University of Ljubljana
Faculty of Computer and
Information Science

FRI Natural language processing course 2021
Project report

Automatic language translation

Erika Stanković and Eva Boneš

Abstract

task selection simple corpus processing/analysis Introduction, existing solutions, initial ideas

Keywords
Keyword1, Keyword2, Keyword3 ...

Advisors: Slavko Žitnik

Introduction

Machine translation (MT) is the problem of automatically translating text from one language to another. It has been researched since the 1950s, but only recently, with the rise of deep learning, did it prove to be solvable, although the possibility of achieving fully automatic machine translations of high quality is still being questioned.

There are three main approaches to solving the problem, all with their advantages and shortcomings. Rule-based machine translation (RBMT) is the oldest of the bunch and it requires expert knowledge about both the source and the target language to develop syntactic, semantic, and morphological rules. In contrast to the other two approaches, RBMT does not need bilingual data and when a set of rules is set for a particular language it can be reused for pairings with different languages. Its main disadvantage is that it requires a lot of expertise to manually set rules and good dictionaries which can take a lot of time and might border on impossible in some languages.

Another approach, which gained popularity in the 1990s uses statistical models based on the analysis of bilingual text corpora. The idea behind statistical machine translation (SMT) as proposed in [1] is if given a sentence T in the target language, we seek the sentence S from which the translator produced T – we want to choose S so as to maximize $Pr(S|T)$. Comparing to RBMT, this approach needs less manual work from linguistic experts and we can achieve more fluent translations (given we have a good language model). STM models are also generally built so that one can be used on more language pairs. But we do need a bilingual corpus which can be hard to find, depending on the language pair we are interested in. It is also less suitable for language pairs with big differences in word order.

As with many computer science fields, current state-of-the-art approaches are based on neural networks. Contrary to building a pipeline of separate tasks with RBMT and STM, we only need to build one network with neural machine translation (NMT). The biggest challenge in building a successful Slovene to English automatic translator is obtaining a big enough bilingual corpus. As with all deep learning approaches, having a large and quality dataset is crucial for the success of the model. To deal with this exact problem, a lot of approaches of pre-training a network on monolingual data (that can be much easier obtained) have been proposed.

Bidirectional Encoder Representations from Transformers (BERT) [2] which was proposed by researchers at Google, uses two strategies to deal with the problem, namely masked language modeling (MLM) and next sentence prediction (NSP). With MLM before feeding sequences into the model, 15% of the words are replaced with a [MASK] token and the model then attempts to reproduce the original value of sequences. This way, along with a bigger dataset we can use this way, the model also achieves more context awareness and when fine-tuned produces more fluent translations. In NSP, the model receives pairs of sentences as input and learns to predict if the second sentence is a subsequent sentence of the first one. Similar to MLM, the model achieves more context-awareness.

In 2020, mRASP [3] that could be very useful in our case has been proposed. The authors built a pretrained NMT model that could be fine-tuned for any language pair. They used 197M pairs of sentences, which is much more than we could obtain for only English-Slovene translations.

Frameworks

fairseq [4]

fairseq is a sequence modeling toolkit written in PyTorch for training models for translation, summarization, and other natural language processing tasks. It provides different neural network architectures, namely convolutional neural networks (CNN), Long-Short-Term Memory (LSTM) networks, and Transformer (self-attention) networks. The architectures can be configured to our own needs and many implementations for different tasks have been proposed since fairseqs introduction in 2019. Alongside different architectures, they also provide pre-trained models and pre-processed test sets for different tasks, but none of them in Slovene, so we won't be able to use them.

Marian NMT [5]

Marian is an efficient Neural Machine Translation framework written in pure C++. It is developed by Microsoft. It provides fast multi-GPU training and batched translation on GPU/CPU. It allows training on raw texts using built-in SentencePiece [6] (an unsupervised text tokenizer and detokenizer).

It provides multiple different models: deep RNNs with Deep Transition Cells [7], transformer models [8], multi-source models [9], RNN and transformer-based language models.

The code is publicly available on https://marian-nmt.github.io/.

T5 [10]

T5 (Text-To-Text Transfer Transformer) is a framework, developed by Google, that trains a single model for a variety of text tasks (translation, summarization, ...). Its performance is comparable to task-specific architectures and was at that time considered state-of-the-art when combined with scale. It uses an encoder-decoder model and it was developed using Tensorflow. For tokenizing, it also uses the SentencePiece model [6].

The code is publicly available on https://github.com/google-research/text-to-text-transfer-transformer.

The framework includes some pre-trained models, but because of the vocabulary they used, it can only process a predetermined, fixed set of languages. This means that for our task of translating from and to Slovenian, we would need to train the model from scratch.

The framework requires a significant amount of preprocessing, because it requires a specific format of input data: ¡input¿
t¡target¿.

XML-RoBERTa [11]

XML-RoBERTa is a Transformer-based masked language model, trained on 2.5TB of filtered CommonCrawl data in 100 languages. It significantly outperforms multilingual BERT (mBERT) in various NLP tasks, especially on low-resource languages. It can even compete with other monolingual models on the GLUE and XNLI benchmarks, proving that multilingual capacity does not sacrifice per-language performance.

Methods

Based on our current analysis, we think the best framework for our task is XML-RoBERTa.

BLEU score

Bilingual evaluation understudy (BLEU) [12] score is a measure, used for comparing a candidate translation of the text to one or more reference translations. Even though it was proposed in 2002, it is still the most commonly used score for comparing different approaches for automatic translation.

The algorithm works by counting matching n-grams between the candidate and reference translations where matches are positionally-independent. The more the matches, the higher the BLEU score. The theoretical maximum value is 1.0, but this is not possible in practice, since the candidate and reference text would have to be identical, and not even human translators can achieve that.

For reference – on a test corpus of 500 sentences, a human translator scored 0.3468. Current state-of-the-art model Admin [13][14] achieves 0.464 on an English-French Europarl dataset [15].

Datasets

OpenSubtitles 2016 [16]

OpenSubtitles 2016 is a corpus of parallel texts in 65 languages, created from translated movie subtitles, found on www.opensubtitles.org. It provides tokenized and untokenized corpus files, as well as sentence alignments in XCES format.

The English corpus contains 2.5G tokens and 337.8M sentences and the Slovene corpus contains 321.4M tokens and 53.3M sentences. Of those, the Slovenian-English parallel texts contain 16.3M sentence pairs and 188.18M words, while the English-Slovenian parallel texts contain 13.7M sentence pairs and 177.06M words.

OpenSubtitles 2018 [17]

OpenSubtitles 2018 is a very similar corpus to OpenSubtitles 2016, but cleaner, uses improved sentence alignment and better language checking. It includes tokenized and untokenized corpus files, as well as sentence alignments in XCES format, in 62 different languages.

360.7M tokens and 59.6M sentences. Of those, the Slovenian-English parallel texts contain 19.6M sentence pairs and 227.48M words, while the English-Slovenian corpus contains 16.4M sentence pairs and 213.00M words.

This corpus also contains intra-lingual sentence alignments between alternative subtitles in the same language.

Europarl [15]

Europarl is a corpus of parallel texts in 21 languages extracted from the proceedings of the European Parliament first collected in 2005. It has since become a benchmark dataset for testing the performances of the models and has been used in many NLP competitions.

The Slovene-English corpus contains 623490 sentences – 12525644 words in Slovene and 15021497 in English.

EMEA [18]

EMEA is a corpus of parallel texts in 22 languages made out of PDF documents from the European Medicines Agency. It provides tokenized and untokenized corpus files in each language with sentence alignments between them, which means that we have translations amongst different European languages which could prove to be useful in future research.

The corpus contains 338195 sentences and 9830000 words.

ELRC [16]

ELRC is a corpus of parallel texts, collected through the European Language Resource Coordination (https://www.elrc-share.eu/). It includes general-domain corpora, as well as domain-specific language resources in the fields of consumer rights, culture, legal domain, social security, health, public procurement, etc. It provides tokenized and untokenized corpus files, parsed corpus files and sentence alignments in XCES format.

The Slovenian-English corpus contains 0.65M tokens and 23.81K sentences.

Results

Discussion

Acknowledgments

References

[1] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Comput. Linguist.*, 16(2):79–85, June 1990.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[3] Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. Pre-training multilingual neural machine translation by leveraging alignment information, 2021.

[4] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[5] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, Melbourne, Australia, 2018.

[6] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226, 2018.

[7] Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. Deep architectures for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 99–107, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 06 2017.

[9] Marcin Junczys-Dowmunt and Roman Grundkiewicz. An exploration of neural sequence-to-sequence architectures for automatic post-editing. 06 2017.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. Bleu: a method for automatic evaluation of machine translation. 10 2002.

[13] Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, 2020.

[14] Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. Very deep transformers for neural machine translation. In *arXiv:2008.07772 [cs]*, 2020.

[15] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. 5, 11 2004.

[16] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[17] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Confer-*

ence on Language Resources and Evaluation (LREC'16), pages 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[18] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).