Learning Translational and Knowledge-based Similarities from Relevance Rankings for CLIR

Shigehiko Schamoni, Felix Hieber, Artem Sokolov, Stefan Riezler

Department of Computational Linguistics, Heidelberg University, Germany

Take Home Message

- Special domains contain structured information capturing cross-lingual relevance.
- Ranking models can be optimized on such cross-lingual relevance data.
- Combining orthogonal information from translation-specific and ranking-specific bilingual word associations outperforms state-of-the-art MT-based CLIR approaches.

Overview

Cross-Language Information Retrieval (CLIR) is the task of finding relevant information in a language different to the query language. Our system *intelligently combines* three complementary model types:

1. systems using *machine translation* and *monolingual retrieval* (MT + IR)
2. recent *word-based linear ranking* models that learn sparse word-correlations across languages
3. dense *domain knowledge models*

We show gains on two *new large-scale datasets*.

State-of-the-Art: MT + IR

Standard MT-based models translate a query and then perform monolingual retrieval, e.g. BM25.

- (**DT**) Direct translation: queries are translated sentence-wise at retrieval time.
- (**PSQ**) Probabilistic structured query:

$$score(E|F) = \sum_{f \in F} BM25(tf(f,E), df(f))$$
$$tf(f,E) = \sum_{e \in E_f} tf(e,E) p(e|f)$$
$$df(f) = \sum_{e \in E_f} df(e) p(e|f)$$

given a source query F, a document E and translation options $E_f = \{e \in E | p(e|f) > p_L\}$.

Word-based Linear Ranking

Let $\mathbf{q} \in \{0,1\}^Q$ be a query and $\mathbf{d} \in \{0,1\}^D$ be a document based on dictionaries of sizes Q and D. A linear ranking model is defined as

$$f(\mathbf{q},\mathbf{d}) = \mathbf{q}^\top W \mathbf{d} = \sum_{i=1}^{Q} \sum_{j=1}^{D} q_i W_{ij} d_j$$

where $W \in \mathbb{R}^{Q \times D}$ encodes a matrix of ranking-specific word associations.

Pairwise Ranking

Finds a weight matrix W such that the inequality $f(\mathbf{q},\mathbf{d}^+) > f(\mathbf{q},\mathbf{d}^-)$ is violated for the fewest number of tuples of a relevant $\mathbf{d}^+$ and an irrelevant $\mathbf{d}^-$ documents for a query $\mathbf{q}$.

- (**BM**) Boosting-based Ranking optimizes an exponential loss weighted by an importance function $\mathcal{D}(\mathbf{q},\mathbf{d}^+,\mathbf{d}^-)$:

$$\mathcal{L}_{exp} = \sum_{(\mathbf{q},\mathbf{d}^+,\mathbf{d}^-) \in \mathcal{R}} \mathcal{D}(\mathbf{q},\mathbf{d}^+,\mathbf{d}^-) e^{f(\mathbf{q},\mathbf{d}^-) - f(\mathbf{q},\mathbf{d}^+)}$$

- (**VW**) Online Stochastic Gradient Descent utilizes the Vowpal Wabbit toolkit optimizing an ℓ_1-regularized hinge loss:

$$\mathcal{L}_{hng} = \sum_{(\mathbf{q},\mathbf{d}^+,\mathbf{d}^-) \in \mathcal{R}} \left(f(\mathbf{q},\mathbf{d}^+) - f(\mathbf{q},\mathbf{d}^-) \right)_+ + \lambda ||W||_1$$

- Memory requirements are reduced by hashing.

Domain Knowledge Models

(**DK**) Domain knowledge models capture domain specific data characteristics:

- *Wikipedia*: features encode article lengths, common images, web links, etc. Intersection between two category sets S and T_n:

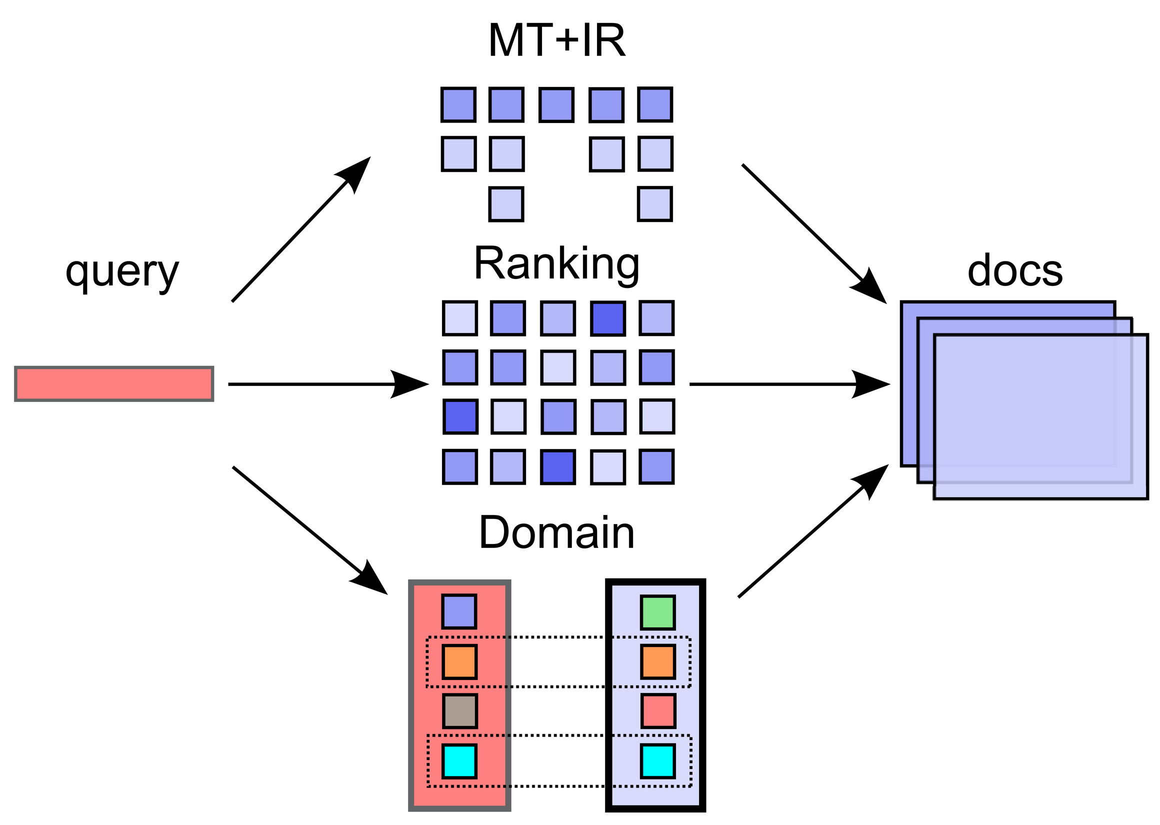
$$\text{score}_n = \frac{1}{2} \left(\frac{|S \cap T_n|}{|S|} + \frac{|S \cap T_n|}{|T_n|} \right)$$

- *Patents*: a feature fires if similar aspects are shared, e.g. common inventor, overlapping International Patent Class codes, etc.

	models	MAP	NDCG	PRES
Patents (JP-EN) *standalone*	DT	0.2554	0.5397	0.5680
	PSQ	0.2659	0.5508	0.5851
	DK	0.2203	0.4874	0.5171
	VW	0.2205	0.4989	0.4911
	BM	0.1669	0.4167	0.4665
Borda	DT+PSQ	0.2747	0.5618	0.5988
	DK+VW	0.3023	0.5980	0.6137
	(DT+PSQ)+(DK+VW)	0.3465	0.6420	0.6858
LinLearn	DT+PSQ	0.2707	0.5578	0.5941
	DK+VW	0.3283	0.6366	0.7104
	DT+PSQ+DK+VW	**0.3739**	**0.6755**	**0.7599**

	models	MAP	NDCG	PRES
Wikipedia (DE-EN) *standalone*	DT	0.3678	0.5691	0.7219
	PSQ	0.3642	0.5671	0.7165
	DK	0.2661	0.4584	0.6717
	VW	0.1249	0.3389	0.6466
	BM	0.1386	0.3418	0.6145
Borda	DT+PSQ	0.3742	0.5777	0.7306
	DK+VW	0.3238	0.5484	0.7736
	(DT+PSQ)+(DK+VW)	**0.4173**	0.6333	0.8031
LinLearn	DT+PSQ	0.3718	0.5751	0.7251
	DK+VW	0.3436	0.5686	0.7914
	DT+PSQ+DK+VW	0.4137	**0.6435**	**0.8233**

Model Combination



- *Borda Counts*: consensus-based voting procedure where a voter distributes a fixed amount of voting points. The aggregated ranking score for two rankings becomes:

$$f_{agg}(\mathbf{q},\mathbf{d}) = \kappa \frac{f_1(\mathbf{q},\mathbf{d})}{\sum_{\mathbf{d}} f_1(\mathbf{q},\mathbf{d})} + (1-\kappa) \frac{f_2(\mathbf{q},\mathbf{d})}{\sum_{\mathbf{d}} f_2(\mathbf{q},\mathbf{d})}$$

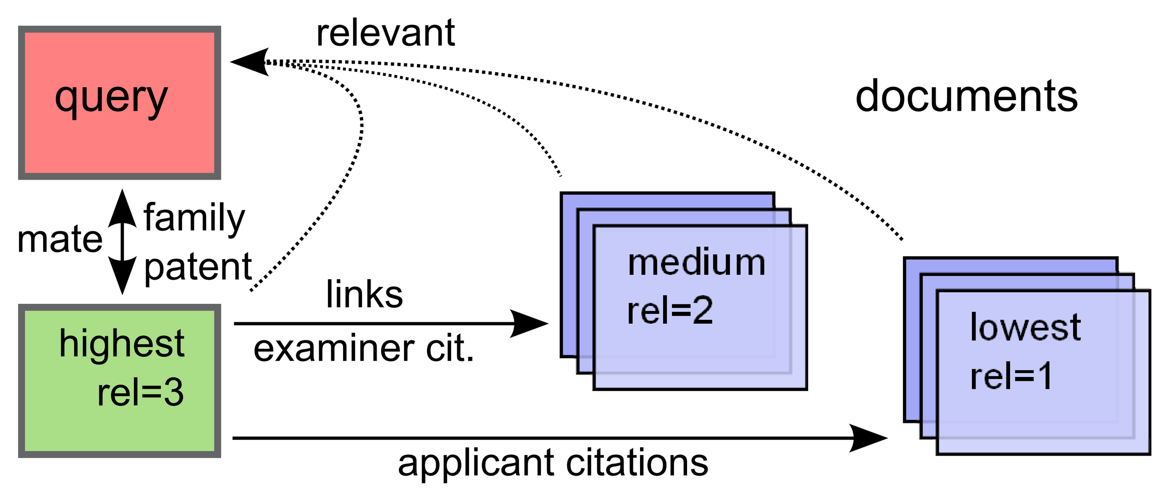
- *Linear Learning*: combination of MT + IR scores, word-based linear ranking scores, and domain knowledge features in a linear model trained with pairwise ranking.

Data

- Japanese-English Patent data (111k + 1,088k) `www.cl.uni-heidelberg.de/boostclir`
- German-English Wikipedia (245k + 1,455k) `www.cl.uni-heidelberg.de/wikiclir`

Tasks

We evaluate models and combinations on two real-world tasks for which data is constructed from cross-lingual patent and Wikipedia data:



- *Patent Prior Art Search*: a patent is relevant if there exists a family relationship (3), it is cited by the examiner (2) or by the applicant (1).
- *Wikipedia Article Retrieval*: an article is considered relevant if it is the cross-language counterpart *mate* (3), or if there exist bidirectional links to/from the mate (2).

In addition to standard preprocessing, correlated feature hashing is applied to ranking data.

Acknowledgements

This research was supported in part by DFG grant RI-2221/1-1 "Cross-language Learning-to-Rank for Patent Retrieval".

DFG

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386