Some Experiments in Mining Named Entity Transliteration Pairs from Comparable Corpora

K Saravanan
Microsoft Research India
Bangalore, India
v-sarak@microsoft.com

A Kumaran
Microsoft Research India
Bangalore, India
kumarana@microsoft.com

Abstract

Parallel Named Entity pairs are important resources in several NLP tasks, such as, CLIR and MT systems. Further, such pairs may also be used for training transliteration systems, if they are transliterations of each other. In this paper, we profile the performance of a mining methodology in mining parallel named entity transliteration pairs in English and an Indian language, Tamil, leveraging linguistic tools in English, and article-aligned comparable corpora in the two languages. We adopt a methodology parallel to that of [Klementiev and Roth, 2006], but we focus instead on mining parallel named entity transliteration pairs, using a well-trained linear classifier to identify transliteration pairs. We profile the performance at several operating parameters of our algorithm and present the results that show the potential of the approach in mining transliterations pairs; in addition, we uncover a host of issues that need to be resolved, for effective mining of parallel named entity transliteration pairs.

1 Introduction & Motivation

Parallel Named Entity (NE) pairs are important resources in several NLP tasks, from supporting Cross-Lingual Information Retrieval (CLIR) systems, to improving Machine Translation (MT) systems. In addition, such pairs may also be used for developing transliteration systems, if they are transliterations of each other. Transliteration of a name, for the purpose of this work, is defined as its transcription in a different language, preserving the phonetics, perhaps in a different orthography [Knight and Graehl, 1997][1]. While traditional transliteration systems have relied on hand-crafted linguistic rules, more recently, statistical machine learning techniques have been shown to be effective in transliteration tasks [Jung et al., 2000] [AbdulJaleel and Larkey, 2003] [Virga and Kudhanpur , 2003] [Haizhou et al., 2004]. However, such data-driven approaches require significant amounts of training data, namely pairs of names in two different languages, possibly in different orthography, referred to as *transliteration pairs*, which are not readily available in many resource-poor languages. It is important to note at this point, that NEs are found typically in news corpora in any given language. In addition, news articles covering the same event in two different languages may reasonably be expected to contain the same NEs in the respective languages. The perpetual availability of news corpora in the world's languages, points to the promise of mining transliteration pairs endlessly, provided an effective identification of such NEs in specific languages and pairing them appropriately, could be devised.

Recently, [Klementiev and Roth, 2006] outlined an approach by leveraging the availability of article-aligned news corpora between English and Russian, and tools in English, for discovering transliteration pairs between the two languages, and progressively refining the discovery process. In this paper, we adopt their basic methodology, but we focus on 3 different issues:

[1] *London* rewritten as லண்டன் in Tamil, or لندن in Arabic (both pronounced as *London*), are considered as transliterations, but not the rewriting of *New Delhi* as புது தில்லி (*puthu thilli*) in Tamil.

1. mining comparable corpora for NE pairs, leveraging a well trained classifier,
2. calibrating the performance of this mining framework, systematically under different parameters for mining, and,
3. uncovering further research issues in mining NE pairs between English and an Indian language, Tamil.

While our analysis points to a promising approach for mining transliteration pairs, it also uncovers several issues that may need to be resolved, to make this process highly effective. As in [Klementiev and Roth, 2006] no language specific knowledge was used to refine our mining process, making the approach broadly applicable.

2 Transliteration Pairs Discovery

In this section, we outline briefly the methodology presented in [Klementiev and Roth, 2006], and refer interested readers to the source for details.

They present a methodology to automatically discover parallel NE transliteration pairs between English and Russian, leveraging the availability of a good-quality Named Entity Recognizer (NER) in English, and article-aligned bilingual comparable corpora, in English and Russian. The key idea of their approach is to extract all NEs in English, and identify a set of potential transliteration pairs in Russian for these NEs using a simple classifier trained on a small seed corpus, and re-ranking the identified pairs using the similarity between the frequency distributions of the NEs in the comparable corpora. Once re-ranked, the candidate pairs, whose scores are above a threshold are used to re-train the classifier, and the process is repeated to make the discovery process more effective.

To discriminate transliteration pairs from other content words, a simple perceptron-based linear classifier, which is trained on *n*-gram features extracted from a small seed list of NE pairs, is employed leveraging the fact that transliteration relies on approximately monotonic alignment between the names in two languages. The potential transliteration pairs identified by this classifier are subsequently re-ranked using a Discrete Fourier Transform based similarity metric, computed based on the frequency of words of the candidate pair, found in the article-aligned comparable corpora. For the frequency analysis, equivalence classes of the words are formed, using a common prefix of 5 characters, to account for the rich morphology of Russian language. The representative prefix of each of the classes are used for classification.

Finally, the high scoring pairs of words are used to re-train the perceptron-based linear classifier, to improve the quality of the subsequent rounds. The quality of the extracted NE pairs is shown to improve, demonstrating viability of such an approach for successful discovery of NE pairs between English and Russian.

3 Adoption for Transliteration Pairs Mining

We adopt the basic methodology presented in [Klementiev and Roth, 2006], but we focus on three specific issues described in the introduction.

3.1 Mining of Transliteration Pairs

We start with comparable corpora in English and Tamil, similar in size to that used in [Klementiev and Roth, 2006], and using the English side of this corpora, first, we extract all the NEs that occur more than a given threshold parameter, F_E, using a standard NER tool. The higher the threshold is, the more will be the evidence for legitimate transliteration pairs, in the comparable corpora, which may be captured by the mining methodology. The extracted list of NEs provides the set of NEs in English, for which we mine for transliteration pairs from the Tamil side of the comparable corpora.

We need to identify all NEs in the Tamil side of the corpora, in order to appropriately pair-up with English NEs. However, given that there is no publicly available NER tool in Tamil (as the case may be in many resource-poor languages) we start with an assumption that all words found in the Tamil corpus are potentially NEs. However, since Tamil is a highly morphologically inflected language, the same NE may occur in its various inflected forms in the Tamil side of the corpora; hence, we collect those words with the same prefix (of fixed size) into a single bucket, called *equivalence class*, and consider a representative prefix, referred to as *signature* of the collection for comparison. The

assumption here is that the common prefix would stand for a Tamil NE, and all the members of the equivalence class are the various inflected forms of the NE. We use such a signature to classify a Tamil word as potential transliteration of an English word. Again, we consider only those signatures that have occurred more than a threshold parameter, F_T, in the Tamil side of the comparable corpora, in order to strengthen support for a meaningful similarity in their frequency of occurrence.

We used a linear Support Vector Machine classifier (details given in a later section) trained on a sizable seed corpus of transliterations between English and Tamil, and use it to identify potential Tamil signatures with any of the NEs extracted from the English side. We try to match each of the NEs extracted from the English side, to every signature from the Tamil side, and produce an ordered list of Tamil signatures that may be potential transliterations for a given English NE. Every Tamil signature, thus, would get a score, which is used to rank the signatures in the decreasing order of similarity. Subsequently, we consider only those above a certain threshold for analysis, and in addition, consider only the top-*n* candidates.

3.2 Quality Refinement

Since a number of such transliteration candidates are culled from the Tamil corpus for a given NE in English, we further cull out unlikely candidates, by re-ranking them using frequency cues from the aligned comparable corpora. For this, we start with the hypothesis, that the NEs will have similar normalized frequency distributions with respect to time, in the two corpora. Given that the news corpora are expected to contain same names in similar time periods in the two different languages, the frequency distribution of words in the two languages provides a strong clue about possible transliteration pairs; however, such potential pairs might also include other content words, such as, சோஷலிஸ்ட் (*soshaliSt*), கவனமாக (*kavanamaaka*), கேட்பது (*keetpathu*), etc., which are common nouns, adjectives or even adverbs and verbs. On the other hand, function words are expected to be uniformly distributed in the corpus, and hence may not have high variability like content words. Note that the NEs in English are not usually inflected. Since Tamil NEs usually have inflections, the frequency of occurrence of a NE in Tamil must be normalized across all forms, to make it reasonably comparable to the frequency of the corresponding English NE. This was taken care of by considering the signature and its equivalence class. Hence the frequency of occurrence of a NE (i.e., its signature) in Tamil is the sum of frequencies of all members in its equivalence class.

For identifying the names between the languages, we first create a frequency distribution of every word in English and Tamil, by creating temporal bins of specific duration, covering the entire timeline of the corpus. The frequency is calculated as the number of occurrences of each signature in the bin interval. Once the frequency distributions are formed, they are normalized for every signature. Given the normalized frequencies, two words are considered to have same (or, similar) pattern of occurrence in the corpus, if the normalized frequency vectors of the two words are the same (or, close within a threshold). Figure 1 shows the frequency of the word *Abishek*, and its Tamil version, அபிஷேக் (*apishek*) as a frequency plot, where a high correlation between the frequencies can be observed.

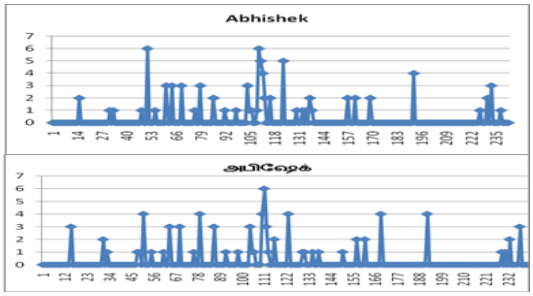

Figure 1: Names Frequency Plot in Comparable Corpora

Hence, to refine the quality of the classifier output, we re-rank the list of candidates, using the distance between the frequency vectors of the English NE, and the Tamil candidate signature. This step moves up those signatures that have similar patterns of occurrence, and moves down those that do not. It is likely that such frequency cues from the comparable corpora will make the quality of matched transliteration pairs better, yielding better mined data.

4 Experimental Setup & Results

In this section, we present the experimental setup and the data that we used for mining transliteration pairs from comparable corpora in two languages: English and the Indian language, Tamil. We evaluate and present the effectiveness of the methodology in extracting NE pairs, between these languages, under various parameters.

4.1 Comparable Corpora

We used a set of news articles from the New Indian Express (in English) and Dinamani (in Tamil) roughly covering similar events in English and Tamil respective, and covering a period of about 8 months, between January and August of 2007. The articles were verified to contain similar set of NEs, though only a fraction of them are expected to be legitimate transliteration pairs. Others related NEs could be translations, for example, *chief minister* in English vs முதல்வர் (*muthalvar*) in Tamil, abbreviation which are not usually transliterated but spelled out , for example, *ICC* in English, and ஐசிசி (*aicici*) in Tamil, or co-references , for example, *New Delhi* in English, and புதுதில்லி (*puthu thilli*) in Tamil. While the number of articles used were roughly the same (~2,400), the number of words in Tamil were only about 70% of that in English. This is partially due to the fact Tamil is a highly agglutinative language, where various affixes (prefixes and suffixes of other content words) stand for function words and prepositions in English, thus do not contribute to the word count. Further, since our focus is on mining names, we expect the same NEs to be covered in both the corpora, and hence we do not expect a severe impact on mining.

Corpus	Time Period	Size	
		Articles	Words
New Indian Express (English)	2007.01.01 to 2007.08.31	2,359	347,050
Dinamani (Tamil)	2007.01.01 to 2007.08.31	2,359	256,456

Table 1: Statistics on Comparable Corpora

From the above corpora, we first extracted all the NEs from the English side, using the Stanford NER tool [Finkel et al, 2005]. No multiword expressions were considered for this experiment. Also, only those NEs that have a frequency count of more than a threshold value of F_E were considered, in order to avoid unusual names that are hard to identify in the comparable corpora. Thus, we extracted from the above corpora, only a subset of NEs found in the English side to be matched with their potential transliteration pairs; for example, for a parameter setting of F_E to 10, we extract only 274 legitimate NEs.

From the Tamil side of the corpora, we extracted all words, and grouped them in to equivalence classes, by considering a prefix of 5 characters. That is, all words that share the same 5 characters were considered to be morphological variations of the same root word or NE in Tamil. After they were grouped, the longest common prefix of the group is extracted, and is used as the signature of the equivalence class. It should be noted here that though the number of unique words in the corpus is about 46,503, the number of equivalence classes to be considered changes depending on the filtering threshold that we use in the Tamil side. For example, at a threshold (F_T) value of 1, the number of equivalence classes is 14,101. It changes to 4,612 at a threshold (F_T) value of 5, to 2,888 at a threshold (F_T) value of 10 and to 1779 at a threshold (F_T) value of 20. However, their signature (i.e., longest common prefix) sizes ranged from 5 to 13 characters. Thus, we had about 14,101 equivalence classes, covering all the words from the Tamil corpus. The equivalence classes thus formed were as shown in Figure 2:

Tamil Signature	Tamil Equiv. Class
ஐஸ்வர்யா (*aiSvaryaa*)	ஐஸ்வர்யா (*aiSvaryaa*), ஐஸ்வர்யாவின் (*aiSvaryaavin*), ஐஸ்வர்யாவுக்கு (*aiSvaryaavukku*), ஐஸ்வர்யாவை (*aiSvaryaavai*), ஐஸ்வர்யாவிற்கும் (*aiSvaryaaviRkum*), ஐஸ்வர்யாவுடன் (*aiSvaryaavutan*)
பிரம் (*piram*)	பிரம்மபுத்திரா (*pirammaputhithiraa*), பிரம்மாண்டமான (*pirammaaNdamaana*), பிரம்பு (*pirampu*), பிரம்மா (*pirammaa*)
காவேரி (*kaaveeri*)	காவேரி (*kaaveeri*)
ஐசிசி (*aicici*)	ஐசிசி (*aicici*), ஐசிசியின் (*aicicyin*), ஐசிசிக்க (*aicici kku*), ஐசிசிதான் (*aicicithaan*), ஐசிசியிடம் (*aiciciyidam*)

Figure 2: Signatures and Equivalence Classes

As can be seen in the table, all elements of an equivalence class share the same signature (by definition). However, some signatures, such as ஐஸ்வர்யா (*aiSvaryaa*), correspond to an equivalence class in which every element is a morphological variation of the signature. Such equivalence classes, we name them *pure*. Some signatures represent only a subset of the members, as this set includes some members unrelated to this stem; for example, the signature பிரம் (*piram*), correctly corresponds to பிரம்மா (*pirammaa*), and incorrectly to the noun பிரம்பு (*pirambu*), as well as incorrectly to the adjective பிரம்மாண்டமான (*piram-maandamaana*). We name such equivalence classes *fuzzy*. Some are well formed, but may not ultimately contribute to our mining, being an abbreviation, such as *ICC* (in Tamil, ஐசிசி), even though they are used similar to any NE in Tamil. While most equivalence classes contained inflections of single stems, we also found morphological variations of several compound names in the same equivalence class such as, அகமத்நகர் (*akamathñakar*), அகமதாபாத் (*akamathaapaath*), with அகமத் (*akamath*).

4.2 Classifier for Transliteration Pair Identification

We used SVM-light [Joachims, 1999], a Support-vector Machine (SVM) from Cornell University, to identify near transliterations between English and Tamil. We used a seed corpus consisting of 5000 transliteration pair samples collected from a different resource, unrelated to the experimental comparable corpora. In addition to the 5000 positive examples from this seed corpus, 5000 negative examples were extracted randomly, but incorrectly, aligned names from this same seed corpus and used for the classifier.

The features used for the classification are binary features based on the length of the pair of strings and all aligned unigram and bigram pairs, in each direction, between the two strings in the seed corpus in English and Tamil. The length features include the difference in lengths between them (up to 3), and a separate binary feature if they differ by more than 3. For unigram pairs, the i^{th} character in a language string is matched to $(i\text{-}1)^{st}$, i^{th} and $(i+1)^{st}$ characters of the other language string. Each string is padded with special characters at the beginning and the end, for appropriately forming the unigrams for the first and the last characters of the string. In the same manner, for binary features, every bigram extracted with a sliding window of size 2 from a language string, is matched with those extracted from the other language string. After the classifier is trained on the seed corpus of hand crafted transliteration pairs, during the mining phase, it compares every English NE extracted from the English corpus, to every signature from the Tamil corpus.

While classifier provided ranked list of all the signatures from Tamil side, we consider only the top-30 signatures (and the words in the equivalence classes) for subsequent steps of our methodology. We hand-verified a random sample of about 100 NEs from English side, and report in Table 5, the fraction of the English NEs for which we found at least one legitimate transliteration in the top-30 candidates (for example, the recall of the classifier is 0.56, in identifying a right signature in the top-30 candidates, when the threshold F_E is 10 & F_T is 1).

It is interesting to note that as the two threshold factors are increased, the number of NEs extracted from the English side decreases (as expected), and the average number of positive classifications per English NE reduces (as shown in Table 2), considering all NEs. This makes sense as the classifier for identifying potential transliterations is trained with sizable corpora and is hence accurate; but, as the thresholds increase, it has less data to work with, and possibly a fraction of legitimate transliterations also gets filtered with noise.

Parameters	Extracted English NEs	Ave. Positive Classifications/ English NE
F_E: 10, F_T: 1	274	79.34
F_E: 5, F_T: 5	588	29.50
F_E: 10, F_T: 10	274	17.49
F_E: 20, F_T: 20	125	10.55

Table 2: Threshold Parameters vs Mining Quantity

Table 3 shows some sample results after the classification step with parameter values as (F_E: 10, F_T: 1). Right signature for *Aishwarya* (corresponding to all correct transliterations) has been ranked 10 and *Gandhi* (with only a subset of the equivalence class

corresponding to the right transliterations) has been ranked at 8. Three different variations of *Argentina* can be found, ranked 2nd, 3rd and 13th. While, in general no abbreviations are found (usually their Tamil equivalents are spelled out), a rare case of abbreviation (*SAARC*) and its right transliteration is ranked 1st.

English Named Entity	Tamil Equivalence Class Signature	Precision	Rank
aishwarya	ஐஸ்வர்யா (*aiSvaryaa*)	1	10
argentina	அர்ஜன்டினாவில (*arjantinaavila*)	1	2
argentina	ஆர்ஜென்டினாவி (*aarjantinaavi*)	1	3
argentina	ஆர்ஜன்டினாவில் (*aarjantinaavil*)	1	13
gandhi	காந்த (*kaañtha*)	0.2121	8
saarc	சார்க் (*saark*)	1	1

Table 3: Ranked List after Classification Step

4.3 Enhancing the Quality of Transliteration-Pairs

For the frequency analysis, we use the frequency distribution of the words in English and Tamil side of the comparable corpora, counting the number of occurrences of NEs in English and the Tamil signatures in each temporal bin spanning the entire corpus. We consider one temporal bin to be equal to two successive days. Thus, each of the English NEs and the Tamil signatures is represented by a vector of dimension approximately 120. We compute the distance between the two vectors, and hypothesize that they may represent the same (or, similar) name, if the difference between them is zero (or, small). Note that, as mentioned earlier, the frequency vector of the Tamil signature will contain the sum of individual frequencies of the elements in the equivalence class corresponding to it. Given that the classifier step outputs a list of English NEs, and associated with each entry, a ranked list of Tamil signatures that are identified as potential transliteration by the classifier, we compute the distance between the frequency vector of every English NE, with each of the top-30 signatures in the ranked list. We re-rank the top-30 candidate strings, using this distance measure. The output is similar to that shown in Table 4, but with possibly a different rank order.

English Named Entity	Tamil Equivalence Class Signature	Precision	Rank
aishwarya	ஐஸ்வர்யா (*aiSvaryaa*)	1	1
argentina	அர்ஜன்டினாவில (*arjantinaavila*)	1	1
argentina	ஆர்ஜென்டினாவி (*aarjantinaavi*)	1	3
argentina	ஆர்ஜன்டினாவில் (*aarjantinaavil*)	1	14
gandhi	காந்த (*kaañtha*)	0.2121	16
saarc	சார்க் (*saark*)	1	1

Table 4: Ranked List after Frequency Analysis Step

On comparing Table 3 and 4, we observe that some of the ranks have moved for the better, and some of them for the worse. It is interesting to note that the ranking of different stems corresponding to Argentina has moved differently. It is quite likely that merging these three equivalence classes corresponding to the English NE *Argentina* might result in a frequency profile that is more closely aligned to that of the English NE.

4.4 Overall Performance of Transliteration Pairs Mining

To find the effectiveness of each step of the mining process in identifying the right signatures (and hence, the equivalence classes) for a given English NE, we computed the Mean Reciprocal Rank (MRR) of the random sample of 100 transliteration pairs mined, in two different ways: First, we computed MRR_{pure}, which corresponded to the first occurrence of a pure equivalence class, and MRR_{fuzzy}, which corresponded to the first occurrence of a fuzzy equivalence class in the random samples. MRR_{fuzzy} captures how successful the mining was in identifying one possible transliteration, MRR_{pure}, captures how successful we were in identifying an equivalence class that contains only right transliterations[2]. In addition, these metrics were computed, corresponding to different frequency thresholds for the occurrence of a English NE (F_E) and a Tamil signature (F_T). The overall quality profile of the mining framework in mining the NE transliteration pairs in English and Tamil is shown in Table 5. Additionally, we also report the *recall* metric (*the fraction of English NEs, for which at least one le-*

[2] However, it should be noted that the current metrics neither capture how pure an equivalence class is (fraction of the set that are correct transliterations), nor the size of the equivalence class. We hope to specify these as part of quality of mining, in our subsequent work.

gitimate Tamil signature was identified) computed on a randomly chosen 100 entity pairs.

Parameters	Classification Step		Frequency Analysis Step		Re-call
	MRR *fuzzy*	MRR *pure*	MRR *fuzzy*	MRR *pure*	
F_E: 10, F_T: 1	0.3579	0.2831	0.3990	0.3145	0.56
F_E: 5, F_T: 5	0.4490	0.3305	0.5064	0.3529	0.61
F_E: 10, F_T: 10	0.4081	0.2731	0.4930	0.3494	0.57
F_E: 20, F_T: 20	0.3489	0.2381	0.4190	0.2779	0.47

Table 5: Quality Profile of NE Pairs Extraction

First, it should be noted that the recalls are the same for both the steps, since Frequency Analysis step merely re-arranges the output of the Classification step. Second, the recall figures drop, as more filtering is applied to the NEs on both sides. This trend makes sense, since the classifier gets less data to work with, as more legitimate words are filtered out with noise. Third, as can be expected, MRR_{pure} is less than the MRR_{fuzzy} at every step of the mining process. Fourth, we see that the MRR_{pure} and the MRR_{fuzzy} improve between the two mining steps, indicating that the time-series analysis has, in general, made the output better.

Finally, we find that the MRR_{pure} and the MRR_{fuzzy} keep dropping with increased filtering of English NEs and Tamil signatures based on their frequency, in both the classification and frequency analysis steps. The fall of the MRRs after the classification steps is due to the fact that the classifier has less and less data with the increasing threshold, and hence some legitimate transliterations may be filtered out as noise. However, the frequency analysis step critically depends on availability of sufficient words from the Tamil side for similarity testing. In frequency analysis step, the fall of MRRs from threshold 5 to 10 is 0.0134 on MRR_{fuzzy} and 0.0035 on MRR_{pure}. This fall is comparatively less to the fall of MRRs from threshold 10 to 20 which is 0.074 on MRR_{fuzzy} and 0.0715 on MRR_{pure}. This may be due to the fact that the number of legitimate transliterations filtered out from threshold 5 to 10 is less when compared to the number of legitimate transliterations filtered out from threshold 10 to 20. These results show that with less number of words filtered, it can get reasonable recall and MRR values. More profiling experiments may be needed to validate this claim.

5 Open Issues in NE pair Mining

In this paper, we outline our experience in mining parallel NEs between English and Tamil, in an approach similar to the one discussed in [Klementiev and Roth, 2006]. Over and above, we made parameter choices, and some procedural modifications to bridge the underspecified methodology given in the above work. While the results are promising, we find several issues that need further research. We outline some of them below:

5.1 Indistinguishable Signatures

Table 7 shows a signature that offers little help in distinguishing a set of words. Both the words, சென்னை (*cennai*) and morphological variations of சென் (*cen*), share the same 5-character signature, namely, சென்ன (*cenna*), affecting the frequency distribution of the signature adversely.

English Named Entity	Tamil Named Entity	Tamil Equivalent Class
chennai	சென்னை (*cennai*)	சென்னை (*cennai*), சென்னையில் (*cennaiyil*), சென்னையிலிருந்து (*cennaiyilirunthu*), சென்னின் (*cennin*), சென்னுக்கு (*cennukku*),சென்னையை (*cennaiyai*)

Table 7: Multiple-Entity Equivalence Class

5.2 Abbreviations

Table 8 shows a set of abbreviations, that are not identified well in our NE pair mining. Between the two languages, the abbreviations may be either expanded, as *BJP* expanded to (the equivalent translation for *Bharatiya Janatha Party* in Tamil), or spelled out, as in *BSNL* referred to as பிஎஸ்என்எல் (*pieSenel*). The last example is very interesting, as each *W* in English is written out as டபிள்யூ (*tapiLyuu*). All these are hard to capture by a simple classifier that is trained on well-formed transliteration pairs.

English Named Entity	Tamil Named Entity
BJP	பாஜக (*paajaka*), பா.ஜ.க. (*paa. ja. ka.*), பாரதீய ஜனதா கட்சி (*paarathiiya janathaa katci*)
BSNL	பிஎஸ்என்எல் (*pieSenel*), பிஎஸ்என்எல்லின் (*pieSenellin*), பிஎஸ்என்எல்லை (*piesenellai*)
WWW	டபிள்யூடபிள்யூடபிள்யூ (*tapiLyuutapiLyuutapiLyuu*)

Table 8: Multiple-Entity Equivalence Class

5.3 Multiword Expressions

This methodology is currently designed for mining only single word expressions. It may be an interesting line of research to mine multiword expressions automatically.

6 Related Work

Our work essentially follows a similar procedure as reported in [Klementiev and Roth, 2006] paper, but applied to English-Tamil language pair. Earlier works, such as [Cucerzan and Yarowsky, 1999] and [Collins and Singer, 1999] addressed identification of NEs from untagged corpora. They relied on significant contextual and morphological clues. [Hetland, 2004] outlined methodologies based on time distribution of terms in a corpus to identify NEs, but only in English. While a large body of literature exists on transliteration, we merely point out that the focus of this work (based on [Klementiev and Roth, 2006]) is not on transliteration, but mining transliteration pairs, which may be used for developing a transliteration system.

7 Conclusions

In this paper, we focused on mining NE transliteration pairs in two different languages, namely English and an Indian language, Tamil. While we adopted a methodology similar to that in [Klementiev and Roth, 2006], our focus was on mining parallel NE transliteration pairs, leveraging the availability of comparable corpora and a well-trained linear classifier to identify transliteration pairs. We profiled the performance of our mining framework on several parameters, and presented the results. Our experiment results are inline with those reported by [Klementiev and Roth, 2006]. Given that the NE pairs are an important resource for several NLP tasks, we hope that such a methodology to mine the comparable corpora may be fruitful, as comparable corpora may be freely available in perpetuity in several of the world's languages.

8 Acknowledgements

We would like to thank Raghavendra Udupa, Chris Quirk, Aasish Pappu, Baskaran Sankaran, Jagadeesh Jagarlamudi and Debapratim De for their help.

References

Nasreen AbdulJaleel and Leah S. Larkey. 2003. Statistical transliteration for English-Arabic cross language information retrieval. In *Proceedings of CIKM*, pages 139–146, New York, NY, USA.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-370.

L Haizhou, Z Min and S Jian. 2004. A Joint Source-Channel Model for Machine Transliteration. In *Proceedings of 42nd Meeting of Assoc. of Computational Linguistics*.

Magnus Lie Hetland. 2004. *Data Mining in Time Series Databases*, a chapter in A Survey of Recent Methods for Efficient Retrieval of Similar Time Sequences. World Scientific.

T. Joachims. 1999. 11 in: Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press.

Sung Young Jung, SungLim Hong, and Eunok Paek. 2000. An English to Korean transliteration model of extended markov window. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 383–389.

Alexandre Klementiev and Dan Roth. 2006. Named Entity Transliteration and Discovery from Multilingual Comparable Corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*, pages 82–88.

Kevin Knight and Jonathan Graehl. 1997. Machine transliteration. In *Proceedings of the Meeting of the European Association of Computational Linguistics*, pages 128–135.

Yusuke Shinyama and Satoshi Sekine. 2004. Named entity discovery using comparable news articles. In *Proceedings the International Conference on Computational Linguistics (COLING)*, pages 848–853.

Richard Sproat, Tao Tao, ChengXiang Zhai. 2006. Named Entity Transliteration with Comparable Corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 73–80, Sydney.

Tao Tao and ChengXiang Zhai. 2005. Mining comparable bilingual text corpora for cross-language information integration. In *KDD'05*, pages 691–696.

Tao Tao, Su-Youn Yoon, Andrew Fister, Richard Sproat, and ChengXiang Zhai. 2006. Unsupervised named entity transliteration using temporal and phonetic correlation. In *EMNLP 2006*, Sydney, July.

Paula Virga and Sanjeev Khudanpur. 2003. Transliteration of Proper Names in Cross-Lingual Information Retrieval. In *Proceedings of Workshop on Multilingual and Mixed-Language Named Entity Recognition*.