seems to plateau rather early: around 30 / 40 for AOC MEERT / AOC XLM



## Conclusion

- Cross-lingual Word Embeddings still competitive in unsupervised CLIR
- Without any task-specific fine-tuning, multilingual encoders fail to outperform static CLWEs
- Performance crucially depends on how one encodes semantic information
- Future work on Multilingual Text Encoders for long documents

#### Thank you for your attention!

