CROSS-LANGUAGE INFORMATION RETRIEVAL (CLIR)

James Mayfield
May 2, 2022
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

An information need is not a query

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

Approach 1

Learn Chinese

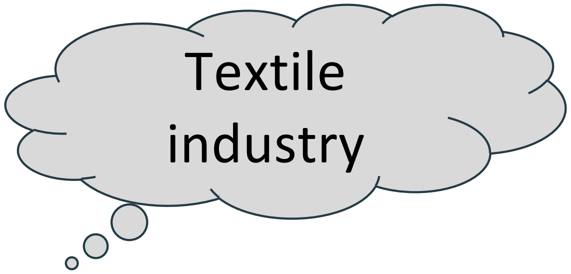
水下编织篮

Formulate Chinese query

Use Chinese query to find documents

Read documents in Chinese

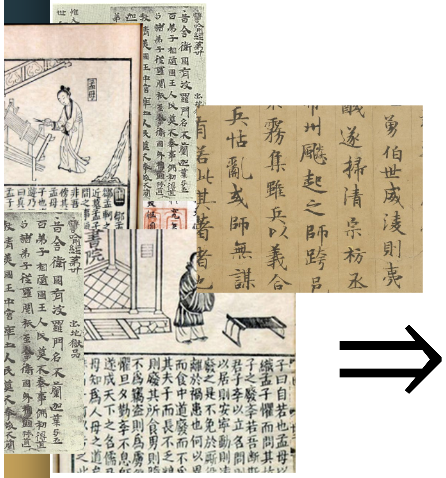
Approach 2



Translate every document into English

Formulate English query

Read documents in English



⇒

When preparing

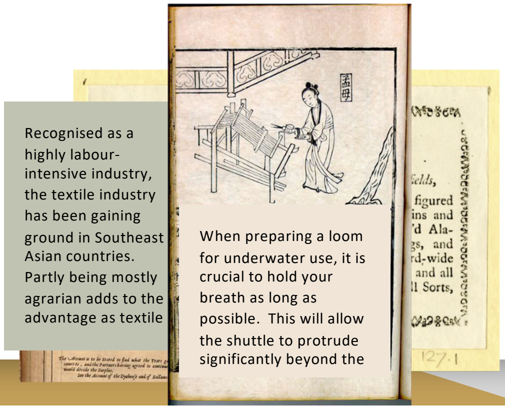
Re
hi
in
th
ha
gr
As
Partly
agraria
advanta

COMMON SENSE;
ADDRESSED TO THE
INHABITANTS
OF
AMERICA,

PROPOSALS
For PRINTING by SUBSCRIPTION,
A FREE and IMPARTIAL WEEKLY
News Paper,
To be entitled—THE
PENNSYLVANIA LEIDGER,
Or the MARYLAND, PENNSYLVANIA, and NEW-JERSEY
WEEKLY ADVERTISER;

Her Majeſties moſt Gracious
LARAT

Use English query to find translated documents



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

figured
ins and
'd Ala-
gs, and
rd-wide
and all
ll Sorts,

Approach 3

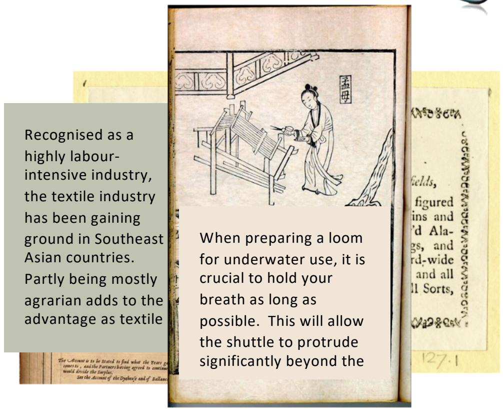
This is Cross-Language Information Retrieval (CLIR)

Use English query to find Chinese documents

Textile industry

Formulate English query

Read documents in English



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

Translate retrieved Chinese documents into English

Search & Multilinguality

Many slides in this presentation stolen from Paul McNamee (A handful of which were previously stolen from me)

- 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

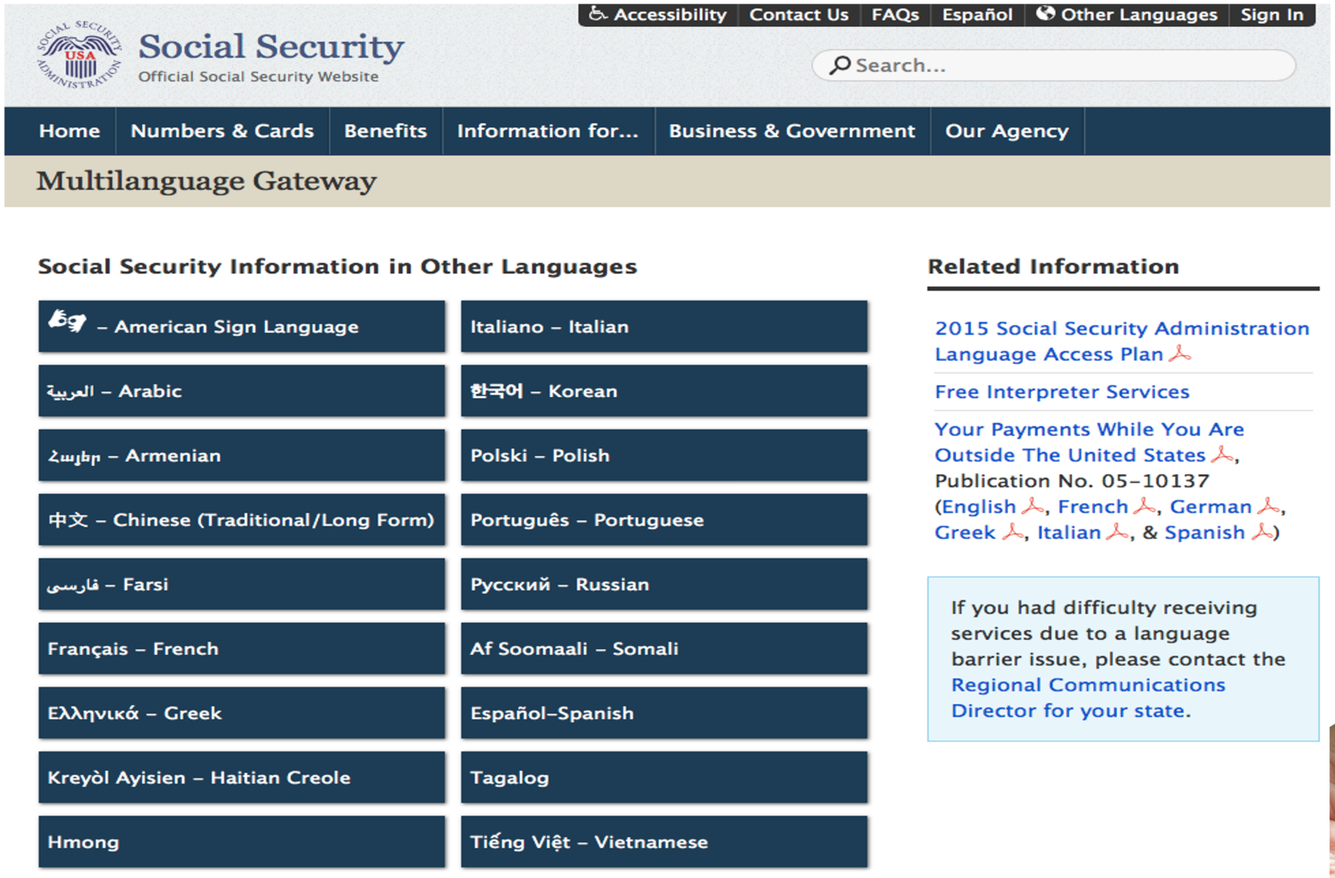
NATIONAL LANGUAGE SERVICE CORPS

Language for the good of all.®

WWW.NLSCORPS.ORG

- 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

Serving All Beneficiaries



Multilanguage Gateway

Social Security Information in Other Languages

– American Sign Language	Italiano – Italian
العربية – Arabic	한국어 – Korean
Հայերր – Armenian	Polski – Polish
中文 – Chinese (Traditional/Long Form)	Português – Portuguese
فارسی – Farsi	Русский – Russian
Français – French	Af Soomaali – Somali
Ελληνικά – Greek	Español–Spanish
Kreyòl Ayisien – Haitian Creole	Tagalog
Hmong	Tiếng Việt – Vietnamese

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

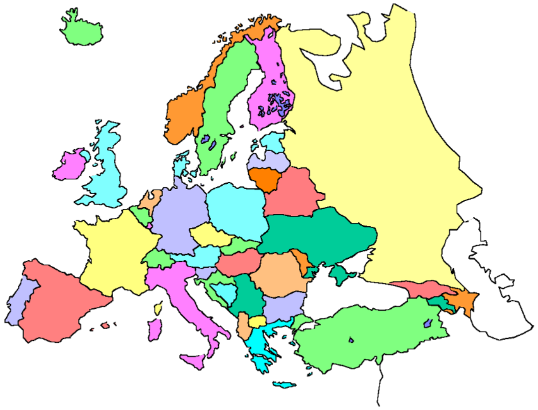
Evaluation of CLIR Search Quality

- 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 2003-2021
  - New at TREC 2022: NeuCLIR track

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)

	#docs	size	00	01	02	03	04	05	06	07	
Bulgarian (BG)	69 k	213 MB						49	50	50	149
Czech (CS)	82 k	178 MB								50	50
Dutch (NL)	190 k	540 MB		50	50	56					156
English (EN)	170 k	580 MB	33	47	42	54	42	50	49	50	367
Finnish (FI)	55 k	137 MB			30	45	45				120
French (FR)	178 k	470 MB	34	49	50	52	49	50	49		333
German (DE)	295 k	660 MB	37	49	50	56					192
Hungarian (HU)	50 k	105 MB						50	48	50	148
Italian (IT)	157 k	363 MB	34	47	49	51					181
Portuguese (PT)	107 k	340 MB					46	50	50		146
Russian (RU)	17 k	68 MB				28	34				62
Spanish (ES)	453 k	1086 MB		49	50	57					156
Swedish (SV)	143 k	352 MB			49	53					102

TREC Spin-offs

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

CLEF INITIATIVE

The CLEF Initiative

Conference and Labs of the Evaluation Forum

NTCIR (NII Testbeds and Community for Information access Research) Project

NTCIR

Forum for Information Retrieval Evaluation
(FIRE)

5th - 7th December 2014
Indian Statistical Institute, Bangalore

РОМИП

СЕМИНАР

TREC 2022 NeuCLIR Track

- English Queries
- Chinese, Russian, and Persian Documents
- neuclir.github.io/
- Easy-to-use baseline system: “Patapsco”
  - Provides basic CLIR with evaluation
  - github.com/hltcoe/patapsco
- Call for participation: trec.nist.gov/pubs/call2022.html

Outline

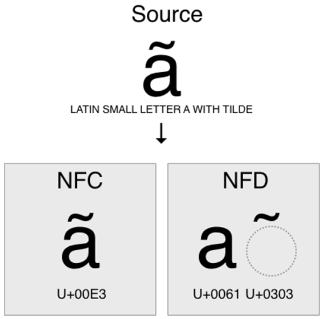
- Introduction
- CLIR Evaluation
- ➡ Attributes of Non-English Text
  - Characters
  - Words
  - Subwords
- Crossing the Language Barrier
- Other Techniques
- Conclusions

Characters

Characters, Code Points, Glyphs and Encodings

- Upper-case-A, lower-case-e and dollar-sign are *characters* (abstract atomic data elements)
- *Code Points* are integers that represent characters
  - ASCII values are code points
- *Unicode* is a particular standard mapping from code points to characters
  - Unicode is a superset of ASCII
- *Glyphs* are graphical representations of characters
  - A, **A**, 𝒜, and 𝓐 are different glyphs for an upper-case-A
- An *encoding* is a way to map a sequence of code points onto a sequence of bytes (suitable for storage on disk, for example)
  - UTF-8 is a common encoding of Unicode

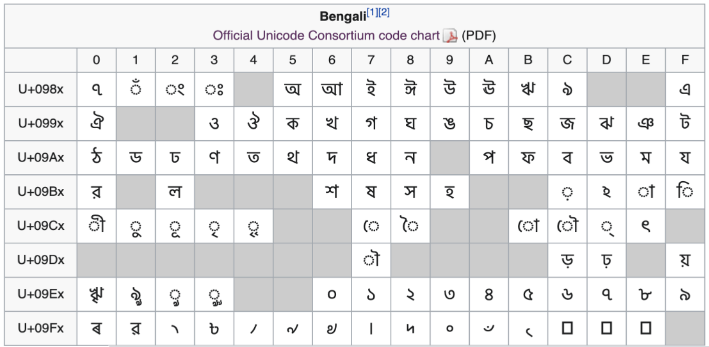
Unicode



- Universal set of code points
- Most common encoding: UTF-8
- Features and issues
  - Normalization
  - Look-alike characters
  - Parallel code blocks
- Handy tools:

  apps.timwhitlock.info/unicode/inspect

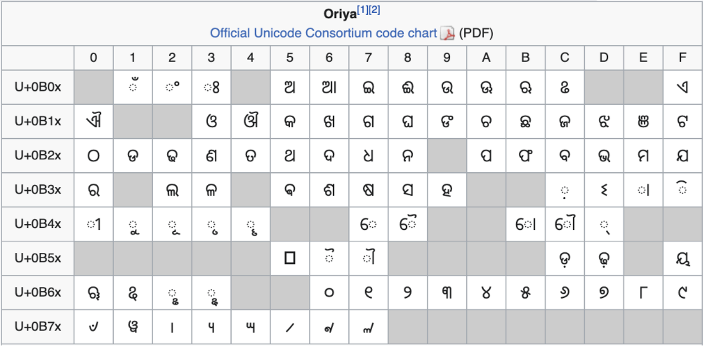
  shapecatcher.com/



Bengali[1][2]

Official Unicode Consortium code chart (PDF)

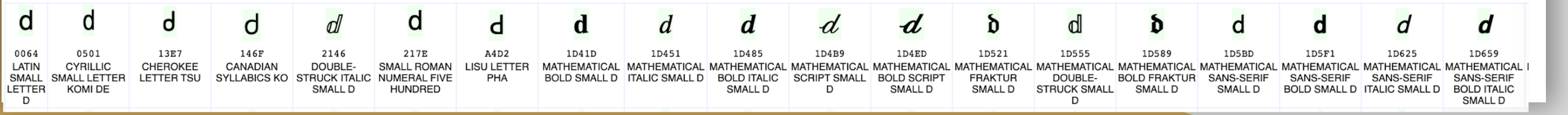
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+098x	ঀ	◌ঁ	◌ং	◌ঃ		অ	আ	ই	ঈ	উ	ঊ	ঋ	ঌ			এ
U+099x	ঐ			ও	ঔ	ক	খ	গ	ঘ	ঙ	চ	ছ	জ	ঝ	ঞ	ট
U+09Ax	ঠ	ড	ঢ	ণ	ত	থ	দ	ধ	ন		প	ফ	ব	ভ	ম	য
U+09Bx	র		ল				শ	ষ	স	হ			◌়	ঽ	◌া	◌ি
U+09Cx	◌ী	◌ু	◌ূ	◌ৃ	◌ৄ			◌ে	◌ৈ			◌ো	◌ৌ	◌্	ৎ	
U+09Dx								◌ৗ					ড়	ঢ়		য়
U+09Ex	ৠ	ৡ	◌ৢ	◌ৣ			০	১	২	৩	৪	৫	৬	৭	৮	৯
U+09Fx	ৰ	ৱ	৲	৳	৴	৵	৶	৷	৸	৹	৺	৻	□	□	□	



Oriya[1][2]

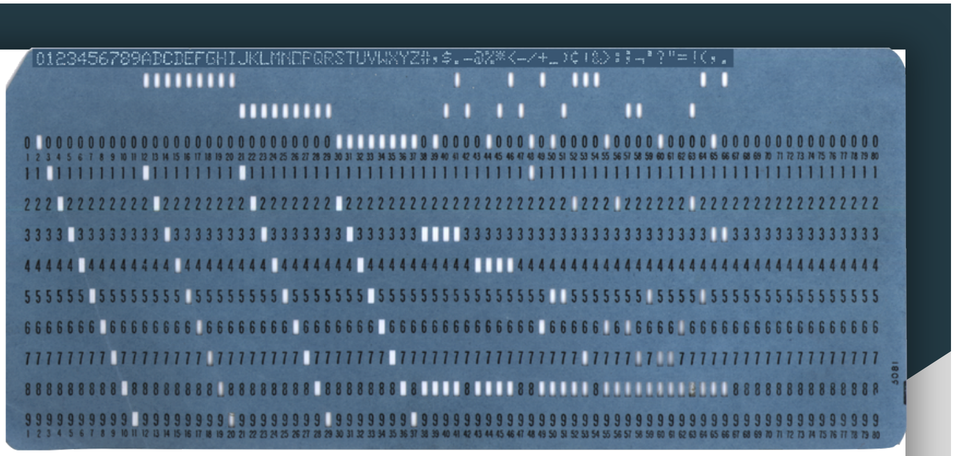
Official Unicode Consortium code chart (PDF)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
U+0B0x		◌ଁ	◌ଂ	◌ଃ		ଅ	ଆ	ଇ	ଈ	ଉ	ଊ	ଋ	ଌ			ଏ
U+0B1x	ଐ			ଓ	ଔ	କ	ଖ	ଗ	ଘ	ଙ	ଚ	ଛ	ଜ	ଝ	ଞ	ଟ
U+0B2x	ଠ	ଡ	ଢ	ଣ	ତ	ଥ	ଦ	ଧ	ନ		ପ	ଫ	ବ	ଭ	ମ	ଯ
U+0B3x	ର		ଲ	ଳ		ଵ	ଶ	ଷ	ସ	ହ			◌଼	ଽ	◌ା	◌ି
U+0B4x	◌ୀ	◌ୁ	◌ୂ	◌ୃ	◌ୄ			◌େ	◌ୈ			◌ୋ	◌ୌ	◌୍		
U+0B5x						□	◌ୖ	◌ୗ					ଡ଼	ଢ଼		ୟ
U+0B6x	ୠ	ୡ	◌ୢ	◌ୣ			୦	୧	୨	୩	୪	୫	୬	୭	୮	୯
U+0B7x	୰	ୱ	୲	୳	୴	୵	୶	୷								



d	0064	LATIN SMALL LETTER D
ԁ	0501	CYRILLIC SMALL LETTER KOMI DE
Ꮷ	13E7	CHEROKEE LETTER TSU
ᑯ	146F	CANADIAN SYLLABICS KO
ⅆ	2146	DOUBLE-STRUCK ITALIC SMALL D
ⅾ	217E	SMALL ROMAN NUMERAL FIVE HUNDRED
ꓒ	A4D2	LISU LETTER PHA
𝐝	1D41D	MATHEMATICAL BOLD SMALL D
𝑑	1D451	MATHEMATICAL ITALIC SMALL D
𝒅	1D485	MATHEMATICAL BOLD ITALIC SMALL D
𝒹	1D4B9	MATHEMATICAL SCRIPT SMALL D
𝓭	1D4ED	MATHEMATICAL BOLD SCRIPT SMALL D
𝔡	1D521	MATHEMATICAL FRAKTUR SMALL D
𝕕	1D555	MATHEMATICAL DOUBLE-STRUCK SMALL D
𝖉	1D589	MATHEMATICAL BOLD FRAKTUR SMALL D
𝖽	1D5BD	MATHEMATICAL SANS-SERIF SMALL D
𝗱	1D5F1	MATHEMATICAL SANS-SERIF BOLD SMALL D
𝘥	1D625	MATHEMATICAL SANS-SERIF ITALIC SMALL D
𝙙	1D659	MATHEMATICAL SANS-SERIF BOLD ITALIC SMALL D

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

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

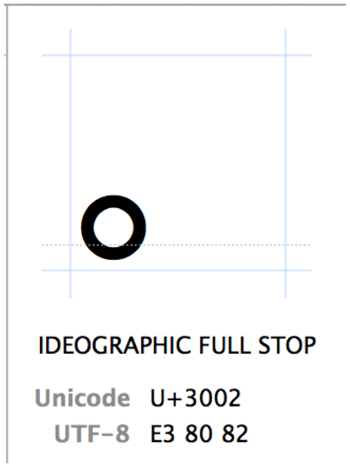
Source: Wikipedia List of writing systems 8/28/2020
<en.wikipedia.org/wiki/List_of_writing_systems>

WORDS

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



IDEOGRAPHIC FULL STOP

Unicode U+3002
UTF-8 E3 80 82

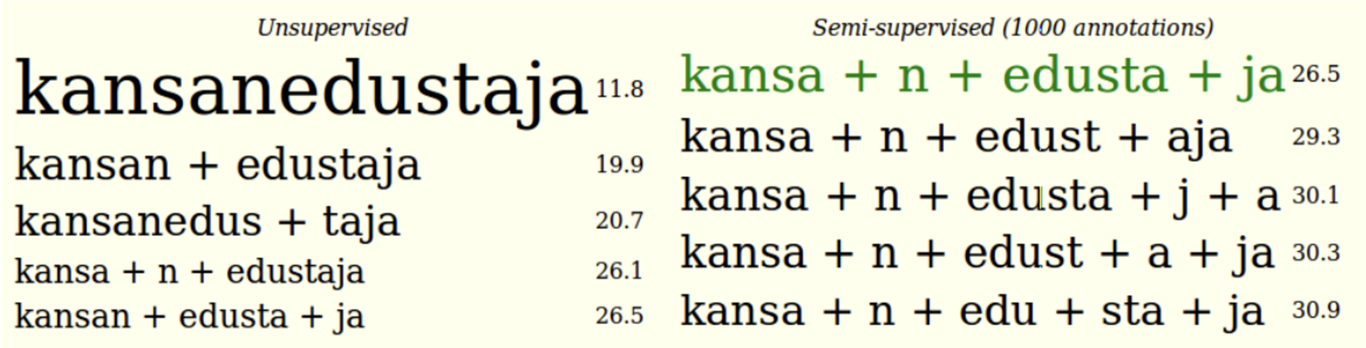
- Some languages have unambiguous end-of-sentence markers
  - E.g., Chinese full stop
- Dr. Mulholland of Mulholland Dr. says “In other langs., sentence segmentation is not so easy.” Dr. Mulholland is right.
  - Six periods, two sentences
- Two main approaches:
  - Rule-based
  - Machine learning

Word Boundary Detection

- In some languages (e.g., English), blank space and punctuation are strong predictors of word boundaries
- In others (e.g., Chinese), wordsaresimplyruntogetherwithoutbreaks.
- 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



Unsupervised		*Semi-supervised (1000 annotations)*	
kansanedustaja	11.8	kansa + n + edusta + ja	26.5
kansan + edustaja	19.9	kansa + n + edust + aja	29.3
kansanedus + taja	20.7	kansa + n + edusta + j + a	30.1
kansa + n + edustaja	26.1	kansa + n + edust + a + ja	30.3
kansan + edusta + ja	26.5	kansa + n + edu + sta + ja	30.9

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 => εjuggling => 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

Look Ma, no sentence splitter, lemmatizer, stopword list, lexicon, thesaurus, or other language-specific customization!

- 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

	s	w	i	m	m	e	r	s	
_	s	w	i	m					
	s	w	i	m	m				
		w	i	m	m	e			
			i	m	m	e	r		
				m	m	e	r	s	
					m	e	r	s	_

Advantages:

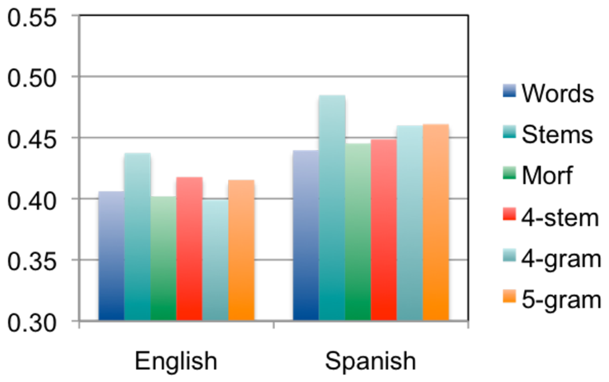
- simple
- address morphology
- surrogate for short phrases
- robust against spelling & diacritical errors
- Language-independent

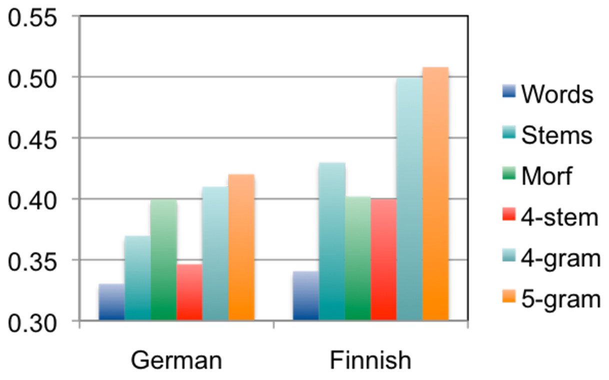
Disadvantages:

- conflation (e.g., simmer, polymers)
- speed and disk usage penalties

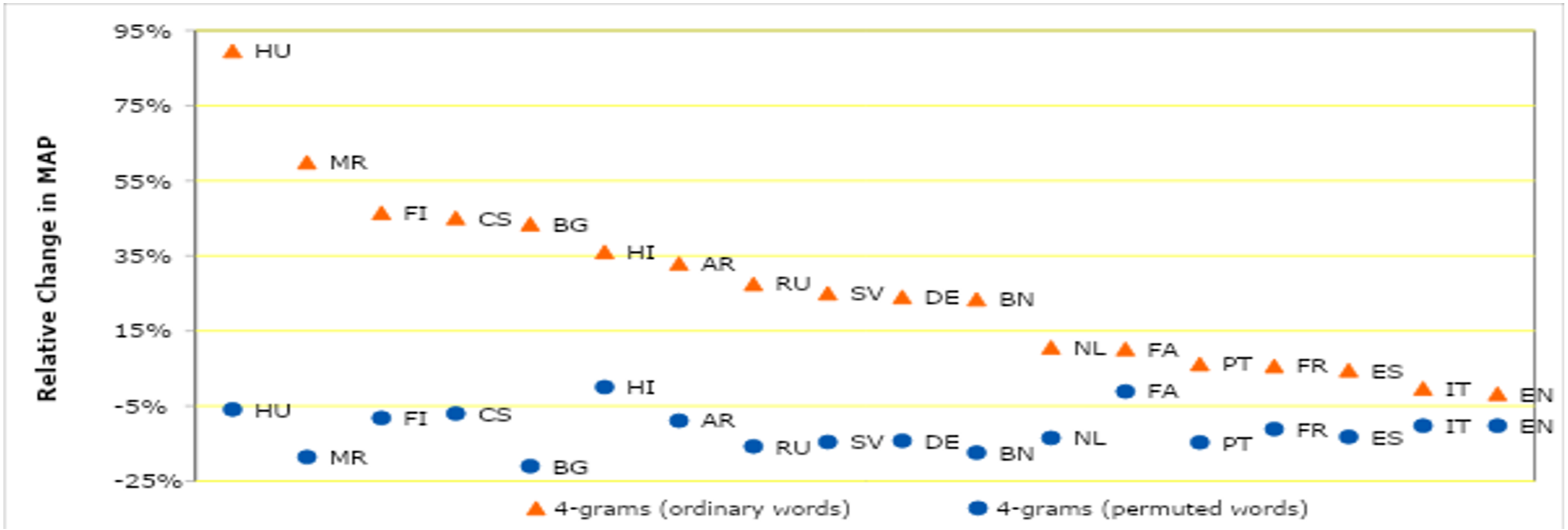
Tokenization Comparison

- 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

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

```
_ s h e _ s e l l s _ s e a s h e l l s _
_s h e _s e l l s _s e a s h e l l s _
_s h e _s e ll s _s e a s h e ll s _
_s h e _s e lls _s e a s h e lls _
_s h e _se lls _se a s h e lls _
```

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

WordPiece Tokenization

ord

- 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

```
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
- Example
  - "L'appartement est grand & vraiment bien situe en plein centre"
  - "▁L" "'" "app" "ar" "tement" "▁est" "▁grand" "▁" "&" "▁v" "r" "ai" "ment" "▁bien" "▁situe" "▁en" "▁plein" "▁centre"

Outline

- Introduction
- CLIR Evaluation
- Attributes of Non-English Text
- ➡ Crossing the Language Barrier
  - Do Nothing
  - Transliteration
  - Machine Translation
  - Dictionary Lookup
  - Multilingual Embeddings
  - Pivoting
  - End-to-end Retrieval
- Other Techniques
- Conclusions

Translation: What Should Be Translated?

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
  - Pro: Translation is fast (per query)
- Translate both to an interlingua
  - Con: More translation required
  - Pro: Interlingua might better support retrieval than hu
  - Pro: Supports multi-way CLIR

Crossing the Language Barrier

Question 2: How should we cross the language barrier?

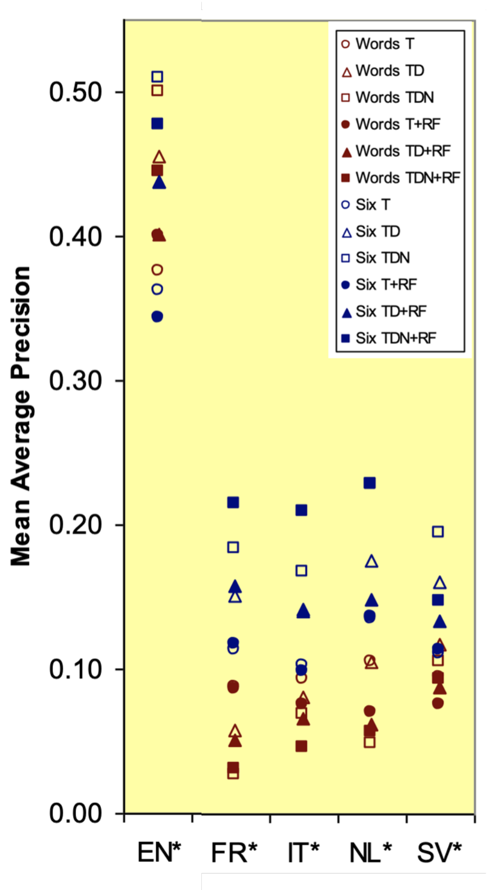
- Do nothing
- Transliteration
- Machine Translation
- Dictionary Lookup
- Multilingual Embeddings
- Pivoting
- End-to-End Retrieval

Crossing the Language Barrier

Do Nothing

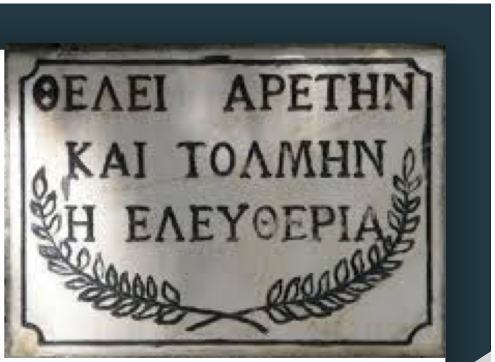
- Sometimes called *cognate matching*
- Buckley et al., 1997: French is misspelled English
  - Applied spelling correction to convert English query to French, then used monolingual retrieval
  - Outperformed many systems at TREC-6

- McNamee & Mayfield 2002: Dutch *is* English
  - Character n-gram tokenization
  - CLEF-2001 English documents, non-English queries



Crossing the Language Barrier

Transliteration



- Transliteration is mapping from the characters of one script to those of a different script in a way that preserves sounds
- Greek word: Ελευθερία
  - *Translation: Freedom*
  - *Transliteration: Eleutheria*
- 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

ಯಂತ್ರ

inneal

मशीन

යන්ත්‍රය

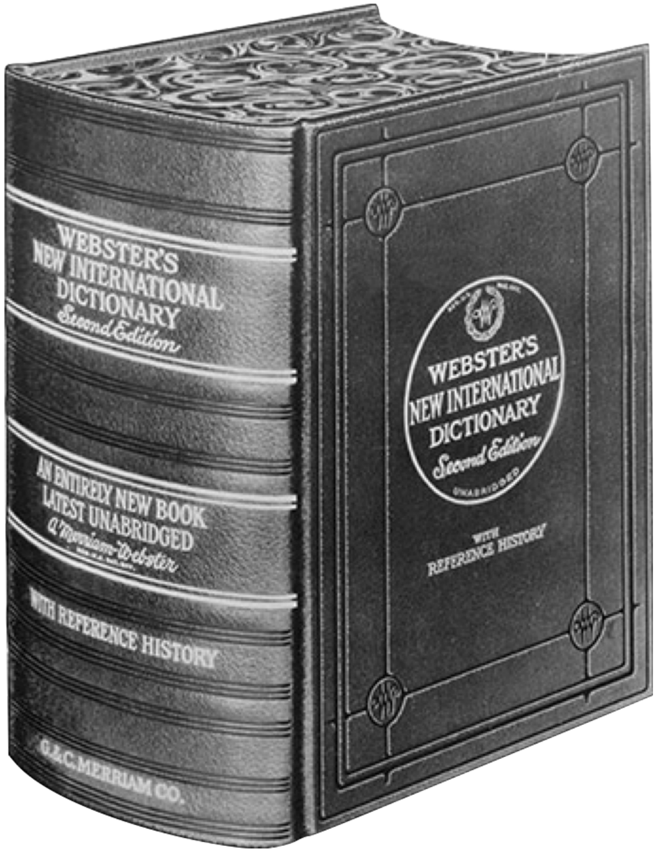
ماشنا

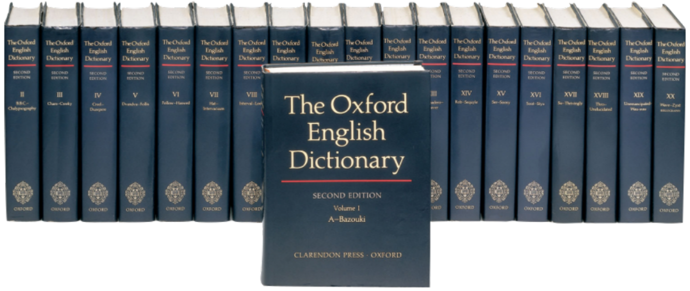
- Most straightforward approach to CLIR
- Radical improvement in machine translation over past four years
  - But much of the gain from using neural approaches comes from improved fluency
  - Not clear how improved fluency can help IR
  - Correlation between machine translation performance and retrieval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

Crossing the Language Barrier

Dictionary Lookup

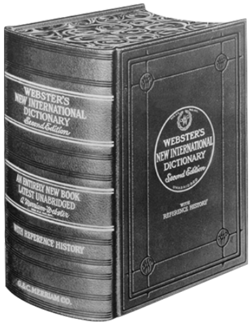
- Word-by-word machine translation
- Keys to success
  - Comprehensive dictionary
    - Matches domain of query
  - Method to select translation(s)
  - Query augmentation





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

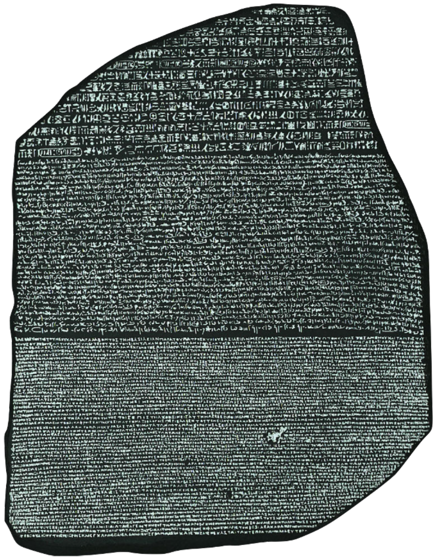


Corpus-based Translation

Given aligned parallel texts and a particular term to translate:

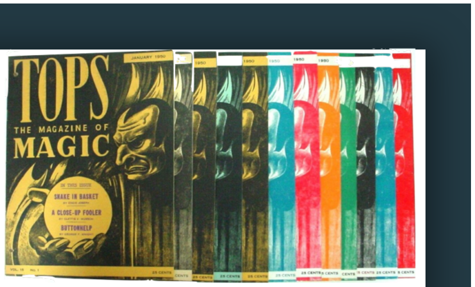
- Find set of documents (sentences) in the source language containing the term
- Examine corresponding foreign documents
- Extract 'good' candidate translation(s)
- Goodness can be based on term similarity measures (Dice, PMI, IBM Model 1, etc.)

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

poisson	pêche	eaux	islandais	cee	baisse
fish	fishing	waters	iceland	eec	decline
freshwater	fisheries	water	icelandic	programme	drop
fishermen	fishermen	sewage	denmark	european	prices
fishing	fishery	pollution	norway	nations	price



Issues in Dictionary-Based CLIR

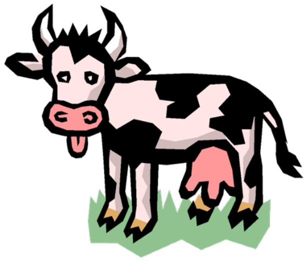
“The main [translation] problems associated with dictionary-based CLIR are (1) untranslatable search keys due to the limitations of general dictionaries, (2) the processing of inflected words, (3) phrase identification and translation, and (4) lexical ambiguity in source and target languages.” - Pirkola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

Subwords can help two (and a half) of these:

- Out-of-Vocabulary words (OOV)
- Morphological Variation
- (Surrogate) Phrase Translation

A. Pirkola, T. Hedlund, H. Keskusalo, and K. Järvelin, ‘Dictionary-Based Cross-Language Information Retrieval: Problems, Methods, and Research Findings.’ *Information Retrieval*, 4:209-230, 2001.

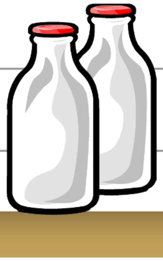
Translating Character N-grams



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 Países Bajos)

	German	Italian	French	Dutch
Word	milch	latte	lait	melk
Stem	milch	latt	lait	melk
4-grams	milc	latt	lait	melk
	ilch	latt		
5-grams	_milc	_latt	_lait	_melk
	milch	latte	lait_	melk_
	ilch_	atte_		



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

		Acquis Corpus			Europarl Corpus		
		words	stems	5-grams	words	stems	5-grams
BG	Bulgarian	0.0591	x	**0.0898**	x	x	x
CS	Czech	0.1107	x	**0.2479**	x	x	x
DE	German	0.1802	0.2097	**0.2952**	0.2427	0.2646	**0.3519**
ES	Spanish	0.2583	0.3072	**0.3661**	0.3509	0.3721	**0.4294**
FI	Finnish	0.1286	0.1755	**0.3552**	0.2135	0.2488	**0.3744**
FR	French	0.2508	0.2733	**0.3013**	0.2942	0.3233	**0.3523**
HU	Hungarian	0.1087	x	**0.2224**	x	x	x
IT	Italian	0.2365	0.2656	**0.2920**	0.2913	0.3132	**0.3395**
NL	Dutch	0.2474	0.2249	**0.3060**	0.2974	0.2897	**0.3603**
PT	Portuguese	0.2009	x	**0.2544**	0.2365	x	**0.2931**
SV	Swedish	0.2111	0.2270	**0.3016**	0.2447	0.2534	**0.3203**
PMAP		0.1811		**0.2756**	0.2714		**0.3527**
% change				**63.5%**			**31.9%**
PMAP-7		0.2161	0.2405	**0.3168**	0.2764	0.2950	**0.3612**
% change			13.1%	**56.0%**		7.1%	**33.0%**

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)

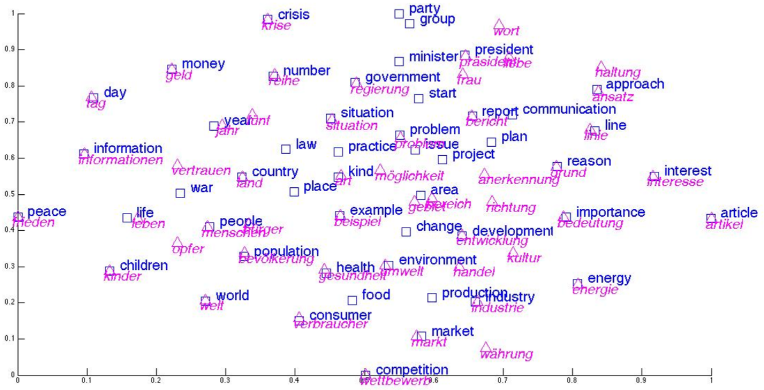
Two Forms of Multilingual Embedding

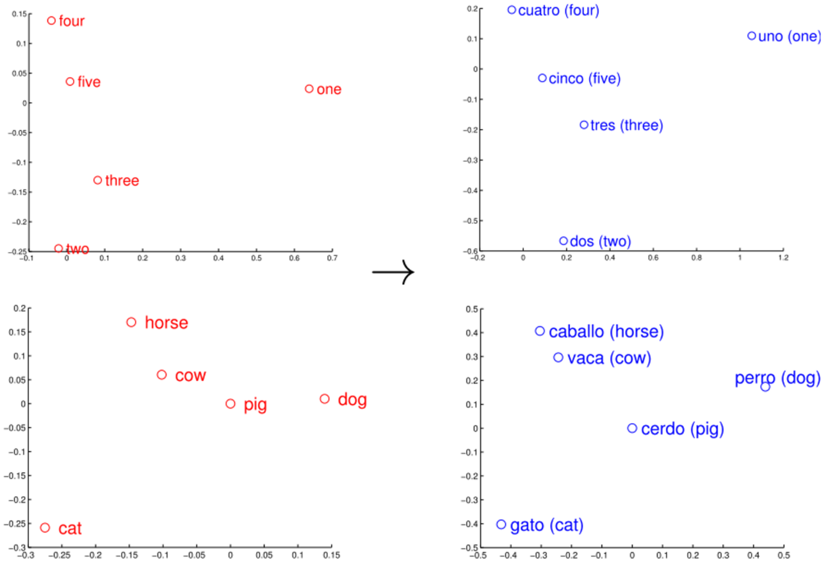
Shared embedding space

- 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

Crossing the Language Barrier

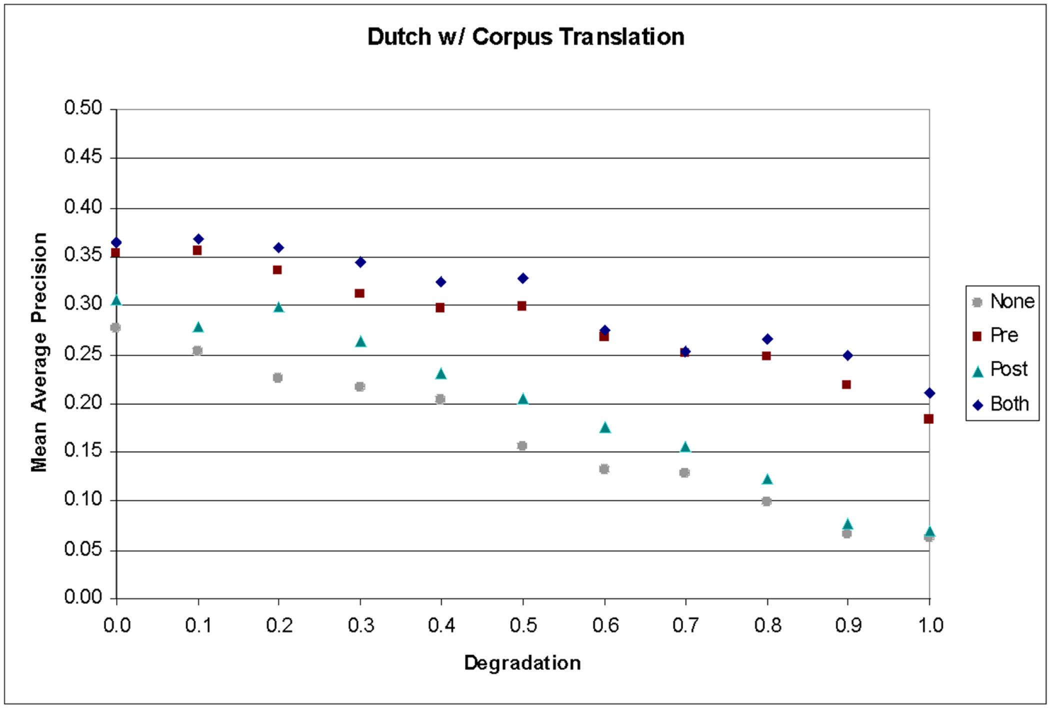
End-to-End Retrieval

- 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 Datasets
    - Translated MS MARCO
      - github.com/unicamp-dl/mMARCO
      - Others available off NeuCLIR page
    - WikiCLIR
      - Uses 2.8M first sentence of Wikipedia articles as queries
      - Automated relevance judgments in 25 languages
      - cs.jhu.edu/~kevinduh/a/wikiclir2018/

Outline

- Introduction
- CLIR Evaluation
- Attributes of Non-English Text
- Crossing the Language Barrier
- Other Techniques
  - Pre- and Post-Translation Expansion
  - Probabilistic Structured Queries
- Conclusions

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

McNamee and Mayfield, *Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources*, SIGIR-2002.

Slide by Doug Oard

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:

$$TF(e, D_k) = \sum_{f_i} p(e|f_i) \times TF(f_i, D_k)$$

$$DF(e) = \sum_{f_i} p(e|f_i) \times DF(f_i)$$

PSQ II
One-Step Cleaner Disinfectant - Pine Scent
KEEP OUT OF REACH OF CHILDREN
DANGER PELIGRO

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

Pirkola, Puolamäki, and Järvelin, ‘Applying Query Structuring in Cross-Language Retrieval’, IPM 39(3), 2003

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

  Ballesteros and Croft, ‘Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval’, SIGIR-97

- Poor Translation Resources

  Demner-Fushman and Oard, ‘The Effect of Bilingual Term List Size on Dictionary-Based Cross-Language Information Retrieval’, HICSS-03

  McNamee and Mayfield, ‘Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources’, SIGIR-02

Thank you

Questions?