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.
<table>
<thead>
<tr>
<th>Score</th>
<th>Id</th>
<th>Links</th>
<th>English Excerpts</th>
<th>Swahili Excerpts</th>
</tr>
</thead>
</table>
| 1     | 83588864 | en, sw | • tmnj 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 kidney problems.  
• the acting director of tmnj, doctor taye bel jaffierie said the launch of the website will enable tanzanians to get various information about the infection. | • akizindua kitabu hicho, neibu waziri wa afya, msandele ya jamii, jinsia, wasawa na waroto, dix hamis kigwagwala, aliasema, “upatikanaji wa kitabu cha figo na tovuti cha kiswahili, kinachoongewa na watu wengi ni hatua kubwa katika kuzuzuzwa na kubu kuponya wa figo.”  
• kiswahili aliye kuwasaidia vitaweniilisha mamili ya wagonjwa wa figo na takribani watumiaji wa kiwara hili milioni 100 kupa bana na ugonjwa wa figo.  
• kigwangala alisema kwamba kitabu hicho kina lengan maarifa habari za magizi kwa wagonjwa na familia zao kuwahusu mafaka inayotolea yote wa figo. |
| 0.59  | 44727963 | en, sw | • but we see that after lowering the price the citizens are no longer benefiting from that because they are charged very highly specifically, because the contractors themselves are using to go to the villages to do the jobs.  
• after finding it i explained to yousif 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 kwa baada ya kushusha bel wananchi hawafaidiki kwa sababu ya hivyo wapi kwa sababu sana kwa sababu wale miamine wamejulikana umbaji wa wanamitaka kushughula vijiji kwa fikishwa zile kazi.  
• baada ya kukipata nikameleza vugufu anipipe kama waki mimi nitinga kwenye hicho kitabu  
• nambo moja kwenye kitabu cha contract tunatake utumili nguzo za miti kuni na tatu. |
| 0.55  | 90947015 | en, sw | • the religious leaders of linda district they have not been 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 linda 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 wiliaya ya lindi, shabu ndemana aliyasema hayo jana wakati aliwainishua na watendasi wa ngazi wa kata ya majengo na mtama, kwenye kikao cha kazi.  
• katibu tawala wa wiliaya ya lindi, thomas safari alisema mada nyotingi dini, wa kisabuni wanatuma kitabu kioko cha cheki cha kwao badala ya seerkali na kuangiza vitumike vyeti uya seerkali 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.
Multilingual Word Embedding

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

<table>
<thead>
<tr>
<th>Arabic: text</th>
<th>Bulgarian: text</th>
<th>Catalan: text</th>
<th>Croatian: text</th>
<th>Czech: text</th>
<th>Danish: text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greek: text</td>
<td>Hebrew: text</td>
<td>Hungarian: text</td>
<td>Indonesian: text</td>
<td>Italian: text</td>
<td>Macedonian: text</td>
</tr>
<tr>
<td>Norwegian: text</td>
<td>Polish: text</td>
<td>Portuguese: text</td>
<td>Romanian: text</td>
<td>Russian: text</td>
<td>Slovak: text</td>
</tr>
<tr>
<td>Slovenian: text</td>
<td>Spanish: text</td>
<td>Swedish: text</td>
<td>Turkish: text</td>
<td>Ukrainian: text</td>
<td>Vietnamese: text</td>
</tr>
</tbody>
</table>

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
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