Multiword Named Entities Extraction from Cross-Language Text Re-use

Parth Gupta\textsuperscript{1}, Khushboo Singhal\textsuperscript{2}, Paolo Rosso\textsuperscript{1}

\textsuperscript{1}NLE Lab, UPV, Spain
\textsuperscript{2}IR-Lab, DA-IICT, India

May 27, 2012
Outline

Introduction

Approach

Results

Remarks & Future Work

References
What is cross-language text re-use?

Text in one language is generated from taking reference of another text written in a different language.

- At document level - Noisy parallel text
- At fragment level - Comparable text

Examples

- Wikipedia articles, News stories, Student reports etc.
Hindi - English Language Pair

- is not extremely disconnected **But**
- necessary resources including sufficient parallel data is absent
- technological inadequacies like
  - still majority of Indian people use computers (including Web) in English
  - use of Hindi keyboards is even lesser (probably negligible)
Multiword Named Entities (MWNEs)

- Named entities are quite often multiword units
- Tagging 2275 English news articles, out of total 21,208 unique NEs: 9,079 (43%) were single word and 12,129 (57%) were multiword NEs.

<table>
<thead>
<tr>
<th>Type</th>
<th>Frequent Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>2</td>
</tr>
<tr>
<td>Location</td>
<td>2</td>
</tr>
<tr>
<td>Organisation</td>
<td>4</td>
</tr>
</tbody>
</table>
Approach

1. Identify the text re-use document pairs
2. *Mine* MWNEs from these pairs

Text re-use document pairs identification [Gupta and Singhal 2011]
Mining Module

- **Input:** Candidate MW units $S$ and $T$, where $S$ and $T$ are source and target MW units respectively.
  - where, $S = \{s_1, s_2, \cdots, s_N\}$ and $T = \{t_1, t_2, \cdots, t_M\}$

- **Output:** $<s', t'>$ which maximises $\zeta(.,.)$
  - where, $s'$ and $t'$ are sub-strings of $S$ and $T$ respectively

- **Problem formulation:**
  - $\max \ zeta(S, T)$ instead of $\max \ zeta(S|T)$

- Joint estimation by the means of transliteration, translation and language models
  - i.e. the longest matching subsequence validated by language models
Challenges

Localisation - candidate selection

- without localisation it becomes extremely computation intensive (brute force)
- we handle localisation with transliteration mapping
- i.e. pass the $S$ and $T$ to the MWNE mining module if there is a transliteration match
How it works

Consider

- \( S = \ldots \) Government of Madras ...
- \( T = \ldots \text{Madras Sarkar} \ldots \) (\textit{Madras Sarkar})
Motivation for Joint Estimation

- Conditional estimation [Bhole et al. 2011] $\zeta(T|S)$ depends on prior knowledge of source MWNE
- Example of some English MWNEs partially identified by NE tagger
  - “Bayes” ("Bayes Theorem")
  - “Sundereshwara Temple” ("Meenakshi Sundereshwara Temple")

P. Gupta, K. Singhal, P. Rosso
1 NLE Lab, UPV, Spain, 2 IR-Lab, DA-IICT, India
Multiword Named Entities Extraction from Cross-Language Text Re-use
CL!TR-2011-Dataset

- Wikipedia articles of English (5032) and Hindi (190) related to “computer science” and “tourism” domain
- For more details [Barrón-Cedeño et al.2011]

Utilities

- Translation Model - Universal Word (UW) Bilingual Dictionary
- Transliteration Model - Google Transliterate API

---

Results

<table>
<thead>
<tr>
<th>Type</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>FM</td>
<td>0.57</td>
<td>0.38</td>
<td>0.49</td>
</tr>
<tr>
<td>FM+PM</td>
<td>0.86</td>
<td>0.57</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Table: Performance evaluation of MWNE extraction algorithm on Hindi-English language pair. FM is full match and FM+PM is full and partial match.
Examples

Correctly Identified

<table>
<thead>
<tr>
<th>English</th>
<th>Hindi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sawai Madhopur</td>
<td>स्वाई माधोपुर</td>
</tr>
<tr>
<td>DSIR model</td>
<td>डीएसआईआईआर मॉडल</td>
</tr>
<tr>
<td>Government of Madras</td>
<td>मद्रास सरकार</td>
</tr>
<tr>
<td>Kashgar Ladakh</td>
<td>कशगर लदाख</td>
</tr>
<tr>
<td>Medical Board</td>
<td>मेडिकल बोर्ड</td>
</tr>
</tbody>
</table>

Partially Identified

<table>
<thead>
<tr>
<th>English</th>
<th>Hindi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranthambore National Park</td>
<td>रणथम्बोर राष्ट्रीय पark</td>
</tr>
<tr>
<td>Meenakshi Amman</td>
<td>मीनाक्षी अम्मान</td>
</tr>
<tr>
<td>India Company</td>
<td>इन्डिया कंपनी</td>
</tr>
<tr>
<td>administered Gilgit</td>
<td>प्रशासित गिलगित</td>
</tr>
</tbody>
</table>

False Positive

<table>
<thead>
<tr>
<th>English</th>
<th>Hindi</th>
</tr>
</thead>
<tbody>
<tr>
<td>computing these values</td>
<td>मूल्यों कंप्यूटिंग</td>
</tr>
<tr>
<td>year the Sikhs</td>
<td>वर्ष सिखों</td>
</tr>
<tr>
<td>probability of b</td>
<td>प्रयक्तता बी</td>
</tr>
<tr>
<td>where b</td>
<td>जहां बी</td>
</tr>
</tbody>
</table>
Remarks

- Joint estimation does not require prior knowledge of NEs in the source language, hence suits more naturally to the problem.
- This module is very suitable for language pairs connected with a pivot language.
- The preliminary results on the noisy parallel text are encouraging.

Future Work

- We want to test the model on a larger scale and with a wide variety of languages and comparable corpora.
- We would also test the model in the pivot architecture.
Thank You! 😊

Relevant Venues

- **PAN** @ FIRE - *held in conjunction with FIRE 2012*
  - Task of cross-language high similarity search on Indian language news stories
  - 17-19 December, Kolkata, India.

- **PAN** @ CLEF - *held in conjunction with CLEF 2012*
  - Task of plagiarism detection (also from cross-language perspective)
  - 17-20 September, Rome, Italy.
References I

Alberto Barrón-Cedeño, Paolo Rosso, Sobha Lalitha Devi, Paul Clough, and Mark Stevenson.
2011.
Pan@fire: Overview of the cross-language !ndian text re-use detection competition.

Abhijit Bhole, Goutham Tholpadi, and Raghavendra Udupa.
2011.
Mining multi-word named entity equivalents from comparable corpora.

Parth Gupta and Khushboo Singhal.
2011.
Mapping hindi-english text re-use document pairs.