CROSS-LANGUAGE INFORMATION RETRIEVAL (CLIR)

James Mayfield
April 12, 2021
CMSC 476/676
Outline

Introduction
CLIR Evaluation
Attributes of Non-English Text
Crossing the Language Barrier
Other Techniques
Conclusions
Our Problem
Huge document collection
Must use automated methods

Must infer information need behind analyst’s query
Goal: find documents that satisfy information need
Analyst speaks English
Must span language barrier

Documents written in foreign language (e.g., Chinese)

I need to know about recent developments in the Southeast Asian textile industry
An information need is not a query
Approach 1

Learn Chinese

水下编织篮

Formulate Chinese query

Use Chinese query to find documents

Read documents in Chinese
Recognised as a highly labour-intensive industry, the textile industry has been gaining ground in Southeast Asian countries. Partly being mostly agrarian adds to the advantage as textile

When preparing a loom for underwater use, it is crucial to hold your breath as long as possible. This will allow the shuttle to protrude significantly beyond the

Approach 2

Translate every document into English

Formulate English query

Read documents in English

Use English query to find translated documents
Approach 3

This is Cross-Language Information Retrieval (CLIR)

Read documents in English

Use English query to find Chinese documents

Formulate English query

Textile industry

Recognised as a highly labour-intensive industry, the textile industry has been gaining ground in Southeast Asian countries. Partly being mostly agrarian adds to the advantage as textile

When preparing a loom for underwater use, it is crucial to hold your breath as long as possible. This will allow the shuttle to protrude significantly beyond the
Search & Multilinguality

- Official Languages
  - EU: 23, India: 22, UN: 6, Switzerland: 4, Belgium: 3

- National Security
  - DoD National Language Service Corps: Chinese, Hausa, Hindi, Indonesian, Marshallese, Russian, Somali, Swahili, Thai, and Vietnamese

- E-Commerce
  - “To reach 80% of the world's Internet users, a Web site needs to support a minimum of 10 languages” – Byte Level Research, 2007
  - “One-fourth of Hispanics must be served in Spanish if retailers want their business.” - Forrester Research, 2008
## Social Security Information in Other Languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Sign Language</td>
<td>Italiano – Italian</td>
</tr>
<tr>
<td>العربية – Arabic</td>
<td>한국어 – Korean</td>
</tr>
<tr>
<td>Հայերեն – Armenian</td>
<td>Polski – Polish</td>
</tr>
<tr>
<td>中文 – Chinese (Traditional/Long Form)</td>
<td>Português – Portuguese</td>
</tr>
<tr>
<td>فارسی – Farsi</td>
<td>Русский – Russian</td>
</tr>
<tr>
<td>Français – French</td>
<td>Af Soomaali – Somali</td>
</tr>
<tr>
<td>Ελληνικά – Greek</td>
<td>Español–Spanish</td>
</tr>
<tr>
<td>Kreyòl Ayisyen – Haitian Creole</td>
<td>Tagalog</td>
</tr>
<tr>
<td>Hmong</td>
<td>Tiếng Việt – Vietnamese</td>
</tr>
</tbody>
</table>

## Related Information

- 2015 Social Security Administration Language Access Plan
- Free Interpreter Services
- Your Payments While You Are Outside The United States
- Publication No. 05–10137
  (English, French, German, Greek, Italian, & Spanish)

If you had difficulty receiving services due to a language barrier issue, please contact the Regional Communications Director for your state.
Outline

Introduction
CLIR Evaluation
Attributes of Non-English Text
Crossing the Language Barrier
Other Techniques
Conclusions
Evaluation of CLIR Search Quality

- CLIR at Text REtrieval Conference (TREC)
  - Spanish and Chinese monolingual, bilingual (TREC 4-6)
  - French, German, & Italian bilingual, multilingual (TREC 6-8)
  - Chinese (TREC-9)
  - Arabic (TREC 2001 & TREC 2002)
  - No CLIR at TREC since 2003

http://trec.nist.gov/
Cross-Language Evaluation Forum

- Patterned after TREC
- Focus on European languages
  - Bulgarian, Czech, Dutch, English, Finnish, French, German, Hungarian, Italian, Portuguese, Russian, Spanish, Swedish (added Farsi in 2008)
- Tasks
  - Monolingual & Bilingual Retrieval
  - Cross-Language Spoken Document Retrieval
  - Human-interactive CLIR
  - Question Answering
  - Web Retrieval
  - Cross-Language Image Search

http://www.clef-campaign.org/
## CLEF Ad Hoc Test Sets (2000 – 2007)

<table>
<thead>
<tr>
<th>Language (ID)</th>
<th>#docs</th>
<th>size</th>
<th>00</th>
<th>01</th>
<th>02</th>
<th>03</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>07</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian (BG)</td>
<td>69 k</td>
<td>213 MB</td>
<td>49</td>
<td>50</td>
<td>50</td>
<td>49</td>
<td>50</td>
<td>50</td>
<td>149</td>
<td></td>
</tr>
<tr>
<td>Czech (CS)</td>
<td>82 k</td>
<td>178 MB</td>
<td>50</td>
<td>50</td>
<td>56</td>
<td>50</td>
<td>49</td>
<td>50</td>
<td>156</td>
<td></td>
</tr>
<tr>
<td>Dutch (NL)</td>
<td>190 k</td>
<td>540 MB</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>156</td>
<td></td>
</tr>
<tr>
<td>English (EN)</td>
<td>170 k</td>
<td>580 MB</td>
<td>34</td>
<td>47</td>
<td>42</td>
<td>54</td>
<td>42</td>
<td>50</td>
<td>49</td>
<td>367</td>
</tr>
<tr>
<td>Finnish (FI)</td>
<td>55 k</td>
<td>137 MB</td>
<td>30</td>
<td>45</td>
<td>45</td>
<td>50</td>
<td>49</td>
<td>50</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>French (FR)</td>
<td>178 k</td>
<td>470 MB</td>
<td>34</td>
<td>49</td>
<td>50</td>
<td>52</td>
<td>49</td>
<td>50</td>
<td>49</td>
<td>333</td>
</tr>
<tr>
<td>German (DE)</td>
<td>295 k</td>
<td>660 MB</td>
<td>37</td>
<td>49</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>192</td>
<td></td>
</tr>
<tr>
<td>Hungarian (HU)</td>
<td>50 k</td>
<td>105 MB</td>
<td>50</td>
<td>48</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>148</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian (IT)</td>
<td>157 k</td>
<td>363 MB</td>
<td>34</td>
<td>47</td>
<td>49</td>
<td>51</td>
<td>51</td>
<td>51</td>
<td>181</td>
<td></td>
</tr>
<tr>
<td>Portuguese (PT)</td>
<td>107 k</td>
<td>340 MB</td>
<td>46</td>
<td>50</td>
<td>50</td>
<td>46</td>
<td>50</td>
<td>50</td>
<td>146</td>
<td></td>
</tr>
<tr>
<td>Russian (RU)</td>
<td>17 k</td>
<td>68 MB</td>
<td>28</td>
<td>34</td>
<td>34</td>
<td>28</td>
<td>34</td>
<td>34</td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>Spanish (ES)</td>
<td>453 k</td>
<td>1086 MB</td>
<td>49</td>
<td>50</td>
<td>57</td>
<td>49</td>
<td>50</td>
<td>57</td>
<td>156</td>
<td></td>
</tr>
<tr>
<td>Swedish (SV)</td>
<td>143 k</td>
<td>352 MB</td>
<td>49</td>
<td>53</td>
<td>53</td>
<td>49</td>
<td>53</td>
<td>53</td>
<td>102</td>
<td></td>
</tr>
</tbody>
</table>
TREC Spin-offs

- Europe (CLEF)
  - 2000 – present
- Japan (NTCIR)
  - 1999 - present
- India (FIRE)
  - 2008 - present
- Russia (ROMIP)
  - 2003 - 2014
Outline

Introduction
CLIR Evaluation
Attributes of Non-English Text
  Characters
  Words
  Subwords
Crossing the Language Barrier
Other Techniques
Conclusions
Characters
Unicode

- Universal set of code points
- Most common encoding: UTF-8
- Features and issues
  - Normalization
  - Look-alike characters
  - Parallel code blocks
- Handy tool: https://shapecatcher.com/
Other Encodings

There are scores of encodings beyond UTF-8 that can still be found on the Internet

- UTF-16, UTF-32 – Unicode encodings
- ASCII
- ISO8859-1 (Latin-1) – ASCII variants
  - -2 through -16
- EBCDIC – IBM mainframes
- CP-437 – IBM PC
  - -720 through -822
- Windows-1252 – Windows encodings
  - -1250-1258
- MacOS Roman

- GBT-2312 – Simplified Chinese
  - GBK, GB-18030
- Big5 – Traditional Chinese
- JIS X-0208 – Japanese
  - JIS X-0213
- KS X-1001 – Korean
  - EUC-KR
  - ISO 2022-KR
Writing Systems

● The world’s languages are written in many different scripts
● Some languages use different scripts for different words
  ○ Japanese: Kanji, Katakana, Hiragana, Romanji
● Some languages are even written in multiple writing systems (Digraphia)
  ○ Serbian: Cyrillic, Latin
● Many languages that use writing systems other than Latin have transliterations into Latin script
  ○ Chinese: Pinyin
● Transliteration into Latin characters often necessitated by lack of keyboards for other writing systems
### Number of Speakers Worldwide by Script

<table>
<thead>
<tr>
<th>Name</th>
<th>Active Speakers (millions)</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latin</td>
<td>4,900</td>
<td>English, Spanish, French, Portuguese, Romanian, etc.</td>
</tr>
<tr>
<td>Chinese</td>
<td>1,340</td>
<td>Chinese, Japanese (Kanji), Korean (Hanja), etc.</td>
</tr>
<tr>
<td>Arabic</td>
<td>660</td>
<td>Arabic, Persian, Urdu, Punjabi, Pashto, etc.</td>
</tr>
<tr>
<td>Devanagari</td>
<td>608</td>
<td>Hindi, Marathi, Konkani, Nepali, Sanskrit, etc.</td>
</tr>
<tr>
<td>Bengali</td>
<td>265</td>
<td>Assamese, Bengali, Bishnupriya Manipuri, Meitei Manipuri</td>
</tr>
<tr>
<td>Cyrillic</td>
<td>250</td>
<td>Bulgarian, Russian, Serbian, Ukrainian, etc.</td>
</tr>
<tr>
<td>Kana</td>
<td>120</td>
<td>Japanese, Okinawan, Ainu</td>
</tr>
<tr>
<td>Javanese</td>
<td>80</td>
<td>Javanese</td>
</tr>
<tr>
<td>Hangul</td>
<td>79</td>
<td>Korean</td>
</tr>
<tr>
<td>Telugu</td>
<td>74</td>
<td>Telugu</td>
</tr>
</tbody>
</table>

<en.wikipedia.org/wiki/List_of_writing_systems>
Segmentation

● (At least) three levels of segmentation:
  ○ Sentence segmentation: where are the sentence boundaries?
  ○ Word segmentation: where are the word boundaries?
  ○ Morphological segmentation: where are the morphemes?
Sentence Boundary Detection

- Some languages have unambiguous end-of-sentence markers
  - E.g., Chinese full stop
- Dr. Mulholland of Mulholland Dr. says “In other langs., sentence segmentation is not so easy.” Dr. Mulholland is right.
  - Six periods, two sentences
- Two main approaches:
  - Rule-based
  - Machine learning
Word Boundary Detection

- In some languages (e.g., English), blank space and punctuation are strong predictors of word boundaries.
- In others (e.g., Chinese), words are simply run together without breaks.
- Main approaches:
  - Rule-based
  - Machine learning
  - Ignore problem through use of subwords
Morphological Segmentation

- Goal: identify morphological components of a word
- Handling morphology is critical for avoiding OOV in morphologically complex languages
- Morfessor: statistical approach
  - Mines large text collection
  - Identifies most likely break points
Morphological Processes

- **Abbreviations:** BTW, FYI, w/o, Dr.
- **Acronyms:** NASA, MIT, IBM
- **Blending:** breakfast/lunch ➔ brunch; turducken
- **Borrowing:** ombrelli (umbrella), quiche, kindergarten
- **Clipping:** professor ➔ prof; gymnasium ➔ gym
- **Compounding:** airport, girlfriend, father-in-law
- **Conjugation:** swim/swims/swam/swum
- **Contractions:** do not ➔ don’t
- **Declension:** I/me/my/mine
- **Derivation:** compute(v), computer(n); boy(n), boyish(adj)
- **Doubling:** bye-bye; night-night
- **Inflection:** number or gender: fox+es; act+or/act+ress
- **Military:** Pacific/Command ➔ PACOM (clipping + compounding)
- **Miscellaneous:** H2O, i18n (internationalization)
- **Texting:** 4 (for), CUL8R, RUOK
Stemming

- Applicable to alphabetic languages
- An approximation to lemmatization
- Identify a root morpheme by chopping off prefixes and suffixes

Most stemmers are rule-based:
- ing => ε juggling => juggl
- es => ε juggles => juggl
- le => -l juggle => juggl

The Snowball project provides high quality, rule-based stemmers for many European languages

http://snowball.tartarus.org/
SUBWORDS
Subword Representations of Language

- Use pieces of words for indexing
- Two main flavors
  - Character N-Grams
  - Byte Pair Encoding (BPE)
- Advantages
  - Counteracts data sparseness
  - Reduces OOVs (out-of-vocabulary)
- Disadvantages
  - Larger indexes
  - Doesn’t play well with word-based processes
Character N-Grams

- Represent text as overlapping substrings of n characters
- Fixed length of n of 4 or 5 is effective in alphabetic languages
- For text of length m, there are m-n+1 n-grams

Advantages:
- simple
- address morphology
- surrogate for short phrases
- robust against spelling & diacritical errors
- Language-independent

Disadvantages:
- conflation (e.g., simmer, polymers)
- speed and disk usage penalties
Tokenization Comparison

- **Words**
  - Straightforward for most languages
  - Generally produce poor performance

- **Stems**
  - Effective in Romance languages
  - Not always available

- **Character N-grams**
  - Language-neutral
  - Large performance gains in complex languages
Source of N-gram Power

- Idea: remove morphology
- Letter order of words was randomly permuted (consistently)
  - golfer -> legfro, team-> eamt
  - golfing, golfer, golfed no longer share a morpheme
Byte Pair Encoding (BPE)

- Originally a compression technique

```python
vocab = {letters}
while (|vocab| < TARGET_SIZE)
    Form a new token T by concatenating most common token pair
    vocab = vocab U {T}
```

- In some applications, this allows words never seen before (Out Of Vocabulary, or OOV) to be processed appropriately
WordPiece Tokenization

- BERT uses WordPiece tokenization
  - Based on BPE: Start with alphabet, merge until desired number of tokens achieved
  - New tokens may not cross word boundaries
  - English BERT has a vocabulary of 30,000 tokens
  - Multilingual BERT has a vocabulary of 119,547 tokens

- WordPiece Algorithm

```python
code
vocab = {letters}
while (\|vocab\| < TARGET_SIZE):
    Use training data to create language_model(vocab)
    Form a new token T by concatenating the pair of tokens to that maximizes
    the likelihood of training data when added to the language model
    break if likelihood increase < threshold
vocab = vocab U {T}
```
WordPiece Tokenization cont.

- Special tokens for sentence prediction objective
  - [CLS] Beginning of sentence(s)
  - [SEP] End of each sentence
  - [CLS] i’ve had a perfectly wonderful evening [SEP] but this wasn’t it [SEP]

- Example: embeddings => [em ##bed ##ding ##s]
  - The double pound sign means that the previous token is part of the same word

- Word embeddings
  - WordPiece embeddings do not encode most complete words
  - Two approaches:
    - Average vectors for component word pieces
    - Use just first or last subword
SentencePiece Tokenization

- Open source analog to WordPiece
- Does not require prior word segmentation
- Available from [https://github.com/google/sentencepiece](https://github.com/google/sentencepiece)
- Example
  - "L'appartement est grand & vraiment bien situé en plein centre"
  - "_L" """"app" "ar" "tement" "est" "grand" 
  - "&" "v" "r" "ai" "ment" "bien" "situe" "en"
  - "plein" "centre"
Question 1: In which direction should we cross the language barrier?

- **Translate the documents**
  - Pro: Provides lots of context to get accurate word translations
  - Con: Translating millions of documents is time-consuming and computationally expensive

- **Translate the queries**
  - Con: Not much context in query itself
  - Pro: Might have other information about user that assists in translation
  - Pro: Translation is fast (per query)

- **Translate both to an interlingua**
  - Con: More translation required
  - Pro: Interlingua might better support retrieval than human language
  - Pro: Supports multi-way CLIR
Crossing the Language Barrier

Question 2: How should we cross the language barrier?

- Do nothing
- Transliteration
- Machine Translation
- Dictionary Lookup
- Multilingual Embeddings
- Pivoting
- End-to-End Retrieval
Crossing the Language Barrier

Do Nothing

- Sometimes called *cognate matching*
- Buckley et al., 1997: French is misspelled English
  - Applied spelling correction to convert English query to French, then used monolingual retrieval
  - Outperformed many systems at TREC-6
- McNamee & Mayfield 2002: Dutch is English
  - Character n-gram tokenization
  - CLEF-2001 English documents, non-English queries
Transliteration is mapping from the characters of one script to those of a different script in a way that preserves sounds.

Greek word: Ελευθερία
- Transliteration: Eleutheria
- Translation: Freedom

Names are often transliterated rather than translated when mapping to a different language.

Several approaches to transliteration:
- Rule-based (usually hand-coded)
- Grapheme-based translation
- Phoneme-based translation
- Alignment
Crossing the Language Barrier
Machine Translation

- Most straightforward approach to CLIR
- Radical improvement in machine translation over past three years
  - But much of the gain from using neural approaches comes from improved fluency
  - Not clear how improved fluency can help IR
  - Correlation between machine translation performance and retrieval performance has been inconsistent
Crossing the Language Barrier
Dictionary Lookup

- Word-by-word machine translation
- Keys to success
  - Comprehensive dictionary
    - Matches domain of query
  - Method to select translation(s)
  - Query augmentation
Two Types of Dictionary

- Manually-generated
  - Commercial dictionaries expensive (~$10K / language pair)
  - Unclear how to pick the right word(s) from possible translations

- Corpus-based (MT translation tables)
  - In-domain aligned Parallel Corpora are uncommon
  - Translation results may be biased by domain of source text
Corpus-based Translation

Given aligned parallel texts and a particular term to translate:

- Find set of documents (sentences) in the source language containing the term
- Examine corresponding foreign documents
- Extract ‘good’ candidate translation(s)
- Goodness can be based on term similarity measures (Dice, MI, IBM Model 1, etc.)

The Rosetta Stone was discovered in 1799 by Napoleonic forces in Egypt. British physicist Thomas Young determined that cartouches were names of royalty. In 1821 Jean François Champollion began deciphering hieroglyphics using parallel data in Demotic and Greek
## Sample Corpus-based Translations

<table>
<thead>
<tr>
<th>poisson</th>
<th>pêche</th>
<th>eaux</th>
<th>islandais</th>
<th>cee</th>
<th>baisse</th>
</tr>
</thead>
<tbody>
<tr>
<td>fish</td>
<td>fishing</td>
<td>waters</td>
<td>iceland</td>
<td>eec</td>
<td>decline</td>
</tr>
<tr>
<td>freshwater</td>
<td>fisheries</td>
<td>water</td>
<td>icelandic</td>
<td>programme</td>
<td>drop</td>
</tr>
<tr>
<td>fishermen</td>
<td>fishermen</td>
<td>sewage</td>
<td>denmark</td>
<td>european</td>
<td>prices</td>
</tr>
<tr>
<td>fishing</td>
<td>fishery</td>
<td>pollution</td>
<td>norway</td>
<td>nations</td>
<td>price</td>
</tr>
</tbody>
</table>
Issues in Dictionary-Based CLIR

“The main problems associated with dictionary-based CLIR are (1) untranslatable search keys due to the limitations of general dictionaries, (2) the processing of inflected words, (3) phrase identification and translation, and (4) lexical ambiguity in source and target languages.” - Pirkola et al.

Subwords can help two (and a half) of these:
- Out-of-Vocabulary words (OOV)
- Morphological Variation
- (Surrogate) Phrase Translation

Corpus-based translation can be applied to character n-grams!

- ‘work’ (from *working*) maps to ‘abaj’ (as in *trabajaba*)
- ‘yrup’ (from *syrup*) maps to ‘rabe’ (as in *jarabe*)
- ‘therl’ (from *Netherlands*) to ‘ses b’ (as in *Paises Bajos*)

<table>
<thead>
<tr>
<th></th>
<th>German</th>
<th>Italian</th>
<th>French</th>
<th>Dutch</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td>milch</td>
<td>latte</td>
<td>lait</td>
<td>melk</td>
</tr>
<tr>
<td><strong>Stem</strong></td>
<td>milch</td>
<td>latt</td>
<td>lait</td>
<td>melk</td>
</tr>
<tr>
<td><strong>4-grams</strong></td>
<td>milc</td>
<td>latt</td>
<td>lait</td>
<td>melk</td>
</tr>
<tr>
<td></td>
<td>ilch</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>5-grams</strong></td>
<td>_milc</td>
<td>_latt</td>
<td>_lait</td>
<td>_melk</td>
</tr>
<tr>
<td></td>
<td>milch</td>
<td>latte</td>
<td>lait_</td>
<td>melk_</td>
</tr>
<tr>
<td></td>
<td>ilch_</td>
<td>atte_</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Advantages of Character N-gram Translation

- Almost no such thing as an OOV n-gram
- Quality of alignments more important than corpus size
- Less data sparseness
- With 5% of Europarl n-grams outperform words with any amount of (Europarl) parallel data
## CLEF Bilingual English to X

<table>
<thead>
<tr>
<th>Language</th>
<th>Acquis Corpus</th>
<th>Europarl Corpus</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>words</td>
<td>stems</td>
<td>5-grams</td>
</tr>
<tr>
<td>BG</td>
<td>0.0591</td>
<td>x</td>
<td>0.0898</td>
</tr>
<tr>
<td>CS</td>
<td>0.1107</td>
<td>x</td>
<td>0.2479</td>
</tr>
<tr>
<td>DE</td>
<td>0.1802</td>
<td>0.2097</td>
<td>0.2952</td>
</tr>
<tr>
<td>ES</td>
<td>0.2583</td>
<td>0.3072</td>
<td>0.3661</td>
</tr>
<tr>
<td>FI</td>
<td>0.1286</td>
<td>0.1755</td>
<td>0.3552</td>
</tr>
<tr>
<td>FR</td>
<td>0.2508</td>
<td>0.2733</td>
<td>0.3013</td>
</tr>
<tr>
<td>HU</td>
<td>0.1087</td>
<td>x</td>
<td>0.2224</td>
</tr>
<tr>
<td>IT</td>
<td>0.2365</td>
<td>0.2656</td>
<td>0.2920</td>
</tr>
<tr>
<td>NL</td>
<td>0.2474</td>
<td>0.2249</td>
<td>0.3060</td>
</tr>
<tr>
<td>PT</td>
<td>0.2009</td>
<td>x</td>
<td>0.2544</td>
</tr>
<tr>
<td>SV</td>
<td>0.2111</td>
<td>0.2270</td>
<td>0.3016</td>
</tr>
<tr>
<td>PMAP</td>
<td>0.1811</td>
<td></td>
<td>0.2756</td>
</tr>
<tr>
<td>% change</td>
<td>63.5%</td>
<td></td>
<td>64.0%</td>
</tr>
<tr>
<td>PMAP-7</td>
<td>0.2161</td>
<td>0.2405</td>
<td>0.3168</td>
</tr>
<tr>
<td>% change</td>
<td>13.1%</td>
<td></td>
<td>13.1%</td>
</tr>
</tbody>
</table>
Crossing the Language Barrier

Multilingual Embeddings

- Embeddings: placement of indexing tokens in high (300-1000) dimensional vector space
- Preserves relationships among terms
- Often called CLWEs (cross-language word embeddings) or CLEs (cross-language embeddings)
- Commonly evaluated on bilingual lexicon induction
- Can identify possible translations using approximate nearest neighbors algorithms
- Embeddings can be static (e.g., Word2Vec or GloVe) or contextual (e.g., BERT)
Two Forms of Multilingual Embedding

Shared embedding space
- Supervised using sentence-aligned corpora
- Supervised using document-aligned corpora
- Pseudo-mixing
  - Embeddings built from documents where some words have been replaced by translations

Embedding space alignment
- Unsupervised
- Shared term-based alignment
  - E.g., identical strings, cognates, numerals
- Dictionary-based alignment
Crossing the Language Barrier

Pivoting

- Jump from source to target language through third “pivot” language
- Useful for low resource languages
- Can use different techniques for the two jumps
- Typically:
  - First language pair is high resource (e.g., English/Russian)
  - Second language pair comprises closely-related languages (e.g., Russian/Ukrainian)
- Or, pivot through English
  - Often, English/Language1 and English/Language2 resources are readily available, where Language1/Language2 resources are not
In end-to-end retrieval, the system is trained directly on query-document training pairs
- Monolingually, the MS MARCO datasets have served this purpose
- A key barrier to training end-to-end neural CLIR systems is a lack of such query-document training pairs
- Large-Scale CLIR Dataset
  - Uses 2.8M first sentence of Wikipedia articles as queries
  - Automated relevance judgments in 25 languages
  - [http://cs.jhu.edu/~kevinduh/a/wikiclr2018/](http://cs.jhu.edu/~kevinduh/a/wikiclr2018/)
Pre- and Post-Translation Expansion

- Pre-translation expansion: add new terms to query before translating it
- Post-translation expansion: add new terms to query after translating it
- X-axis: Reduction in size of translation dictionary
- Y-axis: Performance

Translation as Synonymy

White House

(blanca OR clara)  (casa)

\[ \frac{1}{3} \left[ \frac{TF_1}{DF_1} + \frac{TF_2}{DF_2} + \frac{TF_3}{DF_3} \right] \]

- **Unbalanced:**
  Overweights query terms that have many translations

\[ \frac{1}{2} \left[ \frac{TF_1 + TF_2}{DF_1 \cup DF_2} + \frac{TF_3}{DF_3} \right] \]

- **Structured:**
  Deemphasizes query terms with any common translation
Probabilistic Structured Queries

- Many possible translations, learned from parallel text
- Each with an estimated translation probability
- Term frequency and document frequency of query term $e$ computed using term frequency and document frequency of its translations:

\[
\begin{align*}
TF(e, D_k) &= \sum_{f_i} p(e | f_i) \times TF(f_i, D_k) \\
DF(e) &= \sum_{f_i} p(e | f_i) \times DF(f_i)
\end{align*}
\]
Outline

Introduction
CLIR Evaluation
Attributes of Non-English Text
Crossing the Language Barrier
Other Techniques
Conclusions
Paul McNamee’s List of Foundational CLIR Literature

- **Translate Documents or Queries**
  - McCarley, ‘Should we Translate the Documents or the Queries in Cross-Language Information Retrieval’, ACL-99

- **Translation Ambiguity**
  - Gollins and Sanderson, ‘Improving Cross-Language Retrieval with Triangulated Translation’, SIGIR-01
  - Wang and Oard, ‘Combining bidirectional translation and synonymy for cross-language information retrieval’, SIGIR-06
Paul McNamee’s List of Foundational CLIR Literature cont.

- **Query Expansion and CLIR**
  

- **Poor Translation Resources**
  
  Demner-Fushman and Oard, ‘The Effect of Bilingual Term List Size on Dictionary-Based Cross-Language Information Retrieval’, HICSS-03
  
  McNameee and Mayfield, ‘Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources’, SIGIR-02
Thank you

Questions?