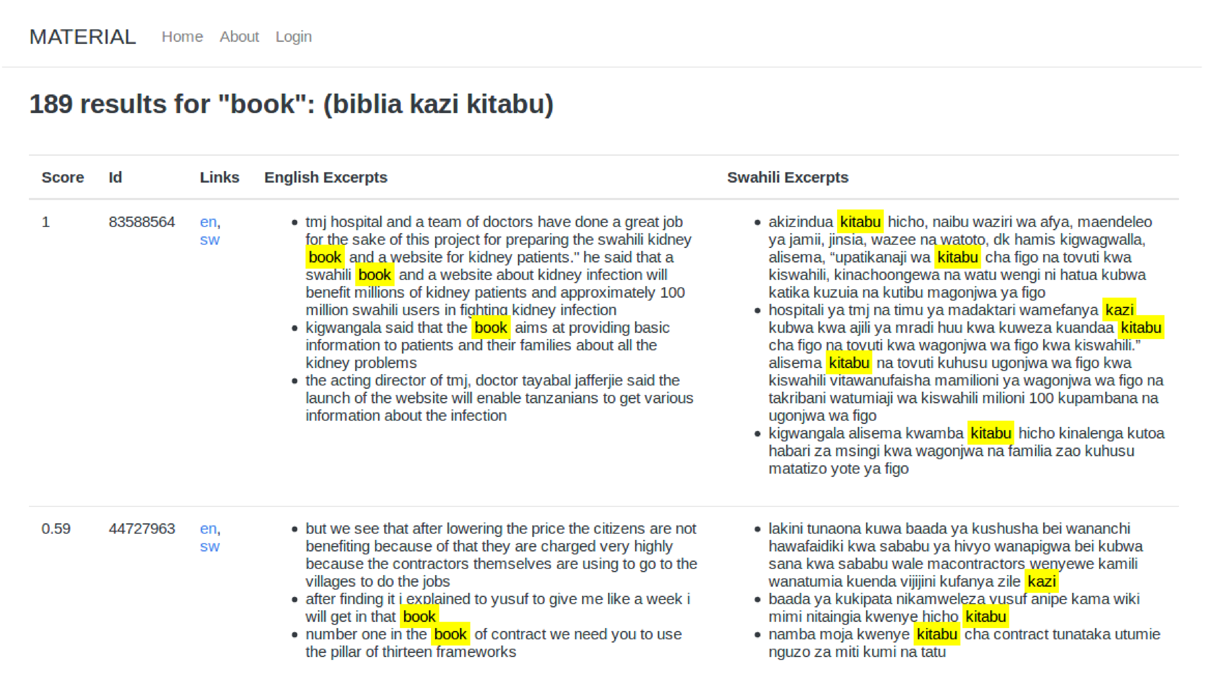
Yale

English Retrieval for other Languages

J. Dadashkarimi, R. Zhang, J. Kasai, W. T. Ting,
C. Westerfield, H. Taggarse, and D. Radev

Problem



1. Search English queries
2. Documents are indexed in other languages
3. The most common approaches are query translation and document translation
4. query translation by Dictionary and document translation by Machine Translation

QT and Dictionary

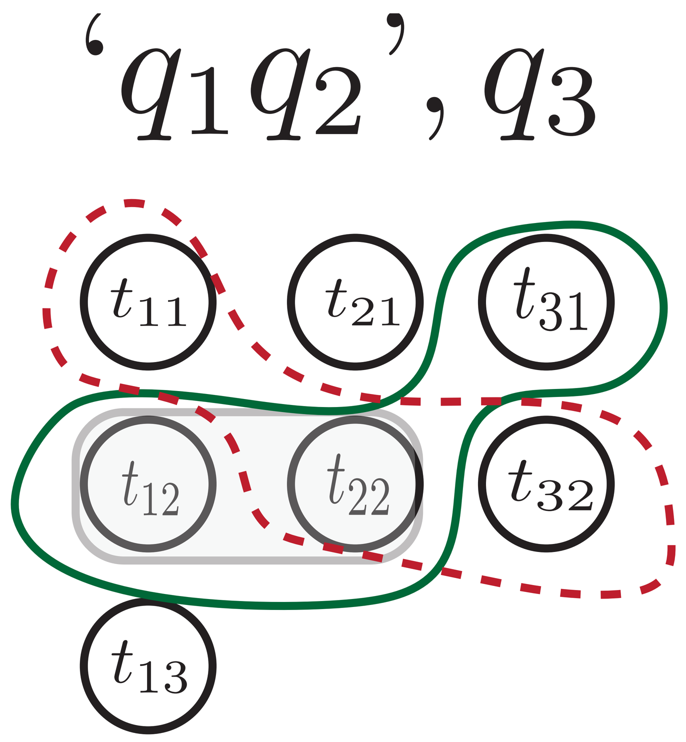


Fig. 1 The proposed phrase-based query translation. t_{ij} is the $j-$th translation for q_i. t_{12} and t_{22} are phrases in target language.

1. build all combinations and take the maximum
2. $\#\max([\#1(t_{11}, t_{22}), t_{32}], [\#1(t_{12}, t_{22}), t_{31}])$

DT and Proximity

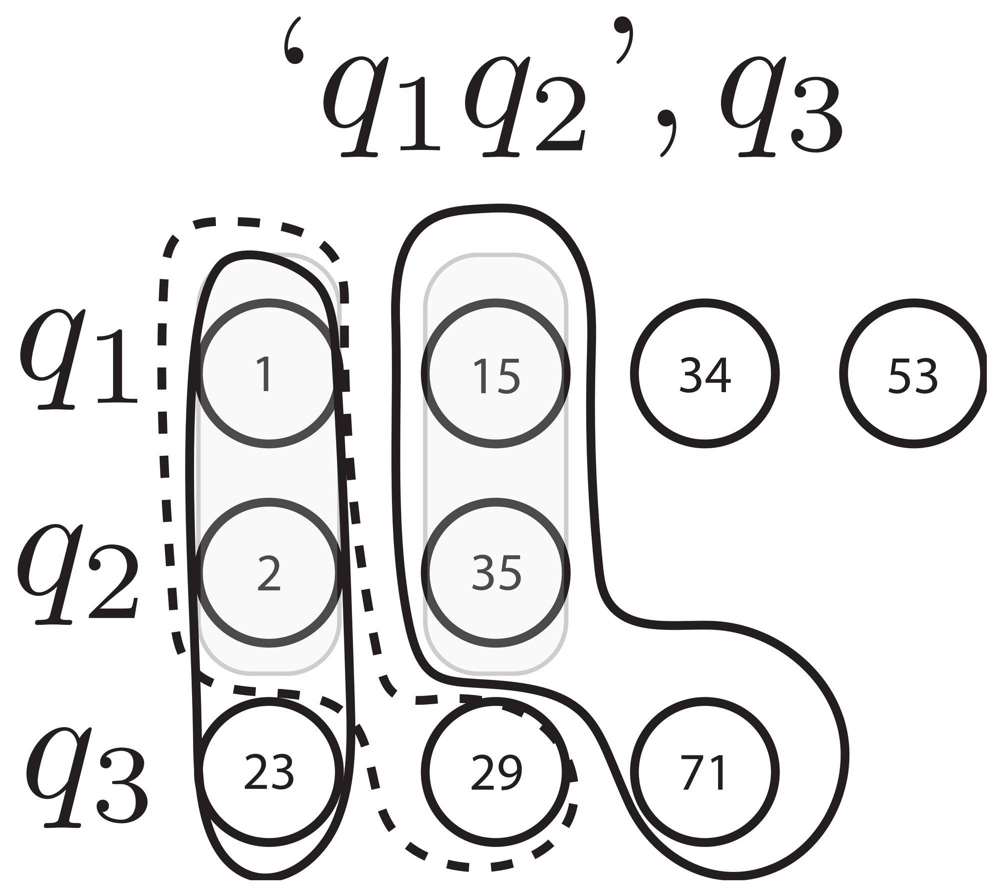


Fig. 2 The proposed query hit proximity approach. A query hit should have q_1 and q_2 adjacent to each other. Based on how far are the query terms are located in each document $f(q,d)$ gets lower score. This metric keeps the relative order of the query terms for matching.

1. Find different combinations of query terms' positions
2. Find query hits that keep the order of the query
3. take the maximum possible hits
4. $\#\max([1, 2, 23], [1, 2, 29], [15, 35, 71])$
5. we can also have partial matching where a part of the query appears in the document
6. then we can combine $p(q|d)$ based on language modeling:

$$s(d) = \alpha * p(q|d) + (1-\alpha) * f(q,d) \tag{1}$$

$$f(q,d) = -\log\left(\sum_{i=0}^{|q|} \|q_i, q_{i+1}\|_d\right) \tag{2}$$

Experimental Setup

1. We used Indri retrieval system
2. Europarl for CLEF collections and Bitext for surprise languages as training data for MT
3. We used hard cutoff of 20 for surprise languages and then ran a tuning from 1 to 20 to find a suitable cutoff for the collection
4. Wiktionary is used as a bilingual dictionary
5. Neural Machine Translation is used for document translation
6. Indri scores, word embedding, IDF, document length, and quey length are featues for soft cutoff
7. We used normalization and lower casing for all collections

Table 1. Collection Statistics

ID	Lang.	Collection	Queries	#docs	#qrels
SP	Spanish	EFE 1994	CLEF 2002, topics 91-140	215,738	2,854
de	German	Frankfurter Rundschau 94, SDA 94, Der Spiegel 94-95	CLEF 2002-03, topics 91-140	225,371	1,938
fr	French	Le Monde 94, SDA French 94-95	CLEF 2002-03, topics 251-350	129,806	3,524
sw	Swahili	Analysis	300 constrained queries	471	390
tl	Tagalog	Analysis	300 constrained queries	462	233

Experiments

1. Document translation get's consistently better results than Query translation
2. Phrase based retrieval get's as same result as synonym operator
3. Cutoff is very important for AQWV but not MAP
4. Wiktionary provides multiple choices for matching but MT provides only one option
5. System combination is a necessary step
6. We should consider variance around the best cutoff for each collection
7. Dictionary coverage is very important since a lot of queries have not translations

Table 2. Experimental Results

Lang	ANALYSIS/DEV	QT/DT	Phrase	Morph	Reranking	MAP	P@10	AQWV
SW	ANALYSIS	QT	no	no	no	0.213	0.0805	0.1025
SW	ANALYSIS	DT	no	no	no	0.3162	0.1085	0.1877
SW	ANALYSIS	QT	yes	no	no	0.212	0.0727	0.1313
SW	ANALYSIS	DT	yes	no	no	**0.3303**	**0.107**	**0.2606**
TL	ANALYSIS	QT	no	no	no	0.2469	0.0656	0.202
TL	ANALYSIS	DT	no	no	no	0.5285	0.1286	0.3389
TL	ANALYSIS	QT	yes	no	no	0.24	0.0641	0.2324
TL	ANALYSIS	DT	yes	no	no	**0.5621**	**0.1296**	**0.4282**