Mining Named Entity Transliteration Equivalents from Comparable Corpora

Raghavendra Udupa K Saravanan A Kumaran Jagadeesh Jagarlamudi
Microsoft Research India
Bangalore, INDIA 560080
+91.80.6658.6000
raghavu@microsoft.com

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval.

General Terms

Algorithms, Experimentation.

Keywords

Mining, Named Entities, Named Entity Translation Equivalents, Cross-Language Information Retrieval, Comparable Corpora.

1. INTRODUCTION

Named Entities (NEs) form a significant fraction of query terms in Information Retrieval (IR) systems and have a substantial impact on their retrieval performance. NEs are even more important in Cross Language Information Retrieval (CLIR), as in addition to being a significant component of query terms, any errors in their translations act as noise affecting adversely the retrieval performance (Mandl and Womser-Hacker, 2005, Xu and Weischedel, 2005). From the resource side for CLIR, bilingual dictionaries typically offer only limited support as they do not have sufficient coverage of NEs, as new NEs are introduced to the vocabulary of a language every day. On the other hand, machine transliteration systems often produce misspelled or incorrect transliterations affecting the CLIR retrieval performance.

In recent times, the large quantity and the perpetual availability of news corpora in many of the world's languages simultaneously, have spurred interest in a promising alternative to NE translation or transliteration, namely, the *mining* of Named Entity Transliteration Equivalents (NETEs) from such news corpora (Klementiev and Roth, 2006; Tao et al., 2006). Formally, comparable news corpora are time-aligned news stories in a pair of languages from a reasonably long period in time. NETEs mined from comparable news corpora could be valuable in many tasks such as CLIR and MT, to effectively complement the bilingual dictionaries and the machine transliteration systems. This opportunity is precisely what we address in our work.

We introduce a novel method, called **MINT** (**MI**ning **N**amed-entity **T**ransliteration equivalents), with the following innovations for effective mining of NETEs from comparable corpora:

- MINT relies on little linguistic resources, requiring a Named Entity Recognizer (NER) in only one language; hence NETEs from even a resource poor language may be mined, when paired with a language where an NER is available.
- MINT does not require any language-specific knowledge, and hence may be employed effectively in many language pairs.
- MINT does not rely on frequency statistics of NEs, and hence may be applied to even the infrequent named entities, which form the vast majority of NEs found in news corpora.
- Finally, MINT is computationally efficient and may be used for mining large comparable corpora.

In this paper, we outline the details of the MINT method and our evaluation process to demonstrate its effectiveness and robustness on comparable corpora between a diverse set of languages, namely, English (En), Hindi (Hi) and Kannada (Ka), from three distinct language families.

2. MINT: MINING COMPARABLE DATA

The MINT method is based on a key insight that news articles in multiple languages with similar content contain highly overlapping set of NEs, and hence, may be expected to yield NETE richly. MINT has two stages, as shown in Figure 1; in the first stage, documents from the comparable corpora are compared to identify article pairs of similar content and in the second stage, NETEs are mined from these article pairs.

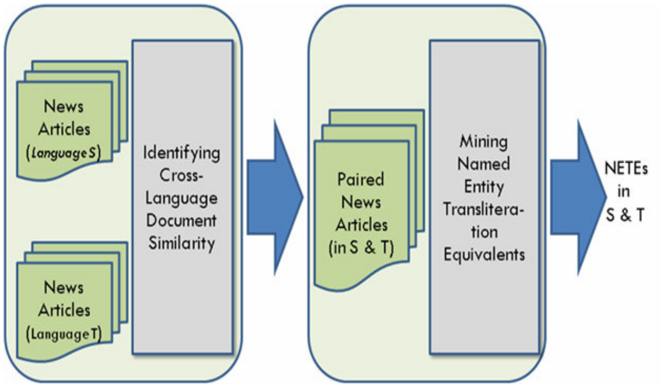
News Articles (Language S)
News Articles (Language T)
Identifying Cross-Language Document Similarity
Paired News Articles (in S & T)
Mining Named Entity Translitera-tion Equivalents
NETEs in S & T

Figure 1. Two Mining Stages of MINT Method

CIKM'08, October 26-30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

2.1 Identifying Similar Cross-language Documents

In the first stage of the MINT method, documents from the comparable corpora (C_S, C_T) in languages S and T are compared for content similarity to produce a collection $A_{S,T}$ of article pairs, namely (D_S, D_T), that have similar content. The cross-language document similarity model uses negative KL-divergence between the source and target document probability distributions, and measures the similarity between a given pair of documents. Given two documents D_S, D_T in source and target languages respectively, the cross-language document similarity between the two documents is given by

$$\sum_{w_T \in D_T} \sum_{w_S \in D_S} p(w_S \mid D_S) p(w_T \mid w_S) \log p(w_T \mid D_T) \cdot$$

2.2 Mining NETE from Similar Documents

The second stage of the MINT method works on each pair of articles (D_S, D_T) in the collection $A_{S,T}$ and produces a set $P_{S,T}$ consisting of ($\boldsymbol{\varepsilon_S}$, $\boldsymbol{\varepsilon_T}$) that are transliteration equivalents. The transliteration similarity between $\boldsymbol{\varepsilon_S}$ and $\boldsymbol{\varepsilon_T}$, is measured by the transliteration similarity model employing a logistic function, $\frac{1}{1+e^{-<w, \phi(\varepsilon_S, \varepsilon_T)>}}$, where $\phi(\boldsymbol{\varepsilon_S}, \boldsymbol{\varepsilon_T})$ is the feature vector that captures cross-language associations for the pair ($\boldsymbol{\varepsilon_S}$, $\boldsymbol{\varepsilon_T}$) (such as, the character sequences, couplings of substrings, monotonicity of alignment, lengths, etc.) and w is the weights vector, which is learnt discriminatively over a corpus of known transliterations.

3. EMPIRICAL EVALUATION

Our empirical investigation consists of experiments in three data environments, each with a different focus:

1. IDEAL: An environment to measure the effectiveness of Stage 2, in mining NETE from oracle-aligned news articles.
2. NEAR-IDEAL: An environment in which articles are aligned automatically in the stage 1 of MINT from a set of multilingual news articles that are known to have comparable articles, and such aligned articles are mined for NETEs. This environment defines an upper bound on MINT performance on realistic scenarios.
3. REAL: An environment in which NETEs are mined from large comparable corpora consisting of hundreds of thousands of news articles in a pair of languages, where even the existence of a comparable target language article for an article in the source language is not guaranteed.

The test bed for evaluating the IDEAL data environment consisted of 200 articles from a corpus consisting of ~2500 pairs of aligned articles. In NEAR-IDEAL, the same corpus without the pairing information was used. Finally, in REAL data environment, a test bed of 100 articles from a corpus of ~200K published articles in a pair of languages was used. Table 1 provides the Mean Reciprocal Rank (MRR) of the mined NETE in each of the three environments. We implemented a baseline system, CoRanking, along the lines of (Klementiev and Roth, 2006). We extracted NEs from the English articles using Stanford named entity recognizer (Finkel et al., 2005).

Table 1. Results of MINT Method

Environments	Language Pairs	MINT		CoRanking	
		MRR@1	MRR@5	MRR@1	MRR@5
IDEAL	En-Ka	0.94	0.95	0.26	0.26
	En-Hi	0.93	0.95	-	-
NEAR IDEAL	En-Ka	0.92	0.94	0.26	0.26
	En-Hi	0.82	0.87	-	-
REAL	En-Ka	0.86	0.88	-	-

The comparative results for the IDEAL and NEAR IDEAL environments shown in Table 1 indicate vast improvement over the baseline. In the IDEAL and NEAR IDEAL environments, the high MRRs of MINT indicate that nearly all the mined NETE were among the very first candidates output by the stage 2 of MINT method. The MRRs have been computed only for the source side named entities that have a valid transliteration in the target side article.

For identifying similar documents in Stage 1 of MINT we used a time window of 3 days for En-Ka language pair, in addition to the document similarity model; in En-Hi language pair, no time window was used, as the data set lacked publication dates for the articles. The lack of timestamp information prevented us from running the CoRanking algorithm on the En-Hi language pair. Even in the REAL environment, the MINT method is highly effective, even though it was run on corpora that are two orders of magnitude larger than that used in the IDEAL environment.

4. CONCLUSION

This paper shows that MINT, a simple and intuitive mining method employing cross-language document similarity and transliteration similarity models, is capable of mining NETEs effectively from comparable corpora. Our empirical investigation showed that MINT performs close to optimal on comparable corpora consisting of pairs of similar articles when the pairings are known a priori; MINT induces fairly good pairings and performs exceedingly well even when the pairings are not known a priori.

5. REFERENCES

[1] Ballesteros, L. and Croft, B. 1998. Dictionary Methods for Cross-Lingual Information Retrieval. *Proc. of DEXA'96.*

[2] Finkel, J. R., Grenager, T. and Manning, C. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proc. of the 43nd Annual Meeting of the ACL.*

[3] Klementiev, A. and Roth, D. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. *Proc. of the 44th Annual Meeting of the ACL.*

[4] Mandl, T., and Womser-Hacker, C. 2005. The Effect of named entities on effectiveness in cross-language information retrieval evaluation. *ACM Symposium on Applied Computing*.

[5] Tao, T., Yoon, S., Fister, A., Sproat, R. and Zhai, C. 2006. Unsupervised named entity transliteration using temporal and phonetic correlation. *Proc. of EMNLP 2006.*

[6] Xu, J. and Weischedel, R. 2005. Empirical studies on the impact of lexical resources on CLIR performance. *Information Processing and Management.*