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

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

[1] Natural Language Engineering Lab - ELiRF
Department of Information Systems and Computation
Universidad Politécnica de Valencia, Spain
http://users.dsic.upv.es/grupos/nle
{pgupta,prosso}@dsic.upv.es

[2] IR-Lab, DA-IICT, India.
http://irlab.daiict.ac.in
khushboo_singhal@daiict.ac.in

Abstract

In practice, many named entities (NEs) are multiword. Most of the research, done on mining the NEs from the comparable corpora, is focused on the single word transliterated NEs. This work presents an approach to mine Multiword Named Entities (MWNEs) from the text re-use document pairs. Text re-use, at document level, can be seen as noisy parallel or comparable text based on the level of obfuscation. Results, reported for Hindi-English language pair, are very encouraging. The approach can easily be extended to any language pair.

1. Introduction

Text re-use refers to using the text again from its original source. There are different situations which fall under the category of text re-use e.g. paraphrasing, quotation and copying (plagiarism). Moreover, text re-use is not limited to a single language, it can be cross-lingual in case of translated documents and cross-language plagiarism. Detection of such text re-use helps in various applications, e.g. checking the authenticity of the text, identifying near duplicates. Moreover, the identified document pairs can also be exploited for mining natural language resources. The difficulty of detection of re-use even increases when the source and target texts are in different languages which is called cross-language text re-use. There are two levels of text re-use:

1. Document level: The entire text of the document is re-used from some source, and

2. Fragment level: One or some of the sections of the document are containing re-used text.

Irrespective of the types and levels of the re-use, both, the source and target texts talk about the same concept with a high overlap in semantics and paraphrasing compared to an independent original work on the same topic. From now onward, we would talk in context of cross-language text re-use which can also be seen as noisy parallel or comparable text based on the level of obfuscation. This makes it more exploitable for mining the various cross-language resources like named entities, multiwords expression units, translation and transliteration probabilities.

Multiword units are very useful in many natural language processing (NLP) applications like multiword expressions for phrase based statistical machine translation (SMT) (Lambert and Banchs, 2005), MWNEs for cross-language news aggregation, finding NE equivalents in multilingual environment and measuring cross-language similarity for finding potential near-translation of the document from the multilingual corpora (Steinberger et al., 2002).

Named entities are very efficient elements in the cross language information retrieval (CLIR) and NLP applications like machine translation, machine transliteration, mention detection (Zitouni and Florian, 2008), news aggregation (Liu and Birnbaum, 2008) and plagiarism detection (Gupta et al., 2010). There have been many approaches for machine transliteration in order to find and use NEs in respective applications (Karimi et al., 2011). As suggested by Oh et al. (2006), the transliterations generated by the statistical methods are not often accurate, moreover, there can be more than one transliterations possible for a particular term. Therefore, it makes more sense to $mine$ the NEs from the readily available multilingual resources like parallel and comparable text. On a similar note, Udupa et al. (2008) and Klementiev and Roth (2006), both, attempt to mine NEs from a comparable multilingual corpora. A considerable amount of research has been done on the extraction of NEs from the comparable corpora, but most of the methods at the core are meant for the single word NEs and more specifically transliterated single word NEs. Bhole et al. (2011) suggested an approach to mine MWNEs from a comparable corpus, which can be seen as very close to the approach we propose in this paper. The key difference lies in the prior knowledge and the problem formulation, the former tries to formulate the problem as a conditional probability of target language MWNE alignment for the given source language MWNE, while we do not assume any prior knowledge of source language MWNE and pose the problem as joint probability estimation.

Though $enough$ mono-lingual, and to some extent cross-lingual, resources are available for Hindi-English, they are still not abundant to solve the general problems of NLP with high accuracy, compared to that achieved for some

peer English language pairs e.g. English-Spanish. This lag is due to the absence of sufficient parallel data and, to an extent, technological and cultural inadequacy for Hindi resource creation environment e.g. (less Hindi speakers prefer to use computers in Hindi and even less people use a Hindi keyboard). This makes it more important to exploit the present *poor* resources to the fullest.

The rest of the paper is structured as follows. Section 2 talks about the importance and challenges involved with MWNE. Section 3 presents the proposed approach in detail. In Section 4 we describe the corpus and report results with analysis. Finally in Section 5, we conclude the work with some future directions.

2. Motivation

A general observation gives an insight to the nature of NEs that many NEs are multiword e.g. name of a person with the surname (e.g. Barrak Obama), full name of an organization (e.g. Technical University of Valencia), city name with the country name (e.g. Valencia, Spain) and so on.

In order to understand the distribution and amount of the MWNEs, we tagged 2275 English news articles[1] using an English NE recogniser[2] (NER). Out of total 21,208 unique NEs: 9,079 (43%) were single word and 12,129 (57%) were multi-word NEs. This demonstrates the importance for explicit handling of the MWNEs.

Bhole et al. (2011) report the issues involved in finding the MWNEs and the nature of MWNEs. MWNEs are not merely a transliteration of terms, rather they may include translation, sequential shuffling, acronyms, one-to-many and many-to-one correspondence among the terms and so on.

The limitation of the conditional probability estimation based method is that the performance is dependent on the accuracy and efficiency of the source language NE recognition. To understand this phenomenon, we carried an experiment where we tagged the English documents using an English NER. We noticed that the NER identifies some of the NEs partly, for example "Bayes" instead of "Bayes Theorem", "Sundereshwara Temple" instead of "Meenakshi Sundereshwara Temple". In addition, there were many false positives. Finding the MWNEs in target language based on these source MWNEs will lead to a very noisy identification which needs to be handled by pruning. Therefore, we pose the problem of MWNE identification as the estimation of a joint probability for two string sequences being an MWNE pair.

3. Algorithm

First, the text re-use document pairs from the non-comparable source collection are found based on the standard CLIR methods of query translation. We consider the Hindi document as the query and retrieve the most similar English document from the indexed source collection. After fetching such pairs, we mine them to extract the MWNEs. For finding re-used document pairs, we use the system reported in (Gupta and Singhal, 2011).

3.1. Multiword Named Entity Extraction

First of all we find the transliteration match between the source and the target document. Suppose the terms s_{match} and t_{match} represent the corresponding terms of the transliteration match in the source and target documents respectively. Let $S = \{s_1, \cdots, s_N\}$ be a multiword unit including and around s_{match} of the source language (English) of length N and similarly, let $T = \{t_1, \cdots, t_M\}$ be the target language (Hindi) multiword unit including and around t_{match} of length M. The multiword pair <S,T> which maximises the Eq. 2 as shown below, is considered as MWNE.

$$\begin{aligned} &\max \zeta(S,T) \text{ subject to} \\ &\phi(S,T) = \min(N,M)\ , \\ &|N-M| \leq 1 \text{ and} \\ &\sum_{i=1}^{N}\sum_{j=1}^{M} T(s_i,t_j) \geq \theta \end{aligned} \tag{1}$$

where,

$$\zeta(S,T) = \ell_s(S) * \ell_t(T) * \sum_{i=1}^{N}\sum_{j=1}^{M} \delta(s_i,t_j) \tag{2}$$

$$\phi(S,T) = \sum_{i=1}^{N}\sum_{j=1}^{M} \psi(s_i,t_j) \tag{3}$$

$$\delta(s_i,t_j) = \begin{cases} D(s_i,t_j) & \text{if translation} \\ T(s_i,t_j) & \text{if transliteration} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$\psi(s_i,t_j) = 1 \qquad \text{if } \delta(s_i,t_j) \neq 0 \tag{5}$$

ℓ_s and ℓ_t define source and target language model respectively. $D(s_i,t_i) \neq 0$ when dictionary translation for term s_i is t_j, and accordingly $T(s_i,t_j)$ signifies the same for transliteration engine. We consider it an exception and ignore the term s_i when $D(s_i,t_j) = T(s_i,t_j)$, i.e. the translation and transliteration of term s_i is the term t_j. The values taken by $\delta(s_i,t_j)$ are normalised values. For a multiword unit to be a NE, at least one of its terms has to be a transliteration which is maintained by assigning the third condition in Eq. (1) where θ can be set accordingly.

Basically s_{match} and t_{match} help to locate the area of the document pair where the chances of finding an NE is very high. Then after, the approach selects the longest substring pair around s_{match} and t_{match} to be an MWNE pair using the above formulation.

4. Results and Analysis

We report the results of our proposed algorithm on the recently developed corpus called CL!TR-2011 which contains the cross-language text re-use documents of Hindi and English.

[1] Articles are taken from the year 2007 crawl of "The Telegraph" of section "frontpage", which can also be accessed through English Corpus of Forum for Information Retrieval Evaluation (FIRE) 2010.

[2] Alias-i. 2008. LingPipe 4.1.0. We use `cmd_ne_en_news_muc6.sh` script. `http://alias-i.com/lingpipe` (accessed June 25, 2010)

4.1. Corpus

The CL!TR-2011-Dataset[3] contains 190 Hindi documents and 5032 English documents. The documents in the corpus are generated from Wikipedia[4] and are related to the "tourism" and "computer science" domains. Table 1 contains the basic information about the corpus in terms of the size. More details about the corpus can be found in (Barrón-Cedeño et al., 2011).

Partition	$\|D\|$	$\|D_{tokens}\|$	$\|D_{types}\|$
D_{hi}	388	216 k	5 k
D_{en}	5032	9.3 M	644 k

Table 1: Statistics of the CL!TR-2011-Dataset. D_{hi} represents the Hindi document set and D_{en} represents the English document set. $|.|$ is the $size - of$ function.

The corpus contains four types of Hindi documents, mainly categorized by the amount of obfuscation of re-use, namely "Exact", "Heavy", "Light" and "None". The text re-use is through the machine translation with manual corrections. The "Exact" refers to the exact re-use of the text without any modifications, "Heavy" refers to the re-use with very less modifications, "Light" refers to the re-use with high modifications and "None" refers to no re-use.

4.2. Evaluation

Tables 2 and 3 summarize the performance of the text re-use document pair finding module. The configuration (morphological analyser (M) + bilingual dictionary (D) + transliteration (T)), which produced the best results, is used to retrieve the text re-use pair.

Method	**Precision**	**Recall**	**F-Measure**
M+D+T	0.695	0.904	0.786

Table 2: Performance of finding text re-use document pairs on test data. M+D+T signifies the combination of morphological analyser, bilingual dictionary and transliteration for query translation.

Type	Exact	Heavy	Light
Recall	1.0000	0.9070	0.8551

Table 3: Performance evaluation based on different levels of re-use.

In order to extract the MWNEs from the identified document pairs, we give the re-use document pairs of type "Exact", which are 34 in total, to the MWNE extraction module. For the evaluation of MWNE extraction module, we manually identify MWNE pairs in these 34 re-used document pairs, which serves as the gold standard. We limit our evaluation to only type "Exact" because in this preliminary study we wish to investigate the behaviour of the algorithm in a smaller controlled environment.

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

[4] http://www.wikipedia.org/

4.3. Analysis

We use the Universal Word (UW)[5] Version 3.1 Hindi-English bilingual dictionary to represent the $D(s_i, t_j)$ and use Google Transliterate API[6] to represent the $T(s_i, t_j)$ in Eq. (4). The language model for the source and the target languages are computed on the respective language subsets of the CL!TR-2011-Dataset. Results obtained for the MWNE extraction algorithm, are reported in Table 4. We consider two types of results, full match (FM): where a complete MWNE is identified and, partial match (PM): where a part of the MWNE is identified.

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

Table 4: Performance evaluation of MWNE extraction algorithm on Hindi-English language pair. FM is full match and FM+PM is full and partial match.

The corpus contains some of the documents related to "Computer Science" domain, which have some small scientific notations in the text, such as, $P(b|a)$. These notations are identified by the algorithm as an MWNE, and hence hurt the precision. Some examples of this phenomenon along with other false positives are depicted in 7. We take the transliteration engine as a binary model i.e the exact transliteration is considered, though the algorithm is capable to handle continuous values. Hence, near transliterations are missed, which in turn, hurts the recall. The reported results are for multiword NEs and we do not consider single-word NEs for the experimentation. $\phi(S, T)$ in Eq. (1) keeps an account of the number of term pairs contributing to the final score, applying this as a condition in Eq. (1) helps to determine the boundary of the MWNE. The language model helps in voting out the false positives in terms of unnecessary translation match, which is not a part of the MWNE, for example, in case of (English: "Indo Aryan and", Hindi: "इन्डो आर्यन और") the trailing "the" is removed providing the tighter boundaries.

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

Table 5: Examples of correctly identified full MWNEs.

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

Table 6: Examples of partially identified MWNEs.

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

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

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

Table 7: Examples of false positive MWNEs.

Tables 5, 6 and 7 depict some of the fully, partially and falsely identified MWNEs respectively. We further investigated the language model voting for the partially identified MWNEs and learnt that the performance can be increased if the language model is trained on a larger but related domain corpora. This algorithm can also easily be extended to other languages, provided the translation model and the transliteration model between the desired language pair and the language models for both the languages are available. Moreover, in the absence of the translation and transliteration models between the desired pair of languages, pivot language based strategy can very easily be incorporated in this approach.

5. Conclusion and Future Work

We are able to suggest a new approach to mine MWNE equivalents from text re-use pair of documents successfully. The approach of jointly estimating MWNEs for a language pair, without having prior knowledge of MWNEs in either of them. The preliminary investigation gives encouraging results for Hindi-English. The algorithm can easily be adapted for the distant language pairs, for which, many cross-language resources are not available directly, but which share a common resource rich pivot language. In future, we intend to evaluate this approach on such language pairs. Though the evaluation is carried on the text re-use document pairs without obfuscation which in nature is noisy parallel text, we believe the algorithm can be extended to the comparable corpora, which we intend to investigate in future.

6. Acknowledgment

The work of the researchers of the Technical University of Valencia has been done in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems and it has been partially funded by the European Commission as part of the WIQ-EI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework, and by MICINN as part of the Text-Enterprise 2.0 project (TIN2009-13391-C04-03) within the Plan I+D+i.

7. References

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.

Parth Gupta, Sameer Rao, and Prasenjit Majumder. 2010. External plagiarism detection: N-gram approach using named entity recognizer - lab report for pan at clef 2010. In *CLEF (Notebook Papers/LABs/Workshops)*.

Sarvnaz Karimi, Falk Scholer, and Andrew Turpin. 2011. Machine transliteration survey. *ACM Comput. Surv.*, 43(3):17:1–17:46, April.

Alexandre Klementiev and Dan Roth. 2006. Named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 82–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Patrik Lambert and Rafael Banchs. 2005. Data inferred multi-word expressions for statistical machine translation. In *Proc. of Machine Translation Summit X*, pages 396–403, Phuket, Thailand.

Jiahui Liu and Larry Birnbaum. 2008. What do they think?: aggregating local views about news events and topics. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 1021–1022, New York, NY, USA. ACM.

Jong-Hoon Oh, Key-Sun Choi, and Hitoshi Isahara. 2006. A comparison of different machine transliteration models. *J. Artif. Int. Res.*, 27:119–151, October.

Ralf Steinberger, Bruno Pouliquen, and Johan Hagman. 2002. Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc. In *CICLing*, pages 415–424.

Raghavendra Udupa, K. Saravanan, A. Kumaran, and Jagadeesh Jagarlamudi. 2008. Mining named entity transliteration equivalents from comparable corpora. In *CIKM*, pages 1423–1424.

Imed Zitouni and Radu Florian. 2008. Mention detection crossing the language barrier. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 600–609, Stroudsburg, PA, USA. Association for Computational Linguistics.