

# **ENGLISH RETRIEVAL FOR OTHER LANGUAGES** J. DADASHKARIMI, R. ZHANG, J. KASAI, W. T. TING, C. WESTERFIELD, H. TAGGARSE, AND D. RADEV

#### PROBLEM MATERIAL Home About Login 189 results for "book": (biblia kazi kitabu) Swahili Excerpts spital and a team of doctors have done a great job or the sake of this project for preparing the swahili kidney ya jamii, jinsia, wazee na watoto, dk hamis kigwagwalla. **book** and a website for kidney patients." he said that a alisema. "upatikanaii wa <mark>kitabu</mark> cha figo na tovuti kwa swahili **book** and a website about kidney infection will kiswahili, kinachoongewa na watu wengi ni hatua kubwa benefit millions of kidney patients and approximately 100 katika kuzuia na kutibu maqoniwa va fiq million swahili users in fighting kidney infection 🔹 hospitali ya tmi na timu ya madaktari wamefanya 🛛 🗛 kigwangala said that the book aims at providing basic kubwa kwa aiili ya mradi huu kwa kuweza kuandaa **kita**l formation to patients and their families about all the cha figo <u>na tovu</u>ti kwa wagonjwa wa figo kwa kiswahili lisema <mark>kitabu</mark> na tovuti kuhusu ugonjwa wa figo kwa kidnev problem kiswahili vitawanufaisha mamilioni ya wagonjwa wa figo na the acting director of tmj, doctor tayabal jafferjie said the launch of the website will enable tanzanians to get various akribani watumiaji wa kiswahili milioni 100 kupambana na information about the infection uqoniwa wa fiqo 🛛 kigwangala alisema kwamba <mark>kitabu</mark> hicho kinalenga kutoa abari za msingi kwa wagonjwa na familia zao kuhusu natatizo vote va fig benefiting because of that they are charged very highly awafaidiki kwa sababu ya hivyo wanapigwa bei kubwa sana kwa sababu wale macontractors wenyewe kamili because the contractors themselves are using to go to th illages to do the i wanatumia kuenda viiiiini kufanva zile <mark>kaz</mark> after finding it i explained to yusuf to give me like a week baada va kukipata nikamweleza vusuf anipe kama wil mimi nitaingia kwenye hicho **kital** will get in that bo number one in the book of contract we need you to use namba moja kwenye kitabu cha contract tunataka utumie the pillar of thirteen frameworks nguzo za miti kumi na tati

- 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

## 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



$$(q_1 q_2', q_3)$$
  
 $(t_{11}, t_{21}, t_{31})$   
 $(t_{12}, t_{22}, t_{32})$   
 $(t_{13})$ 

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

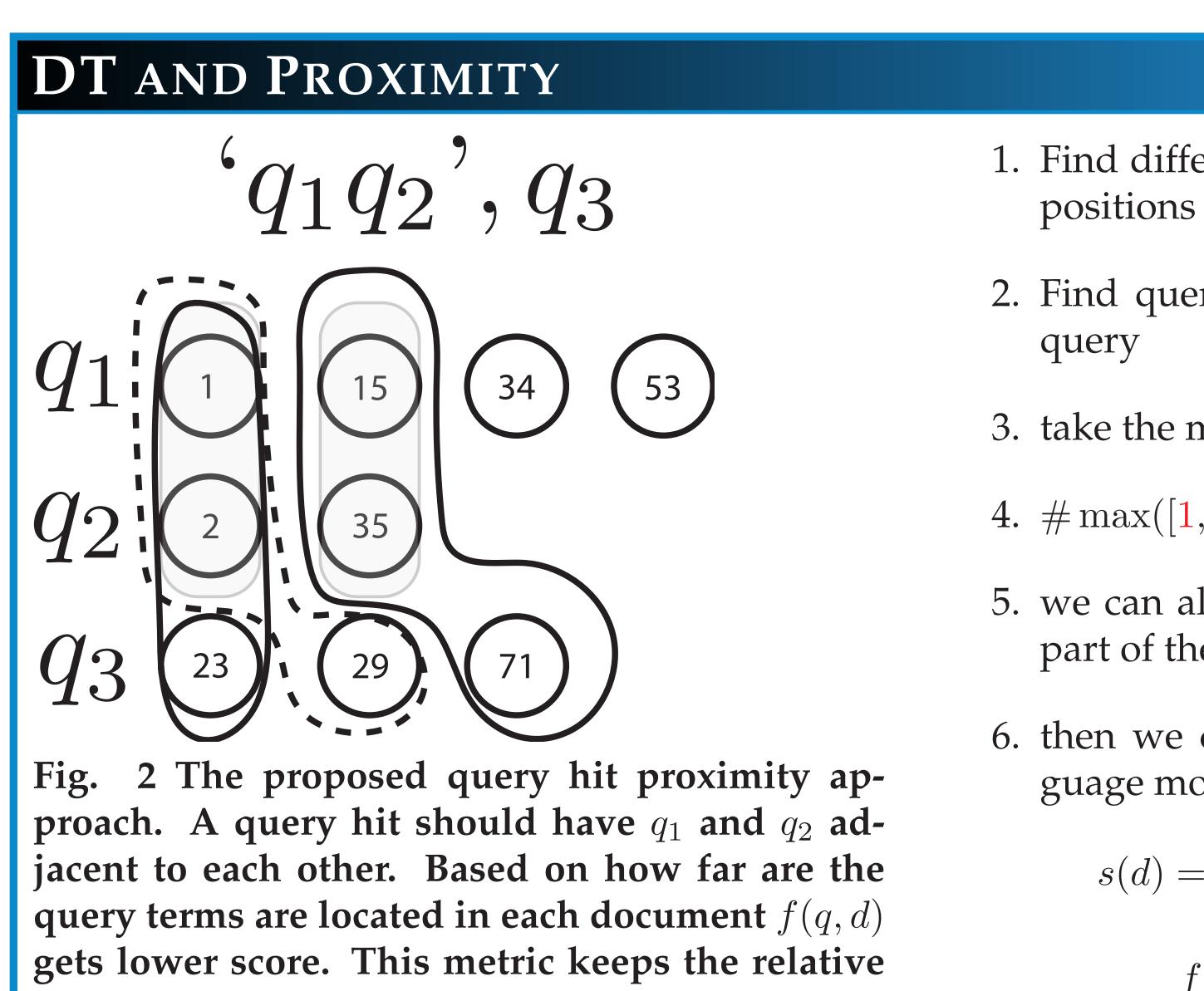
- . build all combinations and take the maximum
- 2.  $\# \max([\#1(t_{11}, t_{22}), t_{32}], [\#1(t_{12}, t_{22}), t_{31}])$

 $Q_3$ 

order of the query terms for matching.

F	X	P

- Lan SW SW SW SW TL

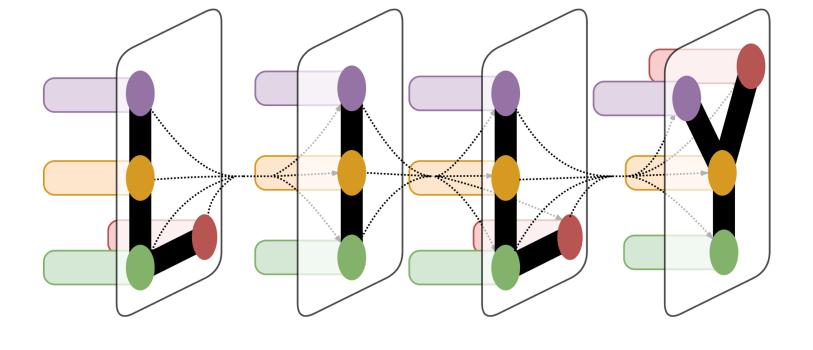


## PERIMENTS

- 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

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



1. Find different combinations of query terms'

2. Find query hits that keep the order of the

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:

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

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