End-to-End Neural CLIR by Sharing Representation

LILY Spring 2018 Workshop

Rui Zhang

Cross-lingual Information Retrieval (CLIR)

Information Retrieval

- Retrieve relevant documents from a corpus for a given user query.
- e.g., Google Search
- Usually monolingual, i.e., documents and queries are in the same language.
- TF-IDF, BM25

Cross-lingual Information Retrieval (CLIR)

- The documents are in a language different from that of the user's query.
- e.g., an investor wish to monitor the consumer sentiment from tweets around the world.

189 results for "book": (biblia kazi kitabu)

Score	Id	Links	English Excerpts	Swahili Excerpts
1	83588564	en, sw	• tmj hospital and a team of doctors have done a great job for the sake of this project for preparing the swahili kidney book and a website for kidney patients." he said that a swahili book and a website about kidney infection will benefit millions of kidney patients and approximately 100 million swahili users in fighting kidney infection • kigwangala said that the book aims at providing basic information to patients and their families about all the kidney problems • the acting director of tmj, doctor tayabal jafferjie said the launch of the website will enable tanzanians to get various information about the infection	• akizindua kitabu hicho, naibu waziri wa afya, maendeleo ya jamii, jinsia, wazee na watoto. dk hamis kigwagwalla, alisema, "upatikanaji wa kitabu cha figo na tovuti kwa kiswahili, kinachoongewa na watu wengi ni hatua kubwa katika kuzuia na kutibu magonjwa ya figo • hospitali ya tmj na timu ya madaktari wamefanya kazi kubwa kwa ajili ya mradi huu kwa kuweza kuandaa kitabu cha figo na tovuti kwa wagonjwa wa figo kwa kiswahili." alisema kitabu na tovuti kuhusu ugonjwa wa figo kwa kiswahili vitawanufaisha mamilioni ya wagonjwa wa figo na takribani watumiaji wa kiswahili milioni 100 kupambana na ugoniwa wa figo • kigwangala alisema kwamba kitabu hicho kinalenga kutoa habari za msingi kwa wagonjwa na familia zao kuhusu matatizo yote ya figo
0.59	44727963	en, sw	• but we see that after lowering the price the citizens are not benefiting because of that they are charged very highly because the contractors themselves are using to go to the villages to do the jobs • after finding it i explained to yusuf to give me like a week i will get in that book • number one in the book of contract we need you to use the pillar of thirteen frameworks	• lakini tunaona kuwa baada ya kushusha bei wananchi hawafaidiki kwa sababu ya hivyo wanapigwa bei kubwa sana kwa sababu wale macontractors wenyewe kamili wanatumia kuenda vijijini kufanya zile kazi • baada ya kukipata nikamweleza yusuf anipe kama wiki mimi nitaingia kwenye hicho kitabu • namba moja kwenye kitabu cha contract tunataka utumie nguzo za miti kumi na tatu
0.55	90947015	en, sw	• the religious leaders of linda district they are have not being allowed to preside over weeding without making it public for 21 days before so that if there is mistakes can be detected as well as controlling early marriage of students • the administrative secretary of lindi district, thomas safari said many times the islamic religion they use their own one book of certificate instead of the government and he ordered the government certificate and religious certificate to be used to remove disturbance	• mkuu wilaya ya lindi, shaibu ndemanga aliyasema hayo jana wakati alipozungumza na watendaji wa ngazi ya kata ya majengo na mtama, kwenye kikao cha kazi • katibu tawala wa wilaya ya lindi, thomas safari alisema mara nyingi dini ya kiislamu wanatumia kitabu kimoja cha cheti cha kwao badala ya serikali na kuagiza vitumike vyeti vya serikali na vile vya dini, ili kuondoa usumbufu

Methods for CLIR

Translation-based approach

- A pipeline of two components: translation + monolingual IR
- Can be further divided into document translation and query translation

e.g., the query is in English and documents are in Swahili

- Query translation from English to Swahili using a bilingual dictionary.
- Document translation from Swahili to English using a machine translation system.

Methods for CLIR

Translation-based approach is difficult

- Query Translation
  - rely on a comprehensive bilingual dictionary
  - Hard to translate short text queries and phrases
- Document Translation
  - Need to build a reliable machine translation system
- Especially for low-resource languages

Neural (Monolingual) Information Retrieval

Many successful neural IR systems have emerged:

- DUET (Mitra et al., 2017)
- PACRR (Hui et al., 2017)
- DSSM (Huang et al., 2013)
- DESM (Mitra et al., 2016)
- MatchPyramid (Pang et al., 2016)
- DRMM (Guo et al., 2016)

… ...

But, they are evaluated in Monolingual IR settings.

Research Goal and Challenges

Goal: Build an end-to-end neural CLIR that

- models local information
  - unigram term match
  - position-dependent information such as proximity and term positions.
- models global information
  - semantic matching in distributed representation space
- directly learns from (query,document,relevance) supervisions
- performs better than the pipeline translation-based approach because it avoids cascading errors

Research Goal and Challenges

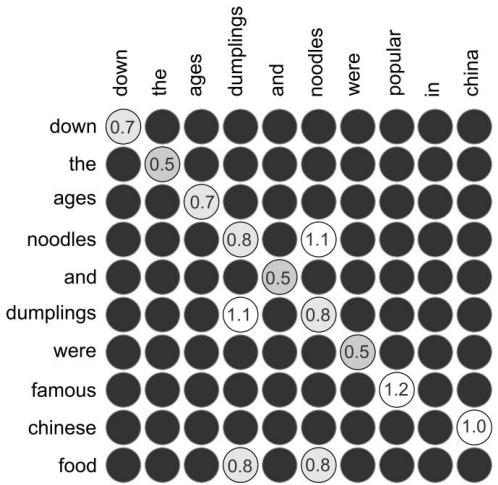
Challenges

- How can we capture local information and global information when query language and document language are different?
- How can we use and learn shared representation for multiple languages?

Proposed Method

1) Use multilingual word embeddings to build a similarity matrix.

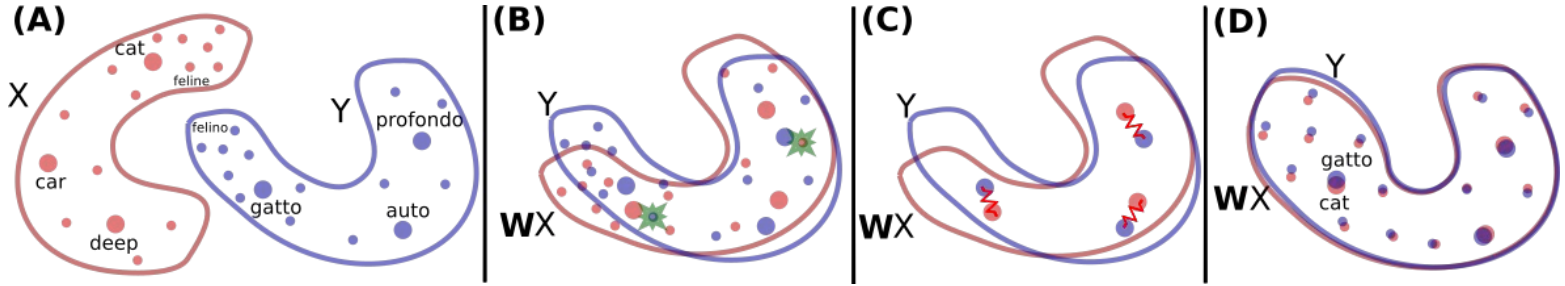
- This models local information.



(b) Dot Product

MatchPyramid (Pang et al., 2016)

Multilingual Word Embedding



Multilingual word Embeddings

We release fastText Wikipedia **supervised** word embeddings for **30** languages, aligned in a **single vector space**.

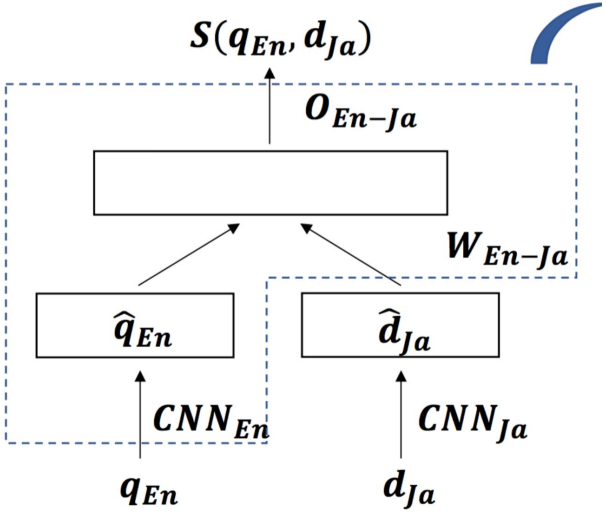
Arabic: *text*	Bulgarian: *text*	Catalan: *text*	Croatian: *text*	Czech: *text*	Danish: *text*
Dutch: *text*	English: *text*	Estonian: *text*	Finnish: *text*	French: *text*	German: *text*
Greek: *text*	Hebrew: *text*	Hungarian: *text*	Indonesian: *text*	Italian: *text*	Macedonian: *text*
Norwegian: *text*	Polish: *text*	Portuguese: *text*	Romanian: *text*	Russian: *text*	Slovak: *text*
Slovenian: *text*	Spanish: *text*	Swedish: *text*	Turkish: *text*	Ukrainian: *text*	Vietnamese: *text*

https://github.com/facebookresearch/MUSE

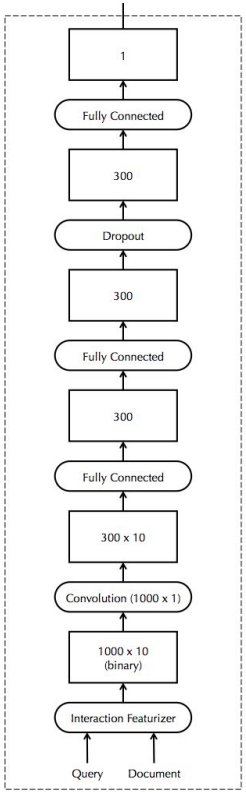
Proposed Method

2) Use monolingual or multilingual embedding to learn a shared distributed representation

- This models global information.

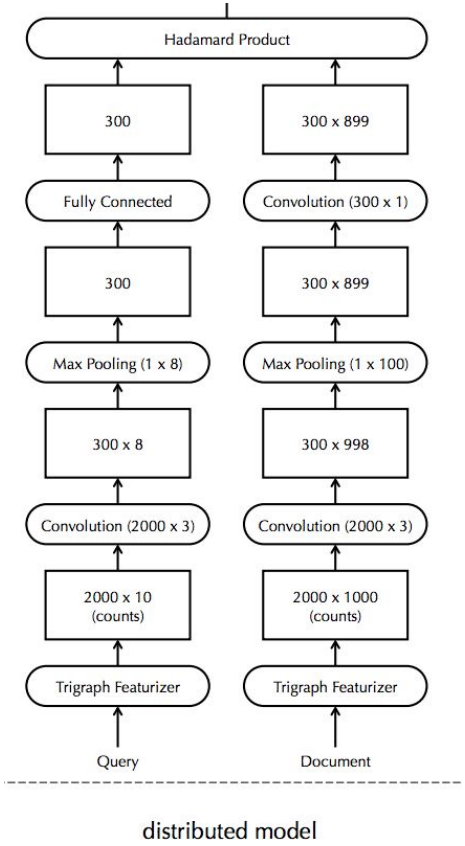


DUET for CLIR - Local Model



local model

DUET for CLIR - Global Model



Data Sets

WikiCLIR (Sasaki et al., 2018)

- Automatically created from parallel wiki pages
- Large-scale, 25 languages

Standard CLIR task

- CLEF
- NTCIR
- TREC

thank you
danke
謝謝
ngiyabonga
شكراً جزيلاً
рахмат
teşekkür ederim
gracias
merci
obrigado
dziękuję
bedankt
спасибо
Баярлалаа
go raibh maith agat
mochchakkeram
sukriya
kop khun krap
terima kasih
감사합니다
ありがとう
ευχαριστώ
arigatô
dakujem
grazie
takk
taiku
sagolun
tapadh leat
dank je
mersi
barka
welalin
tack
spas
kia ora
faafetai lava
vinaka
спасибі
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
tenki
murakoze
kiitos
dankie
nanni
nandri
dhanyavad
maururuu
hvala
köszönöm
gracje
bayarlalaa
enkosi
dankon
aciú
akun
djiere dieuf
tau
дякую
mamnun
chokrane
chnorakaloutioun
gratias ago
gràcies
sulpáy
sobodi
děkuji
mèsi
didi madloba
kam sah hamnida
najis tuke
rahmat
tanemirt
rahmet
diolch
dhanyavadagalu
shukriya
mercê
мерси
trugarez
xièxie
তোমাকে ধন্যবাদ