Information Retrieval

BITS Pilani
Pilani Campus

Abhishek
April 2020

innovate achieve lead

BITS Pilani
Pilani Campus

CS F469, Information Retrieval

Lecture topics: Cross-Lingual IR

- **Cross-Lingual IR or Cross Language IR**
- **Multi-Lingual IR**

Cross-lingual IR (CLIR):

- Retrieval of documents in a language different from that of a query
  - Query is in language X
  - Documents are in language Y

Multilingual IR (MLIR):

- Query can be in different languages
- Documents can be in different languages

Example of CLIR and MLIR

- E.g.: query: “major earthquakes in recent year”
- We want to retrieve both passages below:
  - There is a major earthquake in Wenchuan, China in 2008 (EN)
  - Un tremblement de terre violent `a Wenchuan secoue la Chine en 2008 (FR)

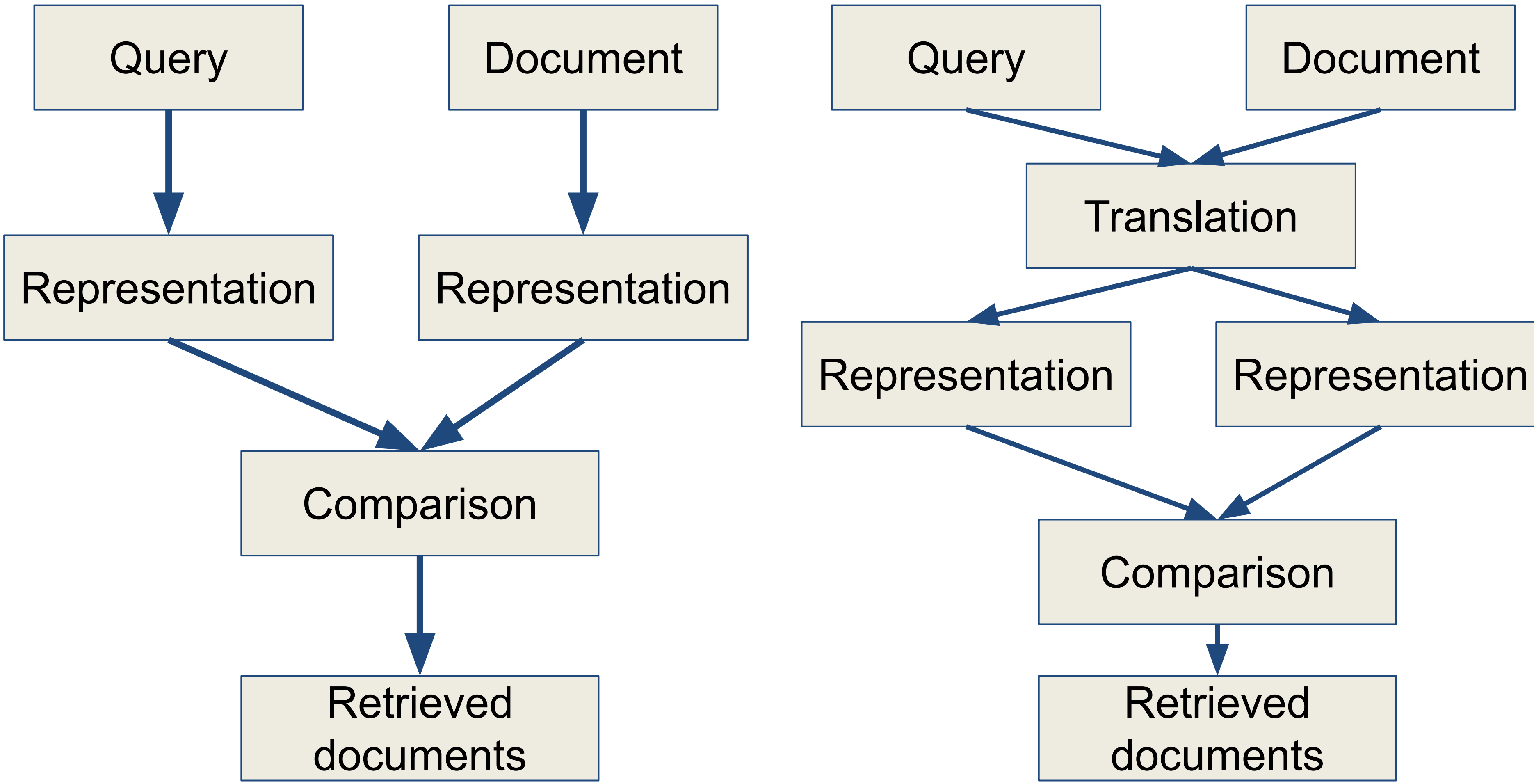
Need for CLIR and MLIR

- Among top 10 million websites, the content languages distribution is as follows: 59% English, 41% non-English.
- An information searcher might wants to retrieve relevant documents in whatever language
  - Intelligence:
    - Govt. Intelligence agencies,
    - companies (finding competing companies, finding calls for tenders, …)
- A user speaking several languages also may want an MLIR to avoid typing the same query several times in different languages.

Problems in CLIR

- CLIR and MLIR are based on monolingual IR: all the problems of monolingual IR
- Problems due to the differences in languages

Monolingual vs Cross lingual



The Translation module

Query translation :

- Mapping the query representation into the document representation
  - **Pro:** flexible, more interaction with the user (who could choose the languages of interest, can correct the translation.)
  - **Cons:** translation ambiguity amplified by the lack of context.

Document translation:

- Mapping the document representation into the query rep.
  - **Pro:** more context
  - **Cons**: one has to determine in advance to which language each document should be translated, all the translated versions should be stored.

The Translation module

Inter-lingua translation

- Mapping document and query representation to a 3rd language
  - **Pro:** useful if there is no resource for a direct translation.
  - **Cons:** lower performance than the direct translation
- **Most used approach in CLIR: Query translation.**

Machine Translation (MT)

- Automatically translating from one language to another language.

Examples:

Information Retrieval→ Récupération de l'information (F)

→ 信息检索 (C)

→ Informationsrückgewinnung (G)

→ सूचना पुनर्प्राप्ति

(H)

Machine Translation Challenges

Lexical Ambiguity

Example: (English → Spanish)

- **book** the flight → reservar
- Read the **book** → libro

Example:

- **Kill** a man → matar
- **Kill** a process → acabar

Examples from Michael Collins slides

Differing word orders

- English word order: subject - verb - object
- Japanese word order: subject - object - verb

English:	IBM bought Lotus
Japanese:	IBM Lotus bought

English:	Source said that IBM bought Lotus yesterday
Japanese:	Source yesterday IBM Lotus bought that said

Examples from Michael Collins slides

Syntactic Ambiguity

John hit the dog with the stick.

John golpeo el perro con el palo / que tenia el palo

Examples from Michael Collins slides

Machine Translation Methods

Translation Methods

- Dictionary Based
- Statistical Methods

Dictionary Based Query Translation: Overview

- This approach tries to identify and select the possible translations of each source word from a bilingual dictionary.
- English-French dictionary examples:
  - access: attaque, accéder, intelligence, entrée, accès
  - branch: branche, bifurquer, succursale
  - data: données, matériau, data
- For each word, there are several candidates. Thus, for multi-word query there are several possible sequences.

Dictionary Based Query Translation: Approach

- For each query word
  - Determine all the possible translations (through a dict.)
- Selection
  - Select the set of translation words that produce the highest ***cohesion***

Cohesion

- Frequency of two translation words together

E.g. For translating “data access”

  - data: données, matériau, data
  - access: attaque, accéder, intelligence, entrée, accès

(accès, données)	152 *
(accéder, données)	31
(données, entrée)	21
(entrée, matériau)	3

- Frequency from a document collection or from the Web

Statistical Machine Translation

Summary of CLIR

- CLIR = Query Translation + IR
    - Integrate QT with IR
    - QT is one step in the global IR process

Multilingual IR

- MLIR = CLIR + merging
  - Translate the query into different languages
  - Retrieve doc. in each language
  - Merge the results into a single list.

Multilingual IR

Merging

- Round-robin
  - Take the first from the list of F, E, I, …
  - Take the second from the list of F, E, I, …
  - Assumption: similar number of rel. doc., ranked similarly
- Raw score
  - Mix all the lists together
  - Sort according to the similarity score
  - Assumption: similar IR method, collection statistics

Introduction to Statistical MT

innovate achieve lead

- Parallel corpus is available in several language pairs.
- Basic idea: use parallel corpus as a training set of translation examples.
- Example of parallel corpus collection:
  - OPUS: http://opus.nlpl.eu/

Noisy Channel Model

- Goal:
  - translation system from **Source to Target** language.
  - Have a model **p(t | s)** which estimates conditional probability of any target language sentence **t** given the source language sentence **s**. Use the training corpus to set the parameters.
- A Noisy Channel Model has two components:
  - **p(t)** the language model
  - **p(s | t)** the translation model
- Using the above two, we can estimate:
  - Learn a distribution **p(t | s)** = $\mathbf{\mathit{arg\ max}_t}$ **p(t)p(s | t)**

More about Noisy Channel Model

- The language model **p(t)** could be a trigram model, estimated from any data (parallel corpus not needed to estimate the parameters)
- The translation model **p(s | t)** is trained from a parallel corpus of Source/Target pairs.
- Note:
  - The translation model is backwards!
  - The language model can make up for deficiencies of the translation model.
  - Later we'll talk about how to build **p(s | t)**
  - Decoding, i.e., finding

$$\mathbf{argmax_t\, p(t)p(s \mid t)}$$

  - is also a challenging problem.

Language Modeling Problem

- We have some (finite) vocabulary,

  Say V = {the, a, book, read, bank, two, … }

- We have infinite set of strings, V'

  the STOP

  a STOP

  a book STOP

  a two the book read STOP

  …

  …

Language Modeling Problem (Continued)

- We have a training sentence of example sentences in English.
  - Billion or more words

- We need to learn a probability distribution p, i.e., a function that satisfies

$$\Sigma_{x \in V'} \; p(x) = 1 \quad p(x) \geq 0 \text{ for } x \in V'$$

p(the STOP) = 10^{-12}

p(the fan STOP) = 10^{-8}

p(the fan saw Sachin STOP) = 10^{-11}

p(the the fan saw saw STOP) = 10^{-15}

...

Why we want to model p(x)?

- Speech Recognition was the original motivation.
  - Map input analog signal to sequence of words.
  - Confusing sound/words
    - Wreck a nice beach
    - Recognize speech
- Machine translation

A Naive Method

- We have N training sentences
- For any sentence $x_1 \ldots x_n$, $c(x_1 \ldots x_n)$ is the number of times the sentence is seen in our training data
- A naive estimate:

$$p(x_1 \ldots x_n) = c(x_1 \ldots x_n) / N$$

Markov Process

- Consider a sequence of random variables $X_1, X_2, \ldots X_n$. Each random variable can take any value in finite set V. For now, we assume that the length n is fixed (e.g., n = 100).
- Our goal: model

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

First-Order Markov Process

$$P(X_1 = x_1, X_2 = x_2, \dots , X_n = x_n)$$

$=$

Second-Order Markov Process

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

$=$

Modeling Variable Length Sequences

- We would like the length of the sequence, n, to also be a random variable.
- A simple solution: always define X_n = STOP, where STOP is a special symbol.

Trigram Language Models

- A trigram language model consists of:
  a. A finite set V.
  b. A parameter q(w | u, v) for each trigram (u, v, w) such that w ∈ V U {STOP} , and u, v ∈ V U {*}

- For any sentence $x_1 \ldots x_n$, where $x_i \in V$ for i = 1 … (n-1), and x_n = STOP, the probability of the sentence under the trigram language model is

$$p(x_1 \ldots x_n) = \prod_{i=1}^{n} q(x_i | x_{i-2}, x_{i-1})$$

  where we define $x_0 = x_{-1} = *$

An Example

- For a sentence:

  The dog barks STOP

We would have:

$$p(\text{the dog barks STOP}) = q(\text{the} \mid *, *) \times q(\text{dog} \mid *, \text{the}) \times q(\text{barks} \mid \text{the}, \text{dog}) \times q(\text{STOP} \mid \text{dog}, \text{barks})$$

The Trigram Estimation Problem

$q(w_i | w_{i-2}, w_{i-1})$

For example: q(barks | the, dog)

A natural estimate (the “maximal likelihood estimate”)

$$q(w_i | w_{i-2}, w_{i-1}) = Count(w_{i-2}, w_{i-1}, w_i) / Count(w_{i-2}, w_{i-1})$$

q(barks | the, dog) = Count(the dog barks) / Count(the dog)

Sparse Data Problems

$$q(w_i | w_{i-2}, w_{i-1}) = Count(w_{i-2}, w_{i-1}, w_i) / Count(w_{i-2}, w_{i-1})$$

q(barks | the, dog) = Count(the dog barks) / Count(the dog)

Say our vocabulary size is $N = |V|$, then there are N^3 parameters in the model.

E.g. $N = 20{,}000 \rightarrow 20000^3 = 8 \times 10^{12}$ parameters

The Bias-Variance Trade-off

- Trigram maximum-likelihood estimate

  $q_{ML}(w_i | w_{i-2}, w_{i-1}) = Count(w_{i-2}, w_{i-1}, w_i) / Count(w_{i-2}, w_{i-1})$

- Bigram maximum-likelihood estimate

  $q_{ML}(w_i | w_{i-1}) = Count(w_{i-1}, w_i) / Count(w_{i-1})$

- Unigram maximum-likelihood estimate

  $q_{ML}(w_i) = Count(w_i) / Count()$

Linear Interpolation

- Take our estimate $q(w_i | w_{i-2}, w_{i-1})$ to be

$$q(w_i | w_{i-2}, w_{i-1}) = \lambda_1 \times q_{ML}(w_i | w_{i-2}, w_{i-1}) + \lambda_2 \times q_{ML}(w_i | w_{i-1}) + \lambda_3 \times q_{ML}(w_i)$$

where $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and $\lambda_i \geq 0$ for all i.

Example:

q(barks| the, dog) = ⅓ q_{ML}(barks| the, dog) + ⅓ q_{ML}(barks | dog) + ⅓ q_{ML}(barks)

Assuming all lambdas values are equal.

Linear Interpolation (Continued)

- Is $q(w_i| w_{i-2}, w_{i-1})$ a valid estimator?

$$\sum_{w \in V'} q(w|u, v) = 1$$

References

Language Modeling:
http://www.cs.columbia.edu/~mcollins/lm-spring2013.pdf