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

Parth Gupta[1], Khushboo Singhal[2], Paolo Rosso[1]

[1]NLE Lab, UPV, Spain

[2]IR-Lab, DA-IICT, India

May 27, 2012

Outline

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)

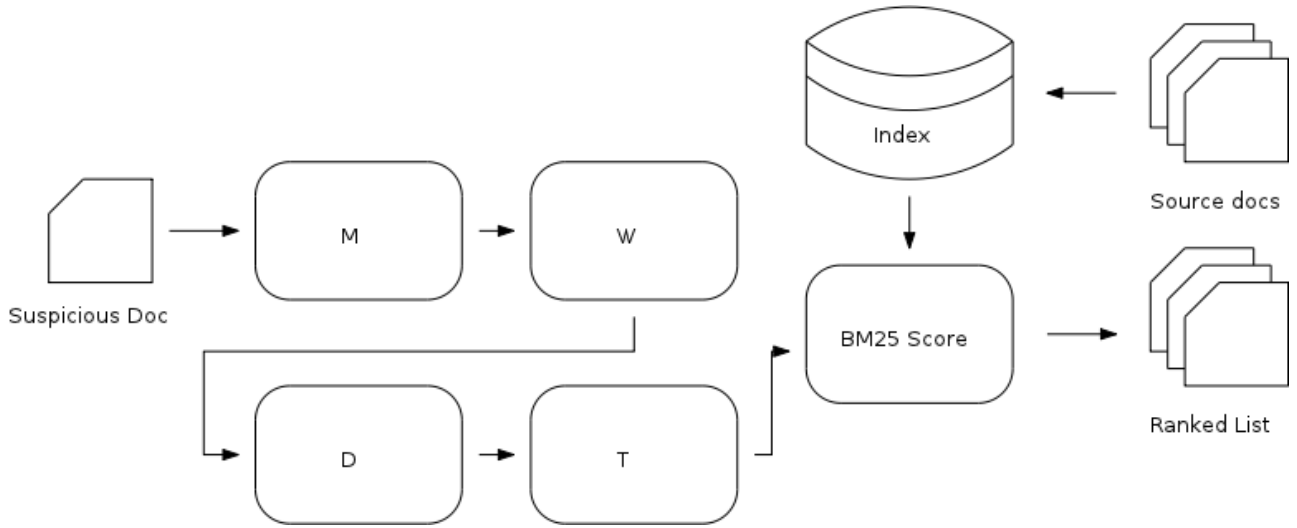
- Named entities are quite often multiword units
- Tagging 2275 English news articles, out of total 21,208 unique NEs: 9,079 (43%) were single word and 12,129 (57%) were multiword NEs.

Type	Frequent Length
Person	2
Location	2
Organisation	4

Approach

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

Text re-use document pairs identification [Gupta and Singhal2011]



Mining Module

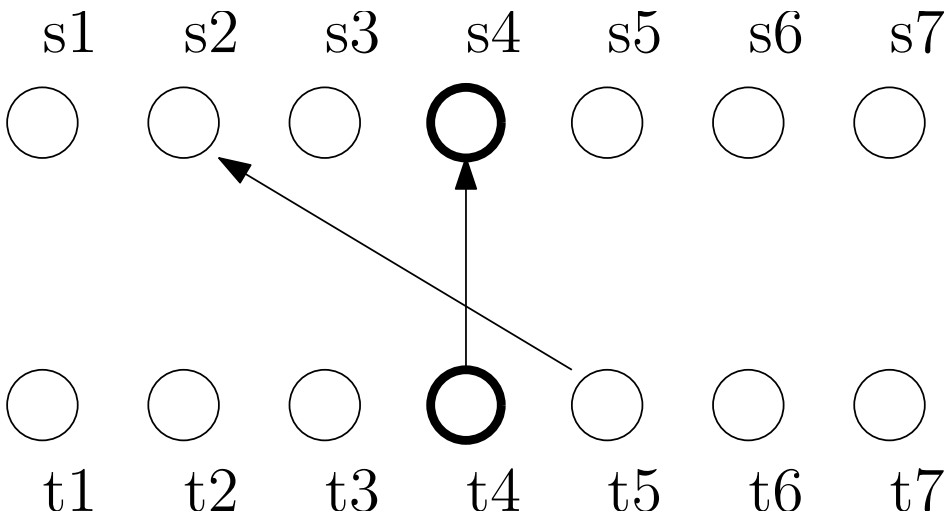
- **Input:** Candidate MW units S and T, where S and T are source and target MW units respectively.
  - where, $S = \{s_1, s_2, \cdots, s_N\}$ and $T = \{t_1, t_2, \cdots, t_M\}$
- **Output:** $< s', t' >$ which maximises $\zeta(.,.)$
  - where, s' and t' are sub-strings of S and T respectively
- Problem 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 = ... Government of Madras ...
  - T = ... मद्रास सरकार ... (*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"*)

CL!TR-2011-Dataset[1]

- 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[2]
- Transliteration Model - Google Transliterate API[3]

[1] http://users.dsic.upv.es/grupos/nle/fire-workshop-clitr.html

[2] http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/index.php

[3] http://code.google.com/apis/language/

Results

Type	Precision	Recall	F-Measure
FM	0.57	0.38	0.49
FM+PM	0.86	0.57	0.69

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

English	Hindi
Sawai Madhopur	स्वाई माधोपुर
DSIR model	डीएसआईआर मॉडल
Government of Madras	मद्रास सरकार
Kashgar Ladakh	कशगर लद्दाख
Medical Board	मेडिकल बोर्ड

Partially Identified

English	Hindi
Ranthambore National Park	रणथंभोर राष्ट्रीय
Meenakshi Amman	मीनाक्शी अम्मान
India Company	इन्डिया कंपनी
administered Gilgit	प्रशासित गिलगित

False Positive

English	Hindi
computing these values	मूल्यों कंप्यूटिंग
year the Sikhs	वर्ष सिखों
probability of b	प्रयिकता बी
where b	जहां बी

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.

Language Langue Linguaggio Языка Sprache Lengua Linguagem

NLEL Natural Language Engineering Lab

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.
In *Working notes of Forum for Information Retrieval Evaluation (FIRE 2011)*, pages 131–139, Mumbai, India, December.

Abhijit Bhole, Goutham Tholpadi, and Raghavendra Udupa.
2011.
Mining multi-word named entity equivalents from comparable corpora.
In *Proceedings of the 3rd Named Entities Workshop (NEWS 2011)*, pages 65–72, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Parth Gupta and Khushboo Singhal.
2011.
Mapping hindi-english text re-use document pairs.
In *Working notes of Forum for Information Retrieval Evaluation (FIRE 2011)*, pages 141–146, Mumbai, India, December.