Neural Comput & Applic (2019) 31:2455–2467
DOI 10.1007/s00521-017-3206-2

ORIGINAL ARTICLE

Statistical machine translation of Indian languages: a survey

Nadeem Khan Jadoon[1] · Waqas Anwar[2] · Usama Ijaz Bajwa[2] · Farooq Ahmad[2]

Received: 12 July 2015 / Accepted: 21 May 2016 / Published online: 17 November 2017

Abstract In this study, performance analysis of a state-of-art phrase-based statistical machine translation (SMT) system is presented on eight Indian languages. State of the art in SMT on different Indian languages to English language has also been discussed briefly. The motivation of this study was to promote the development of SMT and linguistic resources for these Indian language pairs, as the current systems are in infancy stage due to sparse data resources. EMILLE and crowdsourcing parallel corpora have been used in this study for experimental purposes. The study is concluded by presenting the performance of baseline SMT system for Indian languages (Bengali, Gujarati, Hindi, Malayalam, Punjabi, Tamil, Telugu and Urdu) into English with average 10–20 % accurate results for all the language pairs. As a result of this study, both of these annotated parallel corpora resources and SMT system will serve as benchmarks for future approaches to SMT in Hindi → English, Urdu → English, Punjabi → English, Telugu → English, Tamil → English, Gujarati → English, Bengali → English and Malayalam → English.

Keywords Statistical machine translation (SMT) · Parallel corpus · Natural language processing (NLP) · Phrase-based translation

1 Introduction

In this section, a brief background of machine translation is given. An overview of machine translation (MT) approaches is also discussed with the SMT approach being used in this research work. Indian languages selected for this work are also discussed briefly.

1.1 Machine translation

Machine translation (MT) can be defined as an automated system that analyses text from a source language (SL), by applying some computation on that input, and produces equivalent text in a required target language (TL) ideally without any kind of human intervention.

It is one of the most interesting and the hard problem in the field of NLP [1]. The two challenges in machine translation are adequacy and fluency. The former is to develop a system that adequately represents the ideas expressed in the source language into the target language. The latter is to represent those ideas grammatically. The common approaches to machine translation are the rule-based approach and corpus-based approach.

In the rule-based approach, the text in the source language is analyzed using various tools such as: a morphological parser and analyzer and then transformed into an intermediate representation. A set of rules are used to generate the text in target language of this intermediate representation. A large number of rules are necessary to capture the phenomena of natural language. These rules transfer the grammatical structure of the source language into target language. As the number of rules increases, the system become more complicated [2] and slow to translate. Formulation of a large number of rules is a tedious process and requires years of effort and linguistic analysis.

✉ Waqas Anwar
waqas@ciit.net.pk

[1] Department of Computer Science, COMSATS Institute of Information Technology, Abbottabad, Pakistan

[2] Department of Computer Science, COMSATS Institute of Information Technology, Lahore, Pakistan

In another approach, large parallel and monolingual corpora are used as source of knowledge. This approach can be further divided into statistical approach and example-based approach. In the statistical approaches, target text is generated and scored through a statistical model, from parallel corpus. Here, MT is also identified as a decision problem, and a better target language phrase id is decided from the given source language. Further, Bayes rule and statistical decision theory are used to solve this decision problem. Statistical decision theory and Bayesian decision rules are used to minimize errors of decision. SMT [1] gives better results if additional training data are available.

SMT is superior to rule-based and example-based systems in that it does not require human interpenetration and can build a translation system in an unsupervised manner directly from the training data. With the rapid proliferation of internet and increasing availability of data, SMT is currently the most popular and prevalent paradigm. SMT can be represented by different models, and a basic architecture of simple SMT system model is shown in Fig. 1. An arrow from translation model to language model shows that the language model contains the target side corpora as well. The arrow from language model to translation text shows that the fluency of the translation depends upon the quality of language model. In this study, we use phrase-based SMT model, and an overview of this model is given in the next section.

1.1.1 Phrase-based model

In this work, the phrase-based SMT models [3, 4] are used and their performance is evaluated on the morphologically rich Indian languages. Phrase-based models are used to translate phrases of one or more words as atomic units [1].

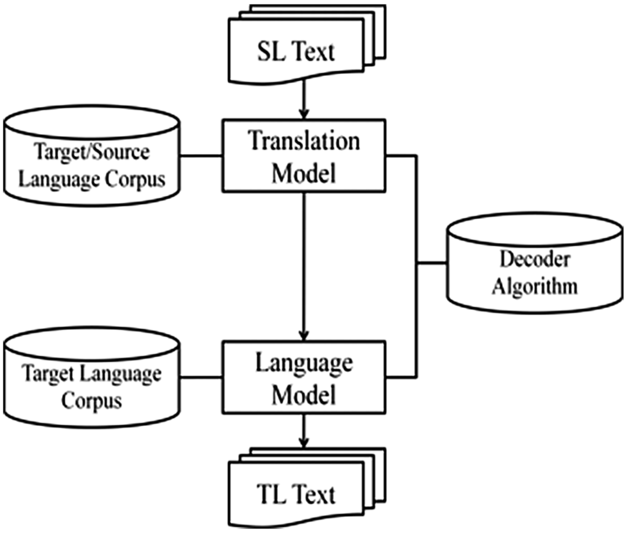

Fig. 1 Architecture of a typical SMT system

These models divide the input sentence into phrases and produce the target phrases, and at the end reordering of these phrases is done. Phrase-based models memorize local dependencies such as short reordering, idiomatic collocations, insertions and deletions.

In addition, phrase-based models are based on the noisy channel model introduced by [5] in the information theory. Given a source sentence *F,* the objective is to find a target sentence *E*, which maximizes the likelihood of two components, the translation (or adequacy) and the language (or fluency model).

Every sentence *F* is an arrangement of words symbolized as $f_1^J = f_{1\ldots}f_{j\ldots}f_J$ is decoded into a sentence E of target language, and symbolized as $e_1^I = e_{1\ldots}e_{i\ldots}e_I$. The objective is to find a target sentence that maximizes the model:

$$\hat{e}_1^I = \text{argmax}\, P\left(e_1^I|f_1^J\right) \tag{1}$$

For decoding sentence f_1^J into sentence e_1^I, we require to calculate $P\left(e_1^I|f_1^J\right)$, the translation model probability. Using Bayes theorem, we can decompose the above equation as:

$$P\left(e_1^I|f_1^J\right) = \frac{P\left(f_1^J|e_1^I\right)\cdot P\left(e_1^I\right)}{P\left(f_1^J\right)} \tag{2}$$

Subsequently, the goal is to get the most out of general probable translation hypotheses for the specified source sentence f_1^J. Equation 2 will be computed for every sentence in Language *E*. But *P* (f_1^J) is not modified for every translation hypothesis. Therefore, we can neglect the denominator *P* (f_1^J) from Eq. 2.

$$\hat{e}_1^I = \text{argmax} P\left(f_1^J|e_1^I\right)\cdot P\left(e_1^I\right) \tag{3}$$

The model of the likelihood distributed for the first term in Eq. 3 $\left(P\left(f_1^J|e_1^I\right)\right)$, probability of translation (*f,e*) is called *translation model*, and the distribution of $P\left(e_1^I\right)$ is called the *language model*.

1.2 Language selection

In this work, eight common spoken languages in the subcontinent are selected. Parallel corpus of all the languages is available for the experiment.

Bengali (Bangla) Bengali is the national language of Bangladesh and one of the officially spoken languages of India. More than 21 million people speak Bengali as their either first or second language [6]. There are roughly 10 million native speakers of Bengali in Bangladesh and around 85 million in India in the states like West Bengal, Assam and Tripura. Bengali is also known as Bangla, and it is associated with Indo-Iranian family. Like most languages it is also written from left to right. Its sentence structure is similar to English, i.e., subject object verb (SOV). All letters are written in same case, and there are no capital letters. The source of its punctuation is English language of nineteenth century.

Gujarati It is a member of Indo-Aryan branch of languages. Forty-six million people in the Indian state of Gujarat speak Gujarati [7]. Evolution of Gujarati language took place in twelfth century. Gujarati declension is considerably complicated. It contains three genders masculine, feminine, and neuter and two numbers singular and plural. For nouns it has three cases nominative, oblique and agentive locative. It is written from left to right with writing style SOV.

Hindi It is the national and official language of India. Four hundred twenty-five million people speak Hindi as their first language and more than 12 million people as their second language [8]. Outside India, some communities in South Africa, Mauritius, Bangladesh, Yemen and Uganda also communicate in Hindi language. Hindi is a member of the Indo-Aryan group within the Indo-Iranian branch of the Indo-European language family. Like in Persian, Hindi adjectives do not change as a result of number change in noun. Its preposition is similar to English. Unlike other Sanskrit-based languages like Gujarati, it has only two genders, i.e., masculine and feminine. Case marking in Hindi is simple due to Persian influence and reduces it to direct form and an oblique form. Case relations are shown postpositions. Like many languages it is also written from left to right, but its writing style is SOV. Modern standard Hindi evolved from the interaction of Muslim from Afghanistan, Iran, Turkey, Central Asia and elsewhere.

Due to Persian influence Hindi borrowed some part of vocabulary from Persian language such as dresses [e.g., پاجامہ, pajama (trouser); ، چادر chadar (sheet)], cuisine [(e.g., قورمہ, korma; کباب, kebab)], cosmetics [e.g., صابن, sabun (soap); حنا, hina, hen-na], furniture [e.g., کرسی, kursi (chair); میز, maiz (table)], construction [e.g., دیوار (wall)].

A large number of adjectives and their nominal derivatives (e.g.,-abad-inhabitedand-abadi-population) and a wide range of other items and concepts are so much a part of the Hindi language that purists of the post-independence period have been unsuccessful in purging them. While borrowing Persian and Arabic words, Hindi also borrowed phonemes, such as /f/ and /z/, though these were sometimes replaced by /ph/ and /j/. For instance, Hindi renders the word for force as either zor or jor and the word for sight as nazar or najar. In most cases the sounds /g/ and /x/ were replaced by /k/ and /kh/, respectively. Contact with the English language has also enriched Hindi. Many English words, such as button, pencil, petrol and college are fully assimilated in the Hindi lexicon.

Malayalam Malayalam is also a widely spoken language in India, mainly in the state of Kerala where it is an official language. In Tamil Nadu and Karnataka, few societies communicate in Malayalam language. It belongs to South Dravidian which is subpart of Dravidian language. Around 35 million people speak this language [9]. There exist different slangs between social caste lines which causes diglossia, i.e., difference between formal, literary and colloquial forms of speech. Like other Dravidian languages it also has a series of retroflex constants **(/ḍ/, /ṇ/, and /ṭ/)** pronounce by touching the tip of tongue to the roof of the mouth. Its writing style is SOV and has nominative accusative case marking pattern. It has three genders, i.e., masculine, feminine and neuter. Inflection is generally marked via suffixation. Unlike other Dravidian languages, Malayalam inflects its finite verb only for tense—not for person, number or gender.

Punjabi (Panjabi) It is a member of the Indo-Aryan subgroup of the Indo-European language family. More than 10 million people speak this language [10] in the domain that was discordant between Pakistan and India during cleave. This language is officially added in Indian constitution. Some small societies in UAE, UK, USA, Canada, South Africa and Malaysia speak Punjabi. It is of two miscellanies; one is western which is known as Lahnda and second is eastern known as Gurmukhi. There are two ways to write Punjabi, one is by Perso-Arabic script and other is by Gurmukhi alphabets which were conceived by Sikh Guru Angad (1539-52) rules for scriptural use. Its writing style is SOV and written from left to right (Gurmukhi) and right to left (Perso-Arabic).

Tamil Tamil is the member of Dravidian language and is the official language of the Tamil Nadu state. It is also the official language in Sri Lanka and Singapore and is also spoken by many people is Malaysia, Mauritius, Fiji and South Africa. In 2004, it was declared as classical language of India which means it met three criteria; its origins are ancient; it has an independent tradition; and it possesses a considerable body of ancient literature. Around 66 million people speak Tamil language [11].

Three times, changes occurred in grammatical and lexical form of this language, Old Tamil (from about 450 BCE to 700 CE), Middle Tamil (700 CE 1600 CE) and Modern Tamil (from 160 CE onwards). Its writing system developed from Brahmi script. Over the time its letters changed shapes until sixteenth century CE when printing was introduced and its shape stabilized. The major addition to the alphabet was the incorporation of Grantha letters to write unassimilated Sanskrit words, although a few letters with irregular shapes were standardized during the modern period. A script known as Vatteluttu (round script) is also in common use. With time, changes in the way of speaking this language occurred. Tamil language spoken in India is different from that which is spoken in Sri Lanka. Its writing style is SOV, and within Tamil Nadu there are phonological differences between the northern, western and southern speech.

Telugu Telugu is one of the most spoken languages among the Dravidian language family. In southeastern part of India, people communicate in this language. In Andhra Pradesh it is the official language. Worldwide, 75 million people speak Telugu language [12]. The oldest material belonging to this language is of 575 CE. The Telugu script is used for writing Telugu, which is derived from Calukya Dynasty. Its writing style is SOV and written from left to right. Visually, it differs from many of the North Indian scripts in that the letters have a rounded base.

Urdu is also a member of the Indo-Aryan group within the Indo-European family of languages. Urdu is the national language of Pakistan, while it is officially recognized language in Indian constitution as well. More than 100 million people [13] within Pakistan and India speak in Urdu. Apart from these two nations Urdu is also spoken by the immigrants and in small societies in UK, USA and UAE. Urdu and Hindi are bilaterally audible. This language developed and stemmed from Indian subcontinent; therefore, it is similar to Hindi. Due to similarity in phonics and grammar, they seem like one language but there sources are different. Urdu is lent from Arabic and Persian, while Hindi is borrowed from Sanskrit that is why they are treated as maverick languages. There is a huge difference in their writing style. Urdu script is an altered and revised form of Perso-Arabic scripts, while Hindi script is a modified form of Devanagari script. Urdu and Hindi sound similar except few variations in short vowel allophones. Urdu withholds a full set of aspirated stops. It is the property of both Indo-Aryan and retroflex stops. Urdu does not retain the complete range of Perso-Arabic consonants, despite its heavy borrowing from that tradition. The largest number of sounds retained is among the spirants; a group of sounds uttered with a friction of breath against some part of the oral passage, in this case /f/, /z/, /zh/, /x/, and /g/. One sound in the stops category, the glottal /q/, has also been retained from Perso-Arabic. Grammatically, Hindi and Urdu are same. Major difference between these two is Urdu is written from right to left, while Hindi is written from left to right. Style of Urdu writing is SOV and exhibit split ergative behavior. In Urdu, Perso-Arabic prefixes and suffixes are more than Hindi. Examples include the prefixes *dar*- “in,” *ba-/baa*- “with,” *be-/bila-/la*- “without” and *bad*- “ill, miss” and the suffixes -*dar* “holder,” -*saz* “maker” (as in *zinsaz* “harness maker”), -*khor* “eater” (as in *mufitkhor* “free eater”) and -*posh* “cover” (as in *mez posh* “table cover”).

1.3 Related work

Initial research has been done to translate Indian languages, mostly focusing Hindi and Bengali. However, most of the focus is still rule based because of the unavailability of parallel data to build SMT systems for these languages.

Dasgupta et al. [14] proposed an approach for English to Bangla MT that uses syntactic transfer of English sentences to Bangla with optimal time complexity. In generation stage of the phrases they used a dictionary to identify subject, object and also other entities like person, number and generate target sentences. Naskar and Bandyopadhyay [15] presented an example-based machine translation system for English to Bangla. Their work identifies the phrases in the input through a shallow analysis, retrieves the target phrases using the example-based approach and finally combines the target phrases using some heuristics based on the phrase reordering rules from Bangla. The authors also discussed some syntactic issues between English and Bangla. Anwar et al. [16] proposed a method to analyze syntactically Bangla sentence using context-sensitive grammar rules which accepts almost all types of Bangla sentences including simple, complex and compound sentences and then interpret input Bangla sentence to English using a NLP conversion unit. The grammar rules employed in the system allow parsing five categories of sentences according to Bangla intonation. The system is based on analyzing an input sentence and converting into a structural representation (SR). Once an SR is created for a particular sentence, it is then converted to corresponding English sentence by NLP conversion unit. For conversion, the NLP conversion utilizes the corpus. Islam et al. [2] proposed a phrase-based statistical machine translation (SMT) system that translates English sentences to Bengali. They added a transliteration module to handle OOV words. A preposition handling module is also incorporated to deal with systematic grammatical differences between English and Bangla. To measure the performance of their system, they used BLEU, NIST and TER scores. Durrani et al. [17] also made use of transliteration to aid translation between Hindi and Urdu which are closely related languages. Roy [18] applied three reordering techniques namely lexicalized, manual and automatic reordering to the source and language in a Bangla–English SMT system. Singh et al. [19] presented a phrase-based model approach to English–Hindi translation. In their work they discussed the simple implementation of default phrase-based model for SMT for English to Hindi and also give an overview of different machine translation applications that are in use nowadays.

Sharma et al. [20] presented English to Hindi SMT system using phrase-based model approach. They used human evaluation metrics as their evaluation measures. These evaluations cost higher than the already available automatic evaluation metrics. Yamada and Knight [21] used methods based on tree-to-string mappings where source language sentences are first parsed and later operations on each node. Eisner [22] presented issues of working with isomorphic trees and presented a new approach of non-isomorphic tree-to-tree mapping translation model using synchronous tree substitution grammar (STSG). Liu et al. [23] first gave idea of using maximum entropy model based on source language parse trees to get n-best syntactic reorderings of each sentence which was further extended to use of lattices.

Bisazza and Federico [24] further explored lattice-based reordering techniques for Arabic–English; they used shallow syntax chunking of the source language to move clause-initial verbs up to the maximum of six chunks where each verb's placement is encoded as separate path in lattice and each path is associated with a feature weight used by the decoder.

Jawaid et al. [25] presented complete study work for English to Urdu MT that uses factored-based MT. In their work they discussed the complete divergence between two languages. Vocabulary difference between Urdu and English has been discussed. The authors showed the importance of factored-based models when we obtained information about the morphology of both source and targeted language.

Khan et al. [26] presented baseline SMT system for English to Urdu translation using hierarchical model given by Chiang [27]. They also made a comparison of simple default phrase-based model with the hierarchical model and showed the performance of simple phrase based is much better for such local language like Urdu than the hierarchical phrase-based approach to SMT.

Singh [28] presented a Punjabi to Hindi machine translation system. The purposed system for Punjabi to Hindi translation has been implemented with various research techniques based on direct MT architecture and language corpus. The output is evaluated in order to get the suitability of the system for the Punjabi–Hindi language pair. Extensive research work can be found in the literature using neural networks technology in the field of MT which is recommended as a good approach by the researchers nowadays. Neural machine translation is a newly proposed approach in MT. The main drawback using the approach is it requires relatively large amount of training corpus as compared to SMT. Khalilov et al. [29] estimated a continuous space language model with a neural network in an Italian to English MT system. Bahdanau et al. [30] presented a neural machine translation by joint learning to align and translate.

In this work, phrase-based SMT models are used and their performance is evaluated on the morphologically rich Indian languages. These languages are low-resource languages in terms of the availability of MT systems (and NLP tools in general) yet together they represent nearly half a billion native speakers. Their speakers are well educated, with many of them speaking English either natively or as a second language. An important phenomenon present in these languages is a high degree of morphological complexity relative to English. Also Indian languages can be highly agglutinative, which means that words are formed by concatenating morphological affixes that convey information such as tense, person, number, gender, mood and voice. Morphological complexity is a considerable hurdle at all stages of the MT pipeline, particularly alignment, where inflectional variations mask patterns from alignment tools that treat words as fragments. Another important factor in these languages is head-finalness, exhibited most obviously in a subject–object–verb (SOV) pattern of sentence structure, in contrast to the general SVO ordering of English sentences.

2 Evaluation

In this section, we adopt two datasets used in the experiments followed by discussion on training, tuning and testing of different model components.

2.1 Dataset

2.1.1 EMILLE corpus

For this work, parallel corpora from diverse domains were collected for all the selected languages. For this purpose the corpus that is selected to use is Enabling Minority Language Engineering (EMILLE). EMILLE is a 63 million word corpus of Indic languages [31] which is distributed by the European Language Resources Association (ELRA). EMILLE contains data from six different categories: consumer, education, health, housing, legal and social documents. These data are based on the information leaflets provided by the UK government and various local authorities. There are 72 parallel files in total for five the source language with each filename consisting of language code, text type (written or spoken), genre and subcategory, connected with hyphen character. The data are encoded in full 2-byte unicode format and marked up in SGML format. The parallel corpus consists of 200,000 words of text in English and its accompanying translations in Hindi, Bengali, Punjabi, Gujarati and Urdu. Its bilingual resources consists of approximately 13,000 sentences for all the available languages from which we were able to sentence-align and extract over 8000 sentence for all languages pairing with English using the sentence alignment algorithm given by Moore [32]. Details about number of parallel sentences that were extracted for each pair are given in Tables 1 and 2.

A sufficiently large English language monolingual corpus is collected for this work. This monolingual corpus is

Table 1 Training and evaluation data for EMILLE

Corpus	Total sentences	Training sentence	Tuning sentence	Testing sentence
Bengali	8520	6816	852	852
Gujarati	8330	6664	833	833
Hindi	9510	7608	951	951
Punjabi	8465	6772	847	846
Urdu	8245	6596	825	824

Table 2 EMILLE vocabulary size for training and test set

Source language	Target language (English)					
	Training size (tokens)		Test size (tokens)		Total sentence pairs (tokens) including tuning sentence tokens	
	Source	Target	Source	Target	Source	Target
Bengali	98,952	90,523	11,073	10,123	124,745	113,923
Gujarati	89,995	86,594	10,328	9785	112,676	107,695
Hindi	137,623	102,754	15,583	11,517	172,352	128,741
Punjabi	110,014	89,136	13,602	10,554	123,616	99,690
Urdu	124,755	86,563	13,465	9222	138,220	95,785

used to build the language model that is used by the decoder to select the most affluent translation from several possible translation options. In this work, it is also tried to gather sufficiently large monolingual data from as many different available online resources as possible like *Europarl* [33]. The next step is to train the language model on the corpus that is suitable to the domain. To fulfill this need, data from diverse domains are collected. The main categories of the collected data are News, Religion, Health, Literature, Science and Education. The WMT 08 News Commentary dataset is used as the main entity for monolingual data, and the target side of the parallel corpora is also added to the monolingual data.

The monolingual corpora collected for this study have around 60 million tokens distributed in nearly 2 million sentences. These figures cumulatively present the number of tokens in all the domains whose data are used to build the language model. It includes monolingual data of the target languages of all parallel corpora collected for this study.

2.1.2 Crowdsourcing parallel corpus

Another notable effort toward creating parallel corpora for Indian languages has been carried out through the use of crowdsourcing [34]. The resource was created by employing large crowd of cheap translators to translate texts in Indian languages to English.

It contains parallel data for six languages, namely Bengali, Hindi, Malayalam, Tamil, Telugu and Urdu The following are nine categories: EVENTS, LANGUAGE AND CULTURE, PEOPLE, PLACES, RELIGION, SEX, TECHNOLOGY, THINGS or MISC. The number of segments used for training, tuning and testing of different language pairs is shown in Table 3.

Table 3 Training and evaluation data for Indic corpus

Corpus	Training	Tuning	Testing
Bengali–English	24,000	775	1000
Hindi–English	39,000	1000	1000
Malayalam–English	39,000	1000	1000
Tamil–English	46,000	1000	1000
Telugu–English	45,000	1000	1000
Urdu–English	87,000	980	883

2.2 Experimental setup

2.2.1 Corpus setup

For EMILLE corpus a k-fold cross-validation method is performed for sampling of the corpus for all language pairs. Here, $k = 5$ was selected by taking 4/5 of the total corpus as training and 1/5 as tuning and test set for experiment on all folds. Each fold comprises over 800 segments for tuning and same number of sentences for testing along with above 6500 segments for training for all source languages except Hindi. For Hindi the system got above 9000 segments in total, 7000 + selected for training and about 950 sentences for tuning and testing of Hindi to English translation system.

All these statistics can be seen clearly in Table 1. The first step in this work is sampling of data followed by training, tuning and test sets are tokenized for all folds.

Finally, all datasets are converted to lowercase. This process is repeated for all language pairs using scripts provided by Moses [35] decoder. The lowercase training data are used for word alignment.

2.2.2 Statistical machine translation model

Moses [35], a toolkit for experimenting with different classes of SMT models has been used. In the experiments, phrase-based SMT (PBSMT) for translation from Hindi → English, Urdu → English, Punjabi → English, Telugu → English, Tamil → English, Gujarati → English, Bengali → English, Malayalam → English has also been included. These classes of models are implemented in the Moses toolkit and thus provide a singular framework for carrying out experiments with different types of SMT models.

A Moses toolkit [35] is trained with the following features:

No.	Features	Description
1	Maximum sentence length of 80	
2	GDFA symmetrization of GIZA++ alignments [36]	GIZA++ [36] and the heuristics "grow-diag-final-and" are used to generate a word-aligned corpus, where bilingual phrases with maximum length 80 are extracted
3	Interpolated Kneser–Ney smoothed 5-g language model with SRILM [37] used at runtime	SRILM toolkits [37] to train a 5-g language model
4	5-g OSM [38]	
5	msd-bidirectional-fe lexicalized reordering model	The msd-fe reordering model has three features, which represent the probabilities of bilingual phrases in three orientations: monotone, swap or discontinuous. If a msd-bidirectional-fe model is used, then the number of features doubles: one for each direction
6	Sparse lexical and domain features [39]	
7	Distortion limit of 6	
8	100-Best translation options	
9	MBR decoding [40]	
10	Cube pruning [41] with a stack size of 1000 during tuning and 5000 during test	
11	No reordering over punctuation heuristic	

The system tuned with the *k*-best batch MIRA algorithm [42].

Language model is built on the available monolingual English corpus. This language model is implemented as an n-gram model using the SRILM [37] toolkit. For all the experiments in all languages, the same language model is used for all folds of the source languages as translation is being performed from Indian languages into English. For crowdsourcing parallel corpus experiments, the language model is trained using the monolingual WMT-13 shared task data which is built from 148 M English sentences.

2.3 Results

2.3.1 EMILLE corpus

As the languages used in this work are sparse-resourced, relatively lower scores for BLEU [43] were achieved with a mean of 0.12 and a standard deviation of 0.06 on the given test sets using the fivefold cross-validation method. Table 4 presents the results of experiments for all language pairs. The results are composed of BLEU and NIST score evaluated over the test corpora and also the UNK (OOV words) count over that test corpus for all the selected language pairs. The subsequent subsections present evaluation results for all language pairs for both seen, i.e., data taken from the training set and the unseen, i.e., testing data.

Bangla–English For Bengali–English language pair, a decent BLEU scores is achieved with a mean $X = 0.118$ and a standard deviation $\sigma = 0.043$ on unseen data and $X = 0.364$ with a standard deviation $\sigma = 0.018$ on seen data. For NIST obtained, $X = 3.786$ and a standard deviation $\sigma = 0.522$ on unseen data and $X = 7.878$ with a standard deviation $\sigma = 0.328$ on seen data.

When counting the unknown words in translation of this SMT system achieved $X = 610$ and a standard deviation $\sigma = 59$ on unseen data and $X = 130$ with standard deviation $\sigma = 8$ on seen data. An example of translation output from the trained system is given below. The example is composed of source segment with its reference translation from

Table 4 Evaluation results of developed SMT system for all language pairs

Language pair	BLEU		NIST		UNK count	
	Mean X	σ	Mean X	σ	Mean X	σ
Bengali–English	0.118	0.043	3.786	0.522	203	20
Gujarati–English	0.119	0.059	3.674	0.701	226	25
Hindi–English	0.115	0.068	3.779	0.804	224	30
Punjabi–English	0.150	0.09	4.185	1.158	197	36
Urdu–English	0.140	0.038	4.260	0.535	183	15

Table 5 Bangla–English phrase table for given example

S.No	Input Phrase	Reference Phrase
1	ডিপার ◌্ টমেন ◌্ ট অফ	the department of
2	দি এনভায়রণমেন ◌্ ট	of the environment
3	ট ◌্ রান ◌্ সপোর	Transport
4	ট এণ ◌্ ড দি রিজিওনস	and the regions

test corpus. A segmented output of translation output is also given.

Example
Source: ডিপার ◌্ টমনে ◌্ ট অফ দি এনভায়রণমনে ◌্ ট ট ◌্ রান ◌্ সপে ার ◌্ ট এণ ◌্ ড দি রিজিওনস

Reference: department of the environment transport and the regions

Output: the department of |0–5| the environment |6–9| transport |10–16| and the regions |17–22|

The indexes in the output represent which source words produced this output; for example, “the department of” was produced by a source phrase containing source words indexed between 0 and 5.

Table 5 presents input phrases along with corresponding reference phrases for the example mentioned above. A clear difference can be observed between the reference translation and the one achieved from the developed system. The translation output is segmented into different phrases, and decoder fetches the translation from the developed phrase table. The reordering model also gave poor result for such small amount of data.

In output the first six words of source are translated to “The department of,” then next three to “the environment,” then next five to just a single output “transport” and so on. Here, it can be noted that how sparseness affect the output, the phrase table contains only one single output word for five input words. Table 6 shows the actual BLEU, NIST score for all the folds along with the OOV words count.

Gujarati–English For this pair, again decent BLEU scores were achieved as compared to small amount of training corpus with a mean of $X = 0.119$ and a standard deviation $\sigma = 0.059$ on unseen data and $X = 0.403$ and a standard deviation $\sigma = 0.012$ on seen data.

For NIST we obtained, $X = 3.674$ and a standard deviation $\sigma = 0.701$ on unseen data and $X = 8.136$ and a standard deviation $\sigma = 0.153$ on seen training corpus.

When counting the unknown words in translation of this SMT system, $X = 678$ and a standard deviation $\sigma = 77$ on unseen data and $X = 117$ and a standard deviation $\sigma = 16$ on seen data were achieved.

An example of translation output from the trained system is given below. The example is composed of source segment with its reference translation from test corpus. A segmented output of translation output is also given.

Table 6 Evaluation results for Bangla–English translation

Folds	BLEU		NIST		UNK count	
	Seen	Unseen	Seen	Unseen	Seen	Unseen
F1	0.403	0.075	8.284	3.157	129	630
F2	0.342	0.082	7.590	3.36	120	670
F3	0.347	0.098	7.517	3.826	141	617
F4	0.363	0.153	7.899	4.249	122	621
F5	0.375	0.182	8.101	4.338	138	512

Table 7 Gujarati–English phrase table for given example

S. no	Input phrase	Reference phrase
1	અમુક બેનિફિટો માટે તમે નેશનલ	for some benefits you
2	ઇનશ	NULL
3	યોરન	No
4	શકોન	Your
5	ટ ◌્ રીબ ◌્ યુશનો	Contributions
6	ભરેલાં હોવાં	Have to have paid
7	જ જોઇએ	Must
8	અથવા	Or
9	માની લેવામાં આવશે	Be deemed to have ceased

Example
Source:

અમુક બેનિફિટો માટે તમે નેશનલ ઇનશ ◌્ યોરન ◌્ શકોન ◌્ ટ ◌્ રીબ ◌્ યુશનો ભરેલાં હોવાં જ જોઇએ અથવા એવું માની લેવામાં આવશે કે તમે તે ભરેલાં છે .

Reference: for some benefits you must have paid or be treated as having paid no contributions.

Output: for some benefits you |0–4| ઇનશ |5–5| your |8–9| no |6–7| contributions |10–16| must |19–20| have paid |17–18| or |21–21| be |22–22| taken |24–24| to.

Table 7 presents input phrases along with corresponding reference phrases for the example mentioned above. A

Table 8 Evaluation results for Gujarati–English translation

Folds	BLEU		NIST		UNK count	
	Seen	Unseen	Seen	Unseen	Seen	Unseen
F1	0.413	0.081	7.942	3.251	144	730
F2	0.397	0.079	8.215	3.146	106	709
F3	0.391	0.089	8.072	3.226	119	729
F4	0.399	0.131	8.104	3.968	108	677
F5	0.420	0.219	8.349	4.338	107	546

clear difference can be observed between the reference translation and the one achieved from the developed system. The translation output is segmented into different phrases, and decoder fetches the translation from the developed phrase table. The reordering model also gave poor result for such small amount of data.

In output the first four words of source are translated to "for some benefits you," then next word could not be translated by the decoder so it becomes an OOV in the translation output. From the phrase table it is seen that many source words translated to just single target output. This is also because of poor tokenization for regional languages as there is no standardized tokenizer available for these languages. Table 8 shows the actual BLEU, NIST score for all the folds along with the OOV words count.

Hindi–English The corpora used for Hindi–English language pair was the most domain relevant and the biggest in size. It resulted in significantly better translation as compared to other language pairs. Hence, it can be concluded that the size and relevance of parallel language corpus have a direct relationship with the quality of translation. For this pair, BLEU scores with a mean of $X = 0.115$ and a standard deviation $\sigma = 0.068$ on unseen data and $X = 0.352$ and a standard deviation $\sigma = 0.025$ on seen data were achieved. For NIST, $X = 3.779$ and a standard deviation $\sigma = 0.804$ on unseen data and $X = 7.634$ and a standard deviation $\sigma = 0.437$ on seen data were attained.

When counting the unknown words in translation of this SMT system, $X = 672$ and a standard deviation $\sigma = 90$ on unseen data and $X = 150$ and a standard deviation $\sigma = 10$ on seen data were noted. Translation output of the developed system is given below in example.

Example

Source: उनसे समंपर ◌़ के लिए पते व टेलीफोन नंबर नीचे दिए हैं ◌़:

Reference: contact addresses and telephone numbers are as follows:

Output: on |0–0| the |2–3| समंपर |1–1| for |4–5| addresses |6–6| and |7–7| telephone |8–8| helpline |9–9| below |10–10| दिए |11–11|:|12–12|

Table 9 presents input phrases along with corresponding reference phrases for the example mentioned above. A clear difference can be observed between the reference translation and the one achieved from the developed system. The translation output is segmented into different phrases, and decoder fetches the translation from the developed phrase table. The reordering model also gave poor result for such small amount of data.

In output the first word of source is translated to "on," then next two words were translated as "the," then again NULL token so it becomes an OOV in the translation output. From the phrase table it is seen that many source words are translated to just single target output. This is also because of poor tokenization for regional languages as there is no standardized tokenizer available for these languages. Table 10 shows the actual BLEU, NIST score for all the folds along with the OOV words count.

Punjabi–English For this pair, again a decent BLEU scores with a mean of $X = 0.15$ and a standard deviation $\sigma = 0.09$ on unseen data and $X = 0.385$ and a standard deviation $\sigma = 0.053$ on seen data were observed.

For NIST, $X = 4.185$ and a standard deviation $\sigma = 1.158$ on unseen data and $X = 7.754$ and a standard deviation $\sigma = 0.242$ on seen data were observed with relatively small amount of training parallel corpus.

Table 9 Hindi–English phrase table for given example

S.No	Input Phrase	Reference Phrase
1	उनसे	On
2	क के	The
3	समंपर	NULL
4	के लिए	For
5	टेलीफोन	Telephone
6	पते व	Addresses
7	नीचे दिए	Of the following
6	हैं◌:	:
7	नंबर	Helpline

Table 10 Evaluation results for Hindi–English translation

Folds	BLEU		NIST		UNK count	
	Seen	Unseen	Seen	Unseen	Seen	Unseen
F1	0.365	0.065	7.765	3.531	134	701
F2	0.381	0.074	8.036	3.076	155	735
F3	0.366	0.068	7.396	3.483	160	754
F4	0.323	0.151	7.677	5.114	154	637
F5	0.328	0.221	7.910	5.721	149	533

When counting the unknown words in translation of this SMT system, $X = 591$ and a standard deviation $\sigma = 110$ on unseen data and $X = 98$ and a standard deviation $\sigma = 13$ on seen data were achieved. The example given below composed of the input source with its reference from the parallel corpus and also the translation output from the developed system.

Example

Source: ਪਹਿਲਾਂ ਇਹ ਪਤਾ ਕਰੋ ਕਿ ਤੁਹਾਨੂੰ ਕਿਹੜੇ ਬੈਨੀਫ਼ਿਟ ਮਿਲ ਸਕਦੇ ਹਨ ।

Reference: check first what benefit or benefits you may be able to get.

Output: check first |0–3| what |6–6| benefits |7–7| that |4–4| you |5–5| can get. |8–11|.

All the segments/phrases of source input are given in above phrase table (see Table 11). Number of differences between the reference and the translation output of the developed system can be found. The translation output is segmented into different phrases, and decoder fetches the translation from the developed phrase table. The reordering model also gave poor result for such small amount of data.

In output the first three words of source are translated to "check first" then all other words were translated to single words in output even the last phrase of over two to four words also translated to single word. From the phrase table it is seen that many source words translated to just single target output. This is also because of poor tokenization in preprocessing for regional languages as there is no standardized tokenizer available for these languages. In Table 12 actual BLEU, NIST score for all the folds along with the OOV word count is presented.

Urdu–English For this language pair, BLEU scores with a mean of $X = 0.14$ and a standard deviation $\sigma = 0.038$ on unseen data and $X = 0.371$ and a standard deviation $\sigma = 0.027$ on seen data were observed. For NIST, $X = 4.26$ and a standard deviation $\sigma = 0.535$ on unseen data and $X = 7.54$ and a standard deviation $\sigma = 0.53$ on seen data were attained with small amount of training parallel corpus.

When counting the unknown words in translation of this SMT system, $X = 550$ and a standard deviation $\sigma = 45$ on unseen data and $X = 117$ and a standard deviation $\sigma = 12$ on seen data were achieved. The example given below shows the different kind of problems faced in getting translation output from the developed system.

Example

Source: .20 - بہتری کی یہ باتیں ایک عمدہ ابتدا ہیں

Reference: 20. These improvements are a good start.

Output: 20. |0–0| the |1–2| these |3–3| things to |4–4| start |7–7| a |5–5| quality |6–6|. |8–9| |||

In output the first word of source and target is same so decoder did nothing with it and its segment from phrase table will be NULL. The next word got totally different output in translation output as compared to the phrase table entry of Table 13. The two source words are translated to four-word phrase in phrase table, but in the translated output a single output translation was obtained. This is because of the n-best translation phrase for a single phrase input. Next, the poorly managed reordering by the baseline phrase-based model can be seen.

All this discussion with given output example leads us to a bottom-line conclusion that if a good tokenizer is there with more corpora for all the selected regional languages, it will lead to decent BLEU scores and fluent translations. Table 14 shows the actual BLEU, NIST score for all the folds along with the OOV word count.

Table 11 Punjab–English phrase table for given example

S.No	Input Phrase	Reference Phrase
1	ਪਹਿਲਾਂ ਇਹ ਪਤਾ ਕਰੋ ਕਿ	Check first
2	ਜਿਹੜੇ	That
3	ਬੈਨੀਫ਼ਿਟ	Benefits
4	ਤੁਹਾਨੂੰ	You
5	ਮਿਲ ਸਕਦੇ ਹਨ ।	can get .

Table 12 Evaluation results for Punjabi–English translation

Folds	BLEU		NIST		UNK count	
	Seen	Unseen	Seen	Unseen	Seen	Unseen
F1	0.409	0.099	7.913	3.094	121	677
F2	0.397	0.071	8.065	3.348	94	703
F3	0.346	0.095	7.871	3.172	93	615
F4	0.369	0.205	7.177	4.445	89	522
F5	0.408	0.283	7.147	4.839	94	440

Table 13 Urdu–English phrase table for given example

S. no.	Input phrase	Reference phrase
1	20.	Null
2	بہتری کی	The need for improvements
3	یہ	These
4	باتیں	Things to
5	ایک	A
6	عمدہ	Good quality
7	ابتدا	Start
8	ہیں -	.

Table 14 Evaluation results for Urdu–English translation

Folds	BLEU		NIST		UNK count	
	Seen	Unseen	Seen	Unseen	Seen	Unseen
F1	0.401	0.110	8.178	3.721	97	563
F2	0.343	0.097	7.219	3.777	113	573
F3	0.383	0.139	7.737	4.174	125	539
F4	0.388	0.161	7.613	4.784	123	597
F5	0.341	0.194	6.953	4.845	127	478

Table 15 Evaluation crowdsourcing parallel corpus

Language	Tuning	Test
Bengali–English	0.197	0.167
Hindi–English	0.193	0.16
Malayalam–English	0.111	0.09
Tamil–English	0.128	0.066
Telugu–English	0.142	0.110
Urdu–English	0.247	0.238

2.3.2 Crowdsourcing parallel corpus

The results from running state-of-the-art baseline systems on crowdsourcing parallel corpus are shown in Table 15. For these experiments, we additionally transliterated OOV words by unsupervised post-decoding transliteration method as described in Durrani et al. [44].

In addition, increasing data can improve BLEU scores in all the language pairs reported. However, the data available for Indian languages are still not enough to reliably estimate translation and reordering models. Table 2 shows that the vocabulary size is not good enough in numbers for training of the SMT system and it is creating data sparseness issue. Further, more data are required to produce better translations. Translation quality can also be improved by studying the similarities between these languages. Data sparseness can be overcome by using methods of triangulation [45, 46] and transliteration [17] which have been shown to be useful for closely related languages.

According to the result discussion given above, it is concluded that tokenizer is major problem in these languages for more accuracy in MT system. Urdu is morphologically rich language with different nature of its characters. Moreover, Urdu text tokenization and sentence boundary disambiguation are difficult as compared to the language like English. Major hurdle for tokenization is improper use of space between words, whereas the absence of case discrimination makes the sentence boundary detection a difficult task.

More specifically, issues of Urdu text tokenization can be divided into two categories: space inclusion issues and space exclusion issues. In Urdu text space is always needed when word ends with non-joiner character or when zero width non-joiner (ZWNJ) is used between two words. For example, "پرانیسڑک" (old road) are two words "پرانی" and "سڑک" without space and without ZWNJ. Space exclusion issues include compound words, for example "عزت و حرمت" (honor) and "طالب علم" (student), reduplication, for example "دھوم دھام" (pomp & show), "دن بدن" (day by day), and "صبح صبح" (early morning), affixation "خوش اخلاق" (polite) and "حیرت انگیز" (amazing), proper nouns "سعودی عرب" (Saudi Arabia) and "صالح بانو" (Sawliha Bano), English words "نیٹ ورک" (network), and abbreviations and acronyms, for example "این ایل پی" (NLP).

3 Conclusion and future work

The developed SMT system takes the Indian language sentences as input, and it generates corresponding closest translation in English. The translation of over 800 sentences was evaluated using automatic evaluation metric, i.e., BLEU evaluation. Due to the low BLUE scores reported in Tables 4 and 15, it is concluded that quality of translation is directly dependent on the scope and quality of parallel language corpora.

In this work all the Indian Languages used got pretty low parallel corpus. As all the eight Indian Languages used in this work exhibit rich morphology, thus resulting in sparse estimates which causes poor translation quality, therefore the results are not as good as the ones reported for the European languages [47] for which parallel and monolingual data are available.

In this study, phrase-based model was employed for training and MIRA was used for tuning of the system. A complete set of experiments is carried out by choosing the training, tuning and test sets from parallel corpus using the fivefold cross-validation method to make up the fact that only a small amount of parallel data were available. It is noted that each of the source Indian language got so much divergence when translating into English and that's why there is significant difference in obtained MT evaluation scores on seen corpus and on unseen test sets.

In future, SMT will be explored by applying other different approaches to develop language models and also the training model for all the South Asian languages whose more parallel corpus is available at the moment or may be available in nearer future. An exhaustive manual qualitative analysis of output translation has been done for all the selected language pairs. Both seen and unseen translation outputs were compared to get proper MT evaluation results as there were UNK (Untranslatable) words occurred in seen data translation as well.

Acknowledgement We would like to thank Dr. Nadir Durrani from University of Edinburgh for his helpful comments and suggestions during the experimentation and proof reading the write up, which has helped us a lot to improve the paper. He also provided examples to be included in the text.

References

1. Koehn P (2010) A book on statistical machine translation. Cambridge University Press, Cambridge
2. Islam Z, Tiedemann J, Eisele A (2010) English to Bangla phrase-based machine translation. In: Proceedings of the 14th annual conference of the European Association for Machine Translation
3. Koehn P, Och F, Marcu D (2003) Statistical Phrase-Based Translation. In: HLT-NAACL: conference combining Human Language Technology conference series and the North American chapter of the Association for Computational Linguistics conference series, pp 48–54
4. Och FJ, Ney H (2000) Improved statistical alignment models. In: Proceedings of the 38th annual meeting of the Association for Computational Linguistics (ACL), pp 440–447
5. Shannon CE (1948) A mathematical theory of communication. Bell System Tech J 27:379–423 and 623–656
6. The Editors of Encyclopædia Britannica (2014) Bengali language. *Encyclopedia Britannica Online*. Encyclopedia Britannica, n.d. Web. 15 June 2014
7. The Editors of Encyclopædia Britannica (2014) Gujarati language. *Encyclopedia Britannica Online*. Encyclopedia Britannica, n.d. Web. 18 June 2014
8. The Editors of Encyclopædia Britannica (2014) Hindi language. *Encyclopedia Britannica Online*. Encyclopedia Britannica, n.d. Web. 18 June 2014
9. The Editors of Encyclopædia Britannica (2014) Malayalam language. *Encyclopedia Britannica Online*. Encyclopedia Britannica, n.d. Web. 18 June 2014
10. The Editors of Encyclopædia Britannica (2014) Punjabi language. *Encyclopedia Britannica Online*. Encyclopedia Britannica, n.d. Web. 18 June 2014
11. The Editors of Encyclopædia Britannica (2014) Tamil language. *Encyclopedia Britannica Online*. Encyclopedia Britannica, n.d. Web. 18 June 2014
12. The Editors of Encyclopædia Britannica (2014) Telugu language. *Encyclopedia Britannica Online*. Encyclopedia Britannica, n.d. Web. 18 June 2014
13. The Editors of Encyclopædia Britannica (2014) Urdu language. *Encyclopedia Britannica Online*. Encyclopedia Britannica, n.d. Web. 18 June 2014
14. Dasgupta S, Wasif A, Azam S (2004) An optimal way towards machine translation from English to Bengali. In: Proceedings of the 7th international conference on computer and information technology (ICCIT)
15. Naskar SK, Bandyopadhyay S (2006) A phrasal EBMT system for translating English to Bengali. In: Workshop on language, artificial intelligence and computer science for natural language processing applications, Bangkok, Thailand, pp 69–72
16. Anwar MM, Anwar MZ, Bhuiyan MA-A (2009) Syntax analysis and machine translation of Bangla sentences. Int J Comput Sci Netw Secur 9:317–326
17. Durrani N, Sajjad H, Fraser A, Schmid H (2010) Hindi-to-Urdu machine translation through transliteration. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp 465–474
18. Roy M (2009) A semi-supervised approach to Bengali–English phrase-based statistical Machine Translation. In: Proceedings of the 22nd Canadian conference on artificial intelligence
19. Singh D et al (2012) Modeling phrase based SMT for English to Hindi language. Int J Res Rev Eng Sci Technol 1:95–99
20. Sharma N (2011) English to Hindi statistical machine translation system. Dissertation, Thapar University, Patiala
21. Yamada K, Knight K (2001) A *syntax*-based statistical translation model. In: Proceedings of the 39th annual meeting of the ACL, pp 523–530
22. Eisner J (2003) Learning non-isomorphic tree mappings for machine translation. In: Proceedings of the ACL interactive poster/demonstration sessions, pp 205–208
23. Liu T, Che W, Li S, Hu Y, Liu H (2005) Semantic role labeling system using maximum entropy classifier. In: Proceedings of CoNLL, pp 189–192
24. Bisazza A, Federico M (2010) Chunk-based verb reordering in VSO sentences for Arabic–English statistical machine translation. In: Proceedings of the joint fifth workshop on statistical machine translation and metrics MATR, WMT'10, pp 235–243
25. Jawaid B, Zeman D (2011) Word-order issues in English-to-Urdu statistical machine translation. Prague Bull Math Linguist 95:87–106 **(ISSN 0032-6585)**
26. Khan N, Anwar W, Bajwa U, Durrani N (2013) English to Urdu hierarchical phrase based SMT system. In: The fourth workshop n South and Southeast Asian NLP (WSSANLP), International joint conference on natural language processing, pp 72–76
27. Chiang D (2005) A hierarchical phrase-based model for statistical machine translation. In: Proceedings of the 43rd annual meeting on Association for Computational Linguistics (ACL)
28. Singh G (2008) A Punjabi to Hindi Machine translation system. In: Proceeding of COLING, 22nd international conference on computational linguistics

29. Khalilov M et al (2008) Neural network language models for translation with limited data. In: Proceedings of 20th IEEE international conference on tools with artificial intelligence, Dayton, Ohio, pp 445–451
30. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473
31. Baker P, EMILLE (2002) A 70-million word corpus of indic languages: data collection, mark-up and harmonization. In: Proceedings of the 3rd language resources and evaluation conference, pp 819–825, LREC'
32. Moore R (2002) Fast and accurate sentence alignment of bilingual corpora. In: Conference of the association for machine translation in the Americas
33. Koehn P (2005) EuroParl: a parallel corpus for statistical machine translation. The tenth Machine Translation Summit, Phuket, Thailand, pp 79–86
34. Post M, Callison-Burch C, Osborne M (2012) Constructing parallel corpora for Six Indian languages via Crowdsourcing. In: Proceedings of the seventh workshop on statistical machine translation, pp 401–409
35. Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th annual meeting of the Association for Computational Linguistics companion volume proceedings of the demo and poster sessions. Association for Computational Linguistics, Prague, Czech Republic, pp 177–180
36. Och F (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51
37. Stolcke A (2002) SRILM-an extensible language modeling toolkit. In: Proceedings of the international conference on spoken language processing, Denver, Colorado, pp 257–286
38. Durrani N, Fraser A, Schmid H, Hoang H, Koehn P (2013) Can Markov models over minimal translation units help phrase-based SMT? In: Proceedings of the 51st annual meeting of the association for computational linguistics
39. Hasler E, Haddow B, Koehn P (2012) Sparse lexicalised features and topic adaptation for SMT. In: Proceedings of the seventh international workshop on spoken language translation, pp 268–275
40. Kumar S, Byrne WJ (2004) Minimum bayes-risk decoding for statistical machine translation. The fifth meeting of the North American Chapter of the ACL, Boston, USA, pp 169–176
41. Huang L, Chiang D (2007) Forest rescoring: faster decoding with integrated language models. In: Proceedings of the 45th annual meeting of the association of computational linguistics, pp 144–151
42. Cherry C, Foster G (2012) Batch tuning strategies for statistical machine translation. In: Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 427–436
43. Papineni K (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of 40th annual meeting of the Association for Computational Linguistics (ACL), pp 311–318
44. Durrani N, Sajjad H, Hoang H, Koehn P (2014) Integrating an unsupervised transliteration model into statistical machine translation. In: Proceedings of the 15th conference of the European chapter of the ACL
45. Cohn T, Lapata M (2007) Machine Translation by triangulation: making effective use of multi-parallel corpora. In: Proceedings of the 45th annual meeting of the association of computational linguistics, Prague, Czech Republic, pp 728–735
46. Bertoldi N, Barbaiani M, Federico M, Cattoni R (2008) Phrase-based statistical machine translation with pivot languages. In: International workshop on spoken language translation evaluation campaign on Spoken Language Translation (IWSLT), Hawaii, USA, pp 143–149
47. Koehn P, Monz C (2005) Shared task: statistical machine translation between European languages. In: Proceedings of the ACL workshop on building and using parallel texts