Sādhanā (2022)47:238
https://doi.org/10.1007/s12046-022-01965-5

Investigation of negation effect for English–Assamese machine translation

SAHINUR RAHMAN LASKAR, ABINASH GOGOI, SAMUDRANIL DUTTA, PROTTAY KUMAR ADHIKARY, PRACHURYA NATH, PARTHA PAKRAY* and SIVAJI BANDYOPADHYAY

Department of Computer Science and Engineering, National Institute of Technology Silchar, Assam, India
e-mail: sahinurlaskar.nits@gmail.com; abinashgogoi_ug@cse.nits.ac.in; samudranil_ug@cse.nits.ac.in; prottay_ug@cse.nits.ac.in; prachurya_ug@ee.nits.ac.in; partha@cse.nits.ac.in; sivaji.cse.ju@gmail.com

MS received 3 April 2022; accepted 2 July 2022

Abstract. Computational linguistics deals with the computational modelling of natural languages, in which machine translation is a popular task. The aim of machine translation is to automatically translate one natural language into another, which minimizes the linguistic barrier of different linguistic backgrounds. The data-driven approach of machine translation, namely, neural machine translation achieves state-of-the-art results on different language pairs, however it needs a sufficient amount of parallel training data to attain reasonable translation performance. In this work, we have explored different machine translation models on a low-resource English–Assamese language pair and investigated different sources of errors, particularly due to negation in English-to-Assamese and Assamese-to-English translation. Negation is a universal, essential feature of human language that has a substantial impact on the semantics of a statement. Moreover, a rule-based approach is proposed in the data preprocessing step which handles modal-verb negation problem that shows significant improvement in translation performance in terms of automatic and manual evaluation scores.

Keywords. English–Assamese; negation; machine translation.

1. Introduction

Machine translation (MT) attempts to improve the quality of automatic translation, relying on data-driven approaches, namely, Statistical Machine Translation (SMT) [1] and Neural Machine Translation (NMT) [2, 3]. The SMT and NMT require a sufficient amount of parallel train data, which is a challenging task in the case of low-resource language pairs translation [4, 5]. Furthermore, low-resource languages which are morphologically rich and the presence of different varieties of inflected words demand more parallel data to attain equivalent results of languages that possess less inflected words [6]. However, [7] considered low-resource pairs when the training data is less than 1 million parallel data. Based on the availability of resources [8, 9], most languages can be categorized into low-resource types. In this work, a low-resource pair, namely, English–Assamese (En–As) [10] is considered for the task of MT. Assamese belongs to the Indo-Aryan language family spoke mostly in Assam, a state in northeast India where it is an official language. The English and Assamese are very different from each other in linguistic aspects like word order, script etc. Assamese follows subject-object-verb (SOV) and morphologically rich language unlike English [10].

We have focused on the negation effect for En–As pair translation. Negation could have a substantial impact on an utterance's semantically [11]. In particular, negation has the potential to cause loss of information or misinformation if mistranslated due to its logical reversal feature. For example, Source: *One can not but admire her determination.* Target: তাইৰ সংকল্পক এজনে প্ৰশংসা নকৰি নোৱাৰে। Bing Translation: এজনে তাইৰ সংকল্পক প্ৰশংসা কৰিব নোৱাৰে। Here, Bing[1] Translation predicts negation word কৰিব instead of নকৰি that results in mistranslated output for English–Assamese translation. There are other linguistic phenomena and analysis directions including pronoun translation [12], anaphora resolution [13], morphological competence [14], and word sense disambiguation [15] that requires significant attention in the evaluation studies of MT. Our contributions are as follows:

- We have explored different possible negation sources for both directions of translation (English-to-Assamese

*For correspondence
Published online: 14 November 2022

[1]https://www.bing.com/translator.

and vice-versa) that affect the quality of translation obtained from various MT models.

- We have proposed a simple rule-based method in data preprocessing that handles model-verb negation problems that show significant improvement in the quality of translation. To the best of our knowledge, we are the first to tackle the negation problem in English–Assamese translation.
- We have achieved state-of-the-art results for English—Assamese MT translational performance in terms of automatic and manual evaluation, focusing on negation effects.

The rest of the paper is structured as follows: Section 2 and 3 discuss the related works and background concept. The tackling of the negation problem is described in section 4. Section 5 reported the experiments and results. The negation effect analysis is discussed in section 6. Lastly, section 7 conclude the paper with future scopes.

2. Related work

In this section, MT works, specifically for the En–As pair, are reported. It is found that all the existing works on this pair, focused on the dataset preparation [10, 16–18]. In [16, 17], SMT systems are developed on a very limited dataset. The En–As parallel corpus, namely, EnAsCorp1.0 [10] is developed and built En–As baseline systems using phrase-based SMT and RNN-based NMT. Recently, Samanantar [18], 11 Indian languages including Assamese, parallel corpora are developed and a transformer-based NMT model is developed for En-to-Indic and Indic-to-En translations. However, [19] improved En–As pair translation on EnAsCorp1.0 [10] using data augmentation approach via augmenting phrase-pairs and leveraging synthetic parallel data. Although, researchers have been exploring the En–As pair for the MT system, there are research gaps, identified as follows:

- Need to analyze En–As pair translation on the linguistic phenomena like the negation effect. Based on the preliminary experiments, it is observed that negation degrades the quality of translation for both directions of translation.
- Need to investigate MT models using more parallel corpus to improve translational performance and evaluate human translation to identify clues that lower the quality of translation.

In this work, we have investigated the negation effect for En–As pair by exploring different MT models. The authors [11] investigated negation problem in the MT task for forward and backward directions of translation in En–Ru (English–Russian), En–Et (English–Estonian), En–De (English–German), En–Tr (English–Turkish), En–Fi (English–Finnish), En–Cs (English–Czech), En–Zh (English–Chinese), En–Lt (English–Lithuanian), En–Gu (English–Gujarati), and En–Kk (English–Kazakh). They utilized NMT approach and studied different sources of error due to negation problem.

3. MT background

This section primarily discusses the background concepts of MT models, mainly SMT and NMT. The fundamental concept of SMT is based upon the Bayes theorem. Prior to NMT, phrase-based SMT attains a state-of-the-art MT approach [1]. It considers the translation problem into a mathematical reasoning problem in which parameters are computed from the parallel corpus. The SMT requires three components, namely, translation model (TM), language model (LM) , and decoder. For example, the translation task of English to Assamese, where the best Assamese translation (a_{best}) for the source English sentence (e) is formulated using Eq. 1.

$$a_{best} = \arg\max_{a} P(a|e) \tag{1}$$

This Eq. 1 can be reformulated into Eq. 2 using the Bayes theorem, as given below.

$$a_{best} = \arg\max_{a} P(e|a) * P(a) \tag{2}$$

We compute $P(e|a)$ and $P(a)$ using TM and LM. Using $\arg\max_a$, we calculate the best possible translation with a decoder. Then, phrase pairs are extracted by TM from parallel corpus and is used to estimate the approximate target words using Eq. 3.

$$P(e|a) = \frac{count(e,a)}{\sum_e count(e,a)} \tag{3}$$

The LM collects words and phrases from the TM. These collected words/phrases are then rearranged by the LM to ensure fluency of translation in predicting correct target sentences. The LM is estimated from monolingual target data, where the target sentence is modelled by the conditional probability of each word given the previous words in the sentence. It is known as n-gram LM. Then, to select the best possible translation decoder is employed by using a beam search strategy. The demerits of SMT include the inability of contextual analysis, long term dependency issues, and system complexity that leads to the development of NMT [2, 3, 20, 21]. In RNN-based NMT, sequential learning is possible via an end-to-end approach which resolves the source-target variable-length problem and adopts long short term memory (LSTM) for encoding and decoding to deal with long-term dependency issues. However, if the sentence is too long, then the encoder losses information. Therefore, the attention mechanism is introduced [2, 3]. The idea behind the attention mechanism is that the decoder focus on various parts of the source sentence globally at different decoding steps. Here, input sequences $s_1, s_2 \ldots s_n$ pass to the encoder and generates a context vector X. Based on conditional probability, as given in Eq. (4), the decoder generates the output sequences $t_1, t_2 \ldots t_m$.

$$P(t \mid s) = \sum_{i=1}^{m} P(t_i \mid t_{<1}, X) \tag{4}$$

Then, an attention vector a is computed by estimating hidden states of source side (h_i) and target side (h_o). Equation (5) presents the computation of an attention vector a.

$$a_v = \frac{\exp(\text{ score } (h_o, h_i'))}{\sum_{i'} \exp(\text{ score } (h_o, h_{i'}))} \tag{5}$$

In current work, we have employed the score function using Eq. (6).

$$score(h_o, h_i') = h_o W_a h_i' \tag{6}$$

The context vector c_l is calculated by considering the weighted average of all the input side hidden states along with the attention vector. An attentional hidden vector generated by the concatenation of h_o and, c_t as shown in Eq. (7).

$$h_o' = \tanh(W_c[c_t, h_o]) \tag{7}$$

Lastly, softmax layer is incorporated to the vector h_o' using Eq. (8) to produce the predicted sentence in the target language.

$$P(t_j \mid t_{<1}, X) = \operatorname{softmax}(W_s h_o') \tag{8}$$

In RNN-based NMT, input sequences are processed based on previous words. However, bidirectional RNN consider two independent RNNs for forward and backward directions [22]. The demerits of RNN-based NMT is lack of parallelization, that leads to development of transformer-based NMT [23]. Here, concept namely, the self-attention mechanism computes attention several times, known as multi-head attention. It helps the model to represent different words through multiple positions. In this work, we have explored different MT models, namely, PBSMT, RNN, BRNN, and transformer models for both directions of translation, i.e., En-to-As and As-to-En.

4. Handling negation in English–Assamese MT

There are different forms of negation sentences in English [11]. Broadly, it can be categorized into two types: a negation that affects the meaning of an entire clause (normal, clausal negation). It includes normal negation used in declarative sentences. For example, *My dog is not sick*, where reversing the truth value of the sentence. Another one affects the meaning of just a single word or phrase (constituent, lexical negation). For example, *The universe is unhappy with us*. Generally, negative particles (*not, no*), negative affixes, or morphemes (*unhappy*), negative pronouns (*nothing, nowhere*), negative adverbs (*hardly, scarcely*), and negative auxiliary verbs (*haven't, do not, doesn't*) are used to represent negation sentence in English. However, negation is indicated in Assamese in three ways in general, as shown in Table 1. In the first example, negation না is affixed with পাওঁ and become নাপাওঁ. The word নোৱাৰো is a negative auxiliary in the second example and বেয়া is a negative verb that is used to represent the negation sentence in the third example.

Table 1. Different negation forms in Assamese language.

Type	Assamese
Negative affixes	মই তাইক ভাল নাপাওঁ। (I do not like her.)
Negative auxiliaries	মই নাচিব নোৱাৰো। (I can't dance.)
Negative verbs	মই তাইক বেয়া পাও। (I dislike her.)

Table 2. Example of handling model-verb negative problems.

Initial	Intermediate	Final
can't	can not	can##not
won't	will not	will##not
shouldn't	should not	should##not

Based on the preliminary experiments (as reported in section 5.3), it is observed that negation sentences having negative auxiliary verbs or model verbs such as "could not," "do not" are not able to produce correct negation sentences in Assamese. It is because for every single negation word in Assamese, there are two or more corresponding tokens in English. For example, "can not" নোৱাৰো. Also, negative model verbs like shouldn't, can't where negation "n't" act as a source of errors in translation. To handle model-verb negation problems, we proposed a two-step solution in the data preprocessing step, as shown in table 2. In the intermediate step,the negative particle "n't" is changed to "not" and then separate the two words or tokens. Furthermore, the separated tokens are combined into a single token in the final step by placing a double hash symbol (##). Our approach brings a solution to encode both token information "can" and "not" as a single token, like single negation tokens in Assamese. The double hash symbols are removed in the post-processing step of

Assamese-to-English translation. The comparative analysis of our approach will be discussed in section 6.

The intermediate and final step of conversion algorithms are presented below:

```
Input1=file1;
Input2=file2;
list_of_lists = 2d-list of words present in each
    line.
l=list of the modal verbs

class REReplacer(object):
  def __init__(self, patterns=R_patterns):
    self.patterns = [(re.compile(regex), repl) for
     (regex, repl) in patterns]

  def replace(self, text):
    for (pattern, repl) in self.patterns:
      text = re.sub(pattern, repl, text)
    return text


list_of_lists = [(line.strip()).split() for line
    in file1]

for j in range(len(list_of_lists)):
  i=0
  ans=""
  while i<len(list_of_lists[j]):
      ans+= list_of_lists[j][i]+" ";
      i+=1
  word=REReplacer()  #call the regex replace
    method inside REReplacer class defined above.
  ans=word.replace(ans)
```

```
res=[]
for j in range(len(list_of_lists)):
  i=0
  ans=""
  while i<len(list_of_lists[j])-1:
    if(list_of_lists[j][i] in l and list_of_lists[
    j][i+1]=="not"):
      ans+= list_of_lists[j][i]+"##"+list_of_lists
    [j][i+1]+" ";
      i+=2
    else:
      ans+= list_of_lists[j][i]+" ";
      i+=1
  if i<len(list_of_lists[j]):
    ans+=list_of_lists[j][i]
  append this string(ans) to a new list res
file2.write('\n'.join(res))

"""
OUTPUT:
can't -> can not->can##not
shouldn't -> should not ->should##not
"""
```

5. Experiment and result

In this section, we briefly discuss the dataset, experimental setup and reported quantitative results with error analysis.

Table 3. Train, validation and test data statistics.

Type	Sentences	Tokens	
		En	As
Train Set-1 [10]	203,315	2,414,172	1,986,270
Train Set-2 [18]	138,353	1,715,435	1,377,336
Total	341,668	4,129,607	3,363,606
Validation Set-1 [10]	4500	74,561	59,677
Validation Set-2 [18]	1000	19,922	16,824
Total	5500	94,483	76,501
Test Set-1 [10]	2500	41,985	34,643
Test Set-2 (Manually prepared)	1000	9253	7745

5.1 *Corpus description*

We have utilized parallel corpora, namely, EnAsCorp1.0 [10] and Samanantar [18]. Table 3 presents the data statistics of train, validation and test data. In this work, more training parallel data of En–As is used as compared to [10, 18] since we have combined both data set as shown in table 3. It is observed that 9% of Train Set-1 and 17% of Train Set-2 contain negation sentences. The total train and validation data contain 13% and 9% negation sentences. For test data, we have used test set (Test Set-1) from [10] only. This is because the source of test data [18] is PMIndia[2] which is already present in the train data of [10]. Test Set-1 contain, 9% negation sentences. To investigate negation effect, we have manually prepared Test Set-2 that contains 100% negation sentences.

5.2 *Experimental setup*

We have used two setups, namely, phrase-based SMT and NMT. For phrase-based SMT, the Moses[3] [24] toolkit is utilized. GIZA++ [25] and IRSTLM [26] are used to extract phrase pairs and language modelling following default settings of Moses. We have explored different NMT models, namely RNN, BRNN and transformer models using the OpenNMT-py [27] toolkit. For RNN and BRNN, 2-layer LSTM based encoder-decoder architecture is used with global attention mechanism and 0.3 drop-out are used. In transformer model, default 6-layer and drop-out of 0.1 are used. Also, Adam optimizer with a learning rate of 0.001 is used in NMT models. We have used byte pair encoding (BPE) [28] with 32k merge operations in the data preprocessing step to handle out-of-vocabulary (OOV) problem. A single NVIDIA Quadro P2000 GPU is used to train the models (tables 4, 5).

[2] http://data.statmt.org/pmindia/v1/parallel/.

[3] http://www.statmt.org/moses/.

5.3 *Result and error analysis*

To evaluate the quantitative results of generated translation obtained from different MT models, automatic evaluation metrics, namely, BLEU [29], RIBES [30], TER [31], METEOR[32] and F-measure scores are used. Except TER, in all other metrics, higher the score, more is the prediction accuracy while in case of TER it is exactly opposite. Table 6, 7, 8, 9, 10 and 11 present the BLEU, RIBES, TER, METEOR and F-measure score results of various MT models by considering the Test Set-1 [10] and Test Set-2. From these tables, it is noticed that the transformer model performs better than other models in both En-to-As and As-to-En translation. Moreover, with our preprocessing approach in transformer model on Test Set-1, we have achieved 0.35+ (En-to-As) and 0.28+ (As-to-En) increment in BLEU, 1.8+(En-to-As) and 1.0+(As-to-En) improvement in TER, 0.00735+(En-to-As) and 0.01470+(As-to-En) increment in RIBES, 0.0007+(En-to-As) and 0.0054+ (As-to-En) increment in METEOR and 0.0030(En-to-As) and 0.0075+(As-to-En) increment in F-measure scores (tables 12, 13), which are reported in table 14, 15, 16, 17 and 18. Similarly, on Test Set-2, we have attained 0.68+ (En-to-As) and 0.46+(As-to-En) increment in BLEU, 0.50+(En-to-As) and 0.40+(As-to-En) improvement in TER, 0.0256+ (En-to-As) and 0.0009+ (As-to-En) increment in RIBES, 0.0045+ (En-to-As) and 0.0017+(As-to-En) in METEOR and 0.0128+(En-to-As) and 0.0019+(As-to-En) increment in F-measure scores, which are reported in table 15, 17 and 19. As compared to [10], we have attained 1.54, 5.10 higher BLEU scores (reported in table 14) on the Test Set-1 for En-to-As and As-to-En translation.

We have performed manual evaluation (also known as human evaluation (HE)). It is used to evaluate the predicted sentence by hiring human evaluators who possess linguistic knowledge of the concerned languages. HE is considered because the automatic evaluation metrics could not assess all the aspects of translation accuracy. Adequacy (AD) and fluency (FL) are the key factors in human evaluation. The adequacy factor is measured by considering the contextual meaning of the predicted sentence with respect to the reference sentence. Whereas, fluency is measured by whether the predicted sentence is well-formed or not, irrespective of the source sentence. The average score of both adequacy and fluency is known as the overall rating[4](OR). Considering an example of a reference sentence as *"John is going to the school"* and the predicted sentence as *"He is a good boy."* In this example, the predicted sentence is inadequate because the contextual meaning changes with respect to the reference sentence. However, the predicted sentence is a well-formed sentence, i.e., fluent. We have measured the assessment criteria on a scale of 1-5 by hiring three human evaluators on 100 sample sentences following [5] for Test Set-1 and Test Set-2 respectively. From table 12, 13, 20 and 21, it is observed that the overall rating scores increases in case of transformer model with our preprocessing step for both directions of translation, 0.15 increments for En-to-As and 0.31 increments for As-to-En on Test Set-1 and 0.29 increments for En-to-As and 0.21 increments for As-to-En on Test Set-2 respectively. In these tables, we have reported average HE scores (adequacy, fluency, and overall rating) of three human evaluators. To measure the reliability of human evaluators, Kappa metric [33, 34] is considered for adequacy and fluency results. The Kappa metric is calculated using Eq. 9.

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{9}$$

Where $P(A)$ indicates the proportion of times that the evaluators agree, and $P(E)$ denotes the proportion of times that evaluators would agree by chance (5 point scale 1/5). The Kappa score, when used in terms of adequacy, tends to be less than the kappa score when used in terms of fluency. It is mainly because in defining the kappa score for adequacy, the human evaluators take into effect the context in which the sentence or a word is being used, which is different for different individuals, while for fluency we see if a translation is fluent, regardless of the correct meaning or context. However, the Kappa scores of adequacy and fluency for the predicted sentences are 1 in case of all the human evaluators agree on the same rating score as shown in table 4 and 5. Table 4 and 5 present sample output of the best model (transformer) with our preprocessing step. Here, we have observed both directions of translation suffer different error types for the same sentence pair. For example, as given in Example-1 (Table 4) the En-to-As translation suffers morphological errors (দৰিদ্ৰ, নাই নে) , omission error or not translated the source word (laugh →হাঁহিব) whereas, As-to-En translation suffers lexical errors (wasted), which results in lower translational adequacy. Another error example is reported in table 5, where En-to-As translation has a morphological error (সন্ধিয়ালৈ), syntactic (word order) error, and lexical negation error (নহয়). And, As-to-En translation suffers omission error of a source word (সন্ধিয়া → in the evening) which lowers both adequacy and fluency of the translated output.

In this quantitative results and error analysis, we have observed that As-to-En translation accuracy is higher than En-to-As translation. It is because En contains more token's frequency as compared to As (as mentioned in table 3), therefore MT model encodes more En tokens and the decoder generates better translation accuracy. Moreover, En-to-As translation suffers morphological errors, unlike As-to-En translation, since As language is a morphologically rich language.

[4] https://nlp.amrita.edu/mtil_cen/#results.

Table 4. Example-1: Error example where source/target, **En**: Do not laugh at the poor. **As**: দৰিদ্ৰলোকৰ ওপ.ৰত হাঁহিব নালাগে৷.

Translation	K Score
En-to-As: দৰিদ্ৰ হৈ থকা নাই নে ।	AD: 1 K (AD)=1 FL: 3 K (FL)= 1
As-to-En: The poor should not be wasted.	AD: 1 K (AD)=1 FL: 4 K (FL)=1

Table 5. Example-2: Error example where source/target, **En**: I don't play cricket in the evening. **As**: মই সন্ধিয়া ক্ৰিকেট নাখেলোঁ৷ .

Translation	K Score
En-to-As: সন্ধিয়ালৈ মোৰ ক্ৰিকেট খেল নহয়৷	AD: 2 K (AD)=1 FL: 3 K (FL)=1
As-to-En: I have no cricket.	AD: 1 K (AD)=1 FL: 2 K (FL)=1

Table 6. Preliminary experimental results of BLEU scores on Test Set-1.

Translation	Model	BLEU
En-to-As	PBSMT	4.20
	NMT (RNN)	6.19
	NMT (BRNN)	6.32
	NMT (Transformer)	6.74
As-to-En	PBSMT	8.41
	NMT (RNN)	12.09
	NMT (BRNN)	12.33
	NMT (Transformer)	12.54

Table 7. Preliminary experimental results of BLEU scores on Test Set-2.

Translation	Model	BLEU
En-to-As	PBSMT	5.89
	NMT (RNN)	7.44
	NMT (BRNN)	7.55
	NMT (Transformer)	7.71
As-to-En	PBSMT	9.68
	NMT (RNN)	14.86
	NMT (BRNN)	14.99
	NMT (Transformer)	15.07

Table 8. Preliminary experimental results of TER and RIBES scores on Test Set-1.

Translation	Model	TER	RIBES
En-to-As	PBSMT	103.6	0.259195
	NMT (RNN)	93.8	0.274114
	NMT (BRNN)	93.5	0.281621
	NMT (Transformer)	93.2	0.293234
As-to-En	PBSMT	90.6	0.293331
	NMT (RNN)	88.9	0.394443
	NMT (BRNN)	88.7	0.408232
	NMT (Transformer)	88.2	0.405376

Table 9. Preliminary experimental results of TER and RIBES scores on Test Set-2.

Translation	Model	TER	RIBES
En-to-As	PBSMT	91.8	0.293872
	NMT (RNN)	87.4	0.384694
	NMT (BRNN)	87.2	0.391403
	NMT (Transformer)	86.8	0.397089
As-to-En	PBSMT	81.8	0.436168
	NMT (RNN)	77.8	0.550128
	NMT (BRNN)	77.5	0.554994
	NMT (Transformer)	76.5	0.558623

Table 10. Preliminary experimental results of METEOR and F-measure scores on Test Set-1.

Translation	Model	METEOR	F-measure
En-to-As	PBSMT	0.0762	0.1677
	NMT (RNN)	0.0964	0.1974
	NMT (BRNN)	0.0994	0.2014
	NMT (Transformer)	0.1021	0.2099
As-to-En	PBSMT	0.1068	0.2092
	NMT (RNN)	0.1262	0.2684
	NMT (BRNN)	0.1321	0.2771
	NMT (Transformer)	0.1354	0.2827

6. Negation effect analysis

In this section, we have analyzed three different varieties of negation effects in both directions of translation which include model-verb negation, model-verb-conjunction negation, and double negation types of sentences. The samples of predicted sentences of the best model (transformer) with and without our approach along with Bing translation are discussed with the following notations:

- ST: Source Test sentence.
- RT: Reference sentence.

Table 11. Preliminary experimental results of METEOR and F-measure scores on Test Set-2.

Translation	Model	METEOR	F-measure
En-to-As	PBSMT	0.1025	0.2186
	NMT (RNN)	0.1177	0.2574
	NMT (BRNN)	0.1185	0.2605
	NMT (Transformer)	0.1192	0.2612
As-to-En	PBSMT	0.1412	0.2723
	NMT (RNN)	0.1800	0.3609
	NMT (BRNN)	0.1824	0.3634
	NMT (Transformer)	0.1829	0.3651

Table 12. Preliminary experimental results of human evaluation scores on Test Set-1.

Translation	Model	AD	FL	OR	K(AD)	K(FL)
En-to-As	PBSMT	1.44	1.74	1.59	0.8286	0.8461
	RNN	1.74	2.32	2.03	0.8433	0.8674
	BRNN	2.18	2.80	2.49	0.8497	0.8698
	Transformer	2.48	3.02	2.75	0.8498	0.8667
As-to-En	PBSMT	2.19	2.53	2.36	0.8644	0.8727
	RNN	2.45	2.96	2.70	0.8856	0.8971
	BRNN	2.52	3.14	2.83	0.8896	0.8998
	Transformer	2.72	3.48	3.10	0.8938	0.9021

Table 13. Preliminary experimental results of human evaluation scores on Test Set-2.

Translation	Model	AD	FL	OR	K(AD)	K(FL)
En-to-As	PBSMT	1.40	2.10	1.75	0.8417	0.8550
	RNN	1.96	2.82	2.39	0.8650	0.8796
	BRNN	2.05	3.14	2.59	0.8698	0.8846
	Transformer	2.32	3.40	2.86	0.8714	0.8905
As-to-En	PBSMT	2.02	2.68	2.35	0.8480	0.8688
	RNN	2.98	3.06	3.02	0.8796	0.8867
	BRNN	3.08	3.21	3.14	0.8817	0.8926
	Transformer	3.15	3.50	3.32	0.8852	0.8914

- PS1: Predicted sentence with our pre-processing step.
- PS2: Predicted sentence without our pre-processing step.
- BT: Bing translation.

1. (a) Type-1. Model-Verb Negation (En-to-As): Example-1
   ST: *We can't hold this for long.*
   RT: আমি ইয়াক বেছি দিন ধৰি ৰাখিব নোৱাৰো।
   PS1: আমি বহুদিন ধৰি থাকিব নোৱাৰো।
   PS2: আমি এই সময় ধৰি ৰাখিব নাই ।
   BT: আমি ইয়াক বেছি দিন ধৰি ৰাখিব নোৱাৰো।

Table 14. With our approach, experimental results of BLEU scores on Test Set-1.

Translation	Model	BLEU
En-to-As	PBSMT	5.32
	NMT (RNN)	7.00
	NMT (BRNN)	7.05
	NMT (Transformer)	7.09
As-to-En	PBSMT	9.28
	NMT (RNN)	12.26
	NMT (BRNN)	12.48
	NMT (Transformer)	12.82

Table 15. With our approach, experimental results of BLEU scores on Test Set-2.

Translation	Model	BLEU
En-to-As	PBSMT	6.56
	NMT (RNN)	8.02
	NMT (BRNN)	8.21
	NMT (Transformer)	8.39
As-to-En	PBSMT	10.41
	NMT (RNN)	15.28
	NMT (BRNN)	15.45
	NMT (Transformer)	15.53

Table 16. With our approach, experimental results of TER and RIBES scores on Test Set-1.

Translation	Model	TER	RIBES
En-to-As	PBSMT	93.9	0.232623
	NMT (RNN)	92.7	0.296057
	NMT (BRNN)	91.8	0.300536
	NMT (Transformer)	91.4	0.300584
As-to-En	PBSMT	90.5	0.354370
	NMT (RNN)	87.9	0.418650
	NMT (BRNN)	87.5	0.419470
	NMT (Transformer)	87.2	0.420078

Discussion: From the above sentences, it is observed that PS2 contains the negation word 'নাই' which is not contextually correct and thus resulting in lower adequacy whereas PS1 and Bing correctly encounters 'নোৱাৰো'. Although, all the predicted sentences are fluent, but adequacy is higher in the case of BT.

1. (a) Type-1. Model-Verb Negation (En-to-As): Example-2

ST: *It is not possible to produce more crops without hard work.*

RT: কঠোৰ পৰিশ্ৰম অবিহনে অধিক শস্য উৎপাদন কৰা সম্ভৱ নহয়।

Table 17. With our approach, experimental results of TER and RIBES scores on Test Set-2.

Translation	Model	TER	RIBES
En-to-As	PBSMT	89.3	0.341498
	NMT (RNN)	87.0	0.405531
	NMT (BRNN)	87.1	0.411171
	NMT (Transformer)	86.3	0.422689
As-to-En	PBSMT	79.6	0.474999
	NMT (RNN)	77.5	0.550305
	NMT (BRNN)	77.0	0.558044
	NMT (Transformer)	76.1	0.559541

Table 18. With our approach, experimental results of METEOR and F-measure scores on Test Set-1.

Translation	Model	METEOR	F-measure
En-to-As	PBSMT	0.0925	0.1918
	NMT (RNN)	0.1014	0.2102
	NMT (BRNN)	0.1016	0.2109
	NMT (Transformer)	0.1028	0.2129
As-to-En	PBSMT	0.1193	0.2430
	NMT (RNN)	0.1382	0.2873
	NMT (BRNN)	0.1399	0.2893
	NMT (Transformer)	0.1408	0.2902

Table 19. With our approach, experimental results of METEOR and F-measure scores on Test Set-2.

Translation	Model	METEOR	F-measure
En-to-As	PBSMT	0.1028	0.2285
	NMT (RNN)	0.1223	0.2712
	NMT (BRNN)	0.1225	0.2720
	NMT (Transformer)	0.1237	0.2740
As-to-En	PBSMT	0.1548	0.3038
	NMT (RNN)	0.1827	0.3618
	NMT (BRNN)	0.1835	0.3661
	NMT (Transformer)	0.1846	0.3670

PS1: কাম অবিহনে বহু শস্য উৎপাদন কৰাটো সম্ভৱ নহয়।

PS2: কাম কৰাৰ অবিহনে অধিক শস্য উৎপন্ন কৰিব নোৱাৰি।

BT: কঠোৰ পৰিশ্ৰম অবিহনে অধিক শস্য উৎপাদন কৰা সম্ভৱ নহয়।

Discussion: In this example, the English phrase 'hard work' means 'কঠোৰ পৰিশ্ৰম' in Assamese. But both PS1 and PS2 indeed are only translating 'work' which means 'কাম' in Assamese. Also, skipped the English word possible which means সম্ভৱ in Assamese language. As a result of this, lower accuracy of PS1 and PS2 unlike BT, though the fluency is good for both of them. In comparison to PS2, PS1 correctly predicts the negation word 'নহয়' like BT.

Table 20. With our approach, experimental results of human evaluation scores on Test Set-1.

Translation	Model	AD	FL	OR	K(AD)	K(FL)
En-to-As	PBSMT	1.83	2.41	2.12	0.8324	0.8502
	RNN	2.42	3.02	2.77	0.8528	0.8705
	BRNN	2.53	3.07	2.80	0.8533	0.8706
	Transformer	2.70	3.11	2.90	0.8541	0.8708
As-to-En	PBSMT	2.42	2.92	2.67	0.8732	0.8832
	RNN	3.08	3.18	3.13	0.8948	0.9034
	BRNN	3.12	3.37	3.24	0.8956	0.9038
	Transformer	3.16	3.66	3.41	0.8958	0.9041

Table 21. With our approach, experimental results of human evaluation scores on Test Set-2.

Translation	Model	AD	FL	OR	K(AD)	K(FL)
En-to-As	PBSMT	1.92	2.23	2.07	0.8507	0.8633
	RNN	2.15	3.26	2.70	0.8716	0.8876
	BRNN	2.20	3.30	2.75	0.8720	0.8883
	Transformer	2.63	3.67	3.15	0.8724	0.8917
As-to-En	PBSMT	2.42	2.96	2.69	0.8578	0.8708
	RNN	3.21	3.48	3.34	0.8856	0.8976
	BRNN	3.26	3.56	3.41	0.8858	0.8985
	Transformer	3.28	3.79	3.53	0.8867	0.8998

1. (b) Type-1. Model-Verb Negation (As-to-En): Example-1

ST: আমি ইয়াক বেছি দিন ধৰি ৰাখিব নোৱাৰো।

RT: *We can't hold this for long.*

PS1: *We can not hold it too much.*

PS2: *We can have much time for that.*

BT: *We can't keep it for long.*

Discussion: In PS1, we notice that the predicted sentence contains the model-verb negation word 'can not' with respect to the negation source word 'নোৱাৰো ' like BT. But PS2 is unable to detect the negation word 'নোৱাৰো'. Thus, the negation word 'not' is missing in the translated text (PS2). Another observation is that 'ধৰি ৰাখিব' means 'hold' in English. BT is translating the very phase to 'keep' which is inappropriate here considering the context of the sentence. However, PS1 correctly encounters 'hold' that results in higher adequacy as compared to BT and PS1. On the other hand, all three outputs have attained good fluency.

1. (b) Type-1. Model-Verb Negation (As-to-En): Example-2

ST: কঠোৰ পৰিশ্ৰম অবিহনে অধিক শস্য উৎপাদন কৰা সম্ভৱ নহয়।

RT:*It is not possible to produce more crops without hard work.*

PS1:*It is not possible to work hard without work.*

PS2:*It can be possible without hard work.*

BT: *It is not possible to produce more crops without hard work.*

Discussion: From the above sentences, it is observed that in PS1, the predicted sentence contains the negation word 'not' but not in PS2 because PS2 is unable to detect the negation 'নহয়' from the source text. Thus, the negation word 'not' is missing from the translated text (PS2). Another observation is that the source text being long, some of its information got lost while in translation, thus resulting in poorer adequacy by both PS1 and PS2, unlike BT. While the fluency is poor in PS1, it is good in PS2, like BT.

2. (a) Type-2. Model-Verb-Conjunction Negation (En-to-As): Example-1

ST: *She could not but congratulate him.*

RT: তাই তেওঁক অভিনন্দন নজনাই নোৱাৰিলে৷

PS1: তেওঁ তেওঁক অভিনন্দন দিব নোৱাৰিলে৷

PS2:তেওঁ তেওঁক অভিনন্দন দিব নোৱাৰে৷

BT: তাই তেওঁক অভিনন্দন জনোৱাৰ বাহিৰে থাকিব নোৱাৰিলে৷

Discussion: From the sentences above, we have observed that the conjunction word plays an important role in negation sentences, as it holds the whole contextual meaning of the sentence. Here, in PS1, PS2 and BT there is a lack of adequacy, as the word 'but' gets lost in the translation and the meaning of the sentence is changed. Moreover, PS1 and PS2 attain higher fluency than BT.

2. (a) Type-2. Model-Verb-Conjunction Negation (En-to-As): Example-2:

ST: *One can not but admire her determination.*

RT: তাইৰ সংকল্পক এজনে প্ৰশংসা নকৰি নোৱাৰে৷

PS1: কোনেও তাইক সন্মান কৰিব নোৱাৰে৷

PS2: তাৰ বাবে তেওঁক সন্মান কৰিব নোৱাৰি৷

BT: এজনে তাইৰ সংকল্পক প্ৰশংসা কৰিব নোৱাৰে৷

Discussion: Based on the above sentences, it is observed that PS1, PS2 and BT generates fluent sentences, but adequacy is very poor in all the three outputs. This is mainly because of the absence of two negation words in the translated sentence, নকৰি and নোৱাৰে in the Assamese context due to the conjunction word 'but' in association with model-verb 'can not'.

2. (b) Type-2. Model-Verb-Conjunction Negation (As-to-En): Example-1

ST: তাই তেওঁক অভিনন্দন নজনাই নোৱাৰিলে৷

RT: *She could not but congratulate him.*

PS1:*She could not congratulate him.*

PS2:*She did not congratulate him.*

BT: *She couldn't congratulate him.*

Discussion: Based on the above sentences, it is noticed that while translating by PS1, PS2, and BT, the conjunction 'but' plays an important role. However, it seems that the modal-verb is not encountering in translation. Thus, the adequacy is quite poor in PS1, PS2 and even in BT. However, fluency is good for all the outputs.

2. (b) Type-2. Model-Verb-Conjunction Negation (As-to-En): Example-2:

ST: তাইৰ সংকল্পক এজনে প্ৰশংসা নকৰি নোৱাৰে৷

RT:*One can not but admire her determination.*

PS1: *One does not want to praise her.*

PS2:*Her commitment is not to be praised.*

BT: *One cannot appreciate her resolve.*

Discussion: Based on the above sentences, it is noticed that PS1, PS2 and BT produce poor adequacy translation due to lack of handling modal-verb-conjunction negation type sentences but, fluency is good for all the outputs.

3. (a) Type-3. Double Negation (En-to-As): Example-1

ST: *I don't disagree with you.*

RT: মই আপোনাৰ সৈতে অসন্মত নহয়৷

PS1: মই তোৰ লগত কথা পতা নাই৷

PS2: মই তোমাৰ লগত কথা পাতিছো৷

BT: মই আপোনাৰ সৈতে অসন্মত নহয়৷

Discussion: The model is unable to translate the negative word 'disagree', which leads to the lower adequacy in PS1 and PS2. The output that we get in PS2 is not even a negative sentence. Overall fluency of all the outputs are on the higher side. However, both adequacy and fluency factors are good in BT.

3. (a) Type-3. Double Negation (En-to-As):Example-2

ST:*I don't dislike you.*

RT: মই তেমাক বেয়া নাপাওঁ৷

PS1: মই তেমাক বেয়া নকৰো৷

PS2: মই তেমাক ঘৃণি নহয়৷

BT: মই তেমাক বেয়া নাপাওঁ৷

Discussion: The negative word dislike has been correctly translated into বেয়া and ঘিণ into Assamese, as evident from PS1 and PS2, respectively. Also, the negative word নাপাওঁ is more contextually correct here instead of নকৰো or নহয় which lowers the adequacy and fluency of PS1 and PS2. In the case of BT, the predicted sentence is fine with respect to the reference sentence.

3. (b) Type-3. Double Negation (As-to-En): Example-1

ST: মই আপোনাৰ সৈতে অসন্মত নহয়৷

RT: *I don't disagree with you.*

PS1: *I do not agree with you.*

PS2: *I dont agree with you.*

BT: *I don't disagree with you.*

Discussion: The model while performing translation is unable to convert the negation word 'অসন্মত' into 'disagree' as evident from the output of PS1 and PS2. The possible reason for this might be that the root word 'সন্মত' is only present in the vocabulary. ('অসন্মত' is actually derived from the root word 'সন্মত'). As the output obtained in PS1 and PS2 don't contain the negation word 'disagree' unlike BT, the adequacy is lower as the meaning of the source text is completely changing in the translated text. Fluency attained by all the outputs is on the higher side.

3. (b) Type-3. Double Negation (As-to-En):Example-2:

ST: মই তোমাক বেয়া নাপাওঁ৷

RT: *I don't dislike you.*

PS1: *I am not bad to you.*

PS2: *I am not bad to you.*

BT: *I don't mind you.*

Discussion: The model while performing translation is unable to convert the negation word 'বেয়া' into 'dislike' as evident from the output of PS1 and PS2. The Assamese word 'বেয়া' most commonly means 'bad' in English. But the context of the source text doesn't permit us to use the English word 'bad' in this case. The negative word 'dislike' also means 'বেয়া' in Assamese, which is more appropriate here. Thus, the adequacy obtained is really low for the output obtained in PS1 and PS2. The adequacy of the output obtained from the BT is somewhat higher than the previous two. Although the fluency is good in BT, but is really poor in PS1 and PS2.

From the above discussion, we have found a few notable points that include our rule-based approach (as discussed in section 4) in data preprocessing significantly improves the modal-verb negation sentences. However, there are research scopes to handle double negation and modal-verb-conjunction negation in both the directions of En–As pair translation. Also, we have observed that the parallel training data needs to be increased to include various types of complex negation sentences in order to increase the accuracy of learning and translating negation sentences.

7. Conclusion and future work

In this paper, we have investigated different types of negation effects for En-to-As and vice-versa by exploring various MT models. To handle modal-verb negation problems, we have proposed a rule-based approach in data preprocessing step that shows significant enhancement in terms of automatic and manual evaluation scores. Furthermore, we have discussed various sources of error involving negation for both directions of translation. In the future works, we will investigate linguistic phenomenon that includes double-negation, modal-verb-conjunction negation, pronoun and anaphora resolution, etc. for the En–As pair translation.

Acknowledgements

We want to thank the Department of Computer Science and Engineering, Center for Natural Language Processing (CNLP), Artificial Intelligence (AI) Lab at the National Institute of Technology, Silchar for providing the requisite support and infrastructure to execute this work.

References

[1] Koehn P (2010) Statistical machine translation. 1st edition. Cambridge University Press, Berlin

[2] Bahdanau D, Cho K and Bengio Y 2015 Neural machine translation by jointly learning to align and translate. In: Bengio Y and LeCun Y (eds) *3rd International Conference on Learning Representations, ICLR 2015, May 7-9, 2015, Conference Track Proceedings*, pp. 1–15, San Diego, CA, USA, 2015. arXiv

[3] Luong T, Pham H Manning C D 2015 Effective approaches to attention-based neural machine translation. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics

[4] Gu J, Hassan H, Devlin J and Li V O K 2018 Universal neural machine translation for extremely low resource languages. In: Walker MA, Ji H and Stent A (eds) *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT USA, June 1–6, 2018, Volume 1 (Long Papers)*, pp. 344–354, New Orleans, Louisiana, 2018. Association for Computational Linguistics

[5] Pathak A, Pakray P and Bentham J 2018 English–mizo machine translation using neural and statistical approaches. *Neural Comput Appl*, 30:1–17

[6] Denkowski M and Neubig G 2017 Stronger baselines for trustable results in neural machine translation. In: *Proceedings of the First Workshop on Neural Machine Translation*, pp. 18–27, Vancouver, August 2017. Association for Computational Linguistics

[7] Kocmi T 2020 Exploring benefits of transfer learning in neural machine translation

[8] Megerdoomian K and Parvaz D 2008 Low-density language bootstrapping: the case of tajiki Persian. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pp. 3293–3298, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA)

[9] Probst K, Brown R, Carbonell J, Lavie A, Levin L S and Peterson E 2001 Design and implementation of controlled elicitation for machine translation of low-density languages. pp. 3293–3298

[10] Rahman Laskar S, Faiz Ur Rahman Khilji A, Pakray P and Bandyopadhyay S 2020 EnAsCorp1.0: English-Assamese corpus. In: *Proceedings of the 3rd Workshop on Technologies for MT of Lowv Resource Languages*, pp. 62–68, Suzhou, China, December 2020. Association for Computational Linguistics

[11] Mosharaf Hossain M, Anastasopoulos A, Blanco E and Palmer A 2020 It's not a non-issue: Negation as a source of error in machine translation. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3869–3885, Online, November 2020. Association for Computational Linguistics

[12] Guillou L and Hardmeier C 2016 PROTEST: A test suite for evaluating pronouns in machine translation. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 636–643, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA)

[13] Voita E, Serdyukov P, Sennrich R and Titov I 2018 Context-aware neural machine translation learns anaphora resolution. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne,*

Australia, July 15-20, 2018, Volume 1: Long Papers, pp. 1264–1274. Association for Computational Linguistics

[14] Burlot F and Yvon F 2017 Evaluating the morphological competence of machine translation systems. In: *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7–8, 2017*, pp. 43–55. Association for Computational Linguistics

[15] Tang G, Sennrich R and Nivre J 2018 An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In: *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31–November 1, 2018*, pp. 26–35. Association for Computational Linguistics

[16] Barman A, Sarmah J and Sarma S 2014 Assamese WordNet based quality enhancement of bilingual machine translation system. In: *Proceedings of the Seventh Global Wordnet Conference*, pp. 256–261, Tartu, Estonia, January 2014. University of Tartu Press

[17] Hannan A, Baruah K K, Das P and Kr Sarma S 2014 Assamese-english bilingual machine translation. *Int J Natural Lang Comput (IJNLC)*

[18] Ramesh G, Doddapaneni S, Bheemaraj A, Jobanputra M, Raghavan A K, Sharma A, Sahoo S, Diddee H, Kakwani D, Kumar N *et al* Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.*Trans Assoc Comput Linguist*, 10: 145–162

[19] Laskar S R, Ur Rahman Khilji A F, Pakray P and Bandyopadhyay S 2022 Improved neural machine translation for low-resource english–assamese pair. *J Intell Fuzzy Syst*, (Preprint): 1–12

[20] Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H and Bengio Y 2014 Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics

[21] Sutskever I, Vinyals O and Le Q V 2014 Sequence to sequence learning with neural networks. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 3104-3112, Cambridge, MA, USA, MIT Press

[22] Ramesh S H and Sankaranarayanan K P 2018 Neural machine translation for low resource languages using bilingual lexicon induced from comparable corpora. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 112–119, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics

[23] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N and Kaiser L 2017 Attention is all you need. In: Guyon I, Luxburg U V, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R, editors, *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc.

[24] 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 and Herbst E 2007 Moses: Open source toolkit for statistical machine translation. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[25] Och F J and Ney H 2003 A systematic comparison of various statistical alignment models. *Comput Linguist*, 29(1): 19–51

[26] Federico M, Bertoldi N and Cettolo M 2008 Irstlm: an open source toolkit for handling large scale language models. In: *INTERSPEECH*, pp. 1618–1621. ISCA

[27] Klein G, Kim Y, Deng Y, Senellart J and Rush A 2017 OpenNMT: Open-source toolkit for neural machine translation. In: *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics

[28] Sennrich R, Haddow B and Birch A 2016 Neural machine translation of rare words with subword units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics

[29] Papineni K, Roukos S, Ward T and Zhu W-J 2002 Bleu: A method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics

[30] Isozaki H, Hirao T, Duh K, Sudoh K and Tsukada H 2010 Automatic evaluation of translation quality for distant language pairs. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 944–952, Cambridge, MA, October 2010. Association for Computational Linguistics

[31] Snover M, Dorr B, Schwartz R, Micciulla L and Makhoul J 2006 A study of translation edit rate with targeted human annotation. In: *In Proceedings of Association for Machine Translation in the Americas*, pp. 223–231

[32] Lavie A and Denkowski M J 2009 The meteor metric for automatic evaluation of machine translation. *Mach Transl*, 23 (2-3):105-115

[33] Cohen J 1960 A coefficient of agreement for nominal scales. *Educ Psychol Meas*, 20(1): 37–46

[34] Seljan S, Brkić M and Vičić T 2012 BLEU evaluation of machine-translated English-Croatian legislation. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 2143–2148, Istanbul, Turkey, European Language Resources Association (ELRA)