Out of Thin Air: Is Zero-Shot Cross-Lingual Keyword Detection Better Than Unsupervised?

Boshko Koloski[1,2] Senja Pollak [1] Blaž Škrlj [1,2] Matej Martinc [1,2]

[1,2]Jožef Stefan Institute, Ljubljana, Slovenia [1,2]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia

Problem description

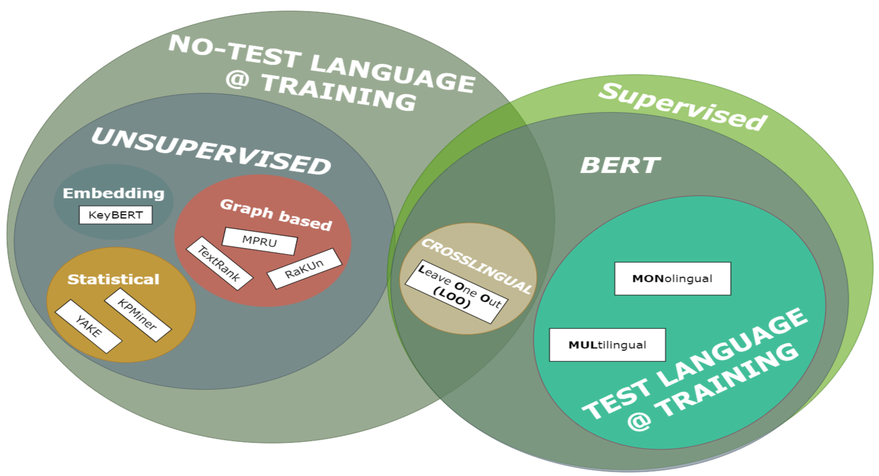
News writers use keywords to link similar articles or to summarize the content of an article.
The problem: While SoA approaches have high precision, they need to use manually annotated data, in order to learn to extract keywords. For some languages and domains there is a lack of learning data.
The solution: We investigate how cross-lingual transfer of keyword extraction performs against unsupervised approaches for such problems, when data from other languages or domains is available.

Data

	Train			Valid			Test		
Language	size	kw_per_doc	kw_present	size	kw_per_doc	kw_present	size	kw_per_doc	kw_present
Latvian	10506	3.2204	0.8691	2627	3.2687	0.8658	11641	3.1964	0.8624
Estonian	8600	3.8244	0.7809	2150	3.7386	0.7785	7747	4 944	0.8073
Slovenian	4796	4.0052	0.5991	1199	4.1643	0.6054	1519	3.8861	0.5995
Croatian	25778	3.5375	0.7047	6445	3.5469	0.6988	3582	3.5274	0.7009
English	207938	5.324	0.4599	51985	5.0350	0.4583	20000	5 349	0.6205
Russian	11064	5.6377	0.7779	2767.0	5.7311	0.7797	11475.0	5.4261	0.7918

Table: Number of documents (size), keywords per document (kw_per_doc) and percentage of keywords present in document's text (kw_present) per split in our experiments.

Methodology



Evaluation

Language		English	Slovenian	Croatian	Latvian	Estonian	Russian
Model	T	F1@10					
Without training data in the target language							
KPMiner	U	0.1584	0.0941	0.1043	0.131	0.0641	0.0578
YAKE	U	0.1449	0.0794	0.1248	0.095	0.0653	0.0966
KeyBert	U	0.1702	0.1153	0.1668	0.1330	0.0923	0.1352
TextRank	G	0.0440	0.0042	0.0041	0.0196	0.0239	0.0392
RaKUn	G	0.1176	0.0875	0.0902	0.0862	0.0605	0.0731
MPRU	G	0.1549	0.0455	0.0683	0.0821	0.0398	0.1171
LOO	C	**0.2856**	**0.2000**	**0.2883**	**0.2844**	**0.2368**	**0.2395**
With training data							
MON	S	0.4658	0.3259	0.4644	**0.6533**	**0.4920**	**0.5979**
MUL	S	**0.4702**	**0.3371**	**0.4674**	0.6532	0.4900	0.5943

Table: Performance of the models according to the F1@10 score.

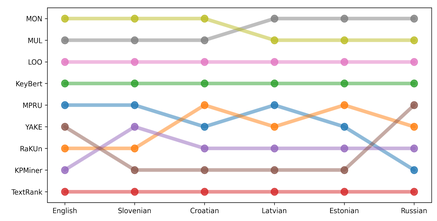


Figure: Ranking of model combinations per language.

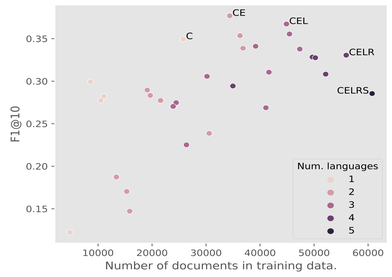


Figure: Performance of various cross-lingual combinations tested on English.

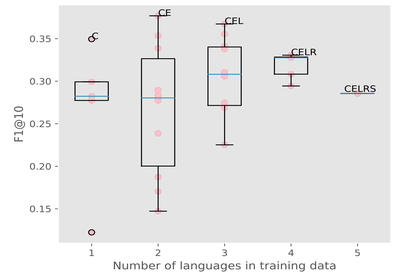


Figure: Correlation between the number of languages and the performance of the model.

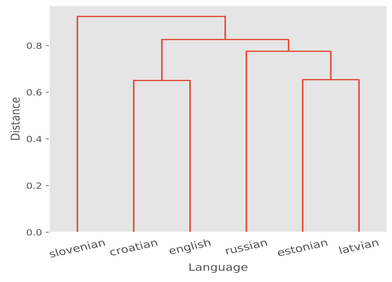


Figure: Dendrogram of the agglomerative clustering of the monolingual models.

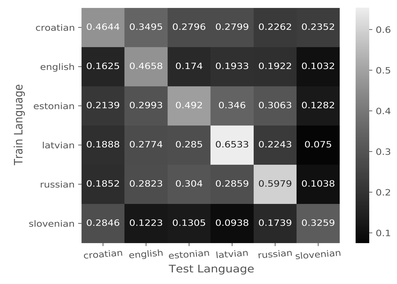


Figure: Monolingual zero-shot cross-lingual evaluation.

Conclusions

YES, Zero-Shot cross-lingual keyword detection **IS BETTER** than unsupervised keyword detection.