The Ubiqus English-Inuktitut System for WMT20

François Hernandez
Ubiqus, Paris, France
fhernandez@ubiqus.com

Vincent Nguyen
Ubiqus, Paris, France
vnguyen@ubiqus.com

Abstract

This paper describes Ubiqus' submission to the WMT20 English-Inuktitut shared news translation task. Our main system, and only submission, is based on a multilingual approach, jointly training a Transformer model on several agglutinative languages. The English-Inuktitut translation task is challenging at every step, from data selection, preparation and tokenization to quality evaluation down the line. Difficulties emerge both because of the peculiarities of the Inuktitut language as well as the low-resource context.

1 Introduction

Ubiqus participated in the English to Inuktitut news translation task of WMT20. We performed a single submission, based on an unconstrained multilingual setup. The approach consists of jointly training a traditional Transformer (Vaswani et al., 2017) model on several agglutinative languages in order to benefit from them for the low-resource English-Inuktitut task (Aharoni et al., 2019).

Though the dataset provided for the task is sizable, with more than a million segments, it's quite narrow domain-wise, as it comes from proceedings of the Nunavut Hansard. The task being translation of news, it's expected to be a much wider domain. For that purpose, we extended the task with datasets of other - linguistically near - languages, as well as in-house datasets introducing more diversity to the domain.

All experiments were performed with the OpenNMT (Klein et al., 2017) toolkit, with *Tokenizer* [1] for data preprocessing and *OpenNMT-py* [2] for model training and inference.

2 Data

2.1 Training corpora

Based on prior internal work on English-Inuktitut translation tasks as well as other low-resource tasks, we focused our experiments on multilingual setups. Inuktitut is an agglutinative language, with a lot of particularities. Some *Uralic languages* like Finnish and Estonian can be considered close to Inuktitut in some linguistic aspects.

Most of our experiments are *unconstrained* with regards to the original WMT task in three ways:

- some datasets are taken from previous WMT tasks (English-Finnish, English-Estonian);
- some datasets are not in the WMT scope (more recent ParaCrawl[3] versions);
- some datasets were built in-house at Ubiqus Labs.

Some Inuktitut resources can easily be found on the internet, mostly from official government of Nunavut websites and initiatives. We performed two sets of data retrieval: a first one based on parallel crawling of multilingual websites, and a second one based on manual retrieval of parallel documents (mostly in PDF format) which then were automatically aligned with a commercial tool. In prior experiments, we also built a set of parallel news articles. Articles were manually retrieved and aligned from both the *Inuktitut* [4] magazine, which provides parallel versions of all its content in English, French, Inuktitut and Inuinnaqtun, and the Nunatsiaq News[5] website, which provides part of its content in both Inuktitut and English. We decided not to include this last dataset because of

[1] https://github.com/OpenNMT/Tokenizer

[2] https://github.com/OpenNMT/OpenNMT-py

[3] https://paracrawl.eu

[4] https://www.itk.ca/category/inuktitut-magazine/

[5] https://nunatsiaq.com

Proceedings of the 5th Conference on Machine Translation (WMT), pages 213–217
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

its high proximity with the *newsdev2020* and *newstest2020* of the task.

A summary of all the datasets used in the experiments is available in Table 1.

2.2 Evaluation sets

During our experiments, we conducted evaluation of the trained models with the provided *newsdev2020-eniu* as well as the *dev*, *devtest* and *test* parts of the Hansard dataset split. The latter were deduplicated prior to evaluation.

As a big part of our experiments revolve around multilingual aspects, we also used *newstest2018-enfi* and *newstest2019-enfi* for English-Finnish, as well as *newstest2018-enet* for English-Estonian.

Finally, we also conducted some evaluation over the test part of our in-house dataset built from *Inuktitut* magazine.

2.3 Data selection and cleaning

Deduplication as well as a few steps of cleaning were applied to every dataset. This consists of removing segments where:

- average token length too short or too long;
- source is strictly equal to target;
- numbers do not match between source and target side;
- source to target character ratio is too extreme.

The difference between raw and selected dataset size is shown in Table 1. It is noticeable that this step is especially important for our in-house datasets, where automatically crawled and aligned data is particularly messy. Also, it seems the Nunavut Hansard dataset is quite clean but contains a lot of duplicates.

2.4 Preprocessing and Tokenization

We decided to work on romanized Inuktitut. This allows straightforward parameter and vocabulary sharing in a basic bilingual English-Inuktitut setup, as well as maximizing the potential benefits of parameter sharing in a multilingual setup. Hence, all the Inuktitut data was romanized prior to any other processing, and we only converted back our *newstest2020-eniu* inferred hypothesis for submission.

All experiments were conducted on data tokenized with a BPE (Sennrich et al., 2016b) model with 12,000 merge operations, learned on the concatenation of all datasets – both source and target – presented in Table 1 (without any particular sampling strategy). This leads to a final vocabulary size of approximately 14k tokens. The choice of a smaller number of BPE merge operations stems from the agglutinative aspect of the language, leading us to think that dividing long tokens into more subwords might be beneficial to learn and share more useful representations. This seems to be also the approach in the baseline system proposed in (Joanis et al., 2020).

3 Experiments

3.1 Mixing languages

The method used to train models on multiple languages relies on the dataset weighting mechanism which is implemented within OpenNMT-py (Klein et al., 2020). When building batches, $weight_A$ examples are sampled from dataset A, then $weight_B$ from dataset B, and so on. This allows to dynamically subsample or oversample any specific dataset or language pair when training.

In order to allow Many-to-Many translation in a single shared model, we need to prepend each source with a tag indicating the target language (Johnson et al., 2017).

3.2 Bilingual only

Since we do not have any internal resource to assess the Inuktitut output, we started some bilingual experiments into English. With the English-Inuktitut datasets only, we realized that even with a base Transformer, the model converged very quickly and gave similar results with several varying hyper parameters. Also, changing the sampling weight of each sub-dataset did not have much impact to the final results. Moreover, English to Inuktitut bilingual experiments gave very poor results on our internal test set based on the Inuktitut Magazine. We hypothesize that there was some kind of overfitting to the Hansard domain. This is why we decided to extend a multilingual set up with more "news" based data.

3.3 Multilingual

We trained a few systems in the following order:

- first, a bilingual (and bidirectional) English-Inuktitut system (base configuration Transformer) using the Nunavut dataset as well as our in-house Web and Documents datasets;

Dataset	Origin	Raw	Selected
Nunavut Hansard v3.0 (Joanis et al., 2020)	WMT EN-IU 2020 Task	2,550,682	737,375
⋆Europarl English-Finnish	WMT EN-FI 2019 Task	1,969,624	1,564,994
⋆Europarl English-Estonian	WMT EN-ET 2018 Task	651,236	566,815
⋆ParaCrawlv6 English-Finnish	ParaCrawl Project	4,286,642	4,207,262
⋆ParaCrawlv6 English-Estonian	ParaCrawl Project	1,785,161	1,755,013
⋆Public Documents	Ubiqus	102,567	66,159
⋆Public Websites	Ubiqus	2,035,594	31,025

Table 1: Characteristics of the datasets used in the experiments. Datasets marked with ⋆ are considered out of the constraints of the WMT English-Inuktitut task.

- next, we added the English-Finnish data;
- then, we added the English-Estonian data;
- finally, we increased the model size.

Results for these systems are summed up in Table 3. We notice that multilingual setups are truly multilingual, in the sense that they provide output in the correct language, even though the scores are not very competitive (approx. 30% below the best scores at the time of the corresponding WMT tasks).

We decided to retain the bigger model (medium Transformer) for the submission. Bigger multilingual models tend to be better with regards to human evaluation, probably because the tasks are better spread across the parameters. This can be a problem in case of overfitting, which does not seem to be the case here as the scores remain in the same range. Also, the bigger model seems to give marginally better results in the additional tasks (Finnish and Estonian), which leads us to think it will be more robust to new test sets.

The configuration used for the final submission is the following:

- **Corpora and weights**: shown in Table 2.
- **Tokenization**: 12,000 BPE merge operations, learned on the concatenation of all datasets.
- **Model**: Transformer Medium (12 $heads$, $d_{model} = 768$, $d_{ff} = 3072$), with Relative Position Representations (Shaw et al., 2018).
- **Training**: Trained with OpenNMT-py on 6 RTX 2080 Ti, using mixed precision. Initial batch size is around 50,000 tokens, final batch size around 200,000 tokens. Training was stopped at 100k steps. Averaging was done continuously through exponential moving average.
- **Inference**: Shown scores are obtained with beam search of size 5 and average length penalty.

Hansard	15
⋆Europarl en-et	2
⋆Europarl en-fi	2
⋆ParaCrawlv6 en-et	10
⋆ParaCrawlv6 en-fi	10
⋆Public Documents (Ubiqus)	5
⋆Public Websites (Ubiqus)	1

Table 2: Dataset weighting used for the submitted system.

4 Future work

Our experiments remain in a rather traditional Neural Machine Translation scope, with the only addition of multiple languages and dataset weighting. Several paths can be explored from this starting point, such as adding more data for the current languages in the setup, authentic or synthetic (e.g. via back-translation (Sennrich et al., 2016a)), or adding other languages that might share some common characteristics, like Hungarian for instance.

Some additional work could also be explored on the tokenization part. For simplicity, our first approach in this paper relies on a very simple shared BPE approach. But, some more sophisticated approaches, maybe language-specific or morphologically adapted (Micher, 2018), may be worth exploring.

System	nd20-eniu	dev	dev-test	test	IM	nt18-enfi	nt19-enfi	nt18-enet
(Joanis et al., 2020)	-	24.2	17.9	19.3	-	-	-	-
en↔iu (base)	15.6	23.9	17.7	19.4	4.7	-	-	-
en↔iu/fi (base)	15.6	23.6	17.5	19.2	7.6	11.8	16.3	2.4
en↔iu/fi/et (base)	15.5	23.3	17.4	18.9	7.6	11.9	16.6	16.9
▷ en↔iu/fi/et (medium)	15.6	23.6	17.3	19.1	7.4	12.1	17.0	17.1

Table 3: BLEU (Papineni et al., 2002) scores for our various experiments, obtained with SacreBLEU (Post, 2018) v1.3.7. The submitted system is marked with ▷. *dev*, *dev-test* and *test* refer to the Hansard dataset evaluation sets. IM stands for *Inuktitut Magazine*.)

Finally, some more novel approaches could be tried, like massive pre-training methods such as BART (Lewis et al., 2019). A similar experimental process could be followed, starting from only the core languages of the task (English and Inuktitut), then extending to other languages and observe the impact.

5 Conclusion

Working on a new, unknown, language is always challenging. Even more so when this language is quite distant from any language you're used to. Also, automated metrics are far from being perfect for such tasks, especially in the context of such a particular language as Inuktitut.

Particularly for this task, human evaluation is key. But, as data, it's quite a scarce resource for Inuktitut. More knowledge of the language would be of tremendous help to better grasp the limits or interesting leads of the various models. One workaround can be to work on the opposite direction (Inuktitut to English), but there is no guarantee the model would have similar behaviour for similar tricks. And, some knowledge about Inuktitut would still be needed to analyze model behavior based on source inputs.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of LREC-2020*, Marseille, France.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The OpenNMT neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109, Virtual. Association for Machine Translation in the Americas.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Jeffrey Micher. 2018. Using the Nunavut hansard data for experiments in morphological analysis and machine translation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 65–72, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.