Śata-Anuvādak : Tackling Multiway Translation of Indian Languages

Anoop Kunchukuttan, Abhijit Mishra, Rajen Chatterjee, Ritesh Shah, Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Bombay
{anoopk, abhijitmishra, rajen, ritesh, pb}@cse.iitb.ac.in

Abstract

We present a compendium of 110 Statistical Machine Translation systems built from parallel corpora of 11 Indian languages belonging to the Indo-Aryan and Dravidian families. We analyze the relationship between translation accuracy and the language families involved. We feel that insights obtained from this analysis will provide guidelines for creating machine translation systems for specific Indian language pairs. For our studies, we built phrase based systems and some extensions. Across multiple languages, we show improvements on the baseline phrase based systems using these extensions: (1) source side reordering for English-Indian language translation, and (2) transliteration of untranslated words for Indian language-Indian language translation. These enhancements harness shared characteristics of Indian languages. To stimulate similar innovation widely in the NLP community, we have made the trained models for these language pairs publicly available.

Keywords: Statistical Machine Translation, Indian Language Machine Translation, Phrase based SMT

1. Introduction

India is a linguistically diverse country with 22 scheduled languages and 30 languages having more than a million native speakers, spanning four language families *(Indo-Aryan, Dravidian, Tibeto-Burman and Austro-Asiatic)* and 10 major scripts. Beneath this diversity, many Indian languages (abbreviated as *IL*) exhibit shared characteristics like: (i) relatively free word order, with SOV being the canonical word order, (ii) similar orthographic systems descended from the *Brahmi* script based on auditory phonetic principles, (iii) vocabulary and grammatical tradition derived from Sanskrit, and (iv) morphological richness.

This diversity calls for translation solutions across a large number of translation pairs to serve government, business and overall social communication needs. While this sheer diversity poses many challenges, the shared characteristics present possible opportunities for exploring novel approaches to machine translation. Statistical Machine Translation (SMT) technology offers the possibility of being able to scale to a large number of translation pairs efficiently. SMT systems have made vast strides in the recent past and are amongst the best performing systems for many language pairs. For Indian languages, the SMT approach has been explored for just a few languages: primarily English and Hindi, along with some other major languages like Urdu, Telugu, Bengali (Ramanathan et al., 2008; Patel et al., 2013; Post et al., 2012).

In this work, we build phrase-based SMT systems and their extensions for 110 language pairs using the *Indian Language Corpora Initiative* (ILCI) corpus (Choudhary and Jha, 2011; Jha, 2012). To the best of our knowledge, this is the largest exercise in building SMT systems for Indian languages in terms of both, the number of language pairs and the corpus size. Our objective is to pursue the following research directions:

- Observe patterns between translation accuracy and the language families involved. Do the patterns suggest that unique SMT system architectures be developed for each language family pair?
- Investigate ways of leveraging shared characteristics of Indian languages to reduce the effort and resources required for building systems involving Indian languages.
- Investigate if learnings from improvement of SMT systems in one language pair can be easily ported to other language pairs, making simultaneous progress of all Indian language SMT systems viable.
- Explore how far phrase based SMT systems for Indian languages can be improved through preprocessing and post-editing extensions.
- Identify the challenges for SMT involving all major Indian languages.
- Determine best principles to build SMT systems for specific language pairs.

This work is the beginning of an effort to answer these questions, and stimulate more experimentation and innovation in SMT techniques for Indian languages.

2. SMT Models Explored

To investigate the questions we enlisted, we built the following SMT models for various language pairs:

2.1. Baseline phrase based system (S1)

Phrase based SMT (PBSMT) systems have been developed for many language pairs and are easily extensible to new language pairs since they don't need linguistic resources. We study the performance of PBSMT systems for translation among Indian languages, specifically the relationship between translation accuracy and language families from the perspective of different language divergences like word order and morphology (Dave et al., 2001), and the effect of corpus size.

Śata-Anuvādak means '100 translators' in Hindi

Language	Code	Tokens	Types	Characters
Hindi	hin	846,251	42,586	4,524,136
Urdu	urd	895,381	42,263	3,885,682
Punjabi	pan	841,188	54,570	4,135,986
Bengali	ben	680,778	57,821	4,318,895
Gujarati	guj	697,583	72,646	4,154,065
Marathi	mar	612,950	88,113	4,315,087
Konkani	kok	640,592	76,944	4,392,063
Tamil	tam	604,541	92,459	5,254,244
Telugu	tel	558,060	108,544	4,311,201
Malayalam	mal	543,516	117,299	5,033,872
English	eng	834,761	36,693	4,405,629
Total		7,755,601	789,938	49,578,169

Table 1: ILCI corpus statistics
ISO-639-2 language codes are shown

2.2. English-IL PBSMT with generic source side reordering rules (S2)

Preprocessing the training and test corpus by reordering the source side sentences to make them conform to target word order has been shown to be beneficial. The improvement occurs for two reasons:

- The decoder's search space can consider candidates with better word order.
- The quality of the phrase table created is better since the alignment template method for phrase extraction can match longer phrases.

We used Ramanathan et al. (2008)'s rule-based reordering system which is based on the following generic transformation principle going from English to Hindi word order:

$$SS_mVV_mOO_mC_m \leftrightarrow C'_mS'_mS'V'_mV'O'_mO'$$

where,
S: Subject, O: Object, V: Verb, C_m: Clause modifier, X': Corresponding constituent in Hindi.
X is S, O or V
X_m: modifier of X
In addition, there are rules like prepositions becoming postpositions. This principle holds across all Indian languages, hence we hypothesize that the same rules will benefit translation from English to any Indian language.

2.3. English-IL PBSMT with Hindi-tuned source side reordering rules (S3)

In this model, we use Patel et al. (2013)'s source side reordering rules. These rules are refinements of *S2* with additional rules found through a focused analysis of word order divergence observed in the English-Hindi translation pair. These include rules for handling interrogative sentences, infinite clauses, adjectival and adverbial phrases, etc. We postulate that the rules resulting from analysis of the English-Hindi system would benefit translations from English to other Indian languages too.

2.4. IL-IL PBSMT with post-editing using transliteration (S4)

Many words in Indian languages have origins in Sanskrit or have been borrowed from it. Languages spoken in adjoining geographical regions, like Gujarati and Marathi, also have many words in common. So, for untranslated words and named entities, transliteration to the target language script can serve the purpose of translation. Therefore, for translation between Indian languages, untranslated words from the baseline phrase based system (S1) are transliterated in an automatic post-editing stage. The post-editing is applied to all language pairs, except those involving Urdu or English. Transliteration between Indic scripts is relatively easy since most Indian scripts are descendants of the *Brahmi* script and are based on phonetic principles. In the ILCI corpus, except Urdu and English, the scripts of all other languages originate from the *Brahmi* script. Hindi, Marathi and Konkani use the same script (*Devanagari*). All these scripts are *abugida* scripts, with vowels, dependent vowel marks (*matraa*) and consonants as basic units. They largely have the same vowel and consonant repository. Therefore, the Unicode standard (Allen et al., 2012) gives corresponding vowels, vowel-signs and consonants the same offsets relative to the start of the Unicode block for the script. We exploit the coordinated Unicode ranges allocated to Indian scripts to develop a simple transliteration method that just maps the Unicode codepoints between scripts.

3. Dataset and Resources

We built 110 machine translation systems from all combinations of the 11 languages in the multilingual ILCI corpus, which contains roughly 50000 parallel sentences.
It represents 7 languages from the **Indo-Aryan family** (Hindi, Urdu, Punjabi, Gujarati, Bengali, Marathi and Konkani), 3 from the **Dravidian family** (Tamil, Telugu and Malayalam) and one from the **West Germanic family** (English).
The sentences are from tourism and health domains (25,000 sentences each). Statistics about the corpus is shown in Table 1. We normalized the corpus to solve issues related to wrong characters, redundant Unicode representation of some Indic characters, etc. Section 3.1. describes the normalization. For every language pair, the corpus was split up as follows: *training set of 46277 sentences, test set of 2000 sentences and tuning set of 500 sentences*. The train, test and tune splits are completely parallel across the 11 languages involved.
Phrase based systems were trained using the *Moses*[1] system, with the *grow-diag-final-and* heuristic for extracting phrases and the *msd-bidirectional-fe* model for lexicalized reordering. We tuned the trained models using Minimum Error Rate Training (MERT) with default parameters (100 best list, max 25 iterations). We trained 5-gram language models on 50000 sentences from the ILCI corpus using the Kneser-Ney smoothing algorithm with *SRILM* [2]. Batch training of multiple SMT systems was done using the *Moses Job Scripts* [3] experiment management system.
The evaluation was done using the BLEU metric (Papineni et al., 2002). BLEU has been used by Koehn et al. (2009) and Koehn (2005) for similar multilingual transla-

[1] http://www.statmt.org/moses/
[2] http://www.speech.sri.com/projects/srilm
[3] https://bitbucket.org/anoopk/moses_job_scripts

	Indo-Aryan							Dravidian			
	hin	**urd**	**pan**	**ben**	**guj**	**mar**	**kok**	**tam**	**tel**	**mal**	**eng**
(A) Phrase based system (S1)											
hin	-	50.30	70.06	36.31	53.29	33.78	36.06	11.36	21.59	10.95	28.15
urd	58.09	-	51.90	26.14	38.92	21.21	25.09	8.13	14.65	7.49	21.00
pan	71.26	44.46	-	30.27	46.24	25.54	29.44	8.96	17.92	7.49	24.01
ben	36.16	24.91	31.84	-	31.24	19.79	23.16	8.88	13.18	8.62	18.34
guj	53.09	34.77	47.60	29.35	-	26.99	29.63	9.95	16.57	7.97	19.58
mar	41.66	25.08	34.75	23.68	33.84	-	27.44	8.34	12.02	7.25	15.87
kok	38.54	25.54	33.53	24.61	31.44	23.69	-	7.96	13.40	8.05	16.92
tam	21.79	15.65	19.32	14.77	17.28	11.10	14.17	-	9.30	6.41	10.90
tel	27.20	19.03	25.14	16.87	22.22	13.47	16.98	7.29	-	6.58	12.09
mal	14.50	10.27	12.53	10.01	10.99	7.01	9.36	4.67	6.25	-	8.36
eng	26.53	18.07	22.86	14.85	17.36	10.17	13.01	4.17	6.43	4.85	-
(B) Phrase based system with source reordering: generic rules (S2)											
eng	29.63	20.42	26.06	16.85	20.11	11.46	15.01	4.97	7.83	5.53	-
(C) Phrase based system with source reordering: Hindi-adapted rules (S3)											
eng	30.86	21.54	27.52	18.20	21.33	12.68	15.73	5.09	8.29	5.68	-

Table 2: %BLEU scores for systems S1, S2 and S3

tion studies. We have also developed extensions to METEOR (Banerjee and Lavie, 2005) for Indian languages, utilizing the IndoWordNet (Bhattacharyya, 2010) as a source for synonyms and an IndoWordnet assisted stemmer (Bhattacharyya et al., 2014). Because of the ability to incorporate synonyms and stemmers, METEOR can potentially be a better evaluation metric for morphologically richer languages. The METEOR scores show the same trend as the BLEU scores for our experiments. Hence, for brevity, we have only reported the BLEU scores.

3.1. Unicode Normalization of Indic Scripts

Text written in Indic scripts suffers from the problem of multiple Unicode codepoints for representation of the same script character. This occurs due to support for compatibility with other standards and use of control characters to supply rendering information. Some examples of this redundant representation are:

- Non-spacing characters like Zero-Width Joiner (ZWJ) and Zero-Width Non-Joiner (ZWNJ), whose role is to control rendering and do not affect the content
- Multiple representations of *Nukta* based characters
- Multiple representations of two part dependent vowel signs
- Typing inconsistencies: e.g. use of pipe (|) instead of *poorna virama* character as sentence delimiter

Multiple representation for the same character causes data sparsity and aggravates the problems of working with a small parallel corpus. Therefore, we convert the corpus to a canonical Unicode representation using the Indic Unicode Normalizer. The normalizer along with the Indian-Indian language transliterator, described in Section 2.4., have been made available as part of the *Indic NLP Library* [4].

	IA	DR	ENG
IA	35.73	10.99	20.55
DR	15.70	6.75	10.45
ENG	17.55	5.15	-

Table 3: Average %BLEU score for language family pairs (IA: Indo-Aryan, DR: Dravidian, ENG: English)

4. Results and Analysis

4.1. Translation Accuracy vis-a-vis Language Families

We analyze the translation accuracy in terms of the language families involved in the translation. Table 2 shows the results of the phrase-based SMT systems (*S1*) for all language pairs. Table 3 shows the average %BLEU score calculated over each language family pair. We see a clear partitioning of language pairs by language family. Translation between Indo-Aryan languages is the easiest, which is not surprising given that these languages have the same word order (SOV), similar case marking systems, and are less inflectional than the Dravidian languages (though more inflectional than English). Though English is morphologically poor, translation between Indo-Aryan and English performs sub-optimally because of the structural divergence between these language families. Translation between Dravidian languages, which are morphologically rich, show very low accuracies. The lowest translation accuracies are reported for English to Dravidian language translation, where morphological richness of Dravidian languages as well as structural divergence between the language families pose severe challenges for translation.

4.2. Effect of corpus size

A pertinent question is - to what extent can translation accuracy be improved by increasing the training corpus size. To investigate this, we trained phrase based SMT systems by varying the training corpus size. Figure 1 shows vari-

[4] `https://bitbucket.org/anoopk/indic_nlp_library`

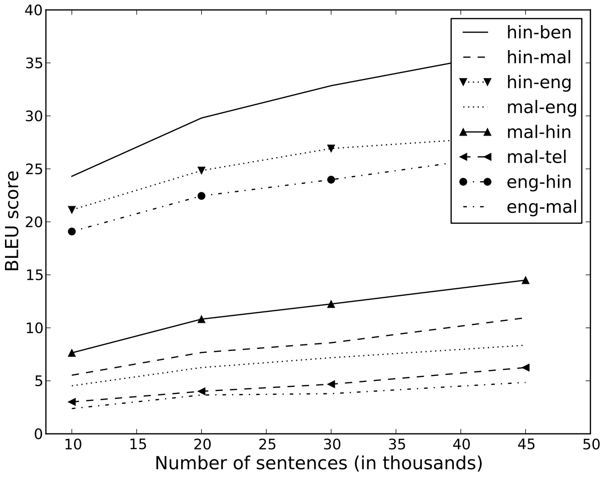

Figure 1: Training set size vs. %BLEU
(shows one language pair from every language family pair)

ation in BLEU scores with training set size for 8 language pairs. Each language pair is a representative of the 8 possible pairs of language families represented in the ILCI corpus. The quality of translation between Indo-Aryan languages improves substantially with increase in training corpus, with further potential for significant improvement as corpus size increases. Language pairs involving morphologically poor languages benefit the most (see the *eng-hin* and *hin-ben* pairs). For other language family pairs, the benefits are not commensurate to the increase in training corpus and the returns due to increased corpus size diminish rapidly (see the *eng-mal* and *mal-tel* pairs).

4.3. The challenge of morphological complexity

The morphological complexity of Indian languages makes translation among them challenging. It is most difficult to translate between Dravidian languages, though they share many characteristics including word order, morphological structure, etc. This suggests that high degree of agglutination and rich morphology are challenging factors for Indian language SMT, a hypothesis reinforced by the following observations:

- Comparatively lower BLEU scores are observed for language pairs involving Marathi and Konkani, which are morphologically the richest Indo-Aryan languages.
- There is a strong inverse correlation (Pearson's coefficient= -0.70) between corpus vocabulary size and average BLEU score translating into a language. The vocabulary size refers to the number of unique words in the corpus. It can be considered to be a proxy for morphological richness, since morphologically richer languages will tend to have higher number of words in a multilingual parallel corpus. Figure 2 illustrates this trend of lower translation accuracy with higher morphological complexity.

Translation Model Entropy

We use the *translation model entropy* measure (Koehn et al., 2009) to study the effect of morphological complexity on uncertainty in translation. *Translation model entropy* expresses the uncertainty involved in selecting a candidate translation of a source phrase from a set of possible translations. The entropy H for a source phrase s is calculated as follows:

$$H(s) = -\sum_{t \in T} P(t|s) * \log_2 P(t|s) \qquad (1)$$

where, T is the set of possible translations of s.

For each language pair in the ILCI corpus, we calculated the entropy of the PBSMT translation models on test sets of 500 sentences. These test sets were not part of the training corpus. Hence, for phrase pairs which are not present in the phrase table, we assumed an extremely small translation probability. We searched through all the possible segmentations of each source sentence and the segmentation resulting in the least average entropy per word was considered. Figure 3 shows the *average sentence entropy* matrix for the ILCI corpus. The entropy values are distributed with a mean of 18.08 and standard deviation of 6.81. The lowest and highest values are 5.1 and 33.2 respectively for Hindi-Punjabi and Telugu-Malayalam.

Increasing morphological complexity leads to data sparsity, which would in turn make the probability estimates unreliable. Thus, we expect the translation model entropy to be high for translation systems involving morphologically rich languages. Indeed, entropy is low for Indo-Aryan language pairs, and high for Dravidian language pairs. Language divergence is an important reason for high entropy between translation pairs. However, Dravidian languages exhibit less divergence amongst themselves. Hence, the high entropy can be best explained by the morphological richness of these languages, resulting from agglutination and syncretism.

Moreover, the translation entropy is higher for translation from morphologically richer languages to poorer languages compared to the other direction. This can be explained by the absence of translations for morphologically richer phrases in the phrase table. Morphological segmentation of the source side corpus before training translation models could overcome this problem. On the other hand, entropy is relatively high for translation from morphologically poor language to rich language. Failure to get reliable probability estimates and uniform distribution of probability mass among possible translation candidates could explain higher translation entropy in this case. Table 4 shows English-Malayalam translation examples to illustrate this behaviour. Many words are incorrectly translated in *English-to-Malayalam* translation, whereas many Malayalam words are not translated in *Malayalam-to-English* translation.

4.4. Source side reordering

To handle word order divergence for English-Indian language translation, we experimented with two source side reordering systems (Ramanathan et al., 2008; Patel et al., 2013). The effectiveness for both these systems for translations to Indian languages other than Hindi has not been studied earlier. Table 2 shows that (1) both the systems

Malayalam→English	
S:	കേരളത്തിന്റെ പ്രകൃതിയെക്കാപ്പം തന്നെ അവിടുത്തെ മതപരമായ ചടങ്ങുകൾക്ക് ആനകളുടെ പങ്ക് വളരെ വലുതാണ്
T:	പ്രകൃതിയെക്കാപ്പം there itself is the very ചടങ്ങുകൾക്ക് religious of elephants
English→Malayalam	
S:	It's booking can be done on the website of IRCTC
T:	ഇതിന്റെ പ്രധാന പാതയിൽ IRCTC ബുക്കിങ്ങും കിട്ടിയത്
G:	{of this} main {of way} IRCTC {also booking} got

Table 4: Malayalam ↔ English translation illustrating the effect of morphological complexity

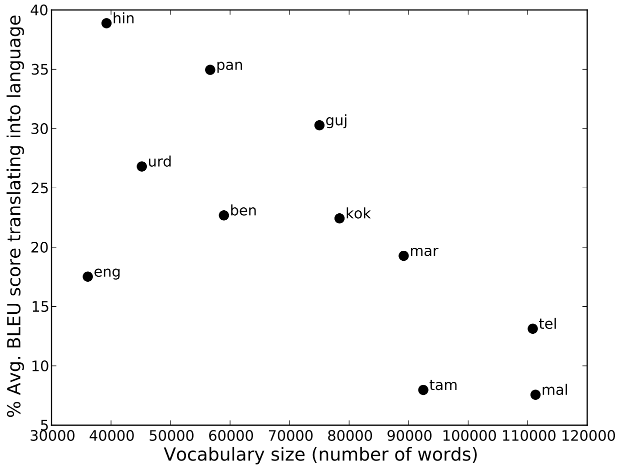

Figure 2: %BLEU vs morphological complexity

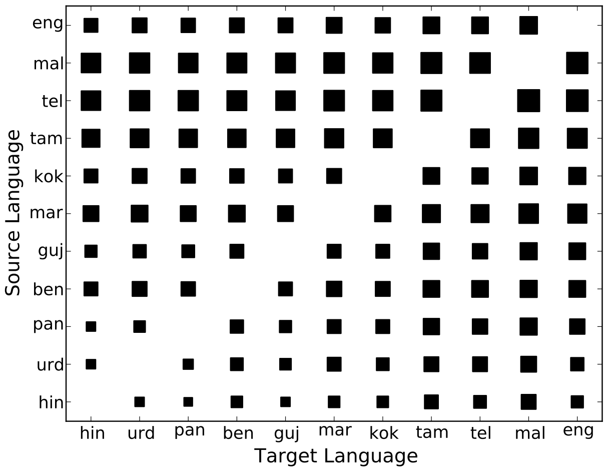

Figure 3: Translation Model Entropy matrix for the ILCI corpus
(The size of the square is a measure of the entropy)

improve the BLEU score for translation to all Indian languages, and (2) the rules guided by analysis of English-Hindi divergence outperforms the generic rules for other languages too. The *S3* system improves the average BLEU score across all target languages by 21.5% compared to 15.1% improvement shown by the *S2* system. Source side reordering helps Dravidian languages slightly more than the Indo-Aryan languages; the *S3* system improves the average BLEU by 22.70% for Dravidian languages.

The *S2* system's pre-ordering rules can be viewed as minimizing the most fundamental and common structural divergences between English and Indian languages. However, the *S3* system's rules consider pre-ordering over deeper constituent structures with English-Hindi pair as the basis for analysis. Table 5 and Table 6 illustrate the improvement of translation quality due to systems *S2* and *S3*. The underlined words are the ones whose reordering is described in this example. In the *English-to-Marathi* translation, the verb movement in the *S2* system output to the end of the sentence hampers understandability. However, better handling of dependent clause reordering improves the verb placement in the *S3* system output. In the *English-to-Gujarati* translation, we clearly see that the movement of the verb *am talking of* towards the end by the *S2* system is undesirable. This constituent movement is corrected by the *S3* system resulting in better word order.

4.5. Post-editing using Transliteration

We evaluate the effect of transliteration using translation recall. The recall scores were obtained using *METEOR*, which helps capture the effect of only the root form being correctly transliterated. Table 7 shows the translation recall, while Table 8 shows the percentage increase in recall over PBSMT after post-editing using transliteration is applied to the output of the baseline phrase based system. Ignoring translation systems between Hindi, Marathi and Konkani (which share the same script), there is a 1.72% average increase in recall due to transliteration post-editing. The recall actually decreases substantially for language pairs with *Devanagari* as source language script and Punjabi as target language (hin-pan, mar-pan, kok-pan). Ignoring these language pairs, the average increase in recall is 2.13%. The corresponding increase in the average BLEU score is a significant 1.06%. Thus, transliteration post-editing significantly improves the translation quality among language pairs using *Brahmi*-derived scripts.

The maximum increase in translation recall (3-5%) is observed for language pairs spoken in geographically adjacent regions, which share a significant part of their vocabulary (guj↔mar, tel↔mar, tel↔mal, tam↔mal). Table 9 shows some examples of shared vocabulary between these language pairs. Named entities are also transliterated correctly between language pairs.

These gains in translation accuracy have been achieved with a very simple transliteration method. We do not handle script specific issues like the final *schwa* deletion in North Indian scripts, *chillu* characters in the Malayalam script,

Src	I am talking of the Tsomoriri lake , which is probably situated at one end of the world
Ref	मी बोलत आहे सो – मोरीरी सरोवराबद्दल , जे कदाचित जगाच्या एका टोकावर वसलेले आहे
S1	मी ठिकाणाचा सो –मोरीरी सरोवर स्थित आहे , जे कदाचित जगाच्या एका टोकावर वसलेले आहे
S2	मी सो – मोरीरी सरोवर , जे कदाचित जगाच्या एका टोकावर वसलेले आहे बोलताना आहे
S3	मी सो – मोरीरी सरोवर बोलत आहे , जे कदाचित जगाच्या एका टोकावर वसलेले आहे

Table 5: English ↔ Marathi translation illustrating the effect of source reordering rules

Src	I am talking of the Tsomoriri lake , which is probably situated at one end of the world
Ref	હું વાત કરી રહી છું સો – મોરીરી સરોવરની, જે કદાચ દુનિયાના એક શિખર પર આવેલ છે
S1	હું છું સો – મોરીરી સરોવર ઉલ્લેખ છે, જે કદાચ એક ભાગ પર આવેલ છે
S2	મને સો–મોરીરી સરોવર છે, જે કદાચ દુનિયાના એક છેડે રહેલું કરી રહ્યો છે
S3	હું સો–મોરીરી સરોવર કરી રહ્યો છે, જે કદાચ દુનિયાના એક ભાગ પર આવેલ છે

Table 6: English ↔ Gujarati translation illustrating the effect of source reordering rules

	hin	**pan**	**ben**	**guj**	**mar**	**kok**	**tam**	**tel**	**mal**
hin	-	82.1	66.4	81.7	66.7	67.4	44.4	54	42.1
pan	87.3	-	60.6	75.4	59.9	62.3	40.3	49.2	36.3
ben	68.4	63.4	-	64.7	52.4	55.8	38.9	41.7	37.4
guj	79.1	73.9	61.1	-	60.8	62	41	47.4	36.3
mar	69.6	59.4	55.3	66.4	-	59.1	38.4	41.2	35.1
kok	67.8	58	55.5	63.8	55.4	-	37.9	41.9	36
tam	51.9	49.2	43	48.5	38.7	43.7	-	35.7	32.9
tel	58.1	54.9	46.6	53.6	42.8	47.1	36.2	-	32.4
mal	45.2	41.1	36	40.6	31.6	35.4	30.1	29.4	-

Table 7: % Recall for system S4: transliteration postediting

	hin	**pan**	**ben**	**guj**	**mar**	**kok**	**tam**	**tel**	**mal**
hin	-	-5.6	1.3	1.9	0	0	1	2	1.7
pan	1.1	-	1	1.4	1.8	1.2	1.2	1.4	1.6
ben	1.3	0.8	-	1.8	2.5	1.8	1.4	1.9	2.3
guj	2.3	1.2	2	-	4.1	2.7	1.6	2.8	3
mar	0	-7.2	2.4	3.7	-	0	1.8	3.7	3.1
kok	0	-7.6	2	2.6	0	-	1.5	2.9	2.7
tam	0.6	0.5	0.7	0.9	1.3	0.9	-	1.7	2.3
tel	2.4	1.3	2.1	3	4.1	2.8	2.7	-	4.2
mal	1.9	1.1	2.4	3.2	4.4	2.7	3.7	5.1	-

Table 8: % Increase in recall for system S4 compared to system S1

absence of aspirated and voiced consonants in Tamil, etc. Named entities are sometimes transliterated in very idiosyncratic ways between Indian languages. A more sophisticated transliteration method can possibly provide substantial improvements in translation recall.

Language Pair	Source	Target	English Meaning
guj↔mar	છત્રીની	छत्रीची	of umbrella
tel↔mar	శరీరంను	शरीराला	body-*OBJ*
tel↔mal	సమానం	സമയമാണ്	is similar

Table 9: Examples of transliteration of shared words

5. Translation Resources and Online Portal

Śata-Anuvādak [5] has been hosted online for public access. The system enables translation between 11 different Indian languages (including English), and provides transliteration support for Indic script input. *Devanagari* transliteration of the translations are displayed for users who can not read the target language script. Users can post-edit the translated text which provides feedback for improvement of the translation systems. To stimulate innovation in Indian language SMT, we have released the phrase tables, language models and the *Indic NLP Library* for academic use.

6. Related Work

Our work is most similar to Koehn (2005) and Koehn et al. (2009), who describe and analyze translation systems for all language pairs in the *Europarl* and *Acquis Communautaire* corpus respectively. While they extended the phrase based systems with pivot language based SMT systems, we have explored pre-processing and post-processing extensions to SMT systems. English to multiple Indian language phrase based SMT systems have been explored by (Post et al., 2012) (6 Indian languages from crowd-generated corpora) and the *Anuvadaksh* project (8 Indian languages)[6](Ramanathan et al., 2008; Patel et al., 2013). To the best of our knowledge, there is no other published work on Indian language to Indian language, and Indian language to English SMT systems.

Most of the work in Indian language MT has involved *transfer*, *example* or *interlingua* based systems. The *Anglabharati* system (Sinha et al., 1995) is an *English-to-Indian* language based pseudo-interlingual system which harnesses the common characteristics of Indian languages in the syntax transfer stage. The *Sampark* system[7] (Bhosale, 2011) is a transfer based system for translation between 9 Indian language pairs that uses a common lexical transfer engine, whereas minimum structural transfer is required between Indian languages. The emphasis is on detailed morphological analysis to enable accurate lexical transfer and target generation.

[5] http://www.cfilt.iitb.ac.in/indic-translator/

[6] http://tdil-dc.in/index.php?option=com_vertical&parentid=72

[7] http://sampark.iiit.ac.in

7. Conclusion

Our SMT experiments involving the largest number of Indian language pairs reveal interesting patterns and points to potential research directions for Indian language SMT:

- Based on translation accuracy, we see clear partitioning of translation pairs by language family. For instance, translations involving Indo-Aryan languages can be done with a high level of accuracy, whereas those involving Dravidian languages are extremely difficult. This suggests that SMT approaches customized to language family pairs must be investigated.
- Common language divergences exist between some language families, therefore common solutions and resources can be utilized for translation between these families. For instance, we show that common source side reordering rules can be used for English to Indian language translation, where *SVO*↔*SOV* divergence exists.
- Common characteristics of Indian languages make it easy to solve a few problems in Indian language SMT. For instance, transliteration between Indian languages is relatively easy since the scripts follow similar principles. We have shown that even a naive transliterator yields gains in translation accuracy.
- We have shown that considerable improvements can be obtained over a baseline phrase-based system with pre-processing (source side reordering) and post-processing stages (transliteration), and more ideas can be explored like re-ranking top-k results, handling OOV words, etc.
- Rich morphology of Indian languages and word order divergence between English and Indian languages are the major factors impacting translation quality. For some language pairs, issues like word order divergence are not crucial. On the other hand, the rich morphology of Indian languages makes it imperative that morphology should be accounted for in SMT models.
- We have analyzed if increase in training corpus size leads to a commensurate improvement in translation accuracy. We see that it helps for translation among Indo-Aryan languages, and points to the fact that large scale acquisition of parallel corpora for these language pairs may yield high quality SMT systems for indicative translation. However, for the morphologically rich Dravidian languages, increase in corpus size will only bring limited gains.

Further, we plan to explore syntax based SMT for English-Indian language translation, factored SMT models and methods to tackle the rich morphology in Indian languages.

8. Acknowledgements

We would like to thank the Technology Development for Indian Languages (TDIL) Programme and the Department of Electronics & Information Technology, Govt. of India for providing the ILCI corpus. We would like to thank all members of the Centre for Indian Language Technology (CFILT) for valuable inputs on various languages which made this work possible.

9. References

Allen, J. D., Anderson, D., Becker, J., and Richard Cook, e., editors, (2012). *The Unicode Standard Version 6.2 - Core Specification*, chapter South Asian Scripts - 1.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Bhattacharyya, P., Bahuguna, A., Talukdar, L., and Phukan, B. (2014). Facilitating Multi-Lingual Sense Annotation: Human Mediated Lemmatizer. In *Global WordNet Conference*.

Bhattacharyya, P. (2010). IndoWordNet. In *Lexical Resources Engineering Conference*.

Bhosale, G. (2011). Marathi to Hindi machine translation. Master's thesis, Indian Institute of Technology Bombay.

Choudhary, N. and Jha, G. N. (2011). Creating Multilingual Parallel Corpora in Indian Languages. In *Proceedings of Language and Technology Conference*.

Dave, S., Parikh, J., and Bhattacharyya, P. (2001). Interlingua-based english--hindi machine translation and language divergence. *Machine Translation*.

Jha, G. N. (2012). The TDIL program and the Indian Language Corpora Initiative. In *Proceedings of Language Resources and Evaluation Conference*.

Koehn, P., Birch, A., and Steinberger, R. (2009). 462 machine translation systems for Europe. *Proceedings of MT Summit XII*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

Patel, R., Gupta, R., Pimpale, P., and Sasikumar, M. (2013). Reordering rules for English-Hindi SMT. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*.

Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*.

Ramanathan, A., Hegde, J., Shah, R., Bhattacharyya, P., and Sasikumar, M. (2008). Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation. In *IJCNLP*.

Sinha, R., Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R., and Jain, A. (1995). ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages. In *IEEE International Conference on Systems, Man and Cybernetics*.