MINT: A Method for Effective and Scalable Mining of Named Entity Transliterations from Large Comparable Corpora

Raghavendra Udupa **K Saravanan** **A Kumaran** **Jagadeesh Jagarlamudi***
Microsoft Research India
Bangalore 560080 INDIA
[raghavu,v-sarak,kumarana,jags}@microsoft.com

Abstract

In this paper, we address the problem of mining transliterations of Named Entities (NEs) from large comparable corpora. We leverage the empirical fact that multilingual news articles with similar news content are rich in Named Entity Transliteration Equivalents (NETEs). Our mining algorithm, MINT, uses a cross-language document similarity model to align multilingual news articles and then mines NETEs from the aligned articles using a transliteration similarity model. We show that our approach is highly effective on 6 different comparable corpora between English and 4 languages from 3 different language families. Furthermore, it performs substantially better than a state-of-the-art competitor.

1 Introduction

Named Entities (NEs) play a critical role in many Natural Language Processing and Information Retrieval (IR) tasks. In Cross-Language Information Retrieval (CLIR) systems, they play an even more important role as the accuracy of their transliterations is shown to correlate highly with the performance of the CLIR systems (Mandl and Womser-Hacker, 2005, Xu and Weischedel, 2005). Traditional methods for transliterations have not proven to be very effective in CLIR. Machine Transliteration systems (AbdulJaleel and Larkey, 2003; Al-Onaizan and Knight, 2002; Virga and Khudanpur, 2003) usually produce incorrect transliterations and translation lexcions such as hand-crafted or statistical dictionaries are too static to have good coverage of NEs[1] occurring in the current news events. Hence, there is a critical need for creating and continually updating multilingual Named Entity transliteration lexicons.

The ubiquitous availability of comparable news corpora in multiple languages suggests a promising alternative to Machine Transliteration, namely, the *mining* of Named Entity Transliteration Equivalents (NETEs) from such corpora. News stories are typically rich in NEs and therefore, comparable news corpora can be expected to contain NETEs (Klementiev and Roth, 2006; Tao et al., 2006). The large quantity and the perpetual availability of news corpora in many of the world's languages, make mining of NETEs a viable alternative to traditional approaches. It is this opportunity that we address in our work.

In this paper, we detail an effective and scalable mining method, called **MINT** (**MI**ning **N**amed-entity **T**ransliteration equivalents), for mining of NETEs from large comparable corpora. MINT addresses several challenges in mining NETEs from large comparable corpora: exhaustiveness (in mining sparse NETEs), computational efficiency (in scaling on corpora size), language independence (in being applicable to many language pairs) and linguistic frugality (in requiring minimal external linguistic resources).

Our contributions are as follows:

- We give empirical evidence for the hypothesis that news articles in different languages with reasonably similar content are rich sources of NETEs (Udupa, et al., 2008).
- We demonstrate that the above insight can be translated into an effective approach for mining NETEs from large comparable corpora even when similar articles are not known a priori.
- We demonstrate MINT's effectiveness on 4 language pairs involving 5 languages (English, Hindi, Kannada, Russian, and Tamil) from 3 different language families, and its scalability on corpora of vastly different sizes (2,000 to 200,000 articles).
- We show that MINT's performance is significantly better than a state of the art method (Klementiev and Roth, 2006).

* Currently with University of Utah.

[1] New NEs are introduced to the vocabulary of a language every day. On an average, 260 and 452 new NEs appeared daily in the XIE and AFE segments of the LDC English Gigaword corpora respectively.

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 799–807,
Athens, Greece, 30 March – 3 April 2009.

Figure 1. Comparable Corpora

We discuss the motivation behind our approach in Section 2 and present the details in Section 3. In Section 4, we describe the evaluation process and in Section 5, we present the results and analysis. We discuss related work in Section 6.

2 Motivation

MINT is based on the hypothesis that news articles in different languages with similar content contain highly overlapping set of NEs. News articles are typically rich in NEs as news is about events involving people, locations, organizations, etc[2]. It is reasonable to expect that multilingual news articles reporting the same news event mention the same NEs in the respective languages. For instance, consider the English and Hindi news reports from the *New York Times* and the *BBC* on the second oath taking of President Barack Obama (Figure 1). The articles are not parallel but discuss the same event. Naturally, they mention the same NEs (such as Barack Obama, John Roberts, White House) in the respective languages, and hence, are rich sources of NETEs.

Our empirical investigation of comparable corpora confirmed the above insight. A study of 200 pairs of similar news articles published by *The New Indian Express* in 2007 in English and Tamil showed that 87% of the single word NEs in the English articles had at least one transliteration equivalent in the conjugate Tamil articles. The MINT method leverages this empirically backed insight to mine NETEs from such comparable corpora.

However, there are several challenges to the mining process: firstly, vast majority of the NEs in comparable corpora are very sparse; our analysis showed that 80% of the NEs in *The New Indian Express* news corpora appear less than 5 times in the entire corpora. Hence, any mining method that depends mainly on repeated occurrences of the NEs in the corpora is likely to miss vast majority of the NETEs. Secondly, the mining method must restrict the candidate NETEs that need to be examined for match to a reasonably small number, not only to minimize false positives but also to be computationally efficient. Thirdly, the use of linguistic tools and resources must be kept to a minimum as resources are available only in a handful of languages. Finally, it is important to use as little language-specific knowledge as possible in order to make the mining method applicable across a vast majority of languages of the world. The MINT method proposed in this paper addresses all the above issues.

[2] News articles from the BBC corpus had, on an average, 12.9 NEs and new articles from the *The New Indian Express,* about 11.8 NEs.

3 The MINT Mining Method

MINT has two stages. In the first stage, for every document in the source language side, the set of documents in the target language side with similar news content are found using a cross-language document similarity model. In the second stage, the NEs in the source language side are extracted using a Named Entity Recognizer (NER) and, subsequently, for each NE in a source language document, its transliterations are mined from the corresponding target language documents. We present the details of the two stages of MINT in the remainder of this section.

3.1 Finding Similar Document Pairs

The first stage of MINT method (Figure 2) works on the documents from the comparable corpora ($\mathcal{C}_S$, $\mathcal{C}_T$) in languages $\mathcal{S}$ and $\mathcal{T}$ and produces a collection $\mathcal{A}_{S,T}$ of similar article pairs ($\mathcal{D}_S$, $\mathcal{D}_T$). Each article pair ($\mathcal{D}_S$, $\mathcal{D}_T$) in $\mathcal{A}_{S,T}$ consists of an article ($\mathcal{D}_S$) in language $\mathcal{S}$ and an article ($\mathcal{D}_T$) in language $\mathcal{T}$, that have similar content. The cross-language similarity between $\mathcal{D}_S$ and $\mathcal{D}_T$, as measured by the cross-language similarity model $\mathcal{MD}$, is at least $\alpha > 0$.

Cross-language Document Similarity Model: The cross-language document similarity model measures the degree of similarity between a pair of documents in source and target languages. We use the negative KL-divergence between source and target document probability distributions as the similarity measure.

Input: Comparable news corpora ($\mathcal{C}_S$, $\mathcal{C}_T$) in languages ($\mathcal{S}$,$\mathcal{T}$)
Crosslanguage Document Similarity Model $\mathcal{MD}$ for ($\mathcal{S}$, $\mathcal{T}$)
Threshold score α.
Output: Set $\mathcal{A}_{S,T}$ of pairs of similar articles ($\mathcal{D}_S$, $\mathcal{D}_T$) from ($\mathcal{C}_S$, $\mathcal{C}_T$).
1 $\mathcal{A}_{S,T} \leftarrow \phi$; // Set of Similar articles ($\mathcal{D}_S$, $\mathcal{D}_T$)
2 **for** each article $\mathcal{D}_S$ in $\mathcal{C}_S$ **do**
3 $\mathcal{X}_S \leftarrow \phi$; // Set of candidates for $\mathcal{D}_S$.
4 **for** each article $\mathcal{d}_T$ in $\mathcal{C}_T$ **do**
5 score = CrossLanguageDocumentSimilarity($\mathcal{D}_S$,$\mathcal{d}_T$,$\mathcal{MD}$);
6 **if** (score ≥ α) **then** $\mathcal{X}_S \leftarrow \mathcal{X}_S \cup (\mathcal{d}_T$, score) ;
7 **end**
8 $\mathcal{D}_T$ = BestScoringCandidate($\mathcal{X}_S$);
9 **if** ($\mathcal{D}_T \neq \phi$) **then** $\mathcal{A}_{S,T} \leftarrow \mathcal{A}_{S,T} \cup (\mathcal{D}_S, \mathcal{D}_T)$;
10 **end**

CrossLanguageSimilarDocumentPairs

Figure 2. Stage 1 of MINT

Given two documents $\mathcal{D}_S$, $\mathcal{D}_T$ in source and target languages respectively, with V_S, V_T denoting the vocabulary of source and target languages, the similarity between the two documents is given by the KL-divergence measure, -KL($\mathcal{D}_S$ || $\mathcal{D}_T$), as:

$$\sum_{w_T \in V_T} p(w_T \mid D_S) \log \frac{p(w_T \mid D_T)}{p(w_T \mid D_S)}$$

where p(w | D) is the likelihood of word w in D. As we are interested in target documents which are similar to a given source document, we can ignore the numerator as it is independent of the target document. Finally, expanding $p(w_T \mid D_s)$ as $\sum_{w_S \in V_S} p(w_S \mid D_S) p(w_T \mid w_S)$ we specify the cross-language similarity score as follows:

Cross-language similarity =
$$\sum_{w_T \in V_T} \sum_{w_S \in V_S} p(w_S \mid D_S) p(w_T \mid w_S) \log p(w_T \mid D_T)$$

3.2 Mining NETEs from Document Pairs

The second stage of the MINT method works on each pair of articles ($\mathcal{D}_S$, $\mathcal{D}_T$) in the collection $\mathcal{A}_{S,T}$ and produces a set $\mathcal{P}_{S,T}$ of NETEs. Each pair (ε_S, ε_T) in $\mathcal{P}_{S,T}$ consists of an NE ε_S in language $\mathcal{S}$, and a token ε_T in language $\mathcal{T}$, that are transliteration equivalents of each other. Furthermore, the transliteration similarity between ε_S and ε_T, as measured by the transliteration similarity model $\mathcal{MT}$, is at least $\beta > 0$. Figure 3 outlines this algorithm.

Discriminative Transliteration Similarity Model:
The transliteration similarity model $\mathcal{MT}$ measures the degree of transliteration equivalence between a source language and a target language term.

Input:
Set $\mathcal{A}_{S,T}$ of similar documents ($\mathcal{D}_S$, $\mathcal{D}_T$) in languages ($\mathcal{S}$,$\mathcal{T}$),
Transliteration Similarity Model $\mathcal{MT}$ for ($\mathcal{S}$, $\mathcal{T}$),
Threshold score β.
Output: Set $\mathcal{P}_{S,T}$ of NETEs (ε_S, ε_T) from $\mathcal{A}_{S,T}$;
1 $\mathcal{P}_{S,T} \leftarrow \phi$;
2 **for** each pair of articles ($\mathcal{D}_S$, $\mathcal{D}_T$) in $\mathcal{A}_{S,T}$ **do**
3 **for** each named entity ε_S in $\mathcal{D}_S$ **do**
4 $\mathcal{Y}_S \leftarrow \phi$; // Set of candidates for ε_S.
5 **for** each candidate $\mathcal{e}_T$ in $\mathcal{D}_T$ **do**
6 score = TransliterationSimilarity(ε_S, $\mathcal{e}_T$,$\mathcal{MT}$) ;
7 **if** (score ≥ β) **then** $\mathcal{Y}_S \leftarrow \mathcal{Y}_S \cup (\mathcal{e}_T$, score) ;
8 **end**
9 ε_T = BestScoringCandidate($\mathcal{Y}_S$) ;
10 **if** ($\varepsilon_T \neq$ null) **then** $\mathcal{P}_{S,T} \leftarrow \mathcal{P}_{S,T} \cup (\varepsilon_S, \varepsilon_T)$;
11 **end**
12 **end**

TransliterationEquivalents

Figure 3. Stage 2 of MINT

We employ a logistic function as our transliteration similarity model $\mathcal{MT}$, as follows:

$$\text{TransliterationSimilarity}(\varepsilon_S, e_T, \mathcal{MT}) = \frac{1}{1+e^{-w^t \cdot \phi(\varepsilon_S, e_T)}}$$

where $\phi(\varepsilon_S, e_T)$ is the feature vector for the pair (ε_S, e_T) and w is the weights vector. Note that the transliteration similarity takes a value in the range [0..1]. The weights vector w is learnt discriminatively over a training corpus of known transliteration equivalents in the given pair of languages.

Features: The features employed by the model capture interesting cross-language associations observed in (ε_S, e_T):

- All unigrams and bigrams from the source and target language strings.
- Pairs of source string n-grams and target string n-grams such that difference in the start positions of the source and target n-grams is at most 2. Here $n \in \{1,2\}$.
- Difference in the lengths of the two strings.

Generative Transliteration Similarity Model: We also experimented with an extension of He's W-HMM model (He, 2007). The transition probability depends on both the jump width and the previous source character as in the W-HMM model. The emission probability depends on the current source character and the previous target character unlike the W-HMM model (Udupa et al., 2009). Instead of using any single alignment of characters in the pair (w_S, w_T), we marginalize over all possible alignments:

$$P(t_1^m \mid s_1^n) = \sum_A \prod_{j=1}^{m} p(a_j \mid a_{j-1}, s_{a_{j-1}}) p(t_j \mid s_{a_j}, t_{j-1})$$

Here, t_j (and resp. s_i) denotes the jth (and resp. ith) character in w_T (and resp. w_S) and $A \equiv a_1^m$ is the hidden alignment between w_T and w_S where t_j is aligned to s_{a_j}, $j = 1, \cdots, m$. We estimate the parameters of the model using the EM algorithm. The transliteration similarity score of a pair (w_S, w_T) is log P(w_T | w_S) appropriately transformed.

4 Experimental Setup

Our empirical investigation consists of experiments in three data environments, with each environment providing answer to specific set of questions, as listed below:

1. **Ideal Environment (IDEAL)**: Given a collection $\mathcal{A}_{S,T}$ of oracle-aligned article pairs ($\mathcal{D}_S$, $\mathcal{D}_T$) in $\mathcal{S}$ and $\mathcal{T}$, how effective is Stage 2 of MINT in mining NETE from $\mathcal{A}_{S,T}$?
2. **Near Ideal Environment (**NEAR-IDEAL**)**: Let $\mathcal{A}_{S,T}$ be a collection of *similar* article pairs ($\mathcal{D}_S$, $\mathcal{D}_T$) in $\mathcal{S}$ and $\mathcal{T}$. Given comparable corpora ($\mathcal{C}_S$, $\mathcal{C}_T$) consisting of only articles from $\mathcal{A}_{S,T}$, but without the knowledge of pairings between the articles,
   a. How effective is Stage 1 of MINT in recovering $\mathcal{A}_{S,T}$ from ($\mathcal{C}_S$, $\mathcal{C}_T$) ?
   b. What is the effect of Stage 1 on the overall effectiveness of MINT?
3. **Real Environment (REAL)**: Given large comparable corpora ($\mathcal{C}_S$, $\mathcal{C}_T$), how effective is MINT, end-to-end?

The IDEAL environment is indeed ideal for MINT since every article in the comparable corpora is paired with exactly one similar article in the other language and the pairing of articles in the comparable corpora is known in advance. We want to emphasize here that such corpora are indeed available in many domains such as technical documents and interlinked multilingual Wikipedia articles. In the IDEAL environment, only Stage 2 of MINT is put to test, as article alignments are given.

In the NEAR-IDEAL data environment, every article in the comparable corpora is known to have exactly one conjugate article in the other language though the pairing itself is not known in advance. In such a setting, MINT needs to discover the article pairing before mining NETEs and therefore, both stages of MINT are put to test. The best performance possible in this environment should ideally be the same as that of IDEAL, and any degradation points to the shortcoming of the Stage 1 of MINT. These two environments quantify the stage-wise performance of the MINT method.

Finally, in the data environment REAL, we test MINT on large comparable corpora, where even the existence of a conjugate article in the target side for a given article in the source side of the comparable corpora is not guaranteed, as in

any normal large multilingual news corpora. In this scenario both the stages of MINT are put to test. This is the toughest, and perhaps the typical setting in which MINT would be used.

4.1 Comparable Corpora

In our experiments, the source language is English whereas the 4 target languages are from three different language families (Hindi from the Indo-Aryan family, Russian from the Slavic family, Kannada and Tamil from the Dravidian family). Note that none of the five languages use a common script and hence identification of cognates, spelling variations, suffix transformations, and other techniques commonly used for closely related languages that have a common script are not applicable for mining NETEs. Table 1 summarizes the 6 different comparable corpora that were used for the empirical investigation; 4 for the IDEAL and NEAR-IDEAL environments (in 4 language pairs), and 2 for the REAL environment (in 2 language pairs).

Corpus	Source - Target	Data Environment	Articles (in Thousands)		Words (in Millions)	
			Src	Tgt	Src	Tgt
EK-S	English-Kannada	IDEAL& NEAR-IDEAL	2.90	2.90	0.42	0.34
ET-S	English-Tamil	IDEAL& NEAR-IDEAL	2.90	2.90	0.42	0.32
ER-S	English-Russian	IDEAL& NEAR-IDEAL	2.30	2.30	1.03	0.40
EH-S	English-Hindi	IDEAL& NEAR-IDEAL	11.9	11.9	3.77	3.57
EK-L	English-Kannada	REAL	103.8	111.0	27.5	18.2
ET-L	English-Tamil	REAL	103.8	144.3	27.5	19.4

Table 1: Comparable Corpora

The corpora can be categorized into two separate groups, group S (for *Small*) consisting of EK-S, ET-S, ER-S, and EH-S and group L (for *Large*) consisting of EK-L and ET-L. Corpora in group S are relatively small in size, and contain pairs of articles that have been judged by human annotators as similar. Corpora in group L are two orders of magnitude larger in size than those in group S and contain a large number of articles that may not have conjugates in the target side. In addition the pairings are unknown even for the articles that have conjugates. All comparable corpora had publication dates, except EH-S, which is known to have been published over the same year.

The EK-S, ET-S, EK-L and ET-L corpora are from *The New Indian Express* news paper, whereas the EH-S corpora are from *Web Dunia* and the ER-S corpora are from *BBC/Lenta News Agency* respectively.

4.2 Cross-language Similarity Model

The cross-language document similarity model requires a bilingual dictionary in the appropriate language pair. Therefore, we generated statistical dictionaries for 3 language pairs (from parallel corpora of the following sizes: 11K sentence pairs in English-Kannada, 54K in English-Hindi, and 14K in English-Tamil) using the GIZA++ statistical alignment tool (Och et al., 2003), with 5 iterations each of IBM Model 1 and HMM. We did not have access to an English-Russian parallel corpus and hence could not generate a dictionary for this language pair. Hence, the NEAR-IDEAL experiments were not run for the English-Russian language pair.

Although the coverage of the dictionaries was low, this turned out to be not a serious issue for our cross-language document similarity model as it might have for topic based CLIR (Ballesteros and Croft, 1998). Unlike CLIR, where the query is typically smaller in length compared to the documents, in our case we are dealing with news articles of comparable size in both source and target languages.

When many translations were available for a source word, we considered only the top-4 translations. Further, we smoothed the document probability distributions with collection frequency as described in (Ponte and Croft, 1998).

4.3 Transliteration Similarity Model

The transliteration similarity models for each of the 4 language pairs were produced by learning over a training corpus consisting of about 16,000 single word NETEs, in each pair of languages. The training corpus in English-Hindi, English-Kannada and English-Tamil were hand-crafted by professionals, the English-Russian name pairs were culled from Wikipedia interwiki links and were cleaned heuristically. Equal number of negative samples was used for training the models. To produce the negative samples, we paired each source language NE with a random non-matching target language NE. No language specific features were used and the same feature set was used in each of the 4 language pairs making MINT language neutral.

In all the experiments, our source side language is English, and the Stanford Named Entity Recognizer (Finkel et al, 2005) was used to extract NEs from the source side article. It should be noted here that while the precision of the NER

used was consistently high, its recall was low, (~40%) especially in the *New Indian Express* corpus, perhaps due to the differences in the data used for training the NER and the data on which we used it.

4.4 Performance Measures

Our intention is to measure the effectiveness of MINT by comparing its performance with the oracular (human annotator) performance. As transliteration equivalents must exist in the paired articles to be found by MINT, we focus only on those NEs that actually have at least one transliteration equivalent in the conjugate article.

Three performance measures are of interest to us: the fraction of distinct NEs from source language for which we found at least one transliteration in the target side (Recall on distinct NEs), the fraction of distinct NETEs (Recall on distinct NETEs) and the Mean Reciprocal Rank (MRR) of the NETEs mined. Since we are interested in mining not only the highly frequent but also the infrequent NETEs, recall metrics measure how effective our method is in mining NETEs exhaustively. The MRR score indicates how effective our method is in preferring the correct ones among candidates.

To measure the performance of MINT, we created a test bed for each of the language pairs. The test beds are summarized in Table 2.

The test beds consist of pairs of similar articles in each of the language pairs. It should be noted here that as transliteration equivalents must exist in the paired articles to be found by MINT, we focus only on those NEs that actually have at least one transliteration equivalent in the conjugate article.

5 Results & Analysis

In this section, we present qualitative and quantitative performance of the MINT algorithm, in mining NETEs from comparable news corpora. All the results in Sections 5.1 to 5.3 were obtained using the discriminative transliteration similarity model described in Section 3.2. The results using the generative transliteration similarity model are discussed in Section 5.4.

5.1 IDEAL Environment

Our first set of experiments investigated the effectiveness of Stage 2 of MINT, namely the mining of NETEs in an IDEAL environment. As MINT is provided with paired articles in this experiment, all experiments for this environment were run on test beds created from group S corpora (Table 2).

Test Bed	Comparable Corpora	Article Pairs	Distinct NEs	Distinct NETEs
EK-ST	EK-S	200	481	710
ET-ST	ET-S	200	449	672
EH-ST	EH-S	200	347	373
ER-ST	ER-S	100	195	347

Table 2: Test Beds for IDEAL & NEAR-IDEAL

Results in the IDEAL Environment:
The recall measures for distinct NEs and distinct NETEs for the IDEAL environment are reported in Table 3.

Test Bed	Recall (%)	
	Distinct NEs	Distinct NETEs
EK-ST	97.30	95.07
ET-ST	99.11	98.06
EH-ST	98.55	98.66
ER-ST	93.33	85.88

Table 3: Recall of MINT in IDEAL

Note that in the first 3 language pairs MINT was able to mine a transliteration equivalent for almost all the distinct NEs. The performance in English-Russian pair was relatively worse, perhaps due to the noisy training data.

In order to compare the effectiveness of MINT with a state-of-the-art NETE mining approach, we implemented the time series based Co-Ranking algorithm based on (Klementiev and Roth, 2006).

Test Bed	MRR@1		MRR@5	
	MINT	CoRanking	MINT	CoRanking
EK-ST	**0.94**	0.26	**0.95**	0.29
ET-ST	**0.91**	0.26	**0.94**	0.29
EH-ST	0.93	-	0.95	-
ER-ST	**0.80**	0.38	**0.85**	0.43

Table 4: MINT & Co-Ranking in IDEAL

Table 4 shows the MRR results in the IDEAL environment – both for MINT and the Co-Ranking baseline: MINT outperformed Co-Ranking on all the language pairs, despite not using time series similarity in the mining process. The high MRRs (@1 and @5) indicate that in almost all the cases, the top-ranked candidate is a correct NETE. Note that Co-Ranking could not be run on the EH-ST test bed as the articles did not have a date stamp. Co-Ranking is crucially dependent on time series and hence requires date stamps for the articles.

5.2 NEAR-IDEAL Environment

The second set of experiments investigated the effectiveness of Stage 1 of MINT on comparable corpora that are constituted by pairs of similar articles, where the pairing information between the articles is with-held. MINT reconstructed the pairings using the cross-language document similarity model and subsequently mined NETEs. As in previous experiments, we ran our experiments on test beds described in Section 4.4.

Results in the NEAR-IDEAL Environment: There are two parts to this set of experiments. In the first part, we investigated the effectiveness of the cross-language document similarity model described in Section 3.1. Since we know the identity of the conjugate article for every article in the test bed, and articles can be ranked according to the cross-language document similarity score, we simply computed the MRR for the documents identified in each of the test beds, considering only the top-2 results. Further, where available, we made use of the publication date of articles to restrict the number of target articles that are considered in lines 4 and 5 of the MINT algorithm in Figure 2. Table 5 shows the results for two date windows – 3 days and 1 year.

Test Bed	MRR@1		MRR@2	
	3 days	1 year	3 days	1 year
EK-ST	**0.99**	0.91	**0.99**	0.93
ET-ST	**0.96**	0.83	**0.97**	0.87
EH-ST	-	0.81	-	0.82

Table 5: MRR of Stage 1 in NEAR-IDEAL

Subsequently, the output of the Stage 1 was given as the input to the Stage 2 of the MINT method. In Table 6 we report the MRR @1 and @5 for the second stage, for both time windows (3 days & 1 year).

Test Bed	MRR@1		MRR@5	
	3 days	1 year	3 days	1 year
EK-ST	**0.92**	0.87	**0.94**	0.90
ET-ST	**0.88**	0.74	**0.91**	0.78
EH-ST	-	0.82	-	0.87

Table 6: MRR of Stage 2 in NEAR-IDEAL

It is interesting to compare the results of MINT in NEAR-IDEAL data environment (Table 6) with MINT's results in IDEAL environment (Table 4). The drop in MRR@1 is small: ~2% for EK-ST and ~3% for ET-ST. For EH-ST the drop is relatively more (~12%) as may be expected since the time window (3 days) could not be applied for this test bed.

5.3 REAL Environment

The third set of experiments investigated the effectiveness of MINT on large comparable corpora. We ran the experiments on test beds created from group L corpora.

Test-beds for the REAL Environment: The test beds for the REAL environment (Table 7) consisted of only English articles since we do not know in advance whether these articles have any similar articles in the target languages.

Test Bed	Comparable Corpora	Articles	Distinct NEs
EK-LT	EK-L	100	306
ET-LT	ET-L	100	228

Table 7: Test Beds for REAL

Results in the REAL Environment: In real environment, we examined the top 2 articles of returned by Stage 1 of MINT, and mined NETEs from them. We used a date window of 3 in Stage 1. Table 8 summarizes the results for the REAL environment.

Test Bed	MRR	
	@1	@5
EK-LT	0.86	0.88
ET-LT	0.82	0.85

Table 8: MRR of Stage 2 in REAL

We observe that the performance of MINT is impressive, considering the fact that the comparable corpora used in the REAL environment is two orders of magnitude larger than those used in IDEAL and NEAR-IDEAL environments. This implies that MINT is able to effectively mine NETEs whenever the Stage 1 algorithm was able to find a good conjugate for each of the source language articles.

5.4 Generative Transliteration Similarity Model

We employed the extended W-HMM transliteration similarity model in MINT and used it in the IDEAL data environment. Table 9 shows the results.

Test Bed	MRR	
	@1	@5
EK-S	0.85	0.86
ET-S	0.81	0.82
EH-S	0.91	0.93

Table 9: MRR of Stage 2 in IDEAL using generative transliteration similarity model

We see that the results for the generative transliteration similarity model are good but not as good as those for the discriminative transliteration similarity model. As we did not stem either the English NEs or the target language words, the generative model made more mistakes on inflected words compared to the discriminative model.

5.5 Examples of Mined NETEs

Table 10 gives some examples of the NETEs mined from the comparable news corpora.

Language Pair	Source NE	Transliteration
English-Kannada	Woolmer	ವೂಲ್ಮರ್
	Kafeel	ಕಫೀಲ್
	Baghdad	ಬಾಗ್ದಾದ್
English-Tamil	Lloyd	லாயிட்
	Mumbai	மும்பையில்
	Manchester	மான்செஸ்டர்
English-Hindi	Vanhanen	वैनहैनन
	Trinidad	त्रिनिदाद
	Ibuprofen	इबूप्रोफेन
English-Russian	Kreuzberg	Крейцберге
	Gaddafi	Каддафи
	Karadzic	Караджич

Table 10: Examples of Mined NETEs

6 Related Work

CLIR systems have been studied in several works (Ballesteros and Croft, 1998; Kraiij et al, 2003). The limited coverage of dictionaries has been recognized as a problem in CLIR and MT (Demner-Fushman & Oard, 2002; Mandl & Womser-hacker, 2005; Xu &Weischedel, 2005).

In order to address this problem, different kinds of approaches have been taken, from learning transformation rules from dictionaries and applying the rules to find cross-lingual spelling variants (Pirkola et al., 2003), to learning translation lexicon from monolingual and/or comparable corpora (Fung, 1995; Al-Onaizan and Knight, 2002; Koehn and Knight, 2002; Rapp, 1996). While these works have focused on finding translation equivalents of all class of words, we focus specifically on transliteration equivalents of NEs. (Munteanu and Marcu, 2006; Quirk et al., 2007) addresses mining of parallel sentences and fragments from nearly parallel sentences. In contrast, our approach mines NETEs from article pairs that may not even have any parallel or nearly parallel sentences.

NETE discovery from comparable corpora using time series and transliteration model was proposed in (Klementiev and Roth, 2006), and extended for NETE mining for several languages in (Saravanan and Kumaran, 2007). However, such methods miss vast majority of the NETEs due to their dependency on frequency signatures. In addition, (Klementiev and Roth, 2006) may not scale for large corpora, as they examine every word in the target side as a potential transliteration equivalent. NETE mining from comparable corpora using phonetic mappings was proposed in (Tao et al., 2006), but the need for language specific knowledge restricts its applicability across languages. We proposed the idea of mining NETEs from multilingual articles with similar content in (Udupa, et al., 2008). In this work, we extend the approach and provide a detailed description of the empirical studies.

7 Conclusion

In this paper, we showed that MINT, a simple and intuitive technique employing cross-language document similarity and transliteration similarity models, is capable of mining NETEs effectively from large comparable news corpora. Our three stage empirical investigation showed that MINT performed close to optimal on comparable corpora consisting of pairs of similar articles when the pairings are known in advance. MINT induced fairly good pairings and performs exceedingly well even when the pairings are not known in advance. Further, MINT outperformed a state-of-the-art baseline and scaled to large comparable corpora. Finally, we demonstrated the language neutrality of MINT, by mining NETEs from 4 language pairs (between English and one of Russian, Hindi, Kannada or Tamil) from 3 vastly different linguistic families.

As a future work, we plan to use the extended W-HMM model to get features for the discriminative transliteration similarity model. We also want to use a combination of the cross-language document similarity score and the transliteration similarity score for scoring the NETEs. Finally, we would like to use the mined NETEs to improve the performance of the first stage of MINT.

Acknowledgments

We thank Abhijit Bhole for his help and Chris Quirk for valuable comments.

References

AbdulJaleel, N. and Larkey, L.S. 2003. Statistical transliteration for English-Arabic cross language information retrieval. *Proceedings of CIKM 2003.*

Al-Onaizan, Y. and Knight, K. 2002. Translating named entities using monolingual and bilingual resources. *Proceedings of the 40th Annual Meeting of ACL.*

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

Chen, H., et al. 1998. Proper Name Translation in Cross-Language Information Retrieval. *Proceedings of the 36th Annual Meeting of the ACL*.

Demner-Fushman, D., and Oard, D. W. 2002. The effect of bilingual term list size on dictionary-based cross-language information retrieval. *Proceedings of the 36th Hawaii International Conference on System Sciences.*

Finkel, J. Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the ACL*.

Fung, P. 1995. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. *Proceedings of the 3rd Workshop on Very Large Corpora*.

Fung, P. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. *Proceedings of ACL 1995*.

He. X. 2007: Using word dependent transition models in HMM based word alignment for statistical machine translation. In *Proceedings of 2nd ACL Workshop on Statistical Machine Translation* .

Hermjakob, U., Knight, K., and Daume, H. 2008. Name translation in statistical machine translation: knowing when to transliterate. *Proceedings ACL 2008*.

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

Knight, K. and Graehl, J. 1998. Machine Transliteration. *Computational Linguistics.*

Koehn, P. and Knight, K. 2002. Learning a translation lexicon from monolingual corpora. *Proceedings of Unsupervised Lexical Acquisition*.

Kraiij, W., Nie, J-Y. and Simard, M. 2003. Emebdding Web-based Statistical Translation Models in Cross-Language Information Retrieval. *Computational Linguistics*., 29(3):381-419.

Mandl, T., and Womser-Hacker, C. 2004. How do named entities contribute to retrieval effectiveness? *Proceedings of the 2004 Cross Language Evaluation Forum Campaign 2004*.

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*.

Munteanu, D. and Marcu D. 2006. Extracting parallel subsentential fragments from non-parallel corpora. *Proceedings of the ACL 2006.*

Och, F. and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*.

Pirkola, A., Toivonen, J., Keskustalo, H., Visala, K. and Jarvelin, K. 2003. Fuzzy translation of cross-lingual spelling variants. *Proceedings of SIGIR 2003.*

Ponte, J. M. and Croft, B. 1998. A Language Modeling Approach to Information Retrieval. *Proceedings of ACM SIGIR 1998.*

Quirk, C., Udupa, R. and Menezes, A. 2007. Generative models of noisy translations with applications to parallel fragments extraction. *Proceedings of the 11th MT Summit*.

Rapp, R. 1996. Automatic identification of word translations from unrelated English and German corpora. *Proceedings of ACL'99*

Saravanan, K. and Kumaran, A. 2007. Some experiments in mining named entity transliteration pairs from comparable corpora. *Proceedings of the 2nd International Workshop on Cross Lingual Information Access*.

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

Udupa, R., Saravanan, K., Kumaran, A. and Jagarlamudi, J. 2008. Mining Named Entity Transliteration Equivalents from Comparable Corpora. *Proceedings of the CIKM 2008*.

Udupa, R., Saravanan, K., Bakalov, A. and Bhole, A. 2009. *"They are out there if you know where to look"*: Mining transliterations of OOV terms in cross-language information retrieval. *Proceedings of the ECIR 2009*.

Virga, P. and Khudanpur, S. 2003. Transliteration of proper names in cross-lingual information retrieval. *Proceedings of the ACL Workshop on Multilingual and Mixed Language Named Entity Recognition.*

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