Experimenting with Different Machine Translation Models in Medium-Resource Settings

Haukur Páll Jónsson[1], Haukur Barri Símonarson[2], Vésteinn Snæbjarnarson[2], Steinþór Steingrímsson[1], and Hrafn Loftsson[1(✉)]

[1] Language and Voice Lab, Reykjavik University, Reykjavik, Iceland
{haukurpj,steinthor18,hrafn}@ru.is
[2] Mieind ehf., Reykjavik, Iceland
{haukur,vesteinn}@mideind.is

Abstract. State-of-the-art machine translation (MT) systems rely on the availability of large parallel corpora, containing millions of sentence pairs. For the Icelandic language, the parallel corpus ParIce exists, consisting of about 3.6 million English-Icelandic sentence pairs. Given that parallel corpora for low-resource languages typically contain sentence pairs in the tens or hundreds of thousands, we classify Icelandic as a medium-resource language for MT purposes. In this paper, we present on-going experiments with different MT models, both statistical and neural, for translating English to Icelandic based on ParIce. We describe the corpus and the filtering process used for removing noisy segments, the different models used for training, and the preliminary automatic and human evaluation. We find that, while using an aggressive filtering approach, the most recent neural MT system (Transformer) performs best, obtaining the highest BLEU score and the highest fluency and adequacy scores from human evaluation for in-domain translation. Our work could be beneficial to other languages for which a similar amount of parallel data is available.

Keywords: Machine translation · Parallel data · Evaluation

1 Introduction

Most work in Machine Translation (MT) through the years has mainly either focused on high-resource or low-resource language pairs. Usually, a language pair is considered high-resource if a parallel corpus exists consisting of millions of sentence pairs. In contrast, a language pair is considered low-resource if either no parallel corpus exists, or the corpus only consists of a few tens or hundreds of thousands of sentence pairs.

H. P. Jónsson, H. B. Símonarson, V. Snæbjarnarson, S. Steingrímsson—Equal contribution.

P. Sojka et al. (Eds.): TSD 2020, LNAI 12284, pp. 95–103, 2020.
https://doi.org/10.1007/978-3-030-58323-1_10

Neural Machine Translation (NMT), in particular sequence-to-sequence models based on attention mechanisms, e.g. the Transformer [22], has in recent years become the dominant paradigm in high-resource settings, replacing the previously long-standing dominance of Statistical Machine Translation (SMT) [10].

One parallel corpus, ParIce [2], containing about 3.6 million English-Icelandic (*en-is*) sentence pairs, currently exists for Icelandic. Given the size of ParIce, and the fact that we have only been able to use about 1.6 million of its sentence pairs for training (see Sect. 3.2), we currently categorize the *en-is* pair as a medium-resource language pair.

In this paper, we present on-going work of experimenting with different MT systems, both based on SMT and NMT, for translating in the *en* → *is* direction. We describe the ParIce corpus and the filtering process used for removing noisy segments, the different models used for training, and the preliminary evaluation – both with regard to BLEU scores and human evaluation. We find that, while using an aggressive filtering approach, the most recent NMT system, based on the Transformer, performs best in our setting, obtaining a BLEU score of 54.71 (6.11 points higher than the next best performing system, Moses). Furthermore, the Transformer system also obtained the highest fluency and adequacy scores from human evaluation, in the in-domain setting. Our work could be beneficial to other languages for which a similar amount of parallel data is available.

2 Related Work

In the last few years, research has shown that the NMT approach has significantly pushed ahead the state-of-the-art in MT, which before belonged to phrase-based SMT (PBSMT) systems. For example, [3] compared and analysed the output of three PBSMT systems and one NMT system for English → German and found, *inter alia*, that *i)* the overall post-edit effort needed on the output from the NMT system is considerably lower compared to the best PBSMT system; *ii)* that the NMT system outperforms the PBSMT on all sentence lengths; and *iii)* that the NMT output contains less morphological errors, less lexical errors and less word order errors.

Even though NMT has emerged as the dominant MT approach, there have also been reports of poor performance when using NMT under low-resource conditions. Compared to SMT, [11] found that NMT systems have lower quality on out-of-domain texts, sacrificing adequacy (how much of the meaning is transferred between the source and the generated target) for the sake of fluency (a rating of how fluent the generated target language is). They also found that the NMT systems performed worse in low-resource settings, but better in high-resource settings.

[5] discuss the quality of NMT vs. SMT. They argue that "so far it would appear that NMT has not fully reached the quality of SMT", based on automatic and human evaluations for three use cases, and that the results depend on the different domains and on the various language pairs.

In a study, using the medium-resource language pair English-Polish, [8] found that an SMT model achieves a slightly better BLEU score than an NMT model

based on an attention mechanism. On the other hand, human evaluation carried out on a sizeable sample of translations (2,000 pairs) revealed the superiority of the NMT approach, particularly in the aspect of output fluency.

Given the mixed findings in the literature regarding comparison between NMT and PBSMT, especially in low or medium-resource settings, we decided to include SMT in our experiments.

The only previously published MT results regarding Icelandic are [4,9], although Icelandic has been included in massively multilingual settings [6]. The results rely either on rule-based systems or variants of transfer learning. In contrast, our work constitutes the first published MT and NMT results for Icelandic based on direct supervised learning.

3 Corpus and Filtering

In this section, we describe the ParIce corpus and explain which parts of it are used for training/testing as well as the filtering process for removing segments not suited for training.

3.1 ParIce

For training, we used ParIce [2], an *en-is* parallel corpus consisting of roughly 3.6 million translation segments. The corpus data is aligned with hunalign [21] and filtered using a sentence scoring algorithm based on a bilingual lexicon bag-of-words method and a comparison between an MT generated translation of a segment and the original segment.

ParIce is a collection of data from different sources, the largest being a collection of EEA regulatory texts (48%), data from OpenSubtitles (37%), published on OPUS [20] but refiltered in the ParIce corpus, and translation segments from the European Medicines Agency (EMA; 11%) published in the Tilde MODEL corpus [17] (other sources amount to 4% of the data). From each of these three corpora, we sampled roughly 2000 segments to serve as test sets.

3.2 Filtering

Starting from the 3.6 million segments compiled in ParIce, we filtered the corpus before training any models. Among the filters we used, many were adapted from the suggestions of [15]. Most of the filters are proxies for alignment errors, OCR errors, encoding errors and general text quality.

Primarily, the filters and post-editing consist of: *1)* empty sentence filter; *2)* identical or approximately identical source and target sequence, measured by absolute and relative edit distance; *3)* sentence length ratio filter, in characters and tokens; *4)* maximum and minimum sequence length filter, in characters and tokens; *5)* maximum token length; *6)* minimum average token length; *7)* character whitelist; *8)* digit mismatch: both sides should have the same set of

number sequences; *9)* unique sequence pair, after removing whitespace, punctuation, capitalization and normalizing all numbers to 0 (all number sequences are equivalent); *10)* case mismatches where one side is all uppercase and the other not; *11)* corrupt symbols, e.g. weird punctuation like ? and " inside words; *12)* many other ad-hoc regular expressions for Icelandic and dataset specific OCR artifacts and encoding errors (e.g. common words where b replaces, i replaces l, missing accents); *13)* normalizing of quotes, bullets, hyphens and other punctuation; *14)* fixing line splits where a word was split due to text reflow.

When applicable, we use the numbers provided in [15]. Otherwise the filters were tuned to fit Icelandic and ParIce specifically. Roughly half of ParIce was filtered out with this approach, leaving 1.6 million translation segments for training, consisting of around 29 million Icelandic tokens and 32 million English tokens.

4 Models

In this section, we describe the key characteristics of PBSMT and NMT models and the three different systems/models we have experimented with: the SMT system Moses, and two NMT models, the first one based on BiLSTM and the second one on the Transformer. Each model attempts to estimate the probability $p(t|s)$, the probability of a sentence t in the target language given a sentence s in the source language.

4.1 PBSMT

In PBSMT, $p(t|s)$ is not modelled explicitly, rather Bayes' theorem is applied and t is reached via a translation model $p(s|t)$ and a language model $p(t)$ by estimating $\text{argmax}_t\ p(s|t)p(t)$. Furthermore, s and t are segmented into smaller phrases, upon which the translation model is defined. The phrases are extracted and their probabilities estimated during training using the underlying parallel corpus. The language model ensures the fluidity of t and can be derived from the training data and/or from a separate monolingual corpus. For further details see [10].

Moses. We used the standard open source implementation of PBSMT, the Moses system[1]. We created a number of different Moses models in order to deal with the morphological richness of Icelandic. For example, we used a large out-of-domain monolingual corpus and tokenizers including subword tokenizers such as SentencePiece [13] with Byte Pair Encoding (BPE) and Unigram for both *is* and *en*, with a 30k vocabulary for each language. For all models we used the default alignment heuristic, the default distortion model, and a 5-gram KenLM [7] language model trained on additional monolingual data, i.e. 6.5 million sentences from the Icelandic Gigaword Corpus [18]. The best performing model, which uses the Moses tokenizer for both *en* and *is*, is evaluated against the NMT based systems in Sect. 5.

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

4.2 NMT

An NMT system attempts to model $p(t|s)$ directly using a large modular neural network that reads s and outputs t, token by token. Instead of representing the tokens symbolically, like PBSMT systems, the tokens are represented using vectors (embeddings). The typical NMT system is based on sequence-to-sequence learning, and consists of two components: an encoder and a decoder. The system is trained to maximize $p(t|s)$ by updating the parameters of the network using stochastic gradient descent to back-propagate the errors from the output layer to the previous layers. The two dominant NMT architectures over the last few years are based on 1) LSTM, and 2) self-attention networks (Transformer).

BiLSTM. The general LSTM model for NMT is described in [19]. In this model, the encoder is an LSTM that converts an input sequence s to a fixed-sized vector v from which the decoder, another LSTM, generates t. Given the embedded tokens of s, $(x_1, \ldots, x_T)$ and v, the model estimates the conditional probability $p(y_1, \ldots, y_{T'}|x_1, \ldots, x_T)$ as follows:

$$p(y_1, \ldots, y_{T'}|x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t|v, y_1, \ldots, y_{t-1}) \tag{1}$$

where $(y_1, \ldots, y_{T'})$ represents the target sentence t, and where T' may be different from T. In other words, the prediction of each target token depends on the encoded version of the whole input sequence, as well as on the previously predicted target words.

The model is further improved by adding an additional LSTM to the encoder which reads the input in the reverse order, i.e. the encoder is bidirectional. Additionally, during decoding, these networks can be augmented with attention [1,14] where alignments between target and source tokens can be modeled more explicitly. We used the standard BiLSTM implementation from OpenNMT[2], medium and large NMT models with Luong attention [14] (4-layer 256 hidden unit encoder, 4 layer 512 hidden unit decoder; large model has 6 layers and double the number of hidden units). We used a 16k joint BPE vocabulary.

Transformer. The Transformer, proposed by [22], builds on previous models in various ways. Its design provides for much better parallelization, and it leverages GPU architecture more so than LSTMs. In general, it achieves better machine translation performance for the same training time and data as compared to LSTMs.

The Transformer consists of stacked transformer blocks, each of which comprises 2–3 sublayers, self-attention, decoder-to-encoder attention, and a fully connected layer. The block operates independently over a sequence of hidden

[2] https://opennmt.net/.

vectors h_i whereby each vector in the sequence can attend to (i.e. receive information from) all other hidden vectors in the sequence before being transformed by the fully connected sublayer. The decoder block has an added attention sublayer that allows it to attend to the encoder in addition to itself. Finally, the last decoder block has a softmax output layer for token probabilties.

The implementation we use is the reference implementation from [22] of the transformer-base architecture which is part of the Tensor2Tensor package[3]. It has 6 layers each for its encoder and decoder with attention head count of 8. We used shared source and target embeddings. The included subword tokenizer provided by Tensor2Tensor was used to build a 16k joint subword vocabulary.

5 Evaluation

In this section, we present the results of automatic and human evaluation of the individual models, Moses, BiLSTM and Transformer, for translating in the $en \rightarrow is$ direction.

Neither NMT model was fine-tuned before evaluation, and the Transformer used checkpoint-averaging (a gain of about 0.5 BLEU). The batch sizes for the Transformer and the BiLSTM were 1700 (subword) tokens and 32 sequences, respectively. No other hyperparameter tuning was performed due to computational restraints.

5.1 BLEU Scores

We use BLEU for automatic evaluation. It is the most widely used MT quality metric and it has reasonably high correlation with human evaluations. Due to possible biases that may be "unfair" to some technologies [16], the BLEU scores cannot be the primary evidence of the quality of our systems. Therefore, we also rely on human evaluation.

As discussed in Sect. 3.1, the test sets consists of about 2000 segments sampled from three parts of the ParIce corpus: EEA, EMA, and OpenSubtitles. Table 1 shows the results for the three system and the different test sets, as well as the combined sets.

Table 1. BLEU scores for the three systems and the different test sets.

Model	EES	EMA	OpenSubtitles	Combined
BiLSTM	38.68	41.60	23.32	38.12
Moses	49.70	54.93	26.11	48.60
Transformer	**56.31**	**58.37**	**34.71**	**54.71**

[3] https://github.com/tensorflow/tensor2tensor.

Table 2. Fluency and adequacy scores from human evaluation.

Test set	Model	Fluency	Adequacy
In-domain	BiLSTM	2.49	2.01
	Moses	3.64	3.84
	Transformer	**4.30**	**4.33**
	Google	3.80	4.16
Out-of-domain	BiLSTM	1.85	1.30
	Moses	2.54	2.32
	Transformer	3.20	2.86
	Google	**3.40**	**3.80**

In [22] it was shown that the Transformer is the dominant model in high-resource settings. Our results indicate that the Transformer also performs best in medium-resource settings. It is, however, noteworthy that the Moses systems performs significantly better than the BiLSTM model.

5.2 Human Evaluation

We recruited three people with translation experience for adequacy evaluation and three Icelandic linguists for fluency evaluation. We randomly chose 100 sentences from our test set for in-domain evaluation, and 100 sentences from news for out-of-domain evaluation. The sentence lengths varied substantially, averaging 18.2 words per sentence, with a standard deviation of 13.7. Each sentence was translated by our three systems as well as by Google Translate, for reference. We used Keops[4] for the evaluation.

The fluency group was given the following instructions: Is the sentence good fluent Icelandic? Rate the sentence on the following scale from 1 to 5. 1 – incomprehensible; 2 – disfluent Icelandic; 3 – non-native Icelandic; 4 – good Icelandic; 5 – flawless Icelandic. The adequacy group was given the following instructions: Does the output convey the same meaning as the input sentence? Rate the sentence on the following scale from 1 to 5. 1 – none; 2 – little meaning; 3 – much meaning; 4 – most meaning; 5 – all meaning.

We calculated the Intraclass Correlation Coefficient (ICC) for both groups. This resulted in ICC of 0.749, with 95% confidence interval (CI) in the range 0.718–0.777 for the fluency group, and ICC of 0.734 and 95% CI in the range 0.705–0.760 for the adequacy group. According to [12], this suggests that inter-rater agreement is moderate to good for both groups.

We calculated adequacy and fluency on our original scale resulting in the values shown in Table 2. The results show that the Transformer is perceived to give more adequate and more fluent translations than our other two systems, both for out-of-domain translations and in-domain, where it even outperforms Google

[4] https://github.com/paracrawl/keops.

Translate, although that may of course be because our in-domain translations are not in Google Translate's domain. Our SMT system performs decently, not as good as the Transformer or Google Translate, but outperforms the BiLSTM system by far.

6 Conclusion

We have described experiments in using three different architectures (Moses, BiLSTM and Transformer) for translating in the $en \rightarrow is$ direction. Automatic and human evaluation shows that the Transformer architecture performs best, followed by Moses and BiLSTM (in that order).

In future work, we intend to experiment with larger model sizes, backtranslation, and bilingual language model pre-training. Explicit handling of named entities is also a problematic issue, as the available parallel data contains very few Icelandic names.

Acknowledgments. This project was funded by the Language Technology Programme for Icelandic 2019–2023. The programme, which is managed and coordinated by Almannarómur, is funded by the Icelandic Ministry of Education, Science and Culture.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (eds.) Proceedings of the 3rd International Conference on Learning Representations, ICLR, San Diego (2015)
2. Barkarson, S., Steingrímsson, S.: Compiling and filtering ParIce: an English-icelandic parallel corpus. In: Proceedings of the 22nd Nordic Conference on Computational Linguistics, NODALIDA, Turku, Finland (2019)
3. Bentivogli, L., Bisazza, A., Cettolo, M., Federico, M.: Neural versus phrase-based machine translation quality: a case study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, Austin, TX, USA (2016)
4. Brandt, M.D., Loftsson, H., Sigurþórsson, H., Tyers, F.M.: Apertium-IceNLP: a rule-based Icelandic to English machine translation system. In: Proceedings of the 15th Annual Conference of the European Association for Machine Translation, EAMT, Leuven, Belgium (2011)
5. Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., Way, A.: Is neural machine translation the new state of the art? Prague Bull. Math. Linguist. **108**(1), 109–120 (2017)
6. Defauw, A., Vanallemeersch, T., Van Winckel, K., Szoc, S., Van den Bogaert, J.: Being generous with sub-words towards small NMT children. In: Proceedings of the 12th Language Resources and Evaluation Conference, LREC, Marseille, France (2020)
7. Heafield, K., Pouzyrevsky, I., Clark, J.H., Koehn, P.: Scalable modified Kneser-Ney language model estimation. In: Proceeedings of 51st Annual Meeting of the Association for Computational Linguistics, ACL, Sofia, Bulgaria (2013)

8. Jassem, K., Dwojak, T.: Statistical versus neural machine translation - a case study for a medium size domain-specific bilingual corpus. Poznan Stud. Contemp. Linguist. **55**(2), 491–515 (2019)
9. Johnson, M., Firat, O., Aharoni, R.: Massively multilingual neural machine translation. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), NAACL, Minneapolis, MN, USA (2019)
10. Koehn, P.: Statistical Machine Translation. Cambridge University Press, Cambridge (2009)
11. Koehn, P., Knowles, R.: Six challenges for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation, Vancouver, Canada (2017)
12. Koo, T., Li, M.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropractic Med. **15**, 155–163 (2016)
13. Kudo, T., Richardson, J.: SentencePiece: a simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP, Brussels, Belgium (2018)
14. Luong, T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, Lisbon, Portugal (2015)
15. Pinnis, M.: Tilde's parallel corpus filtering methods for WMT 2018. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Brussels, Belgium (2018)
16. Reiter, E.: A structured review of the validity of BLEU. Comput. Linguist. **44**(3), 393–401 (2018)
17. Rozis, R., Skadiņš, R.: Tilde MODEL - multilingual open data for EU languages. In: Proceedings of the 21st Nordic Conference on Computational Linguistics, NODALIDA, Gothenburg, Sweden (2017)
18. Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., Gunason, J.: Risamálheild: a very large icelandic text corpus. In: Proceedings of the 11th International Conference on Language Resources and Evaluation, LREC, Miyazaki, Japan, May 2018
19. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS, Montreal, Canada (2014)
20. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey (2012)
21. Varga, D., Németh, L., Halácsy, P., Kornai, A., Viktor Trón, V.N.: Parallel corpora for medium density languages. In: Proceedings of Recent Advances in Natural Language Processing, RANLP, Borovets, Bulgaria (2005)
22. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008. Curran Associates, Inc. (2017)