A Neural Translator Designed to Protect the Eastern Border of the European Union

Nowakowski Artur artur.nowakowski@amu.edu.pl
Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznan, 61-614, Poland

Jassem Krzysztof jassem@amu.edu.pl
Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznan, 61-614, Poland

Abstract

This paper reports on a translation engine designed for the needs of the Polish State Border Guard. The engine is a component of the AI Searcher system, whose aim is to search for Internet texts, written in Polish, Russian, Ukrainian or Belarusian, which may lead to criminal acts at the eastern border of the European Union. The system is intended for Polish users, and the translation engine should serve to assist understanding of non-Polish documents. The engine was trained on general-domain texts. The adaptation for the criminal domain consisted in the appropriate translation of criminal terms and proper names, such as forenames, surnames and geographical objects. The translation process needs to take into account the rich inflection found in all of the languages of interest. To this end, a method based on constrained decoding that incorporates an inflected lexicon into a neural translation process was applied in the engine.

1 Introduction

The Internet, even in its legal form, may be a source of criminal information. Government bodies all over the world search through Internet sites for potentially criminal texts, to prevent certain acts to which such texts may give rise. For example, the Polish State Border Guard, whose function is to protect the eastern border of the European Union, tracks texts that may concern criminal activities such as general smuggling, trafficking of drugs, medicines, alcohol and cigarettes, people trafficking, human organs trafficking, weapons and explosives, sex crime, document fraud, and trafficking of stolen cars and machines. Two factors make this task difficult for employees of the State Border Guard. Firstly, the texts of interest are sparse and not easy to detect. The problem of the detection of such texts is tackled in Nowakowski and Jassem (2021a). Secondly, criminal texts may appear in foreign languages, not known to a particular employee. In such cases a machine translation engine may be of significant help to the user.

This paper describes a neural translator designed for the needs of the Polish State Border Guard. The translator is a component of a system designed to search for and store criminal content. The system is being developed within a research project entitled "Advanced Internet analysis supporting the detection of criminal groups"[1] (the project's short name is AI Searcher). The architecture of the AI Searcher system is described in section 2. Section 3 reports on the data that was used for the training of language pairs applied in the system. Section 4 describes how the translation engine was adapted to the domain of criminal texts. Details on the lexicalized

[1]The project is financed by the Polish National Center for Research and Development.

translation methods applied in the adaptation are presented in section 5. Section 6 gives a few examples that show the difference between adapted and unadapted translation. We conclude the paper with some insights relevant to future work.

2 The AI Searcher project

The AI Searcher project was launched in December 2018. This three-year program has the aim of developing a system to support the protection of the eastern border of the European Union by searching the Internet for criminal texts that may be of interest to employees of the Polish State Border Guard. The user scenario is the following: The employee of the State Border Guard types an inquiry into an edit window. The Query Expansion Module expands the inquiry to a set of queries that are semantically related to the inquiry. The Translation Module translates the set of queries into Russian, Ukrainian, and Belarusian. The Crawler searches the Internet to find texts in Polish, Russian, Ukrainian, and Belarusian related to the queries. The Translation Module translates the foreign texts back to Polish. Finally, the Classifier analyzes the texts to return those with potentially criminal content.

3 Training data

The translator engines designed for the system are trained on the OPUS resources.[2] The sets for training, validation and testing are based on the Tatoeba Challenge[3] (Tiedemann 2020). Statistics on the bilingual corpora used in the project are given in Table 1.

Table 1: Bilingual corpora statistics

Corpus set	Polish–Russian	Polish–Ukrainian	Polish–Belarusian
training set	ca. 19.17m	ca. 1.68m	72,276
validation set	1,000	6,900	287
test set	3,543	2,500	287

The number of sentences for the Polish–Belarusian pair was too low to generate comprehensive translation. A multilingual (Polish–Russian–Ukrainian–Belarusian) model was designed to improve the Polish–Belarusian translation. Its statistics are given in Table 2.

Table 2: Multilingual corpus statistics

Corpus set	Russian–Belarusian	Russian–Ukrainian	Ukrainian–Belarusian
training set	72,870	ca. 1.52m	66,687
validation set	2,743	6,815	1,000
test set	2,500	10,000	2,355

Table 3 shows the BLEU scores of the AI Searcher Translator compared with Google Translate, calculated on the Tatoeba test set.

4 Translation of terminology and personal names

The State Border Guard expects that the translation engine should correctly translate proper names, such as surnames, forenames, geographical locations and objects, brands of cigarettes and alcohol, etc. The lists of such names were created semi-automatically: the names underwent

[2] https://opus.nlpl.eu/

[3] https://github.com/Helsinki-NLP/Tatoeba-Challenge

Table 3: Comparison of BLEU scores

Corpus set	pl -> ru	ru -> pl	pl -> uk	uk -> pl	pl -> be	be -> pl
AI Searcher	47.69	43.06	41.25	43.67	24.75	37.92
Google Translate	42.95	43.05	34.84	38.42	35.39	44.19
difference	+4.74	+0.01	+6.41	+5.25	-10.64	-6.27

automatic transliteration between the Cyrillic and Latin alphabets, and the most frequent names were carefully verified by native speakers. It is worth noting that all verified forenames and surnames were listed and checked together with their inflected forms (there exist 6–7 grammatical cases in all of these languages).

Forenames and surnames in their base Latin form were provided to us by employees of the State Border Guard, names of geographical objects were collected from the available OpenStreetMap resources, and criminal terminology, including brands of cigarettes, cars and alcohol, was gathered from various websites and forums.

Table 4 shows the numbers of base forms for verified proper names.

Table 4: Statistics of proper names

Proper Names	Polish–Russian	Polish–Ukrainian	Polish–Belarusian
male forenames	1,882	1,902	3,477
male surnames	16,142	29,628	17,421
female forenames	2,117	1,962	3,302
female surnames	19,898	26,114	20,170
geographical objects	5,092	7,613	9,460

The adaptation of the translation engine also took into account a lexicon of criminal terms, which consisted of 1203 entries in each of the language pairs.

5 Lexical constraints

The incorporation of lexicon in neural machine translation has been studied thoroughly in recent years (Arthur et al. 2016, Anderson et al. 2017, Hokamp and Liu 2017, Dinu et al. 2019, Song et al. 2019, Exel et al. 2020). The methodology used in the described experiments was based on a constrained decoding and "code-switching" approach. Our approach was focused on constrained decoding, which uses the Grid Beam Search algorithm introduced by Hokamp and Liu (2017) and extended by Post and Vilar (2018) and Hu et al. (2019). We designed an algorithm based on constrained decoding in order to take into account inflected forms of proper names. To evaluate the performance of the algorithm, we carried out experiments in two different scenarios: general and domain-specific. We compared our method with baseline translation, i.e. translation without lexical constraints, in terms of translation speed and translation quality. The lexicalized method resulted in a decrease in translation quality in the general scenario, which shows that augmenting general-domain texts with a specialized lexicon may impair the performance of a neural translator. In the domain-specific scenario the translation showed significant progress, with an increase of over 3 BLEU points. The cost of the algorithm is a decrease in the translation speed. The details of the experiment are reported in Nowakowski and Jassem (2021b). There, several manual metrics for the evaluation of terminology translation were introduced: Placement Rate, Duplication Rate and Inflection Rate. The metrics evaluated the proportions of output sentences in which the target lexicon terms were, respectively, properly placed, not duplicated unnecessarily and correctly inflected. The manual evaluation results showed that our

method ensures terminological adequacy and consistency as well as linguistic correctness when translating into a morphologically rich language in domain-specific scenarios.

6 Examples of lexicalized translation

Tables 5 and 6 show examples of sentences translated with the unadapted and adapted translation engine into Russian and Ukrainian, respectively. The lexicon entries consist of a term in the source language with the equivalent in the target language along with a comma-separated list of its inflectional forms. For each sentence, a manual English translation is given for clarity.

Table 5: Examples of lexicalized translation into Russian

Lexicon entry	Georgy -> Георгий, Георгия, Георгию, Георгием, Георгии
Source sentence	**Georgy** Kuzmin przewozi fajki przez wschodnią granicę.
English translation	Georgy Kuzmin transports cigarettes across the eastern border.
Unadapted MT	Джорджи Кузьмин перевозит сигареты через восточную границу.
Adapted MT	**Георгий** Кузьмин перевозит сигареты через восточную границу.
Lexicon entry	szwarcować -> перебрасывать, перебрасываю, перебрасываешь
Source sentence	Zaczynamy **szwarcować** zioło klientom.
English translation	We are beginning to smuggle the weed to our customers.
Unadapted MT	Мы начинаем портить травы для клиентов.
Adapted MT	Мы начинаем **перебрасывать** траву клиентам.

Table 6: Examples of lexicalized translation into Ukrainian

Lexicon entries	Karpiuk -> Карпюк, Карпюка, Карпюкові, Карпюком hordenina -> горденін горденин гордеїн
Source sentence	Przyniesiemy **hordeninę** do **Karpiuka**.
English translation	We'll bring hordenine to Karpiuk.
Unadapted MT	Ми привеземо гордон до Карпіока.
Adapted MT	Ми принесемо **горденін** до **Карпюка**.
Lexicon entry	przećpać -> накачатись, накачатися, накачати, накачаться
Source sentence	Chcesz okazyjnie **przećpać** w promocyjnej cenie?
English translation	Do you want to get high at a discounted price?
Unadapted MT	Ви хочете побути в промоційній ціні?
Adapted MT	Ви хочете **накачатися** на промоційній ціні?

7 Conclusions

In this case study, a translation engine is part of a system that searches for criminal content in Internet documents written in the Polish, Russian, Ukrainian and Belarusian languages. The adaptation of the translation to the domain of criminal texts consists in the incorporation of lexicon into the neural machine translation engine. The criminal terminology is expected to be translated according to lexical constraints, and the lexical entries should be correctly inflected. An algorithm based on constrained decoding was designed to achieve this goal.

The project described here is ending in December 2021. Work in the near future will focus on further improving Belarusian translation and on increasing efficiency.

References

Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2017). Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–945, Copenhagen, Denmark. Association for Computational Linguistics.

Arthur, P., Neubig, G., and Nakamura, S. (2016). Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

Dinu, G., Mathur, P., Federico, M., and Al-Onaizan, Y. (2019). Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Exel, M., Buschbeck, B., Brandt, L., and Doneva, S. (2020). Terminology-constrained neural machine translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal. European Association for Machine Translation.

Hokamp, C. and Liu, Q. (2017). Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Hu, J. E., Khayrallah, H., Culkin, R., Xia, P., Chen, T., Post, M., and Van Durme, B. (2019). Improved lexically constrained decoding for translation and monolingual rewriting. 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 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

Nowakowski, A. and Jassem, K. (2021a). Detection of criminal texts for the Polish state border guard. In *MIS2-KDD 2021 : The Second International MIS2 Workshop: Misinformation and Misbehavior Mining on the Web*. Association for Computing Machinery. to appear.

Nowakowski, A. and Jassem, K. (2021b). Neural machine translation with inflected lexicon. In *Proceedings of Machine Translation Summit XVIII: Research Track*. Association for Computational Linguistics. to appear.

Post, M. and Vilar, D. (2018). Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Song, K., Zhang, Y., Yu, H., Luo, W., Wang, K., and Zhang, M. (2019). Code-switching for enhancing NMT with pre-specified 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 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

Tiedemann, J. (2020). The Tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182. Association for Computational Linguistics.