Heliyon 5 (2019) e02504

Contents lists available at ScienceDirect

Heliyon

www.elsevier.com/locate/heliyon

Statistical-based system combination approach to gain advantages over different machine translation systems

Debajyoti Banik [a,b], Asif Ekbal [a,b], Pushpak Bhattacharyya [a,b], Siddhartha Bhattacharyya [c,d,*], Jan Platos [c]

[a] Department of Computer Science and Engineering, India
[b] Indian Institute of Technology Patna, India
[c] Faculty of Electrical Engineering and Computer Science, VSB Technical University of Ostrava, Czech Republic
[d] RCC Institute of Information Technology, Kolkata, India

ARTICLE INFO

Keywords:
System combination method
Machine translation
Statistical approach
Neural machine translation (NMT)
Neural network
Hierarchical machine translation (Hiero) systems
Phrase-based statistical machine translation (PBSMT)

ABSTRACT

Every machine translation system has some advantages. We propose an improved statistical system combination approach to achieve the advantages of existing machine translation systems. The primary task is to score all the phrases of the outputs of different machine translation systems selected for combination. Three steps are involved in the proposed statistical system combination approach, viz., alignment, decoding, and scoring. Pair alignment is done in the first step to prevent duplication so that only a single phrase is chosen from various phrases containing the same information. Thus the alignment and scoring strategy are implemented in our approach. Hypotheses are built in the second step. In the third step, we calculate the scores for all the hypotheses. The hypothesis with the highest score is chosen as the final translated output. Wrong scoring can mislead to identify the best part from different systems. It may be noted that a particular phrase may appear in various ways in different translations. To resolve the challenges, we incorporate WordNet in the alignment phase and word2vec in the scoring phase along with the existing factors. We find that the system combination model using WordNet and word2vec injection improves the machine translation accuracy. In this work, we have merged three systems viz., Hierarchical machine translation system, Bing Microsoft Translate, and Google Translate. The broad tests of translation on eight language pairs with benchmark datasets demonstrate that the proposed system achieves better quality than the individual systems and the state-of-the-art system combination models.

1. Introduction

Every machine translation (MT) system has its own strengths and shortcomings [1]. Statistical machine translation (SMT) system [2] is efficient for rare word handling and adequacy preservation whereas neural machine translation (NMT) system suffers from handling rare words, unknown words, domain mismatch, amount of training data, long sentences, and word alignment [3]. Over-translation and under-translation problems exist in the NMT system [4]. Nowadays, the NMT system has improved its performance significantly [5, 6, 7, 8]. NMT system is better to maintain syntactic structure by which it is able to translate with good fluency [9, 10]. NMT system has a limit on the vocabulary size whereas SMT system (like PBSMT system [9] and Hierarchical machine translation (Hiero) systems [10]) do not have this limitation. As a result, it may have a guarantee of the source text's translation coverage since both the adequacy and fluency are important for translation.

- **Source Text:** গণতন্ত্র হলো এক সরকার ব্যবস্থা যা নাগরিকদের ভোট দিতে এবং তাদের পছন্দমত সরকার নির্বাচন করতে দেয়।
- **Transliterated Source Text:** ganatantra halo ek sarkar byabastha ja nagarikder vote dite ebang tader pachandomato sarkar nir-bachan karte dei.
- **Reference text:** Democracy is a system of government that allows the citizens to cast vote and select a government of their choice.
- **System 1:** Democracy is that allows citizens to select the government.
- **System 2:** Democracy a government that allow to cast vote and government of their choice.

* Corresponding author at: Faculty of Electrical Engineering and Computer Science, VSB Technical University of Ostrava, Czech Republic.
E-mail address: dr.siddhartha.bhattacharyya@gmail.com (S. Bhattacharyya).

https://doi.org/10.1016/j.heliyon.2019.e02504
Received 1 November 2018; Received in revised form 11 May 2019; Accepted 18 September 2019

- **System 3:** is government a system citizens allows which select government.
- **Consensus translation:** Democracy is a government system that allows the citizens to cast vote and select a government of their choice.

The proposed system uses various translated outputs of a source text by using various translation systems as its inputs. The output of the proposed system is a consensus translation. From the above example, we have a source text (in Bengali) and there are three translated outputs (in English) by using three systems (System 1, System 2 and System 3). The translated sentence by System 1 is incomplete. The translated sentence by System 2 has a grammatical problem. The verb "is" and the word "select" are missing here. The word "cast vote" is missing in the translated sentence by System 3. Hence, it can be surmised that this is not a fluent sentence. On the other hand, the consensus translation merges the better phrases and provides a better translated output. The output is very much fluent and preserves adequacy. Thus, the proposed approach is efficient in this task.

This paper is organized as follows. The research objectives, motivations, and contributions are introduced in Section 2. Related works are presented in Section 3. Section 4 describes the detailed procedure of the proposed work. Section 5 mentions the description of used benchmark datasets of Indian languages. Prepossessing and the experimental setup are described in Section 6. Section 7 provides results and analysis of the proposed approach. Finally, we conclude in Section 8.

2. Study Area

We propose a statistical approach for combining the outputs of different machine translation systems. Multiple systems are taken as black boxes and only their single best hypothesis is required for the combination. We select only the best phrase among all the multiple systems' translated outputs. Then, all of the best phrases are merged to get the final translation output. Three steps are devoted to combine different systems under consideration, viz., Alignment, Decoding and Scoring. We incorporate WordNet into the alignment algorithm. We also incorporate word2vec [11] into the scoring algorithm. The metrics such as Language Model Probability [12], Mean Length Ratio (MLR), BLEU score [13], Cosine Similarity (CS),[1] Word Mover's Distance (WMD) [14] are used for scoring the hypotheses. With the help of word2vec, we calculate WMD and CS. Our goal is to create a better system for achieving the advantages of the individual systems.

The main focus of this work is to merge the translated outputs of SMT and NMT systems to achieve better accuracy without having any knowledge of the internal architecture. In this paper, we use Hierarchical phrase-based machine translation system as the SMT system and both Google Translate [15] and Bing Microsoft Translate [16] engines as NMT systems. We use Google Translate and Bing Microsoft Translate to show the capability of re-usability. Finally, we find a significant performance improvement on the quality of translated outputs in various language pairs with the help of the proposed system combination model.

The inputs of our system are the translated outputs of the same source text using different systems. The objective of our system is to generate better translated outputs. One part of the translated text (generated by the ith system) could be better than that of the jth system. Some other parts of that translated text (generated by the jth system) could be better than that of the ith system. We select the best phrases from different systems' generated translated outputs. Then they are recombined to increase efficiency and diversity by packing hypotheses. Firstly, all the phrases are divided and then aligned. Beam search is used on top of these alignments to make this search tractable. Then some scores are assigned based on their features. Finally, the hypothesis with the highest score is selected as the final output.

Building a brand new MT system is a time consuming and resource intensive task. But if there are already different MT systems available for the task, then it seems logical to combine their strengths into a single system. The quality of the translated outputs can be improved by using a system combination approach with the help of the existing systems without knowing the detailed systems' architecture. This model is designed based on the behavioral study of individual systems. So, instead of building an MT system from scratch we can take advantages of existing systems by using the proposed system combination model.

Our contributions to produce consensus output depend on the following three steps that improve the alignment and the scoring strategies in order to generate the consensus output.

- We propose the WordNet based strategy for realistic sense wise alignment.
- The primary problem with the previous system combination approach is that it does not have a better scoring strategy [17]. We address this point for better hypothesis selection.
- We identify different important features and incorporate them for better scoring.
- We consider eight language pairs including Indian language pairs in this work.

3. Related Work

Various applications of Natural Language Processing (NLP); like sentiment analysis [18, 19], question answering [20], question routing services in social context [21], medical and health queries [22] are available in the literature and machine translation is one of them. To improve translation quality, researchers have turned their attention to combine various machine translation systems. Firstly, researchers proposed a system combination approach for machine translation in [23], where knowledge-based MT, example-based MT, and lexical transfer MT are combined using the chart manager. The structure of a multi-engine machine translation (MT) system is shown in Figure 1 of [23]. Statistical MT and rule-based MT are combined together to improve the translation accuracy in [24]. Here, researchers used Anusaraaka MT engine on behalf of rule-based MT and generated output enhanced by the resource of phrase table. Moreover, source side re-ordering rules have been used as catalyst in the in-house data set. Researchers injected lexicon after source side re-ordering and used hierarchical framework in [25] to improve the MT accuracy. They used IIT Bombay English-Hindi Corpus[2] for the experiments. To identify the optimal stack size and beam threshold, the parameters of the statistical MT decoder are fixed dynamically using the CN2 unordered algorithm in [26]. HindEnCorp and ILCI datasets are used as the benchmark datasets for this experiment. Multiple string alignment strategies were used to combine various systems in [27]. A confusion network was built after alignment. Finally, the system combined output was determined from the confusion network by using majority voting technique. The primary problem with this approach is word reordering. GIZA++ toolkit was used for alignment to overcome word reordering problem in [28]. Since alignment is the primary concern here, so translation edit rate (TER) [29] scoring was used for alignments [30]. Authors in [31] used hidden Markov models for the hypotheses alignment. Researchers in [32] described a minimum Bayes' risk system combination (a subsequence system combination) approach for machine translation. Inversion transduction grammars were used in [33] for better alignment. A sentence-level system combination model was designed in [34]. Phrase-level and word level system combination models were introduced respectively in [35] and [36]. Among them, confusion network based word-level combination

[1] https://en.wikipedia.org/wiki/Cosine_similarity.

[2] http://www.cfilt.iitb.ac.in/iitb_parallel/.

models were very impressive [37, 38, 39]. To improve machine translation's performance, researchers presented the minimum Bayes' risk based system combination method. This method assembled together the benefits of subsequence-combination and sentence-selection methods in [32]. Two open source toolkits i.e., multi-engine machine translation (MEMT) [17] and Jane [39] were proposed for the system combination.

Nowadays, neural machine translation (NMT) has become a very prevalent topic for automatic language translation. Most of the works mainly focus to handle rare words [40, 41, 42, 43] and make the full monolingual data useful [44, 45, 46]. Some of the models are engaged to improve attention model [4, 47, 48, 49], and integrating SMT [50, 51, 52]. Pre-translation approach was introduced to handle the rare words in a neural network [53]. Here, the authors first translated the input by using the phrase-based statistical machine translation (PBSMT) system and after that a neural network was used for final translation. Different automated MT systems helped us to make it practically more feasible. Recently, a neural-based system combination has been proposed in [1]. Here, the authors used attention based multiple encoders which are horizontally arranged. Every encoder is attached to an MT system. Every encoder flows their information into the decoder without knowing other encoders' behavior. In this architecture, a horizontal encoder receives several systems' information and passes their own information into a common encoder to understand each others' behavior. Finally, their information is inserted into the decoder. Recently, researchers introduced a hybrid approach to combine various MT's outputs [54]. They used neural approach and statistical approach together to combine different MT systems' outputs. The three stack architecture is introduced here. In the first layer, different MT engines work individually for translation. The produced outputs are inserted into neural-based and statistical-based system combination models. A co-ordination model is placed in the third layer. It helps to integrate all the outputs of the first layer taking help from the second layer. This approach helps to improve both the adequacy and fluency of the translated output because it uses the power of neural network with statistical approach. HindEnCorp corpus has been used in the experiments.

In past several years, the statistical-based approaches were introduced for the MT system combination along with word level, phrase level, and sentence level methods [17, 27, 39, 42, 55, 56, 57, 58, 59].

The target of this work is to achieve advantages of various MT systems and systems combination models without knowing their detailed architectures by using their consensus translations. In this work, we use the WordNet [60, 61, 62] and word2vec for the negative-sampling word-embedding method in the proposed system combination approach. We combine various types of neural and statistical machine translation systems using a statistical approach. The source text has been incorporated to disambiguate the word sense [63], [64] along with various systems' translated outputs. We find that the hybrid approach is most powerful till now in [54]. The basic components of this architecture are statistical-based and neural-based system combination models. So, our statistical-based model can be incorporated into the hybrid system combination model to improve the overall accuracy.

4. Methodology

The proposed method has three modules viz., (1) an alignment module which is useful for alignment among strings of various hypotheses generated by different MT systems, (2) a decoding module which is important to build hypotheses by using aligned strings using the beam search algorithm which performs this task in reasonable time and (3) a scoring module for estimating the final hypothesis.

Finally, the hypothesis with the highest score from produced n-best hypotheses is considered as the final output. The detailed architecture of the proposed model is shown in Fig. 1.

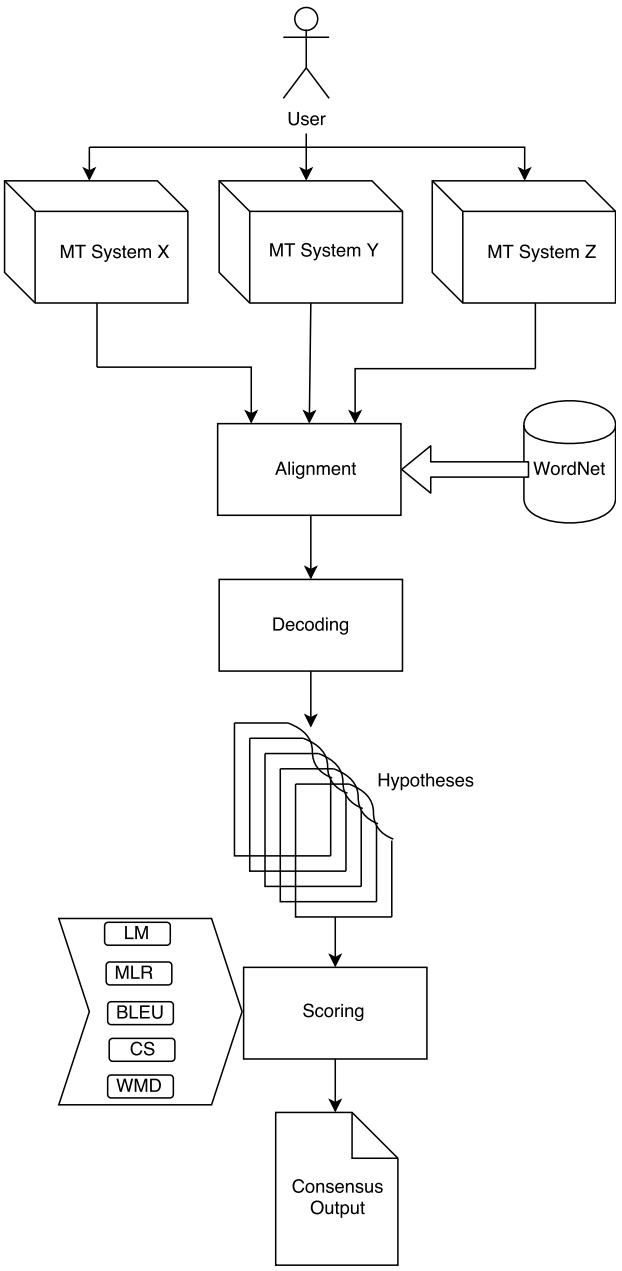

Fig. 1. System combination architecture.

Moreover, MT systems are used for basic translations. WordNet helps to improve the lexical knowledge which is crucial for alignment. The decoding step is used to generate hypotheses. The hypothesis with the best score (which is influenced by LM, MLR, BLEU score, CS, WMD) is selected as the final translation output.

4.1. Alignment

We take the single best outputs $t_1, t_2, ..t_n$ from each of the n participating systems. Then, the sentence pair t_i and t_j are taken and strings (between the sentence pairs) are aligned together. For n sentences, we need to align $\frac{n(n-1)}{2}$ sentences (all possible pairs). A string w_1 in sentence t_i can be aligned to string w_2 in sentence t_j based on the following conditions.

- w_1 and w_2 are same.
- w_1 and w_2 have Wu and PaLMer similarity score $> \quad \delta$

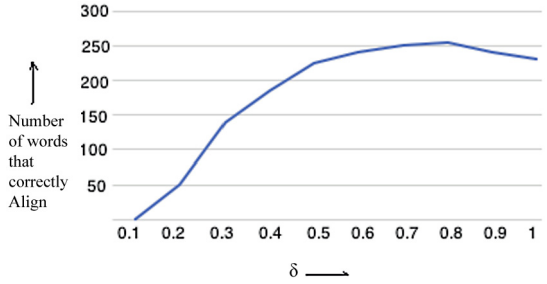

Fig. 2. Optimizing δ value.

After fine-tuning, we achieve the value of δ which is equal to 0.80 for the used dataset. We explore 270 words (which are chosen randomly) and vary the value of δ to watch (manually) the number of words to align correctly. Results of the fine-tuning for δ value selection are shown in Fig. 2. The Wu and PaLMer (WuP) similarity score [65] is used to calculate the similarity score. This considers the situation of ideas i_1 and i_2 in the taxonomy with respect to the position of the Least Common Subsumer (i_1, i_2) (LCS(i_1, i_2)). This similarity score is based of the assumption that the path-based measures are function of the path length, and depth is the similarity between two concepts. The LCS of two nodes (v1, v2) in a directed acyclic graph (or a tree T) is the deepest node that has both same nodes as descendants, where we define each node to be a descendant of itself (so if there is a direct connection from v2 to v1, the lowest common ancestor is v2).

$$Simwup(i_1,i_2) = 2 * \frac{(Dep(LCS(i_1,i_2)))}{(Len(i_1,i_2) + 2 * dep(LCS(i_1,i_2)))}$$

A hypernym of i_1, i_2 which is the lowest node in hierarchy, is represented as LCS (i_1, i_2).

Meteor [66] is used to align the default configuration of English sentences. However, Meteor only supports exact matches for Indian languages. So, the functionality for computing WuP similarity by using Indo-WordNet [67] is added for the Indian languages in our experiments.

4.2. Decoding

Decoding is traversing of the search space on top of the aligned sentences. It is a strategy similar to that described in [17]. We combine parts of single best outputs to generate a set of hypothesis. The hypothesis is formed with some parts of the single best outputs. The steps are detailed below.

- We start with an empty string ϵ in our set of hypothesis.
- Hypothesis h_i always starts and ends with starting and ending words of some single best output.
- We can add the first word of any t_i into ϵ.
- Adding of the words can be continued using words from that sentence or can be switched to a different sentence at any time. The switching between different sentences can happen only at aligned words.
- Whenever a word w is added to hypothesis h_j from some t_i, then word w and any word w' aligned to w in $t_k (k \neq i)$ will become unusable for future additions to h_j. This prevents duplication since aligned words seem to be duplicate.
- Let the word w is added from a single best output t_i to hypothesis h_j. We can add the next word also from t_i or switch to a different t_j. On switching, the first unused word from that sentence is added to the next hypothesis (shown in Fig. 3). In this example, hypotheses building is started from t_2. The first chosen word "Excellent" is aligned to "good" and "prosperous". So these two words will not be part of this hypothesis. There are three possible next words: "students", "scholar", "boy". So, three hypotheses ("excellent boy", "excellent students", "excellent scholar") will be created at this stage. The same procedure will be continued until either traversing of all words is done or the beam search criteria are satisfied.
- The result is an output which is a combination of parts of different systems.

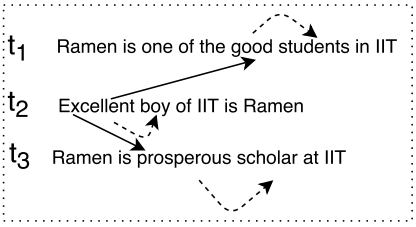

Fig. 3. Participating words for next word selection to build the hypothesis.

The searching is accomplished by using the beam search technique and recombination with a beam size b. The hypotheses that extend the same way are recombined. The search is guided by the following features.

- Hypothesis length in words.
- Language model probability and out of vocabulary (OOV) count of the hypothesis.
- A number of N-gram matches between hypothesis and single best outputs.

4.3. Scoring

The n-best lists are the generated outputs in the decoding step. We define some features by which we calculate the scoring of this n-best list. We represent the n-best list as $h_1, h_2, h_3 ..., h_k$ and calculate the score of each h_i. We select the hypothesis h_k with the maximum score. The following features are used to score the n-best list. The weights for these features are calculated by using the minimum error rate training algorithm [68].

- **Language Model Probability**
  Language model (LM) probability [12] of hypothesis h_i is calculated by using tri-gram language model to ensure fluency. Language can be represented by the conditional probability of the next word in all the previous ones as follows:

$$\bar{P}\left(w_1^T\right) = \prod_{T}^{t=1} \bar{P}\left(w_t | w_1^{t-1}\right)$$

  In sub-sequence $w_i, w_{i+1}, ..., w_{j-1}, w_j$ the tth word is w_t.
- **Mean Length Ratio (MLR)**
  The mean of single best translation's $t_1, t_2, ..t_n$ lengths is calculated as l_t. The lengths of hypotheses are calculated as l_h. The ratio of l_h and l_t is taken as the mean length ratio. Here, the number of words in the sentence is defined as length.
- **BLEU score**
  BLEU score [13] of hypothesis h_i is calculated with respect to each of the single best outputs $t_1, t_2, ..t_n$. This introduces n features, one feature for each participating system. This feature measures the similarity between the hypothesis and the single best output. This feature is also used to calculate the adequacy of the hypothesis with several translated sentences (t_i). The BLEU score is defined as follows.

$$BLEU(t_j, h_i) = BP.exp\left(\sum_{n=1}^{N} w_n log p_n\right)$$

$$Where, BP = \begin{cases} 1 & if \quad t_j > h_i \\ e^{1-\frac{h_i}{t_j}} & if \quad t_j \leq h_i \end{cases}$$

Here, p_n is calculated by using n-grams up to length N and positive weights w_n summing to one.

- **Cosine Similarity (CS)**
  Cosine similarity (CS)[3] of hypothesis (h_i) is calculated with respect to each of the single best outputs $t_1, t_2, ..t_n$. This also introduces n number of features. The following process is carried out to calculate the cosine similarity between two sentences h_i and t_j. The sentence vectors (v_{h_i}, v_{t_j}) of sentences h_i and t_j are calculated and then the cosine between those two vectors are calculated. At first we calculate word vectors followed by summing up them as sentence vector for sentence embedding of sentence S. The cosine similarity for sentence vectors v_{h_i} and v_{t_j} is represented as follows:

$$CS(v_{h_i}, v_{t_j}) = \frac{v_{h_i}^T v_{t_j}}{\left|v_{h_i}\right|\left|v_{t_i}\right|}$$

  The measurement of cosine similarity always lies between -1 to +1. This can be used to prevent some hypotheses that contain repeated words since the hypothesis contains the words of single best output. The sentence vectors of hypothesis and the single best output should be close to each other. There would be unrelated words or word repetitions if these vectors differ by a large amount. Here word order does not matter since we are taking vector sum of word vectors. So, this feature does not penalize word movement.
- **Word Mover's Distance (WMD)**
  WMD [14] of hypothesis h_i is calculated with respect to each of the single best outputs $t_1, t_2, ..t_n$. The cosine similarity calculates the similarity based on same words between hypothesis and the single best output and introduces n features. Word Mover's distance can be calculated as similarity even when there is no common word between the sentences. This can separate hypothesis having good semantic similarity with the single best output from those who do not have this good semantic similarity. This feature ensures adequacy of the hypothesis.

A sample copy of the hypotheses with score is shown here:
0 ||| पकड़ो और परियोजनाओं के परियोजना एक्स " पट्टी के लिए एक तरफ पर ||| LM = -41.0 $BLEU_1 = 0.63$ $BLEU_2 = 0.51$ $BLEU_3 = 0.55$ MLR = 0.87 $W2V_1 = 0.98$ $W2V_2 = 0.99$ $W2V_3 = 0.86$ $WMD_1 = 1.69$ $WMD_2 = 1.56$ $WMD_3 = 6.08$
0 ||| एक्स " पट्टी के लिए पकड़ो एक तरफ पर के और में परियोजना ||| LM = -41.98 $BLEU_1 = 0.58$ $BLEU_2 = 0.54$ $BLEU_3 = 0.45$ MLR = 0.87 $W2V_1 = 0.97$ $W2V_2 = 0.99$ $W2V_3 = 0.87$ $WMD_1 = 2.58$ $WMD_2 = 1.55$ $WMD_3 = 5.84$.
LM refers to the log10 probability of sentence from trigram language model. $BLEU_i$ refers to BLEU [13] score of hypothesis with respect to the system i. MLR is equal to $\frac{\text{Length of hypothesis}}{\text{Mean length of n systems' inputs}}$. $W2V_i$ is cosine of vector of sentence with the respect to the system$_i$. WMD_i is the word mover's distance with respect to the system i.

Publicly available pre-trained word2vec dictionary [69, 70, 71] is used for calculating the cosine similarity and Word Mover's distance. Google word2vec is used for English and word2vec trained with Wikidump data is used for Hindi. The language model score ensures fluency of hypothesis and BLEU score, cosine similarity, Word Mover's distance ensure adequacy. Finally, we combine the scores of the features linearly with weights to calculate the final scores and select the hypothesis with the highest score as the final output. We tune the weights by using ZMERT [72] which is based on the minimum error rate training of machine translation systems. The weights of their corresponding features are tuned using the ZMERT toolkit. Here, we tried to maximize the accuracy of the combined translation. So, the primary task of the ZMERT toolkit is to find the optimal weights to maximize the likelihood of the output as

$$\begin{matrix} Objective: optimize & W_i \\ maximize & T_c \end{matrix}$$

where, W_i and T_c are the weights of ith feature and combined translation. We measure T_c in terms of the BLEU score. The abbreviation study to select the features is shown in Table 3. 1 and 0 refer to considering and not considering the metric, respectively. The accuracy of the combined outputs is compared in terms of the BLEU sore. We find that the feature set with LM, MLR, BLEU, CS, and WMD provides best combined output among the subsets.

Table 1
Statistics of dataset.

Set	#Sentences	#Tokens			
		En	Hi	Ta	Ur
Train	46,996	803,515	844,879	587,735	829,935
Development	1,001	17,958	18,914	13,105	18,016
System combination development	1,000	17,878	18,356	11,837	18,029
Test	1,002	16,061	16,599	11,325	16,226

5. Materials & Methods

The benchmark dataset for Indian languages and English, Indian Language Corpora Initiative (ILCI) corpus [73] are divided into four parts as training set (46996) and development set (1001) for sentence pairs in the Hierarchical statistical phrase-based system. 1000 sentence pairs are used for fine-tuning of system combination model's parameters. 1002 sentences are chosen for the test set. We show the proposed strategy for eight language pairs: Bengali to English (Bn-En), Hindi to English (Hi-En), Bengali to Hindi (Bn-Hi), Tamil to English (Ta-En), English to Hindi (En-Hi), Urdu to English (Ur-En), Tamil to Hindi (Ta-Hi), Urdu to Hindi (Ur-Hi). This dataset consists of tourism and health data. Detailed statistics for the dataset are shown in Table 1.

6. Experimental

For achieving better performance, the raw data is first pre-processed. In this stage, different pre-processing steps like tokenization, truecasing, removing long sentences as well as finding sentences with a length mismatch exceeding certain ratio are carried out.

For the purpose of experimentation, the outputs of three machine translation systems viz., Hierarchical phrase-based machine translation system (Hiero), Bing Microsoft Translate[4] (Bing Microsoft Translate) and Google Translate[5] (Google) are combined. The Hierarchical phase based system is a special type of SMT system whereas Bing Microsoft Translate and Google Translate are special types of NMT system. Google Translate and Bing Microsoft Translate are used to show that our proposed approach has the capability to make better MT system than the existing commercial translators.

We train the Hierarchical phrase-based machine translation system with tri-gram language model by using modified KneserNey smoothing [74] using IRSTLM [75]. The Moses decoder [76] is used for Hierarchical phrase-based machine translation. Our model learns the word alignments with grow-diag-final and heuristics from the parallel training corpus using GIZA++ [77].

The beam size b and n-best list are considered as 500, 300 respectively for decoding and scoring in the proposed system combination approach. The frame parameter r is set to 7 (seven). ZMERT [72], an open source implementation of minimum error rate training is used for fine-tuning.

[3] https://en.wikipedia.org/wiki/Cosine_similarity.

[4] https://www.bing.com/Translate.

[5] https://translate.Google.com/.

Table 2
Performance analysis with respect to BLEU score. Outputs of Google Translator and Bing Microsoft Translate are retrieved as on March 2018.

		Machine translation system			System combination model	
		Hiero	Bing Microsoft Translator	Google Translator	MEMT [17]	Proposed
Language pair	Bn-En	6.48	12.85	14.7	14.74	15.13
	Hi-En	11.78	16.57	23.61	23.10	23.85
	Bn-Hi	17.74	9.92	11.48	16.99	17.85
	Ta-En	0.56	5.56	8.12	8.12	9.03
	En-Hi	13.9	19.35	19.19	19.78	20.23
	Ur-En	9.51	3.29	3.07	9.76	10.98
	Ta-Hi	1.27	5.63	8.15	8.22	8.66
	Ur-Hi	46.05	3.98	2.76	47.05	47.17

Table 3
Abbreviation study for features selection.

	LM	MLR	BLEU	CS	WMD	Output's accuracy (BLEU)
Bn-En	1	1	1	0	0	51.38
	0	1	0	1	1	46.35
	1	0	1	0	1	55.86
	1	1	0	1	0	48.74
	1	1	1	1	1	57.72
En-Hi	1	1	1	0	0	57.48
	0	1	0	1	1	48.79
	1	0	1	0	1	60.31
	1	1	0	1	0	54.77
	1	1	1	1	1	63.74
Hi-En	1	1	1	0	0	61.24
	0	1	0	1	1	51.03
	1	0	1	0	1	65.99
	1	1	0	1	0	59.76
	1	1	1	1	1	66.53

Table 4
Error analysis for Hindi-English translation pair.

Source	कई जगह हमें उनके पंजों के निशान भी दिखे , लेकिन टाइगर का दर्शन न मिलने से मैं उदास था ।
Transliterated	Kaii jagah hameM unke panjoM ke nishaan Bhii dikhe, Lekin taigar ka darshan na milne se maiM udaas tha .
Reference	At many places we saw their paw prints but I was upset for not being able to find the sight of tiger .
Hiero	many places were also their marks of claws us , but Tiger not getting view the I was pensive .
Bing Microsoft Translate	In many places , we could see their toes , but the tiger's philosophy was sad .
Google	In many places we also saw the marks of their claws , but I was sad due to not seeing Tiger's philosophy .
MEMT [17]	In many places we also saw the marks of their claws , but I was sad due to not seeing Tiger's philosophy .
Proposed	In many places we also saw the marks of their claws , but I was sad due to not seeing Tiger .

7. Results

The experiments on eight different language pairs (including Indian language pairs) show that our proposed system combination approach for machine translation has a significant improvement in the quality of MT output. Bilingual Evaluation Understudy (BLEU)[13] and LeBleu [78] are used for quality assessment. The traditional MT evaluation metric BLEU[13] is based on *n*-gram matching that can estimate both fluency and adequacy. It is observed from our experiments that the proposed system combination approach can improve the quality of machine translated outputs. A detailed performance evaluation in terms of BLEU[13] is reported in Table 2. The table shows that the proposed system combination model has better accuracy than every participating systems (such as Moses, Bing Microsoft Translate, Google) and the traditional system combination model. We also use fuzzy based MT evaluation metric (LeBleu)[78] for a more realistic comparison. The detailed statistics of LeBleu[78] based comparison is shown in Table 5. Due to better alignment and scoring, the proposed system helps to improve the accuracy. WordNet injection introduces better alignment. Moreover, the different scoring parameters including Word2Vec, CS, MLR, LM, BLEU, WMD calculate better scores for the hypotheses.

It is observed with the human evaluation that the single best output from Google Translate, Bing Microsoft Translate and Hierarchical sys-

Table 5
Performance analysis based on LeBleu score. Hiero refers to Hierarchical machine translation system's output. Outputs of Google Translator and Bing Microsoft Translate are retrieved as on March 2018.

		Machine translation system			System combination model	
		Hiero	Bing Microsoft Translator	Google Translator	MEMT [17]	Proposed
Language pair	Bn-En	45	51.37	55.69	55.87	57.72
	Hi-En	56.37	58.1	65.94	66.1	66.53
	Bn-Hi	58.91	46.08	49.91	57.61	58.37
	Ta-En	13.52	46.16	46.71	46.71	48.65
	En-Hi	55.04	59.94	62.37	60.43	63.74
	Ur-En	53.81	43.76	42.46	53.95	54.95
	Ta-Hi	23.9	40.6	42.72	45.55	46.74
	Ur-Hi	77.13	41.56	40.48	77.16	78.24

Table 6
Error analysis for Bengali-English translation pair.

Source	হাবেলির বাইরে করা কলাত্মক চিত্রের কারণে শোখাবটিকে বিশ্বের সবথেকে বড়ো মুক্ত কলা গ্যালারিও বলা যায়।
Transliterated	Habelir baire kara kalatmak chitrer karane shokhabtike bishwer sabtheke baro mukta kala gyalario bala jae .
Reference	Due to the artistic pictures made outside the forts Shekhavati is also called as the world 's largest Open Art Gallery .
Hiero	হাবেলির outside done artistic of the pictures due to শোখাবটিকে the most big of the world free art গ্যালারিও be called the .
Bing Microsoft Translate	Outside the building because of the artistic pictures to shokhabtor also known as the world's biggest free art gallery .
Google	Due to fatal images outside of Hablali , Shobhaba is also known as the world's largest free arts gallery .
MEMT [17]	Due to fatal images outside of Hablali , Shobhaba is also known as the world's largest free arts gallery .
Proposed	Due to artistic of the pictures , Shobhaba is also known as the world's largest free arts gallery .

tem contains some untranslated words or translation with extra words. As a result, the adequacy and fluency of the translated outputs are affected whereas the outputs produced by the proposed system combination model contain significantly lesser number of these untranslated words which causes to improve the performance of the proposed model. A snapshot for such a case in Hindi-English translation by using various systems is shown in Table 4. The selection of the best phrase among the outputs produced by the participating systems makes the translation a special essence. WordNet and influencing parameters for scoring take primary role here. So, the accuracy is improved in our proposed model. Bengali-English translation is compared in Table 6. In this case, the translation quality of the Hierarchical system is poor. Google is better than Bing Microsoft Translate but the important phrase "artistic of the pictures" is present in Hierarchical but not in Google Translate. MEMT [17] produces exactly same translation as the Google Translate output. The proposed system intelligently incorporates only the best parts from the Hierarchical system's output into the Google Translate output. We also observe that the outputs produced by Google and Bing Microsoft Translate are very similar most of the times but the output from the SMT system may have different word orders or synonyms. In this case, the system combination model produces translations matching with those of Google and Bing Microsoft Translate. The primary intuitions of this approach in selecting the better features for scoring help to identify the weights of different features using the MERT algorithm. The WordNet and word2vec provide more in-depth knowledge about the syntactic and semantic natures of different MT systems' outputs. Finally, combining the better parts of outputs of different machine translation systems improves the overall accuracy of the translated output.

We have done significance tests, and observe that the results are significant with 95% confidence level (with $p = 0.03$) for English to Hindi. For other languages i.e. Bengali to English, Hindi to English, Bengali to Hindi, Tamil to English, Urdu to English, Tamil to Hindi, Urdu to Hindi the value of p are 0.06, 0.04, 0.06, 0.05, 0.04, 0.06, 0.06, respectively with 95% confidence level. So, the proposed approach stands to be statistically significant.

8. Conclusions

Our paper proposes a new approach to system combination. We combine the single best outputs from Bing Microsoft Translate, Google Translate and SMT systems. Firstly, the similar words in the outputs of different MT systems are aligned by using uni-gram based surface matching, stemmed matching, and WUP similarity score. Then, the hypothesis is created with the help of a beam search algorithm. The hypotheses are ranked by using language model, word2vec and other features. The hypothesis with the highest score is selected as the best translated output. Experiments with eight language pairs help to conclude that WordNet and word2vec based proposed system combination approach gives better translation accuracy. However, it remains to explore the possibility of building a hybrid architecture for system combination using this system combination approach in future.

Declarations

Author contribution statement

Debajyoty Banik: Conceived and designed the experiments; Performed the experiments; Wrote the paper. Asif Ekbal, Pushpak Bhattacharyya: Performed the experiments. Siddhartha Bhattacharyya: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper. Jan Platos: Contributed reagents, materials, analysis tools or data.

Competing interest statement

The authors declare no conflict of interest.

Additional information

The proposed model can be incorporated to prepare complex machine translation toolkit. It has a power of reusability.

References

[1] L. Zhou, W. Hu, J. Zhang, C. Zong, Neural system combination for machine translation, arXiv preprint, arXiv:1704.06393.

[2] P. Koehn, Statistical Machine Translation, Cambridge University Press, 2009.

[3] P. Koehn, R. Knowles, Six challenges for neural machine translation, arXiv preprint, arXiv:1706.03872.

[4] Z. Tu, Z. Lu, Y. Liu, X. Liu, H. Li, Modeling coverage for neural machine translation, arXiv preprint, arXiv:1601.04811.

[5] M. Junczys-Dowmunt, T. Dwojak, H. Hoang, Is neural machine translation ready for deployment? A case study on 30 translation directions, arXiv preprint, arXiv:1610.01108.

[6] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint, arXiv:1409.0473.

[7] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.

[8] N. Kalchbrenner, P. Blunsom, Recurrent continuous translation models, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2017.

[9] P. Koehn, F.J. Och, D. Marcu, Statistical phrase-based translation, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, Association for Computational Linguistics, 2003.

[10] D. Chiang, A hierarchical phrase-based model for statistical machine translation, in: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2005, pp. 263–270.

[11] X. Jin, S. Zhang, J. Liu, Word semantic similarity calculation based on word2vec, in: 2018 International Conference on Control, Automation and Information Sciences (ICCAIS), IEEE, 2018, pp. 12–16.

[12] P.X.F.J. Thorsten Brants, Ashok C. Popat, J. Dean, Large language models in machine translation, 2007.

[13] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 311–318.

[14] M. Kusner, Y. Sun, N. Kolkin, K. Weinberger, From word embeddings to document distances, in: International Conference on Machine Learning, 2015, pp. 957–966.

[15] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: Bridging the gap between human and machine translation, 2016.

[16] W.B. Dolan, J. Pinkham, S.D. Richardson, MSR-MT: the Microsoft research machine translation system, in: Conference of the Association for Machine Translation in the Americas, Springer, 2002, pp. 237–239.

[17] K. Heafield, A. Lavie, Combining machine translation output with open source: the carnegie mellon multi-engine machine translation scheme 93 (2010) 27–36.

[18] Y.-Y. Zhao, B. Qin, T. Liu, et al., Sentiment analysis 21 (2010) 1834–1848.

[19] Y. He, D. Zhou, Self-training from labeled features for sentiment analysis, vol. 47, Elsevier, 2011, pp. 606–616.

[20] J.P. Bufe, D.K. Byron, M.D. Swift, T. Winkler, Establishing user specified interaction modes in a question answering dialogue, US Patent 9,898,170 (Feb. 20 2018).

[21] Z. Liu, B.J. Jansen, Identifying and predicting the desire to help in social question and answering, vol. 53, Elsevier, 2017, pp. 490–504.

[22] A. Spink, Y. Yang, J. Jansen, P. Nykanen, D.P. Lorence, S. Ozmutlu, H.C. Ozmutlu, A study of medical and health queries to web search engines, vol. 21, Wiley Online Library, 2004, pp. 44–51.

[23] S. Nirenburg, R. Frederking, Toward multi-engine machine translation, in: Proceedings of the Workshop on Human Language Technology, Association for Computational Linguistics, 1994, pp. 147–151.

[24] D. Banik, S. Sen, A. Ekbal, P. Bhattacharyya, Can smt and rbmt improve each other's performance?-an experiment with English-Hindi translation, in: Proceedings of the 13th International Conference on Natural Language Processing, 2016, pp. 10–19.

[25] S. Sen, D. Banik, A. Ekbal, P. Bhattacharyya, Iitp English-Hindi machine translation system at wat 2016, in: Proceedings of the 3rd Workshop on Asian Translation (WAT2016), 2016, pp. 216–222.

[26] D. Banik, A. Ekbal, P. Bhattacharyya, Machine learning based optimized pruning approach for decoding in statistical machine translation, IEEE Access 7 (2018) 1736–1751.

[27] B. Bangalore, G. Bordel, G. Riccardi, Computing consensus translation from multiple machine translation systems, in: Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on, IEEE, 2001, pp. 351–354.

[28] E. Matusov, N. Ueffing, H. Ney, Computing consensus translation for multiple machine translation systems using enhanced hypothesis alignment, in: 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006.

[29] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proceedings of Association for Machine Translation in the Americas, vol. 200, 2006.

[30] K.C. Sim, W.J. Byrne, M.J. Gales, H. Sahbi, P.C. Woodland, Consensus network decoding for statistical machine translation system combination, in: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, vol. 4, IEEE, 2007, pp. IV–105.

[31] X. He, M. Yang, J. Gao, P. Nguyen, R. Moore, Indirect-hmm-based hypothesis alignment for combining outputs from machine translation systems, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 98–107.

[32] J. González-Rubio, F. Casacuberta, Minimum Bayes' risk subsequence combination for machine translation, vol. 18, Springer, 2015, pp. 523–533.

[33] D. Karakos, J. Eisner, S. Khudanpur, M. Dreyer, Machine translation system combination using ITG-based alignments, in: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Association for Computational Linguistics, 2008, pp. 81–84.

[34] S. Kumar, W. Byrne, Minimum Bayes-risk decoding for statistical machine translation, in: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004, 2004.

[35] Y. Feng, Y. Liu, H. Mi, Q. Liu, Y. Lü, Lattice-based system combination for statistical machine translation, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3, Association for Computational Linguistics, 2009, pp. 1105–1113.

[36] B. Chen, M. Zhang, H. Li, A. Aw, A comparative study of hypothesis alignment and its improvement for machine translation system combination, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, Association for Computational Linguistics, 2009, pp. 941–948.

[37] A.-V. Rosti, S. Matsoukas, R. Schwartz, Improved word-level system combination for machine translation, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007, pp. 312–319.

[38] N.F. Ayan, J. Zheng, W. Wang, Improving alignments for better confusion networks for combining machine translation systems, in: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics, 2008, pp. 33–40.

[39] M. Freitag, M. Huck, H. Ney, Jane: open source machine translation system combination, in: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 29–32.

[40] M.-T. Luong, I. Sutskever, Q.V. Le, O. Vinyals, W. Zaremba, Addressing the rare word problem in neural machine translation, 2014.

[41] J. Zhang, C. Zong, Bridging neural machine translation and bilingual dictionaries, 2016.

[42] M. Li, J. Zhang, Y. Zhou, C. Zong, The CASIA statistical machine translation system for IWSLT 2009, in: nnnn, 2009.

[43] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, 2015.

[44] Y. Cheng, W. Xu, Z. He, W. He, H. Wu, M. Sun, Y. Liu, Semi-supervised learning for neural machine translation, 2016.

[45] R. Sennrich, B. Haddow, A. Birch, Improving neural machine translation models with monolingual data, 2015.

[46] J. Zhang, C. Zong, Exploiting source-side monolingual data in neural machine translation, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 1535–1545.

[47] M.-T. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, 2015.

[48] H. Mi, B. Sankaran, Z. Wang, A. Ittycheriah, Coverage embedding models for neural machine translation, 2016.

[49] F. Meng, Z. Lu, H. Li, Q. Liu, Interactive attention for neural machine translation, 2016.

[50] S. Shen, Y. Cheng, Z. He, W. He, H. Wu, M. Sun, Y. Liu, Minimum risk training for neural machine translation, 2015.

[51] M. Junczys-Dowmunt, T. Dwojak, R. Sennrich, The AMU-UEDIN submission to the WMT16 news translation task: attention-based NMT models as feature functions in phrase-based SMT, 2016.

[52] W. He, Z. He, H. Wu, H. Wang, Improved neural machine translation with SMT features, in: AAAI, 2016, pp. 151–157.

[53] J. Niehues, E. Cho, T.-L. Ha, A. Waibel, Pre-translation for neural machine translation, 2016.

[54] D. Banik, A. Ekbal, P. Bhattacharyya, S. Bhattacharyya, Assembling translations from multi-engine machine translation outputs, Appl. Soft Comput. 78 (2019) 230–239.

[55] L. Barrault, MANY: open source machine translation system combination 93, Versita (2010) 147–155.

[56] J. Zhu, M. Yang, S. Li, T. Zhao, Sentence-level paraphrasing for machine translation system combination, in: International Conference of Young Computer Scientists, Engineers and Educators, Springer, 2016, pp. 612–620.
[57] A.-V.I. Rosti, B. Zhang, S. Matsoukas, R. Schwartz, Incremental hypothesis alignment for building confusion networks with application to machine translation system combination, in: Proceedings of the Third Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2008, pp. 183–186.
[58] M. Li, C. Zong, Word reordering alignment for combination of statistical machine translation systems, in: Chinese Spoken Language Processing, 2008. ISCSLP'08. 6th International Symposium on, IEEE, 2008, pp. 1–4.
[59] W.-Y. Ma, K. McKeown, System combination for machine translation through paraphrasing, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 1053–1058.
[60] C. Fellbaum, WordNet, Wiley Online Library, 1998.
[61] G.A. Miller, WordNet: a lexical database for English, vol. 38, ACM, 1995, pp. 39–41.
[62] N.S. Dash, P. Bhattacharyya, J.D. Pawar, The WordNet in Indian Languages, Springer, 2017.
[63] F.J. Och, H. Ney, Statistical multi-source translation, in: Proceedings of MT Summit, vol. 8, 2001, pp. 253–258.
[64] O. Firat, K. Cho, Y. Bengio, Multi-way, multilingual neural machine translation with a shared attention mechanism, 2016.
[65] Z. Wu, M. Palmer, Verbs semantics and lexical selection, in: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 1994, pp. 133–138.
[66] S. Banerjee, A. Lavie, Meteor: an automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, pp. 65–72.
[67] S. Bhingardive, H. Redkar, P. Sappadla, D. Singh, P. Bhattacharyya, Indowordnet:: similarity computing semantic similarity and relatedness using indowordnet, in: Global WordNet Conference, 2016, p. 39.
[68] F.J. Och, Minimum error rate training in statistical machine translation, in: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, Association for Computational Linguistics, 2003, pp. 160–167.
[69] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013.
[70] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[71] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013, pp. 746–751.
[72] O. Zaidan, Z-MERT: a fully configurable open source tool for minimum error rate training of machine translation systems 91, Versita (2009) 79–88.
[73] G.N. Jha, The TDIL program and the Indian language corpora initiative (ILCI), in: LREC, 2010.
[74] R. Kneser, H. Ney, Improved backing-off for m-gram language modeling, in: Acoustics, Speech, and Signal Processing, 1995. ICASSP-95, 1995 International Conference on, vol. 1, IEEE, 1995, pp. 181–184.
[75] M. Federico, N. Bertoldi, M. Cettolo, IRSTLM: an open source toolkit for handling large scale language models, in: Ninth Annual Conference of the International Speech Communication Association, 2008.
[76] H. Hoang, P. Koehn, Design of the moses decoder for statistical machine translation, in: Software Engineering, Testing, and Quality Assurance for Natural Language Processing, Association for Computational Linguistics, 2008, pp. 58–65.
[77] F.J. Och, H. Ney, A systematic comparison of various statistical alignment models, vol. 29, MIT Press, 2003, pp. 19–51.
[78] S. Virpioja, S.-A. Grönroos, LeBleu: N-gram-based translation evaluation score for morphologically complex languages, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, 2015, pp. 411–416.