EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION

EAMT2018

Proceedings of the

21st Annual Conference of the European Association for Machine Translation

28–30 May 2018

Universitat d'Alacant

Alacant, Spain

Edited by

Juan Antonio Pérez-Ortiz

Felipe Sánchez-Martínez

Miquel Esplà-Gomis

Maja Popović

Celia Rico

André Martins

Joachim Van den Bogaert

Mikel L. Forcada

Organised by

Universitat d'Alacant

Universidad de Alicante

transducens

research group

ISBN: 978-84-09-01901-4

Translating Short Segments with NMT: A Case Study in English-to-Hindi

Shantipriya Parida **Ondřej Bojar**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
{parida,bojar}@ufal.mff.cuni.cz

Abstract

This paper presents a case study in translating short image captions of the Visual Genome dataset from English into Hindi using out-of-domain data sets of varying size. We experiment with three NMT models: the shallow and deep sequence-to-sequence and the Transformer model as implemented in Marian toolkit. Phrase-based Moses serves as the baseline.

The results indicate that the Transformer model outperforms others in the large data setting in a number of automatic metrics and manual evaluation, and it also produces the fewest truncated sentences. Transformer training is however very sensitive to the hyperparameters, so it requires more experimenting. The deep sequence-to-sequence model produced more flawless outputs in the small data setting and it was generally more stable, at the cost of more training iterations.

1 Introduction

In recent years, neural machine translation (NMT) systems have been gaining more popularity due to their improved accuracy and even more fluency compared with "classical" statistical machine translation systems such as phrase-based MT (PBMT), see e.g. the shared tasks of WMT and IWSLT (Bojar et al., 2017; Cettolo et al., 2017). The major advantages of NMT include the consideration of the entire sentence, capturing similarity of words, and the capacity to learn complex relationships between languages. At the same time, it has been observed that NMT is more sensitive to the shortage of or noise in the parallel training data (Koehn and Knowles, 2017).

Our goal is to create the Hindi version of Visual Genome (Krishna et al., 2017).[1]

Hindi, with 260 million speakers, is the fourth most widely spoken language on the planet (after Chinese, Spanish and English). Hindi is a morphologically rich language (MRL), with e.g. the gender category being reflected in the forms of nouns, verbs and also adjectives (Sreelekha S and Bhattacharyya, 2017). The structural and morphological differences between English and Hindi result in translation difficulties (Tsarfaty et al., 2010).

Visual Genome is a dataset of images, captions and relations. As such, it is potentially useful for many NLP and image processing applications. The Hindi version would allow to exploit this dataset e.g. to create Hindi image labellers or other practical tools.

The textual part of Visual Genome consists primarily of short sentences or noun phrases that were manually attached to rectangular regions in an input image. In the current version, Visual Genome contains 108K distinct images with 5.4 million such labelled regions in total. On average, an image is thus associated with 50 text segments. Text segments can repeat across images and indeed, when de-duplicated, the set of unique strings reduces to 3.15 million unique segments.

Even with this de-duplication, this set remains too big to be translated manually. It is thus natural to attempt to translate this dataset automatically and in this paper, we are trying to find the best base-

[1] http://visualgenome.org/

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 229–238
Alacant, Spain, May 2018.

line translation. In the future, we want to include also information available in the context of each of the labels: either the text descriptions of nearby regions or directly the visual information in a form of multi-modal translation (Matusov et al., 2017; Calixto et al., 2012; Huang et al., 2016).

The paper is organized as follows. Section 2 reviews related work on neural MT and English-Hindi translation. Section 3 describes our experimental setting: data, models and their parameters. Section 4 provides automatic and manual evaluation of the translations and Section 5 discusses the results in closer detail. We conclude in Section 6.

2 Related Work

Singh et al. (2017) have compared two neural machine translation models, convolutional sequence to sequence (ConvS2S) and recurrent sequence to sequence (RNNS2S) for English↔Hindi machine translation task. They have used the IITB corpus for training (see Section 3.1) and also for development and test data. The RNNS2S model was trained using Nematus (Sennrich et al., 2017) and ConvS2S using Fairseq (Gehring et al., 2017), an open source library developed by Facebook. In their evaluation, ConvS2S was better when targetting English (BLEU scores: RNNS2S: 11.55, ConvS2S: 13.76) but RNNS2S was better when targetting Hindi (BLEU scores: RNNS2S: 12.23, ConvS2S: 11.73). As our experiment scope is limited to English to Hindi translation, we have not tried the ConvS2S.

Wang et al. (2017) use the encoder-decoder framework with attention (Bahdanau et al., 2015) for their submission to the Workshop on Asian Translation (WAT) 2017 shared task and observe considerable gains for English-to-Hindi compared to PBMT. Similarly to other works, they benefit from subword units (Sennrich et al., 2016a) and back-translation (Sennrich et al., 2016b), as well as model ensembling.

Agrawal and Misra Sharma (2017) evaluate English-Hindi translation quality using several variants of RNN-based neural network architecture and basic units (LSTMs, Hochreiter and Schmidhuber, 1997, and GRUs, Cho et al., 2014b), including the attention mechanism by Bahdanau et al. (2015) and more layers in the encoder and decoder. The bi-directional LSTM model with four layers and attention performs best.

The early models of NMT have suffered from lower translation quality for long sentences, see e.g. Cho et al. (2014a) and Bahdanau et al. (2015). A recent experiment by Beyer et al. (2017) has however suggested that NMT can perform worse than PBMT also for short segments (insignificantly). It is thus natural to evaluate the effect in our particular setting.

We note that monolingual data plays an important role in boosting the performance of the translation in both PBMT (Brants et al., 2007; Bojar and Tamchyna, 2011) and NMT (Sennrich et al., 2016b; Domhan and Hieber, 2017). We leave these experiments for future work because we would first need to find or select Hindi texts closely matching to the domain of Visual Genome texts.

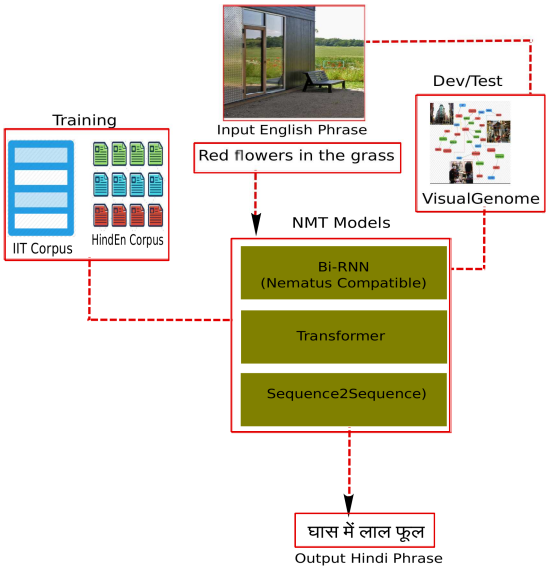


Figure 1: Overall experimental setting.

3 Experiments

The overall framework of our work is shown in Figure 1. The targeted dataset is English text descriptions from Visual Genome but no similar or related data is available in Hindi. So far, we thus used Visual Genome only to select the development and the test set.

We experimented with two parallel corpora as our training data, HindEnCorp and IITB Corpus (see Section 3.1), three NMT models and the PBMT baseline (Section 3.2).

We used the experiment management tool Eman (Bojar and Tamchyna, 2013)[2] for organizing and running the experiments.

[2]http://ufal.mff.cuni.cz/eman

Set	#Sentences	#Tokens	
		En	Hi
Train (HindEnCorp)	273.9k	3.8M	5.6M
Train (IITB)	1492.8k	20.8M	31.4M
Dev (Visual Genome)	898	4519	6219
Test (Visual Genome)	1000	4909	6918

Table 1: Statistics of our data.

3.1 Dataset Description

This section describes the processing and usage of the training and development data. We have used HindEnCorp (Bojar et al., 2014) as the training dataset which contains 274k parallel sentences. Additionally, we have explored the very recent "IIT Bombay English-Hindi Parallel Corpus" (Kunchukuttan et al., 2018) which is supposedly the largest publicly available English-Hindi parallel corpus. This corpus contain 1.49 million parallel segments and it includes HindEnCorp.

The development and test sentences were extracted from the Visual Genome. The original dataset contains images and their region annotations and several other formally captured types of information (objects, attributes, relationships, region graphs, scene graphs and question answer pairs). We built our dataset by extracting only the region descriptions, which are generally short sentences or phrases. We selected the development and test segments randomly and prepared the corresponding Hindi translation by manually correcting Google Translate outputs.

The training and test sets sizes are shown in Table 1. Note that the token counts considerably differ from those reported in the corpus descriptions. Here we report the token counts as obtained by the Moses tokenizer and used in all our experiments.

3.2 MT Models Tested

One of the current most efficient NMT toolkits is Marian[3] (Junczys-Dowmunt et al., 2016), which is a pure C++ implementation of several popular NMT models. All our experiments thus use Marian models.

3.2.1 Marian's nematus Model (Bi-RNN)

The common baseline NMT architecture is the (shallow) attentional encoder-decoder of Bahdanau et al. (2015). A particularly popular implementation of this model is available in the Nematus toolkit (Sennrich et al., 2017),[4] which adds some implementation differences such as a different initial hidden state, a different RNN cell and several others.

Marian implements both the training and inference with the Nematus (Sennrich et al., 2017) model and in fact, it can load models trained by the original Nematus.

We call this setup "Bi-RNN" in the following and use it only in shallow (depth 1) setting.

Parameter	Bi-RNN	S2S	Transformer
beam-size	12	12	12
dec-cell	gru	lstm	–
dec-cell-base-depth	2	4	–
dec-cell-high-depth	1	2	–
dec-depth	1	4	6
decay-inv	–	–	16000
dim-emb	512	512	512
dim-rnn	1024	1024	1024
dropout-rnn	0.2	0.2	–
dropout-src	0.1	0.1	–
dropout-trg	0.1	0.1	–
early-stopping	10	–	–
enc-cell	gru	lstm	–
enc-cell-depth	1	2	–
enc-depth	1	4	6
enc-type	bidirectional	alternating	–
exponential-smoothing	–	0.0001	–
heads	–	–	8
label-smoothing	–	–	0.1
learning-rate	0.0001	0.0001	0.0003
max-length	50	50	100
normalize	–	–	0.6
optimizer	adam	adam	adam
transformer-dim-ffn	–	–	2048
transformer-dropout	–	–	0.1
transformer-dropout-attention	–	–	0
transformer-postprocess	–	–	dhn
warm-up	–	–	16000

Table 2: Model configurations.

3.2.2 Marian's Sequence-to-Sequence (s2s) Model

A more advanced variation of the RNN-based model allows to use deeper layers in both decoder and encoder and it also differs from the original Nematus model in several features, such as a different layer normalization (Sennrich et al., 2017; Junczys-Dowmunt and Grundkiewicz, 2017).

We denote this model "S2S" in the following and use it only in the deep (depth 4) setting.

3.2.3 Marian's transformer Model

The Transformer model (Vaswani et al., 2017) has been recently proposed to avoid the expensive training of RNNs, relying on the attention mechanism.

As explored by Popel and Bojar (2018) with the

[3] http://github.com/marian-nmt/marian
[4] http://github.com/EdinburghNLP/nematus

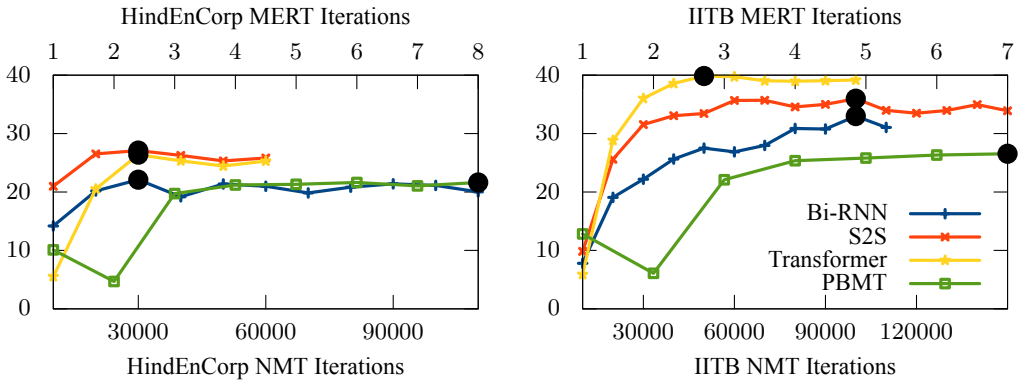


Figure 2: Learning curves in terms of BLEU on dev set. The big black dots indicate which iteration was used for test set translation and evaluation.

original Google implementation,[5] the model can be more difficult to train but it will likely outperform other architectures in both training time and final translation quality. Indeed, we needed to try 9 different configuration settings for Transformer before we got any reasonable performance, compared to just 3 for S2S and 1 for Bi-RNN.

Marian's implementation should be fully compatible with the original Google one.

The configuration parameters used for training of the models are shown in Table 2.

3.2.4 Common Settings

In all NMT experiments, we used the same BPE (Sennrich et al., 2016a), with 30k merges, joint for English and Hindi and extracted from HindEnCorp only. We also tried to extract the BPE from the respective training corpus (i.e. IITB for IITB models) but the performance was lower, perhaps due to domain differences between the corpora. The HindEnCorp BPEs are thus used in all experiments reported here.

3.2.5 Moses PBMT Baseline

For the purposes of comparison, we also train Moses (Koehn et al., 2007) phrase-based MT system with a 5-gram LM and a lexicalized reordering model, trained with the standard MERT optimization towards BLEU. The alignment is based on lowercase tokens, stemmed to the first 4 characters only.

4 Results

Figure 2 presents the learning curves for all the models evaluated on the development set using the BLEU score (Papineni et al., 2002). (PBMT training is displayed in terms of MERT iterations on the secondary x axis.)

For NMT, we validated the model every 10000 iterations and ran the training until the cross-entropy has not improved for 10 consecutive validations. For each model, we selected the iteration where the highest BLEU score was reached and translated the test set with this model.

[5] http://github.com/tensorflow/tensor2tensor

		Bi-RNN	S2S	Transf.	PBMT
HindEnCorp	BLEU	20.68	**26.45**	23.91	20.61
	chrF3	32.30	**39.52**	36.36	36.49
	nCDER	34.04	**40.91**	38.26	32.71
	nCharacTER	12.27	18.47	23.12	**29.05**
	nPER	41.76	49.05	47.01	**50.40**
	nTER	29.63	**35.70**	33.52	24.78
IITB Corpus	BLEU	31.78	32.81	**38.31**	25.06
	chrF3	42.63	44.50	**51.08**	43.09
	nCDER	44.49	44.91	**51.78**	37.54
	nCharacTER	-14.76	-47.00	25.07	**37.55**
	nPER	51.86	52.04	**59.60**	55.17
	nTER	40.62	41.44	**49.05**	32.76

Table 3: Results on the test set, multiplied by 100. Best model according to each automatic metric in bold. Metrics with the prefix "n" were flipped (100 − score) to make better scores higher. The negative numbers for nCharacTER happen when the original CharacTER score is over 1.

4.1 Automatic Evaluation

Table 3 provides automatic scores of the models in several metrics (Papineni et al., 2002; Snover et al., 2006; Leusch and Ney, 2008; Popović, 2015; Wang et al., 2016).[6] We see that on the smaller HindEn-

[6] Note that the exact scores are *heavily* dependent on the tokenization. We collect outputs from all our system after detokenization and tokenize if needed by the metric (chrF3 and CharacTER do not expect tokenized text). We report the scores when Moses tokenizer was used. Using e.g. the Hindi tokenization from IndicNLP, http://github.com/anoopkunchukuttan/indic_nlp_library, leads to sub-

Corp, S2S performs best except in CharacTER and PER where the outputs of PBMT score best. On the larger IITB Corpus, Transformer wins in all metrics except again CharacTER. We suspect that the different evaluation by CharacTER could be an artifact of the Devanagari script used in Hindi.

PER, position-independent error-rate, reflects the overlap of exact word forms used in the reference and the hypothesis, suggesting that PBMT performs reasonably well in terms of preserving words, although the fluency is probably worse.

It should be noted that the automatic scores can be affected by the fact that our test set was created by manual revision of Google Translate outputs. The underlying model of Google Translate is however unknown. Also, we have only one reference translation and it is well known that with more reference translations, automatic evaluations are more reliable (Finch et al., 2004; Bojar et al., 2013).

4.2 Manual Evaluation

To validate the automatic scoring, we manually annotated 100 randomly selected segments as translated by the NMT models.[7]

In this annotation, each annotated segment gets exactly one label from the following set:

Flawless for translations without any error (typesetting issues with diacritic marks due to different tokenization are ignored),

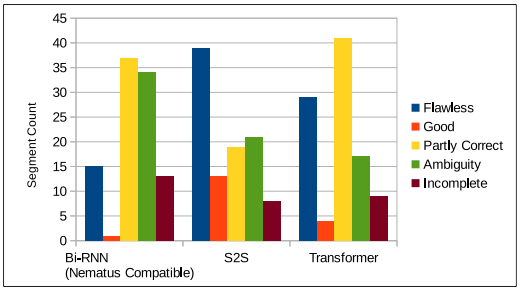
Good for translations which are generally OK and complete but need a small correction,

Partly Correct for cases where a part of the segment is correct but some words are mistranslated,

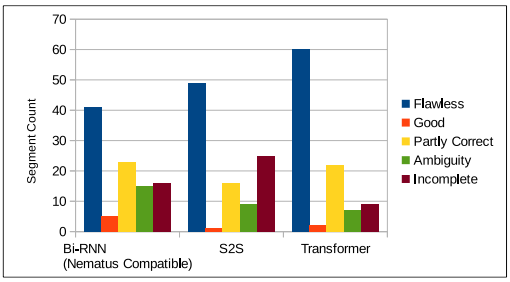
Ambiguity for segments where the MT system "misunderstood" a word's meaning, and

Incomplete for segments that run well but stop too early, missing some content words. This category also includes the relatively rare cases where the NMT model produced just a single word, unrelated to the source.

The results are summarized in Figure 3.



(a) HindEnCorp-trained models



(b) IITB-trained models

Figure 3: Manual evaluation summary.

The manual annotation generally confirms the automatic scores. On HindEnCorp, S2S has the highest number of Flawless segments and Bi-RNN performs worst, having the majority of outputs only Partly Correct and suffering most from Ambiguity.

On IITB, the performance of all the models is generally much better, with 40–60 of the 100 annotated segments falling into the Flawless category. Transformer is a clear winner here and S2S suffers from surprisingly many Incomplete segments.

Some translation samples are shown in Figure 4.

5 Analysis and Discussion

We assumed that PBMT may perform better on short segments. In order to test this assumption, we divided the 1000 test segments into 5 groups based on the source segment length. Group boundaries were chosen to achieve reasonably balance distribution and at least a minimal size for automatic scoring:

Source length:	1–3	4	5	6	7–12
Segment count:	73	380	282	165	100

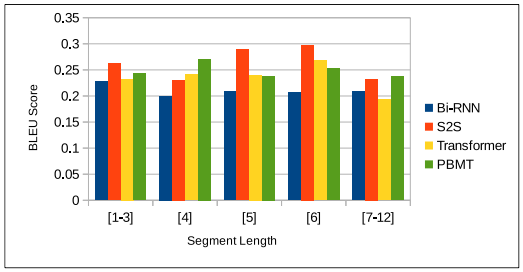
Figure 5 plots BLEU scores evaluated on each group of segments separately. We see that our assumption does not hold and that there is no clear tendency in translation quality based on source sentence length. In the small data setting (HindEnCorp), PBMT scores well sentences of length 4 and

stantially lower scores, e.g. BLEU of 7 instead of 20. Fortunately, these BLEU scores correlate very well (Pearson of 0.94) with our scores.

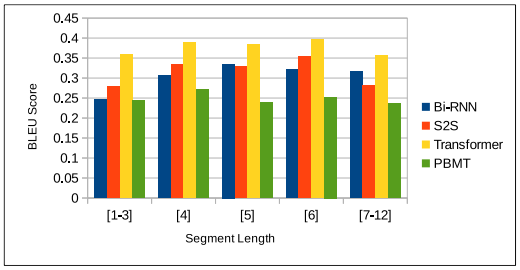
[7]We excluded PBMT from this annotation because its BLEU scores were low; we are now reconsidering this decision given the good performance in PER.

Flawless:
A car on a street
सडक पर एक कार
Gloss: A car on a street
A white and yellow passenger car
एक सफेद और पीला यात ◌ृ री कार
Gloss: A white and yellow passenger car
White part of the chair
कुर ◌ृ सी का सफेद भाग
Gloss: White part of the Chair
Partly Correct:
A man wearing white shorts
एक आदमी सफेद शॉर ◌ृ ट पहनना
Gloss: A man put on white short (output does not convey the intended meaning in the target language)
Dog in a lake
इस झील में कुत ◌ृ ते
Gloss: Dogs in this lake (grammar error: dog vs. dogs)
Ambiguity:
Faucet is above sink
फेसबुक सिंक से ऊपर है
Gloss: Facebook is above sink (bad translation of the word “Faucet’)
Green bean in soup
आत ◌ृ मा में हरा
Gloss: Spirit in green (mis-translated words “bean”, and “soup”)

Figure 4: Sample segment translations and their manual classification.



(a) HindEnCorp-trained models



(b) IITB-trained models

Figure 5: Translation quality for groups of segments based on their source length.

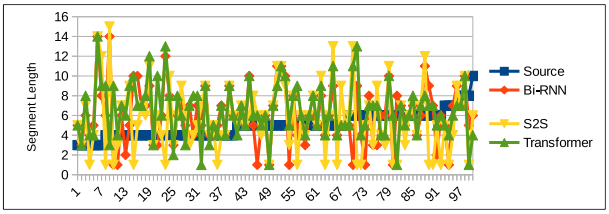


Figure 6: Source and candidate translation lengths for individual segments in the subset of 100 manually-evaluated segments. Segments are sorted by source length. The models were trained on the IITB corpus.

then on sentences over 7 words. In other cases, S2S wins. With the IITB training corpus, Transformer wins and PBMT loses across all lengths.

A generally interesting property of NMT is its ability to correctly predict the sentence length (Shi et al., 2016). We take a look at this by considering both the relation of our candidate translations with the source and with the reference.

Figure 6 plots the length of the translation for individual source segments sorted by length. We see that the target length varies a lot across segments and also different NMT models. In general, outputs are longer than sources but the length of the source is not really followed by any of the models.

We observed on the HindEnCorp training data that some of the NMT models tended to cut off sentences too short in early iterations. To examine this, we checked the difference in length between the candidate and the reference throughout the iterations. The distribution of length differences was however not skewed in any way and the only observable pattern was that the differences get smaller as the training progresses. We plot the differences for the converged runs over the whole 1000 segments in the test set in Figure 7. We see that all the NMT models are very similar, producing output slightly longer (peak at +2) than the reference. The PBMT is optimized well and the peak is located at zero difference between the candidate and reference length. The interesting pattern in NMT outputs of slightly fewer segments with odd differences (+1, +3 and +5) has still to be explained.

6 Conclusion

We have applied the state-of-the-art neural machine translation models and the phrase-based

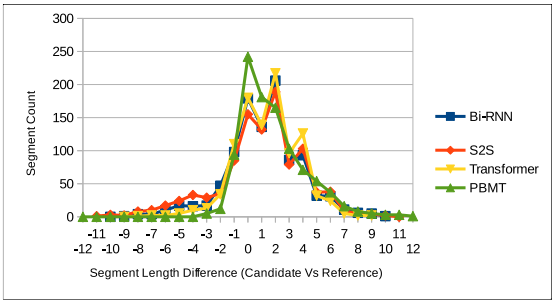


Figure 7: Segment length difference (candidate vs reference) of the IITB-trained models. The positive numbers indicate that candidate is longer than the reference.

baseline to English-to-Hindi translation. Our target domain were relatively short segments appearing in descriptions of image regions in the Visual Genome.

The results indicate that with smaller data (274k parallel segments, 3.8M English tokens), the deep sequence-to-sequence attentional model is the best choice, although the PBMT baseline seemed to perform well in two of the tested automatic metrics, CharacTER and PER. With large parallel data available, Transformer should be preferred and all NMT models clearly outperform PBMT. We have not yet explored the effect of adding monolingual data.

A deeper analysis has not revealed any difference in performance for shorter or longer segments, but the manual annotation suggested that the performance of NMT models varies across individual segments. The overall performance is thus perhaps too crude and it would be suboptimal to decide for a single model.

In the future, we will focus on the possibilities of multi-modal translation (Matusov et al., 2017; Calixto et al., 2012; Huang et al., 2016) to improve translation quality using the Visual Genome images or other contextual information available. Our ultimate plan is to release a machine-translated Hindi version of Visual Genome.

Acknowledgement

This work has been supported by the grants 18-24210S of the Czech Science Foundation, SVV 260 453 and "Progress" Q18+Q48 of Charles University, and using language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (projects LM2015071 and OP VVV VI CZ.02.1.01/0.0/0.0/16 013/0001781).

We thank Dr. Satyaranjan Dash and Miss Sneha Shrivastav for their support in Development and Test Data preparation.

References

Ruchit Agrawal and Dipti Misra Sharma. Building an Effective MT System for English-Hindi Using RNN's. *International Journal of Artificial Intelligence & Applications*, 8:45–58, 09 2017.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR*, 2015.

Anne Beyer, Vivien Macketanz, Aljoscha Burchardt, and Philip Williams. Can Out-of-the-box NMT Beat a Domain-trained Moses on Technical Data? In *Proceedings of EAMT User Studies and Project/Product Descriptions*, pages 41–46, Prague, Czech Republic, 2017.

Ondřej Bojar and Aleš Tamchyna. The Design of Eman, an Experiment Manager. *The Prague Bulletin of Mathematical Linguistics*, 99:39–58, 2013. ISSN 0032-6585.

Ondřej Bojar and Aleš Tamchyna. Improving Translation Model by Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.

Ondřej Bojar, Matouš Macháček, Aleš Tamchyna, and Daniel Zeman. Scratching the Surface of Possible Translations. In *Proc. of TSD 2013*, Lecture Notes in Artificial Intelligence, Berlin / Heidelberg, 2013. Západočeská univerzita v Plzni, Springer Verlag.

Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. HindEnCorp — Hindi-English and Hindi-only Corpus for Machine Translation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3550–3555, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. Large Language Models in Machine Translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, 2007.

Iacer Calixto, Teófilo Emídio de Campos, and Lucia Specia. Images as Context in Statistical Machine Translation. In *In The 2nd Annual Meeting of the EPSRC Network on Vision & Language (VL'12)*, Sheffield, UK, 2012. EPSRC Vision and Language Network.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. Overview of the IWSLT 2017 Evaluation Campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)*, pages 2–14, Tokyo, Japan, 2017.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October 2014a. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014b. Association for Computational Linguistics.

Tobias Domhan and Felix Hieber. Using Target-side Monolingual Data for Neural Machine Translation through Multi-task Learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1500–1505, 2017.

Andrew M. Finch, Yasuhiro Akiba, and Eiichiro Sumita. How Does Automatic Machine Translation Evaluation Correlate with Human Scoring as the Number of Reference Translations Increases? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC*, 2004.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional Sequence to Sequence Learning. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Comput.*, 9(8): 1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. Attention-based Multimodal Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation, WMT*, pages 639–645, 2016.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. An Exploration of Neural Sequence-to-Sequence Architectures for Automatic Post-Editing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP*, pages 120–129, 2017.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In *Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

Philipp Koehn and Rebecca Knowles. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August 2017. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico,

Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL) Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73, May 2017. ISSN 1573-1405. doi: 10.1007/s11263-016-0981-7.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The IIT Bombay English-Hindi Parallel Corpus. In *Proceedings of LREC*, 2018. In print.

Gregor Leusch and Hermann Ney. BLEUSP, INVWER, CDER: Three improved MT evaluation measures. In *NIST Metrics for Machine Translation Challenge*, Waikiki, Honolulu, Hawaii, October 2008.

Evgeny Matusov, Andy Way, Iacer Calixto, Daniel Stein, Pintu Lohar, and Sheila Castilho. Using Images to Improve Machine-Translating E-Commerce Product Listings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pages 637–643, 2017.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Martin Popel and Ondřej Bojar. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70, 2018.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*, 2016a.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*, 2016b.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL; Software Demonstrations*, pages 65–68, 2017.

Xing Shi, Kevin Knight, and Deniz Yuret. Why Neural Translations are the Right Length. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2278–2282, Austin, Texas, November 2016. Association for Computational Linguistics.

Sandhya Singh, Ritesh Panjwani, Anoop Kunchukuttan, and Pushpak Bhattacharyya. Comparing Recurrent and Convolutional Architectures for English-Hindi Neural Machine Translation. In *Proceedings of the 4th Workshop on Asian Translation, WAT@IJCNLP*, pages 167–170, 2017.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings AMTA*, pages 223–231, August 2006.

Sreelekha S and Pushpak Bhattacharyya. Role of Morphology Injection in SMT: A Case Study from Indian Language Perspective. *ACM Trans. Asian & Low-Resource Lang. Inf. Process.*, 17 (1):1:1–1:31, 2017. doi: 10.1145/3129208.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kübler, Yannick Versley, Marie Candito, Jennifer Foster, Ines Rehbein, and Lamia Tounsi. Statistical Parsing of Morphologically Rich Languages (SPMRL) What, How and

Whither. In *Proceedings of the First Workshop on Statistical Parsing of Morphologically-Rich Languages, SPMRL@NAACL-HLT*, pages 1–12, 2010.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 6000–6010, 2017.

Boli Wang, Zhixing Tan, Jinming Hu, Yidong Chen, and Xiaodong Shi. XMU Neural Machine Translation Systems for WAT 2017. In *Proceedings of the 4th Workshop on Asian Translation, WAT@IJCNLP*, pages 95–98, 2017.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. CharacTER: Translation Edit Rate on Character Level. In *ACL First Conference on Machine Translation (WMT)*, Berlin, Germany, August 2016.