Cross-language High Similarity Search

Parth Gupta
pgupta@dsic.upv.es

Natural Language Engineering Lab
Technical University of Valencia (UPV), Spain
http://www.dsic.upv.es/~pgupta
Where is Valencia?
Where is my home?
Background of cross-language information access

- The world has many languages and people do speak and understand them.
- Most of the people in the world are mono/bi/tri-lingual.
- People do travel across the countries where most of the things are in different languages.
- Cross-language information access comes into play where people need information across languages.
- Two cases:
  - Mono-lingual - If a person speaks/understands only one language
  - bi/tri-lingual - If a person speaks/understands multiple languages
• To address the CL information access problem there are two solutions
  ○ Political
    • Teach/Force everyone THE language (mono-lingual)?
  ○ Social
    • Make information across languages accessible to each and everyone in their language.

• Two approaches to address the latter solution (The former is out-of-scope here)
  1. Either have translators to make information accessible
  2. OR Use technology to break the language barrier in information access
Another example

Google - how to get a driving licence in singapore

About 5,280,000 results (0.31 seconds)

Ads related to how to get a driving licence in singapore

Basic Theory Test / FTT - singaporedrivingtests.com
www.singaporedrivingtests.com/
Free 20 Questions Singapore Driving Theory Tests online practice.

Singapore driving license - ExpatFocus.com
www.expatfocus.com/Singapore-driving
Read our free expat guide to driving licenses in Singapore
Singapore Banking - Singapore Property - Singapore Retirement

Singapore Driving License - Pass BTT & Obtain License to Drive
www.singaporetests.com/
Money Back Guarantee as Seen on TV
Test Engine - Full Refund Guaranteed!

eCitizen - Topics - Getting a driving licence for motorcars (Class 3, 3A)
www.ecitizen.gov.sg > Topics > Transport and Motoring
May 7, 2013 - To obtain a Class 3 or 3A driving licence for motorcars, you should take driving lessons by enrolling in a licenced driving school or engaging a...

Conversion of Foreign Driving Licence - Singapore Traffic Police
driving-in-singapore spf.gov.sg/.../driving_in_singapore/.../information...
New PPs who already have a 5-year Singapore driving licence will have to replace it...
7 of 36
Background

Technically there are two approaches to address the issue of cross-language information access

- **Resource based (Manual)** - Here you rely on human efforts
  - hand-crafted dictionaries
  - manual translations
  - manually created word-nets
  - Advantage: Very high quality
  - Disadvantage: Very expensive, slow

- **Statistical** - You just need data
  - train bilingual dictionaries
  - find parallel data and train translation systems
  - derive term-level associations
  - Advantage: Very fast and scalable
  - Disadvantage: Noisy and data-dependent
Cross-language High Similarity Search

• **Statement**
  ○ Find/link documents across languages which contain *almost* identical information out of quasi-comparable collection.
  ○ Quasi-comparable corpus is a collection documents where there are chances to find similar documents across languages but they are not known a priori.

• This is also sometimes referred as cross-language near duplicate identification and cross-language pairwise similarity search.
Motivation

• Cross-language NLP and IR heavily rely on parallel and comparable data
  ◦ Parallel - Where the aligned documents are exact translations of each other
  ◦ Comparable - Where the aligned documents are covering the same information

• Parallel data is precious but scarce

• Most of the available data is quasi-comparable - not topically aligned

• The technologies to extract parallel or comparable fragments from quasi-comparable data will be very useful in such scenarios
Applications based on parallel/comparable corpora

- Cross-language IR
  - Training Cross-language Latent Semantic Indexing models
- Term-level associations
  - Bilingual dictionary
  - Cross-language topic models
- Machine Translation
  - IBM M1 like models
- Cross-language Text Categorisation/Clustering
Current Scene

- All languages don’t have parallel data - and the available data is too small to rely
- Most of the time the available data is domain dependent e.g. Parliament proceedings
- Comparable corpus (Wikipedia) is not reliable in many languages
- In fact many languages do not have enough data
Motivation Contd..

- Two Questions:
  1. What can be considered as a constant source of text across languages?
  2. ... that can contain parallel or comparable fragments?

- Answer
  - Wikipedia articles - often, people create pages by translating English pages!
  - News stories - journalistic text re-use!

- Which languages to work on?
  - Resource Poor Languages - Why?
Background - Web and Languages

Top Ten Languages in the Internet
2010 - in millions of users

- English: 536.6
- Chinese: 444.9
- Spanish: 153.3
- Japanese: 99.1
- Portuguese: 82.5
- German: 75.2
- Arabic: 65.4
- French: 59.8
- Russian: 58.7
- Korean: 39.4
- All the rest: 350.6

Estimated Internet users are 1,966,514,816 on June 30, 2010
Copyright © 2000 - 2010, Miniwatts Marketing Group
Background - Web and Languages

- Language Population\(^1\)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Language</th>
<th>Speakers (millions)</th>
<th>% of world</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mandarin</td>
<td>935</td>
<td>14.1</td>
</tr>
<tr>
<td>2</td>
<td>Spanish</td>
<td>387</td>
<td>5.85</td>
</tr>
<tr>
<td>3</td>
<td>English</td>
<td>365</td>
<td>5.52</td>
</tr>
<tr>
<td>4</td>
<td>Hindi</td>
<td>295</td>
<td>4.46</td>
</tr>
<tr>
<td>5</td>
<td>Arabic</td>
<td>280</td>
<td>4.23</td>
</tr>
<tr>
<td>6</td>
<td>Portuguese</td>
<td>204</td>
<td>3.08</td>
</tr>
<tr>
<td>7</td>
<td>Bengali</td>
<td>202</td>
<td>3.05</td>
</tr>
<tr>
<td>8</td>
<td>Russian</td>
<td>160</td>
<td>2.42</td>
</tr>
<tr>
<td>9</td>
<td>Japanese</td>
<td>127</td>
<td>1.92</td>
</tr>
<tr>
<td>10</td>
<td>Punjabi</td>
<td>95.5</td>
<td>1.44</td>
</tr>
</tbody>
</table>

\(^1\)The estimates used for this list are those of Nationalencyclopededin and is based on estimates published in 2007 - Wikipedia.
Observation

- News stories covering the same event published in different languages may be rich sources of parallel and comparable text.
- Some fragments in these stories are parallel, for example, personal quotes and translated versions of the same content.

- Definitions [Barker and Gaizauskas, 2012]
  - **Focal Event**: The main event or events which provide a focus for the news story
    - e.g. Romney vs. Obama in Ohio: With superior ground operations, the president widens his lead
  - **Background Event**: an event that plays a supporting role in the text, providing context for the focal events
    - e.g. Probable the last encounter between the two
  - **News Event**: a group of related events, broader than and including the focal event, which may be reported over time in different news text installments
    - e.g. Presidential election polls
CLINSS Track at FIRE

- To channelize the community effort we organise a track on this task at FIRE 2013 (Forum for Information Retrieval Evaluation)
- The Track Name: Cross-Language Indian News Story Search (CLINSS)
- Second consecutive year
Task Description

Statement

- For each $t \in T$, find $s \in S$ covering the same focal event and news event

$S = L_1 \cup L_2 \cup \cdots \cup L_n$

$T = \text{English Articles}$
<table>
<thead>
<tr>
<th>Article</th>
<th>Title</th>
<th>Relevance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target</td>
<td>There’s lot more to talk than my 50th Test ton: Tendulkar english-document-00006.txt</td>
<td></td>
</tr>
<tr>
<td>Source1</td>
<td>मेरी 50वीं सेंचुरी के अलवा भी कई बातें हैं: तेंदुलकर (There are many things except my 50th century: Tendulkar hindi-document-24799.txt)</td>
<td>2 (same focal event)</td>
</tr>
<tr>
<td>Source2</td>
<td>सचिन ने बनई सेंचुरी की फिफ्टी (Sachin makes fifty in century hindi-document-08018.txt)</td>
<td>1 (same news event)</td>
</tr>
</tbody>
</table>

**Table:** Example English-Hindi text pairs describing the same news event but different focal events
Corpus Statistics

Table: CL!NSS 2012 corpus statistics. The statistics are shown for the two source partitions, $D_{hi}$ (Hindi) and $D_{gu}$ (Gujarati), and a target collection $D_{en}$. The column headers stand for: $|D|$ number of documents in the corpus (partition), $|D_{tokens}|$ total number of tokens, $|D_{voc}|$ total size of vocabulary (unique terms). $k =$ thousand, $M =$ million.

| Partition | $|D|$ | $|D_{tokens}|$ | $|D_{voc}|$ |
|-----------|------|----------------|-----------|
| $D_{en}$  | 50   | 21k            | 4k        |
| $D_{hi}$  | 50691| 15.6M          | 143k      |
| $D_{gu}$  | 11889| 5.8M           | 282k      |

- Metadata
  - Title of the news story
  - Date of publication
  - Content of the Story
Evaluation Framework

Relevance

- The relevance level of the source news stories for the given test queries will be in 2,1,0 where,
  - $2 = \text{"same news event + same focal event"}$
  - $1 = \text{"same news event + different focal event"}$ and
  - $0 = \text{"different news event"}$

Measures [Järvelin and Kekäläinen, 2002]

- NDCG@k, $k = 1, 5, 10, 20$
Participation Overview

Submission details

• Teams were asked to submit results in terms of rank-list for each language pair.
• Each team could submit up to 3 runs to try different approaches or configurations.

Participation

• 10 Teams registered
• 3 Teams submitted runs
• Total 7 runs for English-Hindi task and 1 run for English-Gujarati task
• All took very different strategies and some runs were without use of MT system!
Results

English-Hindi

**Fig. 4.** Overall evaluation results for English-Hindi partition. The left hand side information corresponds to (run number) and team. The ranking is upon the NDCG@10 values.

English-Gujarati

<table>
<thead>
<tr>
<th>Run</th>
<th>NDCG@1</th>
<th>NDCG@5</th>
<th>NDCG@10</th>
<th>NDCG@20</th>
</tr>
</thead>
<tbody>
<tr>
<td>palkovskii (1)</td>
<td>0.0541</td>
<td>0.0843</td>
<td>0.0955</td>
<td>0.0981</td>
</tr>
</tbody>
</table>

**Table:** Overall evaluation results for English-Gujarati partition.
• Translated the *target* document into *source* language using MT system \(^2\)
• Then similarity estimation between \(S\) and \(T\) was carried out using following components and a rank-list was prepared
  ○ TF-IDF score between title of two documents.
  ○ TF-IDF score between contents of two documents.
  ○ Maximum TF-IDF score between sentences from the contents of the two documents.
  ○ If both documents are published within a window of 10 days, constant 0.5 is added to the final score.

\(^2\)Google Translate: www.translate.google.com
They used Cross-language explicit semantic analysis (CL-ESA) [Potthast et al., 2008]
- Aligned Wikipedia articles of two languages is considered as reference corpus
- $t \in T$ and $s \in S$ is used as a query on corresponding language Wikipedia to generate rank-list (cosine score) called semantic vectors
- Score between $t$ and $s$ is the cosine similarity between these semantic vectors.
- As the similarity is measured using external reference collection, it is called explicit semantic analysis.
• **Run 1** - The search space of $S$ is reduced by a *date-window* of $\pm 2$-days where *date-window* is defined as publication date of $s$ to be $\pm 2$-days of that of $t$.

• **Run 2** - The *date-window* is kept as 7-days.

• **Run 3** - The $S$ is indexed and *translated* $t$ is queried to get top 1000 $s$ to create the search space for CL-ESA.
Team 3 - iiit-h [Reddy and Perumal, 2012]

- Translated $t$ into language of $s$.
- Extracted the key-phrases from $t$ based on $n$-gram filtration and term weighting techniques.
- Indexed the $S$ using Lucene$^3$.
- Queried extracted key-phrases of $t$ on index and retrieved rank-list for each key-phrase.
- The score of $s$ for $t$ is dependent on the number of rank-list it appears in and its rank.

$^3$http://lucene.apache.org/core/
Outlook of the task

Pair(A,B)

Same News Event  Different News Event

Same News Event  Same Focal Event  Same News Event  Different Focal Event

High Parallel Content  Low Parallel Content  No Parallel Content

Parallel Fragments Extraction

Derived Content  Non-derived Content

Year 2012

Task: Story Detection

Year 2013/14

Task: Identifying Parallel Fragments and Extraction

Task: Likely Reuse (or classification)
Current Work

- We are currently exploring the latent space based solution to this problem aka Dimensionality Reduction Techniques.
- Literature survey
  - OPCA: Oriented Principal Component Analysis [Platt et al., 2010]
  - CL-MDS: Cross-language MultiDimensional Scaling [Banchs and Kaltenbrunner, 2008]
  - CL-DBN: Cross-language Deep Belief Networks [Kim et al., 2012]
Latent Space based Framework

<table>
<thead>
<tr>
<th>( t_1 )</th>
<th>( t_2 )</th>
<th>( t_3 )</th>
<th>( \ldots )</th>
<th>( t_m )</th>
<th>( w_1 )</th>
<th>( w_2 )</th>
<th>( w_3 )</th>
<th>( \ldots )</th>
<th>( w_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_1 )</td>
<td>( x_{11} )</td>
<td>( L_1 )</td>
<td>( \vdots )</td>
<td>( \vdots )</td>
<td>( L_2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( d_2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( d_3 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \ldots )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( d_p )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( x_{11} = \{0,1\} \mid [0,1] \mid [0,N] \)

**Figure:** The formulation of the problem to apply different dimensionality reduction techniques. The Documents are part of a comparable or parallel text collection.
The solution

• Solve the matrix to find principal components/abstract representation
• Most of the techniques differ in the manner they solve the matrix
• Some of the methods learn joint representation as shown in previous slide, while others solve languages independently and on top of it there is a dimension mapping module
• These techniques are also referred as Dimensionality Reduction techniques
• Theoretically two categories: Linear and Non-Linear
Dimensionality Reduction Techniques

• CL-LSI (Linear)
  ○ Solves as Principal Component Analysis (PCA) problem
  ○ $Ax = \lambda x$

• OPCA (Linear)
  ○ Solves Generalised Eigen Problem $Ax = \lambda Bx$

• CL-MDS (Non-Linear)
  ○ Uses Multi-Dimensional Scaling to reduce dimensionality
  ○ Rotates the projections to align dimensions

• CL-DBN (Non-Linear)
  ○ Uses Deep Belief Networks to train dimensionality reduction module for each language
  ○ Uses Canonical Correlation to map dimensions in the reduced space
Observations

• Non-linear techniques can find much compact and better representation than their Linear counterparts [Hinton and Salakhutdinov, 2006]
• Non-linear techniques have some convergence issues because of several local-optima
• CL-MDS gives much better representation but does not have an operator to project unseen data
Thank You! 😊

http://www.dsic.upv.es/grupos/nle/clinss.html
References (1)


Learning semantics with deep belief network for cross-language information retrieval.
In COLING (Posters), pages 579–588.

Working note for CL!NSS.
In [FIRE, 2012].

Translingual Document Representations from Discriminative Projections.

A Wikipedia-based Multilingual Retrieval Model.

Working note for CL!NSS.
In [FIRE, 2012].