Computación y Sistemas
ISSN: 1405-5546
computacion-y-sistemas@cic.ipn.mx
Instituto Politécnico Nacional
México

Kumar Arora, Karunesh; Agrawal, Shyam S.
Pre-Processing of English-Hindi Corpus for Statistical Machine Translation
Computación y Sistemas, vol. 21, núm. 4, 2017, pp. 725-737
Instituto Politécnico Nacional
Distrito Federal, México

ISSN 2007-9737

Pre-Processing of English-Hindi Corpus for Statistical Machine Translation

Karunesh Kumar Arora[1], Shyam S. Agrawal[2]

[1] Centre for Development of Advanced Computing, Noida, India

[2] KIIT Group of Institutions, Sohna Road, Bhondsi, Gurugram, India

karunesharora@cdac.in, ss_agrawal@hotmail.com

Abstract. Corpus may be considered as fuel for the data driven approaches of machine translation. Parallel corpus building is a labour intensive task, which makes it a costly and scarce resource. Full potential of available data needs to be exploited and this can be ensured by removing different types of inconsistencies as being faced throughout the NLP domain. The paper presented here describes the experiments carried out on corpus text pre-processing for building the baseline Statistical Machine Translation (SMT) system. Text pre-processing performed here is classified in two stages – i. the first one relates to handling of orthographic representation of content and ii. the second stage relates to handling of non-lexical words. The first stage covers punctuation symbols, casing, word spellings and their normalization while second stage covers handling of numbers and named entities (NEs) applied on the best settings observed in first stage. The motivation behind performing these experiments was to derive a relationship and gauge the extent of pre-processing the corpus, thereby building a considerably optimized baseline SMT system. This baseline system would provide platform for performing further experiments with different syntactic and semantic factors in future. The findings presented here is for English-Hindi language pair, however, the concept of pre-processing is language neutral and can be transcended to any other language pair. The best performance is reported with retaining the punctuation symbols, lower-cased English corpus and spell normalized Hindi corpus for English to Hindi translation. Further to these, in the second stage of experiments, handling numbers and Named Entities have been described wherein these are mapped to unique class labels. The impact of these experiments have been explained with their appropriateness for the concerned language pair.

Keywords. Statistical machine translation, preprocessing, normalization, named entity handling.

1 Introduction

A famous saying by Mercer's reads, "There is no data like more data". This advocates that for Machine Learning problems, the key of success lies in collecting more and more data to solidify the probabilistic evidences. Despite of this belief, obtaining such large volume of data is not feasible at times. When we deal with data scarce language pair, full potential of available data needs to be exploited. Analyzing the corpus reveals that different types of noises are present in the available resources. These include improper and inconsistent usage of punctuation markers, inconsistency in casing and spellings. This problem becomes more severe when we deal with a language like Hindi which is not only morphologically rich but also showcases a lot of spell variations in use. For SMT, consistency in data in the both training and testing scenarios needs to be ensured. The paper presented here describes handling mono-lingual inconsistencies in terms of usage of punctuation markers, casing and spell normalization of the corpus on both source and target sides (Stage-I).

Besides this, non-lexical terms like numbers, dates and named entities (NEs) are much more variable in their presence in the text due to belonging to an open set. These variations present challenge in handling them irrespective of the language concerned. Presence of each NE cannot be ensured in the training corpus. To avoid any unseen NE behaving as Out Of Vocabulary (OOV), these are mapped to unique class labels while

building statistical models. These are replaced with their target language transliterated forms in the final translation. To establish empirical ground for proposed hypothesis, we evaluate and compare different settings. The impact of these preprocessing is observed through SMT quality improvements measured using BLEU scores.

The findings may also help in standardizing the data cleaning process and evaluating the quality of available data resources. English and Hindi corpora are used here as the basis for study. The objective of these experiments is to prepare a most suitable baseline SMT system between these languages, for performing statistical translation with different syntactic and semantic factors later.

The rest of this paper is structured as follows. Section 2 brings a description of related works. Section 3 describes Statistical Machine Translation in brief, the pre-processing in relation to casing, punctuation symbols and Hindi spell normalization is described in Section 4. Section 5 presents the corpus statistics and section 6 details the experiments carried out to evaluate the impact of various settings. Results and observations are described in Section 7. Finally, Section 8 finishes this paper with conclusions and proposals for future work.

2 Related Work

There have been many research efforts on spelling error corrections and it is established as a field of research in itself and in relation to the Machine Translation. Yet, pre-processing of corpus in terms of punctuation markers, casing and spell normalization has not gained much attention.

Sproat et al. [1] has also said that "text normalization is not a problem that has received a great deal of attention, and approaches to it have been mostly ad hoc: to put the issue somewhat bluntly, text normalization seems to be commonly viewed as a messy chore".

Caseli et al. [2] carried out experiments on analyzing the impact of automatic casing and punctuation changes and have shown that these changes have significant impact on translation performance. These experiments have not taken the spell normalization and non-lexical items like numbers, dates and named entities (NEs) into consideration.

Bojar et al. [3] have looked into the data normalization issues in phrase-based machine translation but have not reported any experiment with punctuation and casing handling.

Santanu et al. [4] have described handling of Named Entities but does not talk about any other preprocessing. Various related works on pre-processing have shown that datasets require preprocessing depending upon their intended use.

Lane et al. [5] used class-based translation and language models in speech-to-speech translation in travel domain and presented performance improvement by using a mechanism to handle out-of-vocabulary words.

Markov et al. [6, 7] presented an approach that applies simple pre-processing steps, such as replacing digits, splitting punctuation marks and replacing named entities, before extracting character n-gram features. They examined the effect of preprocessing steps on Authorship Attribution. Similar to the results reported by them, the paper being presented here reports the effect of various pre-processing steps.

Sellami et al. [8] experimented with using a third language as pivot to automatically label multilingual parallel data for Arabic-French pair with NE tags and built lexical database of NEs for facilitating their translation and transliteration.

Okuma et al. [9] proposed a method for replacing the words unseen in the training corpus with high frequency words and have shown gain in translation quality on manual inspection. The paper presented here addresses handling of punctuation, casing, normalization all together and findings with different combinations are reported along with the impact of handling non-lexical entities like numbers, dates and named entities.

3 Statistical Machine Translation

The background papers on this subject [10, 11] describe the statistical machine translation as, that if we are given a source language sentence $S = s_1^I = s_1 \ldots s_i \ldots s_I$, which is to be translated into a target language sentence $T = t_1^J = t_1 \ldots t_j \ldots t_J$.

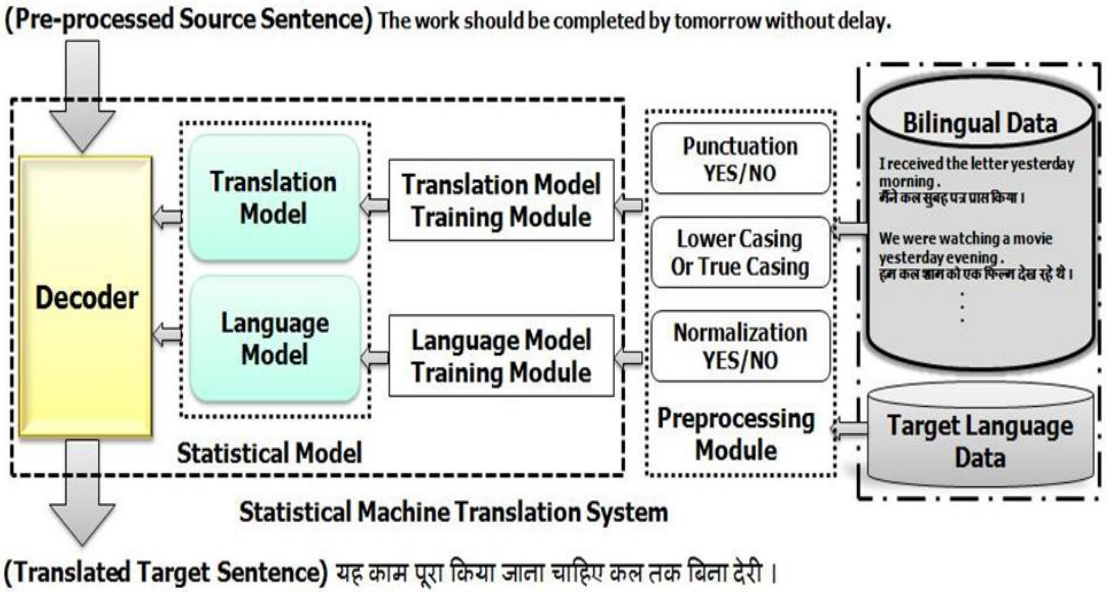

Fig. 1. Phrase-based machine translation with pre-processing

Statistical machine translation is based on a noisy channel model. It considers T to be the target of a communication channel, and its translation S to be the source of the channel.

System may generate multiple translation sentences options and the problem of translation becomes identifying sentence T which fits as the best translation of the source sentence S. Hence, the machine translation task becomes to recover the source from the target.

So, we need to maximize P(T|S). According to the Bayes rule:

$$t^* = \arg\max_t P(t \mid s) = \arg\max_t \frac{P(s \mid t) * P(t)}{P(s)}. \quad (1)$$

As P(S) is constant, then we have:

$$t^* = \arg\max_t P(s \mid t) * P(t). \quad (2)$$

Here, P(s|t) represents, translation model and P(t) represents language model. Translation model plays role of ensuring translation faithfulness and language model is to ensure fluency of translated output.

4 Pre-Processing and Methodology

Pre-processing described in the paper is related to casing, punctuation symbols, spell normalization, numbers and named entities. These are described in the following sub-sections. Fig. 1 shows the steps of a phrase-based SMT system with pre-processing in Stage-I experiments.

The bilingual and monolingual data are pre-processed before preparing translation models and language models. These trained models are used by decoder for translating a given source to target language sentence.

4.1 Casing

Capitalization is language specific feature. English uses capitalization, while Hindi does not have this feature. In English, capitalization is used in the beginning of sentences, to indicate a named entity or a proper noun.

This, in turn, may help to facilitate part-of-speech tagging and Named Entity Recognition (NER). However, capitalization may degrade performance of statistical machine translation, as

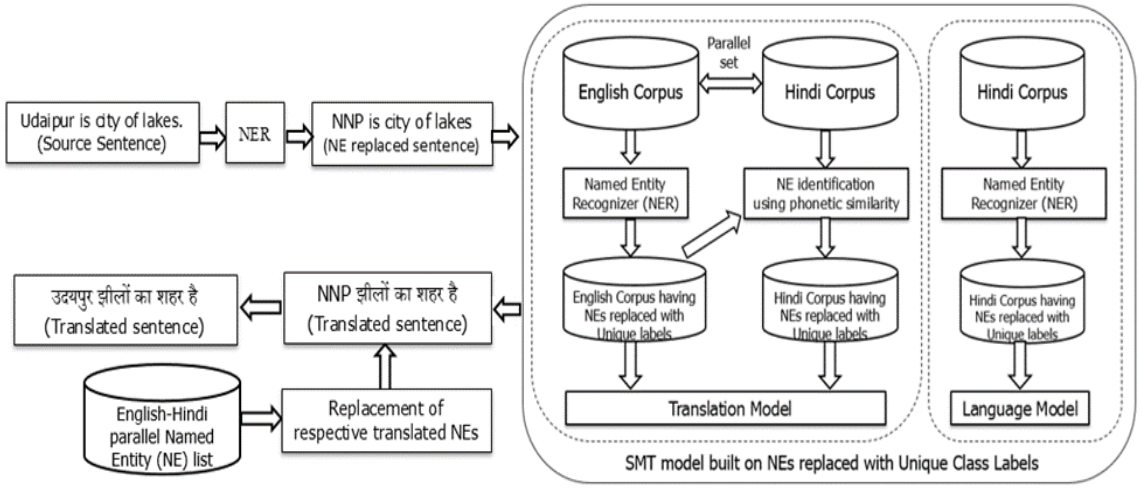

Fig. 2. Flow diagram of parallel NE identification process

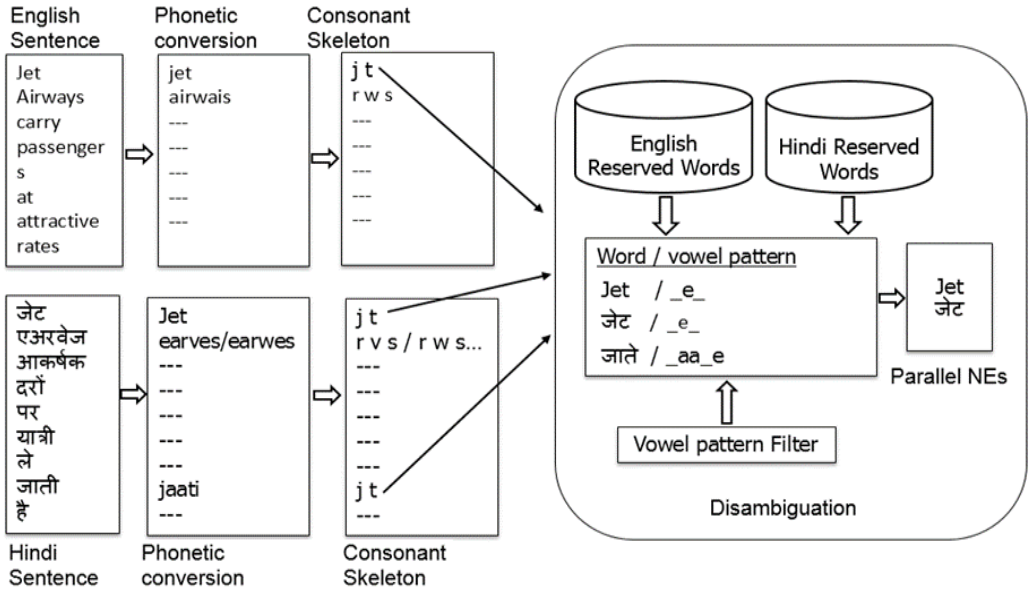

Fig. 3. Flow diagram checking phonetic similarity between an English and Hindi word

the occurrence of a word with and without capitalization would be treated as two different words.

This would reduce their individual count in the corpus. True casing is to keep the word in their natural case and changing the word at the beginning of the sentence to their most frequent form. Lower casing is meant by converting each word to lower cased form irrespective of their position or role in the sentence.

4.2 Punctuations

Punctuation is a mechanism for putting emphasis and for clarity of expression. It helps the reader in terms of readability of an expression.

Additionally, punctuation becomes very important for conveying the intended meaning of an expression, as the placement of punctuation marks also helps in disambiguation of an expression.

4.3 Hindi Language and Spell Normalization

Hindi is official language of India and is the third most spoken language of the world. Spell normalization is a process by which text is transformed in some way to make it consistent in terms of usage of spellings for its words throughout the text. The intention behind this activity is to reduce the lexical redundancy. Different types of normalizations applied to the corpus for our experiments are described below.

a. Same word can be written in multiple ways orthographically e.g. सम्बंध / संबंध / सम्बन्ध / संबन्ध (sambandh, 'relation'). These forms are very productive in nature and almost all are found in the text. In normalization process, these are mapped to one single form with the help of rules. The rule for handling these is - that the `fifth letters' (पंचमाक्षर) of the alphabet sequence and Anuswar (अनुस्वार) can be used interchangeably. If fifth letter of a class of consonants precedes any of the four remaining letters of the same class, the Anuswar can be used in place of that fifth letter; e.g. गंगा (ganga / 'Ganga the river'), घंटा (ghanta / 'bell'), धंधा (dhandha / 'business'), पंप (pamp / 'pump') etc. can be written in place of गङ्गा, घण्टा, धन्धा, पम्प respectively.

b. Chandrabindu (used for nasalized vowels) or Anuswar (used for nasalized consonants) signs are used for showing nasalization. Analyzing the text, it is observed that in place of Chandrabindu, Anuswar is generally being used. However, in very few words, the use of specific nasalization sign makes them two different words like हंस (hans / 'swan') and हँस (hañs / 'laugh') etc. But such words are very limited, so for the MT experiment purpose we have mapped these to one single form with Anuswar.

c. Words of Persio-Arabic origin adopted in Hindi vocabulary occur in the corpus with or without Nukta symbol (the dot below a character) e.g. ज़रूर or जरूर (zaroor or jaroor /'necessarily'). In the normalization process, these are mapped to without Nukta form.

d. Data encoded in Unicode may have more than one way of storage for the same words e.g. Hindi word पहाड़ी (pahari / 'hill') written with pre-composed character ड़, will have character storage as प ह ा ड़ ी, while with combining sequences, it will have character sequence of प ह ा ड ़ ी. To normalize this, we have mapped both representations to single one, the pre-composed character.

e. Same words may be represented in more than one way, depending on the presence or absence of ZWJ or ZWNJ (Zero Width Joiner or Zero Width Non-Joiner) e.g. शक्ति, शक्ति (Shakti / 'power'). In normalization process, we have mapped these to single representation by removing the ZWJ or ZWNJ.

Normalization of the corpus was done on both English and Hindi sides. Besides above, it also covered the following:

– Conversion of Devanagari digits to European digits (०, १, २, ३… to 0, 1, 2, 3…)

– Conversion of punctuation symbol semi-colon ';' to comma ',' and sentence end-markers to dot '.' where it appeared as Devanagari danda (।), as these were interchangeably used in corpus.

– Non-ASCII punctuations are replaced by their ASCII equivalents.

f. English side spell normalization is limited to mapping some spelling variations prevalent in text due to using American English and British English interchangeably e.g. Color vs. Colour, Organization vs. Organisation etc.

Table 1. Corpus statistics

Corpus	#Sent	#Tokens (En)	#Tokens (Hi)
Total corpus	43,977	601,924	615,911
Test corpus (5% of Total)	2,195	29,757	30,265
Development corpus (10% of Total)	4,394	60,471	61,954
Training corpus (75% of Total)	37,388	511,701	523,687

Table 2. Test set of number rich sentences

Test set Name	Test set description	# Sent	# Num
NUM	Number rich sentences extracted from general test set	280	381

Table 3. Test set of NE rich sentences

Test set Name	Test set description	#Sent	#NE
NE-I	NE rich sentences extracted from general test set	1,038	2,134
NE-II	Sub-set of NE-I test set having NEs seen in training corpus	780	1,347
NE-III	Sub-set of NE-I test set having NEs unseen in training corpus	258	787

It is also observed that English words which are written in Devanagari script do not appear in consistent forms. This is due to not having standard dictionary of writing these words. So, these spell variants having un-stable orthography are left un-handled in this experiment.

4.4 Pre-Processing of Numbers

Numbers appear quite frequently in corpus and generally do not contribute towards translation and make the phrase table noisier. However, these need to be passed to the translations for fluency and transfer of information.

In the presented paper, the number entities (including dates, time, monetary values etc.) are mapped to unique labels in parallel corpus, target side corpus to be used for language model and in the source sentence to be translated. After translation, the unique labels are replaced with actual number entities.

4.5 Pre-Processing of Named Entities

Named Entities are expressions that appear quite frequently in text and are much more variable in nature than content words. In translation process, these generally should not get translated and appear in transliterated forms which are commonly phonetic representation in the target language. The appearance of different Named Entities in corpus presents difficulty in learning their translation, as these may be unknown to the training corpus (being an Out of Vocabulary i.e. OOV term) or not having sufficient appearances to learn their translation reliably.

Thus, the nature of problem consequently motivates us to extend the approach similar to

numbers handling, wherein, we adopted a mechanism to map NEs to a unique class label (e.g. NNP). But here, challenge lies in identifying the word appearing as Named Entity in the given context. For recognition of a NE in the sentence, we used Stanford Named Entity Recognizer for English. Different NE classes (e.g. Person, Place etc.) identified by Stanford NE tagger are mapped to single class label (NNP) for our experiment.

The named entities, generally have phonetically similar presence in both sides of English-Hindi parallel corpus. We exploited this feature to filter undesired NEs marked by NE tagger from the training corpus.

Fig 3 below shows the flow diagram for checking phonetic similarity between English and Hindi NEs. The steps are described below –

i. First, words from both sides of parallel corpus are converted to their phonetic forms using letter to sound rules. For example, in Hindi, NE word 'हिडिंबा' is converted to 'hidimbaa". Similarly, English word is also converted to their phonetic form thus bringing it closer to respective Hindi word e.g. 'Julia' to 'juliya', 'Cadbury' to 'cadbari'/'kadbari', 'Hyderabad' to 'haiderabad'.

ii. Then, similarity is checked between consonant skeleton forms of both words, by removing vowels from words excluding the vowel(s) appearing at start of the word. This methodology may return some ambiguous matches and deteriorate precision of NE identification. For example, 'Delhi' got mapped to two Hindi words 'दिल्ली' and 'डाल' and similarly 'Jet' got mapped to two Hindi words 'जेट' and 'जाते' in their corresponding Hindi sentences.

iii. If for a probable NE in one language, there are more than one qualifying candidates in target language, then vowel sequences are compared and words having similar vowel sequences are adjudged best candidates. This method is intended towards achieving high precision e.g. vowel sequence of 'Delhi'/'_e__i' is closer to vowel sequence of 'दिल्ली'/ '_i__i' than of 'डाल'/'_aa_'.

iv. As there are more number of valid forms of words of shorter length, they are more prone to errors in matching e.g. valid words formed by 2 consonants 'j' and 't' are जेट, जाता, जाती, जाते, जात, जाट, जता, जटा etc. To avoid this, the words having less than two consonants are not considered. Similarly, for words having 2-3 consonants, a list of reserved words of both languages (mostly consisting of stop words, pronouns, prepositions, articles and verbs of shorter length) is prepared and words falling in this list are not considered being a valid NE. This helps is filtering out false positives. For example, in the case of 'Jet', the possible wrong alternative 'जाते' gets dropped due to its presence as a small sized Hindi verb word with this methodology.

The parallel NE words decided after these steps in the given sentence pair are replaced with unique class labels.

For both experiments of Stage-II i.e. number pre-processing and Named Entity pre-processing, a pruning mechanism is also applied on phrase table, wherein phrase-pair entries carrying different counts of numbers or NEs between source and target sides, are deleted from phrase table.

5 Corpus Statistics

The experiments described in this paper are carried out using a corpus of 43977 pairs of English-Hindi (en-hi) parallel sentences with 601924 tokens in English and 615911 tokens in Hindi. This corpus contains sentences from the tourism domain (ILCI corpus), parallel sentences from grammar books, travel & tourism domain sentences from web and manually translated sentences. The Indian Languages Corpora Initiative (ILCI) project initiated by the Ministry of Electronics & Information Technology, Govt. of India has developed parallel corpus of the Tourism domain (http://tdil-dc.in). The corpus contains sentences covering travel conversations and information about different visiting places, monuments, temples and parks etc. These sentences have been divided in 3 sets – training, development and test corpus. The size and distribution details are given in the Table 1 shown below.

Table 4. Models and their data description

Model Trained	Punctuation Marks	True-casing	Spell Normalization
M1	YES	YES	YES
M2	YES	YES	NO
M3	YES	NO	YES
M4	YES	NO	NO
M5	NO	YES	YES
M6	NO	YES	NO
M7	NO	NO	YES
M8	NO	NO	NO

Table 5. Pre-processing settings in Training, Test and LM corpus

Experiment Name	Experiment Description	Test Set
Ex-1	Numbers not replaced	NUM Test set
Ex-2	Numbers replaced with unique class label	NUM Test set
Ex-3	Named Entities not replaced	NE-I Test set
Ex-4	Named Entities replaced with Unique Class Label	NE-I Test set
Ex-5	Named Entities not replaced	NE-II Test set
Ex-6	Named Entities replaced with Unique Class Label	NE-II Test set
Ex-7	Named Entities not replaced	NE-III Test set
Ex-8	Named Entities replaced with Unique Class Label	NE-III Test set
Ex-9	Model Combination (Ex-5 + Ex-8)	NE-I Test set

For analyzing the impact of Stage-II experiments, sub-sets from the general test set used in Stage-I are prepared. Table 2 gives the statistics of test set containing number rich sentences (i.e. sentences having at least one number entity) extracted from the general test set. This gives us the information that 280 such sentences (NUM test set) are found out of 2195 general test set sentences and these have 381 number entities.

Similarly, for observing the impact of NEs mapping to unique class labels, following NE rich test sets are prepared.

i. Sentences having at least one NE word (as identified by NE tagger). This test set is further split in two sub-sets as mentioned below.
   a. Sentences having NEs, all of which are present or seen in the training corpus i.e. In Vocabulary (IV) words
   b. Sentences having at least one NE which is not present or unseen in training corpus i.e. treated as OOVs.

Table 3 lists these three test sets. It is observed that NE words cover 13.3% of total words in NE rich sentences extracted from general test set.

The effect of mapping NEs to unique labels is observed by translating these three sets of test sentences in their original form and by translating these sets after mapping their NEs to unique class labels. Corresponding language model is also built after mapping NEs to unique class labels, using in-house developed NE recognizer for Hindi. For more than one NE present in these sentences, the unique class labels are numbered in sequence to facilitate replacing them with their corresponding transliterated words after the experiment. These unique class labels are replaced with their respective transliterated forms after the translation is achieved.

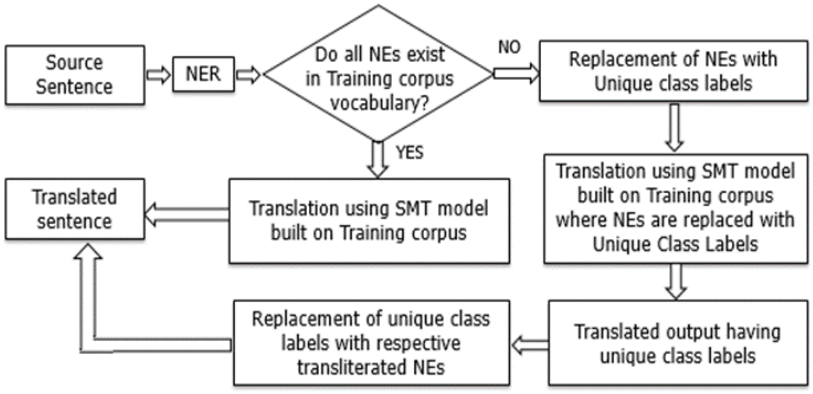

Fig. 4. Flow Diagram for selective use of SMT models (Ex-9)

6 Description of Experiments

For arriving at the most suitable setting in Stage-I pre-processing of the corpus, various combinations have been experimented. These combinations are - First, for handling punctuation symbols, two modes used are, with retaining or removing the punctuation symbols in both (En-Hi) sides of corpus. Second, for dealing with cases, either lowercasing or true-casing have been tried out. In true-casing, the initial word in each sentence is converted to its most probable casing. This requires True-caser model, which is trained on statistics extracted from the training corpus itself. The model is used for English. The third setting is with or without spell normalization of the corpus. This is achieved through in-house developed Hindi and English spell normalizers which handle spell normalization cases as detailed in Section 4.

As last step of pre-processing, to clean-up the parallel corpus, duplicate sentences, empty lines and sentences having more than specified length are removed. Redundant space characters are also removed. For consistent handling of punctuation symbols, spaces have been inserted between words and punctuation symbols.

After arriving at most optimum setting of the punctuation symbols, casing and normalization, the corpus is subjected to the Stage-II pre-processing i.e. number and named entity pre-processing as described in Section 4.

For training the statistical models, standard word alignment GIZA++ [12] and language modeling toolkit KenLM [13] are used. For translation, MOSES phrase based SMT decoder [14] has been used. For evaluation, the automatic evaluation metrics BLEU [15] is applied to the translation output. The main parameters used for Moses configuration are - 5 iterations of IBM-1 and HMM, 3 iterations of IBM-3 and IBM-4 for GIZA++, the maximum phrase length set as 7 and with reordering allowed. The parameters of phrase-based translation systems are tuned on development set using MERT [16].

Eight experiments were performed in Stage-I with different settings of the above described three features namely punctuation marks, true-casing, and spell normalization as listed in Table 4. The settings of training, test and development corpus were kept similar. The language models (LMs) were built on the target language side with the settings of corpus similar to the concerned experiment.

It is assumed that for better performance of the SMT system, both training and test data should be in sync and should use the consistent forms of the words throughout. Not being so, the words seen by the training corpus may be unseen by the test corpus due to their presence in dissimilar form. The pre-processing described in the paper seems simple, but results show that how significantly these impact MT results.

The Table 5 below lists different experiment setups and findings of the experiments conducted in Stage-II, i.e. experiments when number

expressions and NEs are replaced with unique class labels.

The experiment Ex-9 is a special case where combination of both models – First model without any processing of NEs and Second model with processing of NEs (i.e. replacing NEs with unique class labels) are selectively used. The sentences containing NEs seen in training corpus will be sent to First model while the sentences containing any NE which is unseen in training corpus would be directed to the Second model (Fig. 4).

The size of Hindi language model (LM) is 2.3 lakh sentences having 5.2% NE words (as identified by in-house built NE tagger).

7 Results and Observations

Table 6 lists the values of BLEU scores, Papineni et al. [15], for the translations achieved on the test corpus for different SMT models trained using various feature sets (Table 4) of the corpus.

These BLEU scores indicate that in Stage-I pre-processing for English to Hindi translation, best results are obtained with keeping the punctuation symbols intact, lower casing the source (English) side and with spell normalized Hindi text.

Results of experiments show that the spell normalization gives improvements in BLEU scores. This can be observed by comparing the BLEU scores of M1-M2, M3-M4, M5-M6 and M7-M8, as for these pairs, the other two features are kept constant. Similar observation can be seen with punctuation markers, the BLEU scores with having punctuation markers are better than without having punctuation markers (M1-M5, M2-M6, M3-M6 and M4-M8). While with true-casing the reverse phenomenon is observed. The BLEU scores drop down when true-casing is applied (M1-M3, M2-M4, M5-M7 and M6-M8).

It can also be observed by comparing the above pairs that impact of spell normalization is more than the other two factors i.e. casing and punctuations in English to Hindi translation.

The normalization process helps to ensure the maximum possible similarity in training and test corpora. The removal and non-removal of punctuation marks from training and test corpora is performed to test their impact on MT performance.

Table 6. BLEU scores for English (en) to Hindi (hi) translation with different models

Model Trained	BLEU Score (English-Hindi)
M1	25.47
M2	24.85
M3	25.75
M4	24.99
M5	25.07
M6	24.29
M7	25.17
M8	24.44

Table 7. BLEU scores for Hindi (hi) and English (en) translation with different models for NE handling

Experiment Name	English-Hindi BLEU scores
Ex-1	32.14
Ex-2	32.29
Ex-3	27.47
Ex-4	27.68
Ex-5	30.03
Ex-6	28.14
Ex-7	27.77
Ex-8	29.40
Ex-9	29.64

Table 8. Human Evaluation (Ex-6 vs. Ex-5)

# Sent	#Class-S	#Class-B	#Class-P
250 (100%)	196 (78.4%)	38 (15.2%)	16 (6.4%)

For experiments pertaining to pre-processing of NEs, comparing BLEU scores (Table 7) of Ex-7 and Ex-8 show an improvement of +1.63 BLEU points, due to mapping of NEs with unique labels in the sentences having NEs unseen in the training corpus. Similarly, by comparing BLEU scores of Ex-3 and Ex-4 a slight improvement of +0.21 BLEU points is observed for test set NE-I also.

The effect of mapping NEs for in-vocabulary or NEs seen in training corpus is in negative (Ex-5 vs.

Table 9. Example English to Hindi translated and reference translation

SOURCE:	THE TEMPLE OF DEVI BRIJESHWARI IS MOST FAMOUS HERE
REFERENCE:	देवी बृजेश्वरी का मंदिर यहां सबसे अधिक प्रसिद्ध है
EX-5 OUTPUT:	का मंदिर , देवी बृजेश्वरी है सबसे प्रसिद्ध यहां
EX-6 OUTPUT:	देवी बृजेश्वरी का मंदिर सबसे अधिक प्रसिद्ध यहां है
SOURCE:	ONE GETS CONNECTING FLIGHTS FROM THERE FOR MUNICH
REFERENCE:	वहां से म्यूनिख के लिए कनेक्टिंग फ्लाइट्स मिलती है
EX-5 OUTPUT:	लिए उडानें मिलती है , वहां से म्यूनिख के लिए
EX-6 OUTPUT:	कनेक्टिंग उडानें मिलती है , वहां से म्यूनिख के लिए
SOURCE:	A HUGE CROWD HAD GATHERED OUTSIDE THE HOUSE OF MALGUJAAR
REFERENCE:	मालगुजार के घर के बाहर खासी भीड हो गई थी
EX-5 OUTPUT:	भारी भीड जमा थी के घर के बाहर मालगुजार
EX-6 OUTPUT:	भारी भीड जमा थी मालगुजार के घर के बाहर
SOURCE:	THERE IS A ROAD FOR GOING TO RAJRAPPA CHINMASTIKA TEMPLE FROM CHITARPUR ALSO
REFERENCE:	चित्तरपुर से भी रजरप्पा छिन्नमस्तिका मंदिर जाने का मार्ग है
EX-5 OUTPUT:	वहां जाने के लिए एक मार्ग से भी रजरप्पा छिन्नमस्तिका मंदिर चित्तरपुर
EX-6 OUTPUT:	एक सडक रजरप्पा छिन्नमस्तिका मंदिर जाने के लिए चित्तरपुर से भी

Ex-6) with -1.89 BLEU points. For studying the impact on translation quality, we checked the translation outputs manually. For carrying out manual assessment, 250 sentences are randomly selected from this set (NE-II) and the assessment of quality was done by comparing the translation of sentences of Ex-5 and Ex-6.

For this, the observation was recorded by classifying the translation output in three classes namely - S (Same or no change), B (Better) and P (Poor). The Table 8 shows the results of manual evaluation.

It was noticed that despite having a slight dip in the BLEU score, the context resolution and placement of surrounding words is found better and more acceptable for human consumption in Ex-6 output. This can be correlated with some of the example translation outputs given in Table 9 which lists English sentence, Reference translation, Ex-5 translation output and Ex-6 translation output.

The test set for Ex-3 and Ex-9 experiments is same i.e. consisting of all sentences containing NEs, irrespective of seen or unseen in training corpus. Comparing the scores of Ex-3 and Ex-9, we observe a gain of +2.17 BLEU score points (7.9%), which shows methodology of selective use of translation models helps significantly.

The results reported in Caseli et al. [2] show the impact of over 10% between the worst case settings and best case settings for English-Portuguese pair. For English-Hindi pair this is observed as 5% (for Stage-I experiments). The reason for this may be attributed to the fact that in English-Portuguese language pair, casing has impact in both language sides while Hindi does not have casing variation.

Sellami et al. [8] have reported a gain of 2.4% in BLEU score for Arabic-French pair as an impact of NE translation experiment. In our experiments, a higher gain of 7.9% in BLEU score is observed, which shows the effectiveness of proposed methodology.

8 Conclusions and Future Work

The paper presents some experiments pertaining to pre-processing on training and test corpora limited to casing, punctuation markers, spell

normalization, numbers and named entities. The impact of these are observed on translation quality improvement through BLEU score and manual inspection. It is observed that for English-Hindi translation, best results are obtained by keeping the punctuation symbols intact, lower casing the source (English) side and with spell normalized text in Stage-I experiments pertaining to orthographic representation of content.

Spell normalization process had the maximum impact (in Stage-I experiments) on translation improvement. Punctuation markers participate in forming the phrases of the phrase table and their presence impacted positively.

In Stage-II experiments, it is observed that mapping of NEs to a unique class label is effective, especially for the case of NEs unseen in the training corpus. The combination of using both models selectively – without any processing of NEs and with processing of NEs (i.e. replacing NEs with unique class labels) also has significant positive impact. The impact of mapping numbers with unique class labels is not very effective and can be left out.

The best combinations in pre-processing will be used as baseline cases in performing future experiments. Future work will include investigating with more linguistic features like re-ordering source sentences to match the target side word order using source side parse information in the phrase-based SMT. It would also be interesting to see the impact of NEs mapping to unique class label after source side re-ordering.

Acknowledgment

We thank Mukund Kumar Roy for helping in performing experiments. Thanks are also due for the participants of translation evaluation exercise.

References

1. **Sproat, R., Black, A.W., Chen, S., Kumar, S., Ostendrof, M., & Richards, C. (2001).** Normalization of non-standard words. *Computer Speech and Language,* Vol. 15, No. 3, pp. 287–333, DOI: 10.1006/csla.2001.0169.
2. **Caseli, H.M. & Nunes, I.A. (2009).** Statistical Machine Translation: little changes big impacts. *Proceedings of 7th Brazilian Symposium in Information and Human Language Technology*, pp. 1–9.
3. **Bojar, O., Straňák, P., & Zeman, D. (2010).** Data Issues in English-to-Hindi Machine Translation. *Proceedings of the 7th International Language Resources and Evaluation (LREC'10),* pp. 1771–1777.
4. **Santanu, P., Sudip, K.N., Pavel, P., Sivaji, B., & Andy, W. (2010).** Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation. *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications, (ACL´10),* pp. 46–54.
5. **Lane, I.R. & Waibel, A. (2008).** Class-based statistical machine translation for field maintainable speech-to-speech translation. *Proceedings of International Conference on Speech Communications and Technology,* pp. 2362–2365.
6. **Markov, I., Gómez-Adorno, H., Sidorov, G., & Gelbukh, A. (2016).** Adapting Cross-Genre Author Profiling to Language and Corpus. Working Notes Papers of the (CLEF´10), Vol. 1609, pp. 947–955.
7. **Markov, I., Stamatatos, E., & Sidorov, G. (2017).** Improving Cross-Topic Authorship Attribution: The Role of Pre-Processing. *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2017)*, Budapest, Hungary, Springer.
8. **Sellami, R., Deffaf, F., Sadat, F., & Hadrich-Belguith, L.** Improved Statistical Machine Translation by Cross-Linguistic Projection of Named Entities Recognition and Translation. *Computación y Sistemas,* Vol. 19, No. 4 pp. 701–711. DOI: 10.13053/CyS-19-4-2329.
9. **Okuma, H., Yamamoto, H., & Sumita, E. (2008).** Introducing translation dictionary into phrase-based SMT. *IEICE transactions on information and systems,* Vol. E91-10, No. 7, pp. 2051–2057.
10. **Brown, P.F., Pietra, V.J., Pietra, S.A.D., & Mercer, R.L. (1993).** The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics,* Vol. 19, pp. 263–311.
11. **Och, F.J. & Ney, H. (2004).** The Alignment template approach to statistical machine translation. *Computational Linguistics*, Vol. 30, No. 4, pp. 417–449. DOI: 10.1162/0891201042544884.
12. **Och, F.J. & Ney, H. (2003).** A Systematic Comparison of Various Statistical Alignment

Models. *Computational Linguistics,* Vol. 29 No. 1, pp. 19–51. DOI: 10.1162/089120103321337421.

13. **Heafield, K. (2011).** KenLM: faster and smaller language model queries. *Proceedings of the (EMNLP´11) Sixth Workshop on Statistical Machine Translation,* pp. 187–197.
14. **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. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* pp. 177–180.
15. **Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002).** BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual meeting of the Association for Computational Linguistics, pp. 311–318, DOI:10.3115/1073083.1073135.
16. **Och, F. (2003).** Minimum Error Rate Training in Statistical Machine Translation. Proceedings of Association of Computational Linguistics, pp. 160–167. DOI:10.3115/1075096.1075117.

Article received on 23/04/2017; accepted on 06/07/2017.
Corresponding author is Karunesh Kumar Arora.