Evaluating Multilingual Text Encoders for Unsupervised Cross-Lingual Retrieval

Robert Litschko[1], Ivan Vulić[2], Simone Paolo Ponzetto[1], Goran Glavaš[1]

[1]Data and Web Science Group, University of Mannheim, Germany

[2]Language Technology Lab, University of Cambridge, UK

Contact: litschko@informatik.uni-mannheim.de

ECIR 2021

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

EN → { FI, IT, RU, DE }

DE → { FI, IT, RU }

FI → { IT, RU }

Models

A. Baselines

- Machine Translation baseline (**MT-IR**):
  A. Translate query into target language with Google Translate
  B. Monolingual retrieval: Unigram Language Modelling

Google Translate

- 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

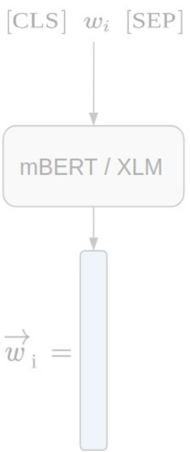
Procrustes (*Mikolov et al. 2013*)

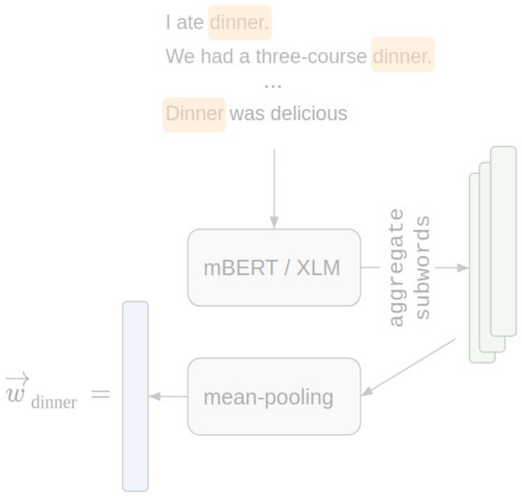
$$W_{L_1} = \arg\min_W ||X_{L_1} W - X_{L_2}||$$

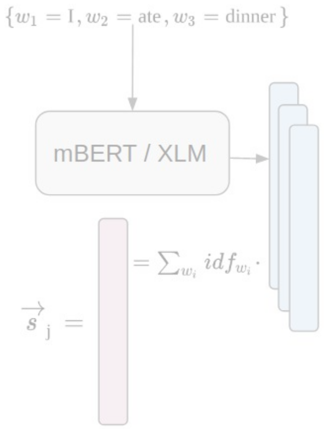
- Adjusted (fair) CLWE Baseline (**Proc-B$_{LEN}$**):
  A. Use first 128 word-pieces for query- and document-embeddings

B. Models based on Multilingual Transformers

Static embeddings *(+procrustes)*		**Dynamic embeddings**
Word in Isolation (**ISO**)	Average Over Contexts (**AOC**)	Sentence EMBeddings (**SEMB**)

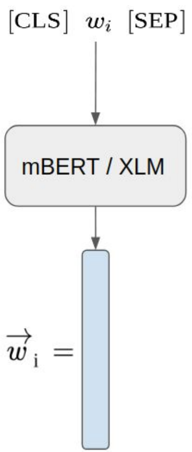


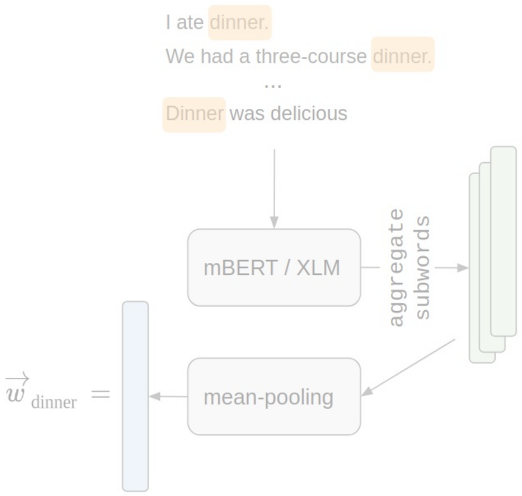


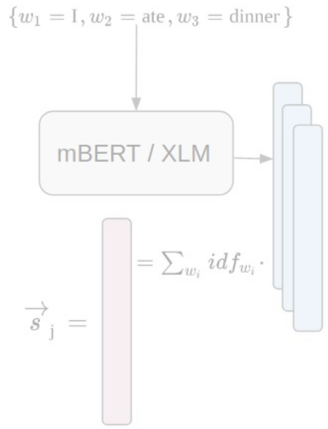


B. Models based on Multilingual Transformers

Static embeddings *(+procrustes)*		**Dynamic embeddings**
Word in Isolation (**ISO**)	Average Over Contexts (**AOC**)	Sentence EMBeddings (**SEMB**)



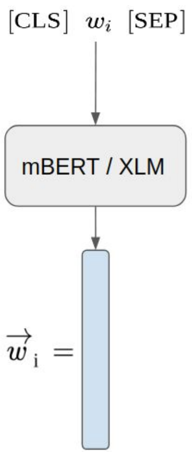




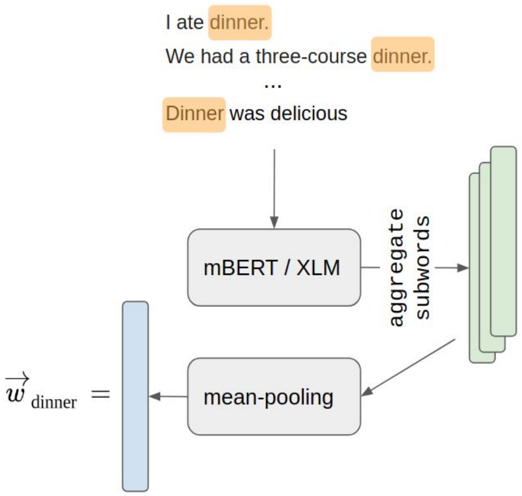
B. Models based on Multilingual Transformers

Static embeddings *(+procrustes)* | **Dynamic embeddings**

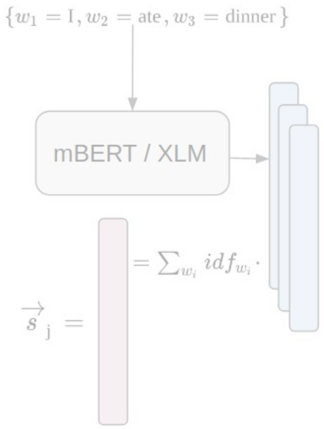
Word in Isolation (**ISO**)



Average Over Contexts (**AOC**)

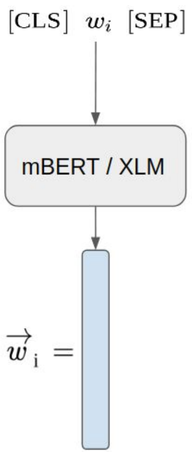


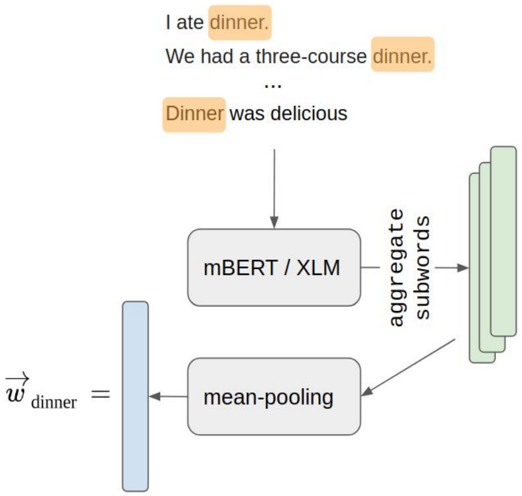
Sentence EMBeddings (**SEMB**)

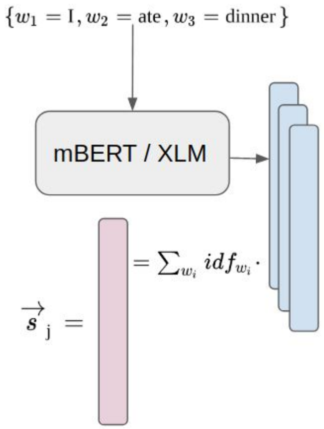


B. Models based on Multilingual Transformers

Static embeddings *(+procrustes)*		**Dynamic embeddings**
Word in Isolation (**ISO**)	Average Over Contexts (**AOC**)	Sentence EMBeddings (**SEMB**)

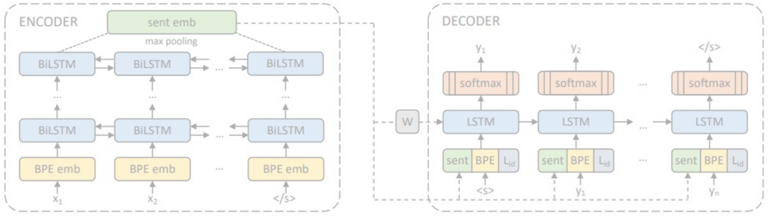






C. Similarity-specialized sentence encoders

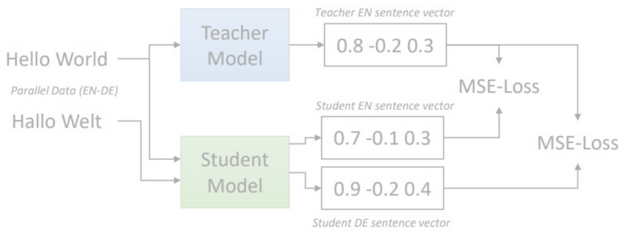
Language Agnostic SEntence Representations (**LASER**)



Artetxe et al. 2019

Teacher-Student Knowledge Distilation (**DISTIL**)

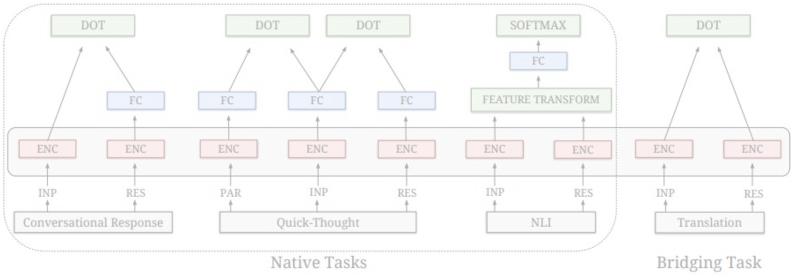
S-BERT



XLM-R
m-USE
DistilmBERT

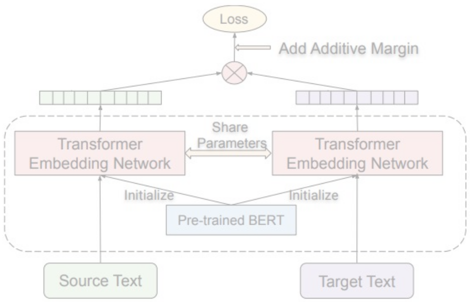
Reimers et al. 2020

multilingual Universal Sentence Encoder (**mUSE**)



Chidambaram et al. 2019

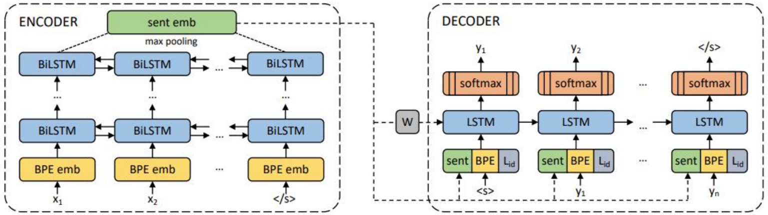
Language-agnostic BERT Sentence Embeddings (**LaBSE**)



Feng et al. 2020

C. Similarity-specialized sentence encoders

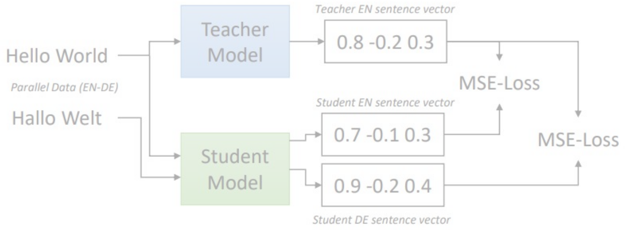
Language Agnostic SEntence Representations (**LASER**)



Artetxe et al. 2019

Teacher-Student Knowledge Distilation (**DISTIL**)

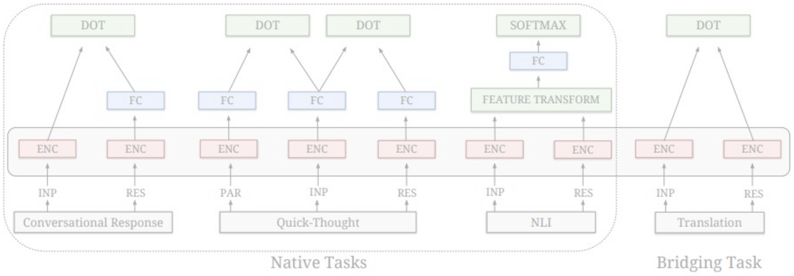
S-BERT



XLM-R
m-USE
DistilmBERT

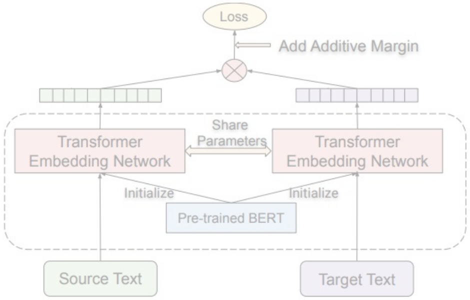
Reimers et al. 2020

multilingual Universal Sentence Encoder (**mUSE**)



Chidambaram et al. 2019

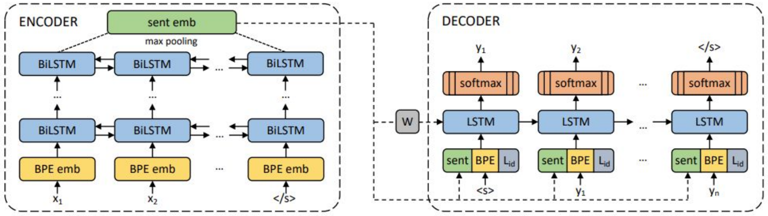
Language-agnostic BERT Sentence Embeddings (**LaBSE**)



Feng et al. 2020

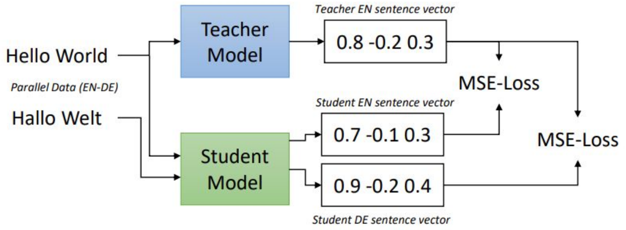
C. Similarity-specialized sentence encoders

Language Agnostic SEntence Representations (**LASER**)	Teacher-Student Knowledge Distilation (**DISTIL**)



Artetxe et al. 2019

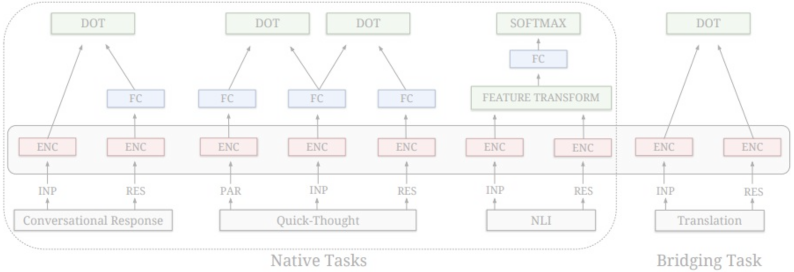
S-BERT



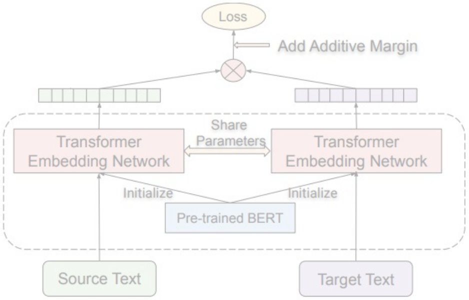
XLM-R
m-USE
DistilmBERT

Reimers et al. 2020

multilingual Universal Sentence Encoder (**mUSE**)	Language-agnostic BERT Sentence Embeddings (**LaBSE**)



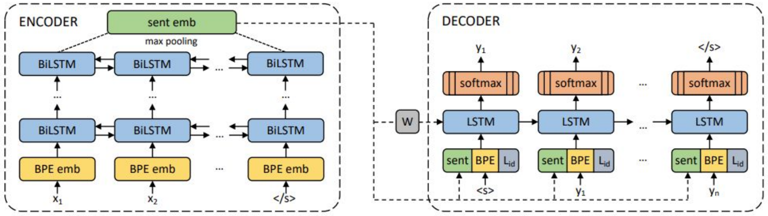
Chidambaram et al. 2019



Feng et al. 2020

C. Similarity-specialized sentence encoders

Language Agnostic SEntence Representations (**LASER**)

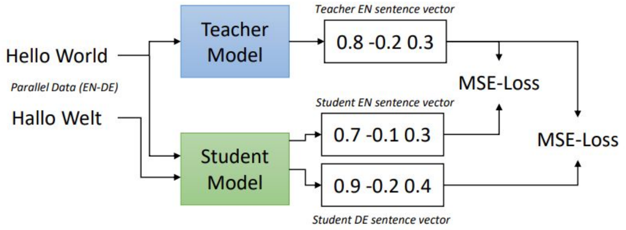


Artetxe et al. 2019

Teacher-Student Knowledge Distilation (**DISTIL**)

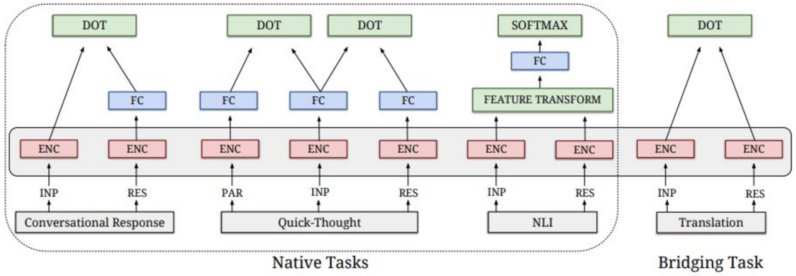
S-BERT

XLM-R
m-USE
DistilmBERT



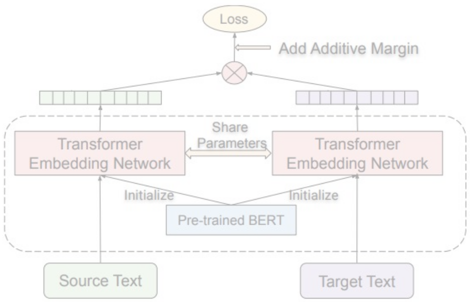
Reimers et al. 2020

multilingual Universal Sentence Encoder (**mUSE**)



Chidambaram et al. 2019

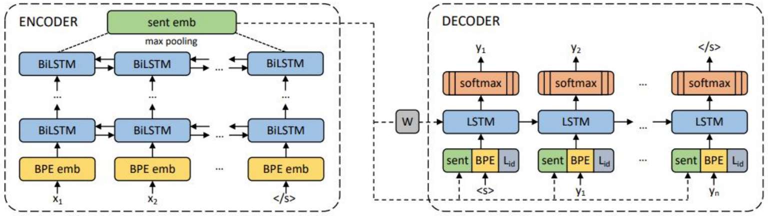
Language-agnostic BERT Sentence Embeddings (**LaBSE**)



Feng et al. 2020

C. Similarity-specialized sentence encoders

Language Agnostic SEntence Representations (LASER)

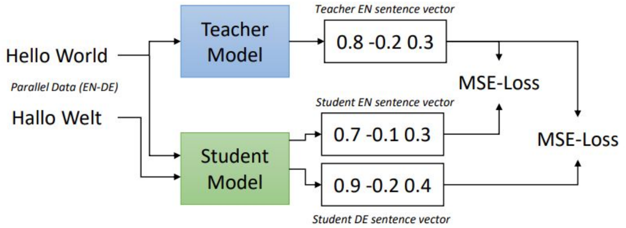


Artetxe et al. 2019

Teacher-Student Knowledge Distilation (DISTIL)

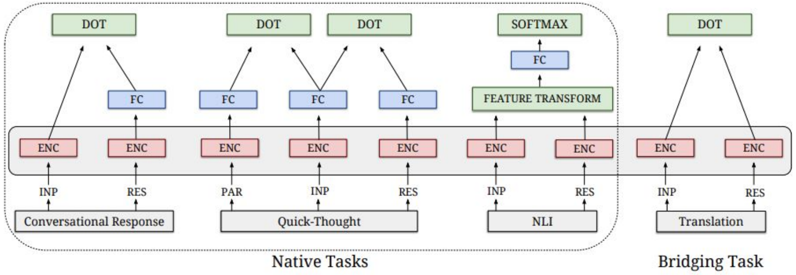
S-BERT

XLM-R
m-USE
DistilmBERT



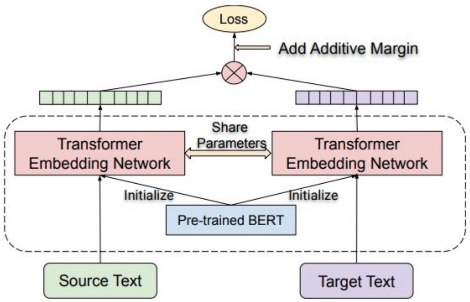
Reimers et al. 2020

multilingual Universal Sentence Encoder (mUSE)



Chidambaram et al. 2019

Language-agnostic BERT Sentence Embeddings (LaBSE)



Feng et al. 2020

Results

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_{LEN}	.165	.232	.176	.194	.207	.186	.192	.126	.154	.181	.196
Models based on multilingual Transformers											
$SEMB_{XLM}$	.199*	.187*	.183	.126*	.156*	.166*	.228	.186*	.139	.174	.178
$SEMB_{mBERT}$	.145*	.146*	.167	.107*	.151*	.116*	.149*	.117	.128*	.136	.137
AOC_{XLM}	.168	.261	.208	.206*	.183	.190	.162	.123	.099	.178	.206
AOC_{mBERT}	.172*	.209*	.167	.193*	.131*	.143*	.143	.104	.132	.155	.171
ISO_{XLM}	.058*	.159*	.050*	.096*	.026*	.077*	.035*	.050*	.055*	.067	.083
ISO_{mBERT}	.075*	.209	.096*	.157*	.061*	.107*	.025*	.051*	.014*	.088	.119
Similarity-specialized sentence encoders (with parallel data supervision)											
$DISTIL_{FILTER}$	.291	.261	.278	.255	.272	.217	.237	.221	.270	.256	.250
$DISTIL_{XLM\text{-}R}$	.216	.190*	.179	.114*	.237	.181	.173	.166	.138	.177	.167
$DISTIL_{USE}$	.141*	.346*	.182	.258	.139*	.324*	.179	.104	.111	.198	.258
$DISTIL_{DistilmBERT}$	**.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_{LEN}**)

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_{LEN}	.165	.232	.176	.194	.207	.186	.192	.126	.154	.181	.196
Models based on multilingual Transformers											
$SEMB_{XLM}$	.199*	.187*	.183	.126*	.156*	.166*	.228	.186*	.139	.174	.178
$SEMB_{mBERT}$	.145*	.146*	.167	.107*	.151*	.116*	.149*	.117	.128*	.136	.137
AOC_{XLM}	.168	.261	.208	.206*	.183	.190	.162	.123	.099	.178	.206
AOC_{mBERT}	.172*	.209*	.167	.193*	.131*	.143*	.143	.104	.132	.155	.171
ISO_{XLM}	.058*	.159*	.050*	.096*	.026*	.077*	.035*	.050*	.055*	.067	.083
ISO_{mBERT}	.075*	.209	.096*	.157*	.061*	.107*	.025*	.051*	.014*	.088	.119
Similarity-specialized sentence encoders (with parallel data supervision)											
$DISTIL_{FILTER}$	.291	.261	.278	.255	.272	.217	.237	.221	.270	.256	.250
$DISTIL_{XLM\text{-}R}$	.216	.190*	.179	.114*	.237	.181	.173	.166	.138	.177	.167
$DISTIL_{USE}$	.141*	.346*	.182	.258	.139*	.324*	.179	.104	.111	.198	.258
$DISTIL_{DistilmBERT}$	**.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

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

Sentence-level CLIR Results

	EN-FI	EN-IT	EN-DE	DE-FI	DE-IT	FI-IT	AVG	w/o FI
Baselines								
MT-IR	.639	.783	.712	.520	.676	.686	.669	.723
Proc-B	.143	.523	.415	.162	.342	.137	.287	.427
Models based on multilingual Transformers								
$SEMB_{XLM}$	.309*	.677*	.465	.391*	.495*	.346*	.447	.545
$SEMB_{mBERT}$	.199*	.570	.355	.231*	.481*	.353*	.365	.469
AOC_{XLM}	.099	.527	.274*	.102*	.282	.070*	.226	.361
AOC_{mBERT}	.095*	.433*	.274*	.088*	.230*	.059*	.197	.312
ISO_{XLM}	.016*	.178*	.053*	.006*	.017*	.002*	.045	.082
ISO_{mBERT}	.010*	.141*	.087*	.005*	.017*	.000*	.043	.082
Similarity-specialized sentence encoders (with parallel data supervision)								
$DISTIL_{XLM-R}$	.924*	.944*	.942*	.911*	.919*	.915*	.849	.882
$DISTIL_{USE}$	.084*	.960*	.952*	.137	.920*	.072*	.521	.944
$DISTIL_{DistilmBERT}$	.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

	EN-FI	EN-IT	EN-DE	DE-FI	DE-IT	FI-IT	AVG	w/o FI
Baselines								
MT-IR	.639	.783	.712	.520	.676	.686	.669	.723
Proc-B	.143	.523	.415	.162	.342	.137	.287	.427
Models based on multilingual Transformers								
SEMB_{XLM}	.309*	.677*	.465	.391*	.495*	.346*	.447	.545
$\text{SEMB}_{\text{mBERT}}$	.199*	.570	.355	.231*	.481*	.353*	.365	.469
AOC_{XLM}	.099	.527	.274*	.102*	.282	.070*	.226	.361
$\text{AOC}_{\text{mBERT}}$	.095*	.433*	.274*	.088*	.230*	.059*	.197	.312
ISO_{XLM}	.016*	.178*	.053*	.006*	.017*	.002*	.045	.082
$\text{ISO}_{\text{mBERT}}$	.010*	.141*	.087*	.005*	.017*	.000*	.043	.082
Similarity-specialized sentence encoders (with parallel data supervision)								
$\text{DISTIL}_{\text{XLM-R}}$	.924*	.944*	.942*	.911*	.919*	.915*	.849	.882
$\text{DISTIL}_{\text{USE}}$	.084*	.960*	.952*	.137	.920*	.072*	.521	.944
$\text{DISTIL}_{\text{DistilmBERT}}$	.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

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

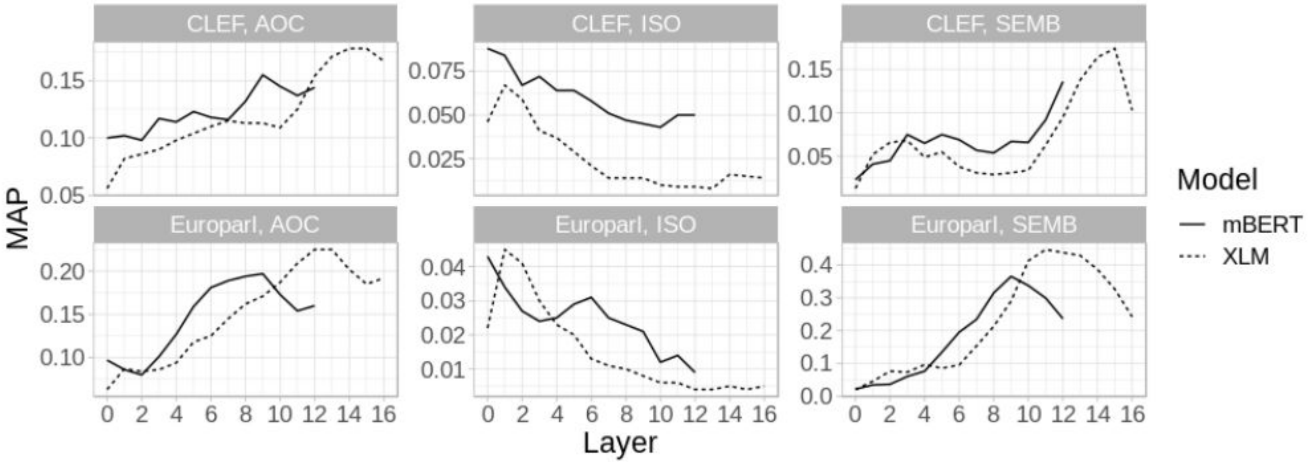
Discussion

Input Sequence-length

Length	$SEMB_{mBERT}$	$SEMB_{XLM}$	$DIST_{use}$	$DIST_{XLM\text{-}R}$	$DIST_{DmBERT}$	mUSE	LaBSE	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

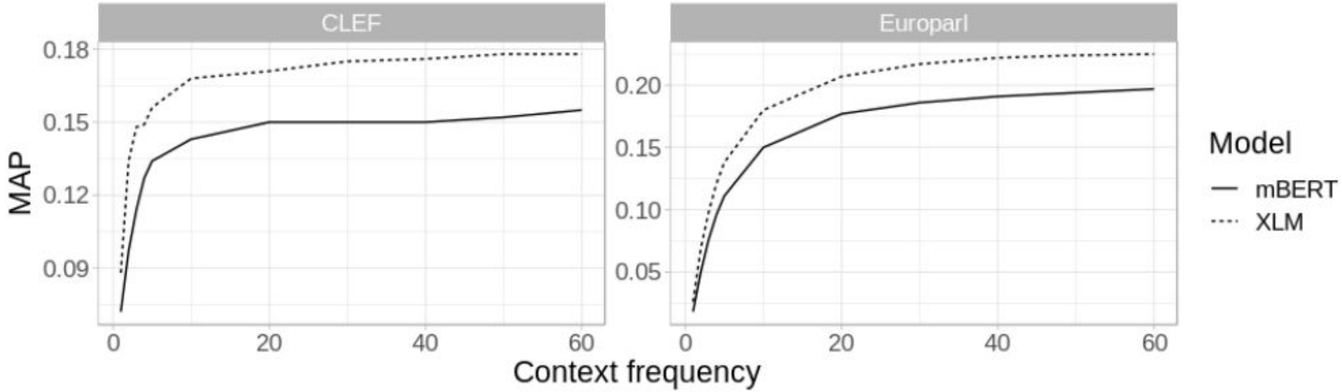
- Multilingual encoders effectively **truncate long documents**
- Encoding longer chunks of documents works slightly worse:
  - More difficult to encode **longer portion of text** semantically accurate
  - If relevance signal not within 128 tokens, it often does not appear beyond

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$_{mBERT}$** / **AOC$_{XLM}$**

Conclusion

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!

github.com/rlitschk/EncoderCLIR