Creating a Corpus for Russian Data-to-Text Generation Using Neural Machine Translation and Post-Editing

Anastasia Shimorina
Lorraine University / LORIA
Nancy, France

Elena Khasanova
Lorraine University
Nancy, France

Claire Gardent
CNRS / LORIA
Nancy, France

{anastasia.shimorina,claire.gardent}@loria.fr
yelena.khas@gmail.com

Abstract

In this paper, we propose an approach for semi-automatically creating a data-to-text (D2T) corpus for Russian that can be used to learn a D2T natural language generation model. An error analysis of the output of an English-to-Russian neural machine translation system shows that 80% of the automatically translated sentences contain an error and that 53% of all translation errors bear on named entities (NE). We therefore focus on named entities and introduce two post-editing techniques for correcting wrongly translated NEs.

1 Introduction

Data-to-text (D2T) generation is a key task in Natural Language Generation (NLG) which focuses on transforming data into text and permits verbalising the data contained in data- or knowledge bases. However, creating the training data necessary to learn a D2T generation model is a major bottleneck as *(i)* naturally occurring parallel data-to-text data does not commonly exist and *(ii)* manually creating such data is highly complex. Moreover, the few parallel corpora that exist for D2T generation have been developed mainly for English. Methods that support the automatic creation of multi-lingual D2T corpora from these existing datasets would therefore be highly valuable.

In this paper, we introduce a semi-automatic method for deriving a parallel data-to-text corpus for Russian from the D2T WebNLG corpus whose texts are in English. Our method includes three main steps. First, we use neural machine translation (NMT) model to translate WebNLG English texts into Russian. Second, we perform a detailed error analysis on the output of the NMT model. Third, we exploit two techniques for automatically post-editing the automatic translations. As 53% of the translation errors bear on named entities, we focus on these in the present paper and leave other error types for further research.

The new corpus, error classification and scripts are available at https://gitlab.com/shimorina/bsnlp-2019.

2 Related work

Our work is related to the domain of automatic post-editing (APE) of machine translation (MT) outputs. The task of APE consists in automatically correcting "black-box" MT output by learning from human corrections. Several WMT APE shared tasks were held focusing on English-German, German-English, and English-Spanish language pairs.[1]

Recent neural approaches to APE include, *inter alia*, multi-source training with original sentences and MT outputs (Junczys-Dowmunt and Grundkiewicz, 2018), encoding corrections by a sequence of post-edit operations (Libovický et al., 2016), as well as standard encoder-decoder architectures (Pal et al., 2016).

Submissions participating in the APE shared tasks extensively use large synthetic corpora (Negri et al., 2018). Despite that fact, a "*do-nothing*" baseline when MT outputs are kept unchanged is hard to beat according to the last year's results of the APE shared task (Chatterjee et al., 2018).

3 The WebNLG D2T Dataset

The WebNLG data-to-text corpus (Gardent et al., 2017) aligns knowledge graphs with textual descriptions verbalising the content of those graphs. The knowledge graphs are extracted from DBpedia (Lehmann et al., 2015) and consist of

[1] For this year round of the shared task, a new English-Russian language pair was added: http://www.statmt.org/wmt19/ape-task.html. We did not make use of the data, since our research started before this recent announcement.

Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 44–49,
Florence, Italy, 2 August 2019. ©2019 Association for Computational Linguistics

RDF triples	<Asterix, creator, René Goscinny> <René Goscinny, nationality, French people>
Original	Rene Goscinny is a French national and also the creator of the comics character Asterix.
MT	Рене Госкино - французский гражданин, а также создатель комического персонажа Астерикс. Rene Goskino French national and also creator comic$_{gen}$ character$_{gen}$ Asterix$_{inan}$
PE	Рене Госинни - французский гражданин, а также создатель персонажа комиксов Астерикса. Rene Goscinny French national and also creator character$_{gen}$ comics$_{gen}$ Asterix$_{anim}$
Errors	named entity, vocabulary, grammar
Links	<René Goscinny, sameAs, Рене Госинни> <French people, sameAs, Французы>

Table 1: WebNLG original instance in the ComicsCharacter category, its Russian translation (MT), and post-edited translation (PE) along with error annotation. Errors are highlighted in blue. Links are RDF triples of the form <*English entity, sameAs, Russian entity*>. However, such links are not available for all entities in DBpedia.

sets of (one to seven) RDF triples of the form <*subject, property, object*>. Textual descriptions are in English, and due to the nature of the knowledge graphs, they have an abundance of named entities. The first two lines of Table 1 show an example of a WebNLG instance.

WebNLG provides textual descriptions for entities in fifteen DBpedia categories (Airport, Artist, Astronaut, Athlete, Building, CelestialBody, City, ComicsCharacter, Food, MeanOfTransportation, Monument, Politician, SportsTeam, University, WrittenWork). The corpus possesses a hierarchical structure: if a set consisting of more than one triple is verbalised, then verbalisations of every single triple are to be found in the corpus. Given the example in Table 1, the pairs {<*Asterix, creator, René Goscinny*>: *René Goscinny created Asterix*} and {<*René Goscinny, nationality, French people*>: *René Goscinny is French*} are also present in the WebNLG data. That structure allows propagating post-edits made in texts describing one triple to those verbalising triple sets of larger sizes.

4 Creating a Russian Version of the WebNLG Dataset

4.1 Neural Machine Translation

Following Castro Ferreira et al. (2018), who created a silver-standard German version of WebNLG, we translated the WebNLG English texts into Russian using the English-Russian NMT system developed by the University of Edinburgh for the WMT17 translation shared task (Sennrich et al., 2017).[2] This system ranks first for the English-Russian News translation task both in automatic metrics[3] and human assessment (Bojar et al., 2017). It is learned using Nematus, an encoder-decoder with attention, based on subword units (byte pair encoding). Since the Edinburgh model was trained on sentence-to-sentence data, we split WebNLG texts into sentences using the WebSplit sentence annotation (Narayan et al., 2017), input each sentence to the NMT system, and then concatenated translations to reconstruct the target texts.

4.2 Manual Post-Editing and Error Analysis

To determine the most common translation errors, we start by manually annotating error types in sentences verbalising one triple.

Error Classification The manual post-editing was done by two experts, native Russian speakers, on a part of the corpus for the categories Astronaut, ComicsCharacter, Monument, University for texts verbalising one triple only. Out of 1,076 machine translation outputs analysed, 856 texts (80%) were post-edited. The experts also classified errors that they identified in a translated text.

To define an error classification, we drew inspiration from various error typologies that were developed in the MT community and applied to different languages. See, for instance, Popović (2018) who provides an overview of different approaches to error classification. We also got some ideas from studies focused on errors made by language learners and non-experienced translators in the Russian-English and English-Russian translation directions (Kunilovskaya, 2013; Rakhilina et al., 2016; Komalova, 2017). That allowed us to extend the classification with some phenomena typical for Russian. Lastly, the classification

[2] `http://data.statmt.org/wmt17_systems/` Specifically, we use their ensemble model consisting of four left-to-right models.

[3] `http://matrix.statmt.org/matrix/systems_list/1875`

Category	Subcategory
Grammar	Case marking
	Copula
	Verbal aspect
	Preposition
	Possessive
	Part-of-speech
	Agreement
	Voice, intentionality
Vocabulary	Ambiguity
	Collocation
	Incorrect translation
Structure	Word Order
	Deletion
	Insertion
Named entity	
Punctuation	

Table 2: Main categories and subcategories of error classification.

was augmented with the notorious errors of the NMT systems: word repetitions, deletions, insertions (partly due to the subword-based nature of the applied NMT), untranslated common words, etc. Main error classes identified for the final classification are shown in Table 2. Named entities were treated as a separate category to highlight problems while applying the NMT system on WebNLG. If a text contained more than one mistake in a particular category, then each mistake was tagged as an error. If a spotted mistake concerned an NE, annotators were allowed to add other categories to specify the error.

Category	Proportion	Agreement
Grammar	17%	0.44
Vocabulary	14%	0.52
Structure	11%	0.32
Named entity	53%	0.67
Punctuation	4%	0.0

Table 3: Proportion of main error types in the manually post-edited data and Cohen's κ scores on the held-out category Athlete.

Error Analysis Table 3 shows the error type distribution in the post-edited texts. Named entities is the largest source of errors with 53% of all corrections. Grammatical and lexical mistakes constitute 17% and 14% of the identified errors respectively, while "Structure" (11%) ranks fourth. In fact, the majority of structural mistakes were spotted in named entities. For example, *the Baku Turkish Martyrs' Memorial* was translated as «Мемориал» «Мемориал» в Баку ('Memorial Memorial in Baku') with the following errors identified: named entity, deletion, deletion, insertion.

The most common errors found in NE translations are:

- copying verbatim English entities into Russian translations (person names, locations);
- wrong transliteration, whereas a standard transliteration exists in Russian. E.g., *Lancashire* translated as Ланкассир ('Lancassir') instead of Ланкашир;
- misinterpretation of a named entity as a common noun. E.g., *Dane Whitman* translated as датчанин Уитмен ('inhabitant of Denmark Whitman') instead of Дейн Уитмен.

It should be noted that since the Edinburgh NMT system used subword units, there were also errors with copying named entities, e.g., *Visvesvaraya Technological University* became *Visvesvaraya Technical University*. In a similar vein, in the example from Table 1, the surname *Goscinny* was misinterpreted as the acronym *Goskino* meaning 'State Committee for Cinematography'.

Inter-annotator Agreement Erroneous words in translations can be attributed to several possible error types. To evaluate consistency between annotators and the appropriateness of the developed error classification, we calculated inter-annotator agreement (Cohen, 1960) on the 86 texts from the DBpedia category Athlete, to which annotators were not exposed before. Table 3 shows the kappa scores. The highest score (0.67) was reached for "Named entity", which corresponds to the substantial agreement. The main source of disagreement for named entities was a decision to perform transliteration or not, e.g., sport club names as *Tennessee Titans* can be kept 'as is' in a Russian text or can be put into Cyrillic. For other categories, agreements range from moderate to fair; as for "Punctuation", the agreement is zero due to the data sparseness in this category (there were two errors only identified by one annotator).

Overall, results show *(i)* consistency in correcting named entities, as well as *(ii)* the importance to perform more annotator training and/or establish clearer guidelines, especially for the "Structure" category.

5 Automatic Post-Editing

To improve the automatic translations, we experiment with two methods: a rule-based method based on the errors found during manual annotation and a neural approach.

5.1 Rule Based Post-Editing

Based on the manual corrections applied to the 1-triple data (WebNLG instances where the input graph consists of a single triple), we extract post-edit rules by building upon the operations used to compute the edit distance (Levenshtein, 1966). For example, given the neural translation (1a) and the manually edited correction (1b), the sequence of edit operations applied to compute the Levenshtein edit distance is (1c), i.e. replace ‘Альба’ by ‘Алба-Юлия’, delete ‘Юлия’, keep ‘–’, keep ‘город’, keep ‘в’, keep ‘Румынии’.

(1) a. ‘Альба Юлия – город в Румынии’
b. ‘Алба-Юлия – город в Румынии’
c. SUB DEL KEEP KEEP KEEP KEEP
d. ‘Alba Julia is a city in Romania’

Based on these edit sequences, we extracted sequences of substitution, deletion, and insertion rules along with the corresponding tokens (e.g., Альба Юлия → Алба-Юлия). We then checked these rules manually and excluded false positives. Lastly, we applied the validated rules to the automatic translations.

That method enabled us to increase the amount of post-edited data: after that procedure the total number of post-edited translations sums up to 4,188 (cf. Table 4).

	1 triple	2-7 triples	All triples
PE	856	3,332	4,188
Total	1,076	4,109	5,185

Table 4: Corpus statistics: number of post-edited (PE) texts. Total corresponds to both PE and non-PE texts.

5.2 Automatic Post-Editing Model

To see to which extent corrections can be learned automatically, we built a corpus of (MT, RPE) pairs where MT is an automatic translation and RPE is its correction using the rule-based system described in the preceding section and trained an APE model on it.

The baseline system is a “*do-nothing*” baseline where MT outputs are left unmodified. In our case, that baseline gives 82.4 BLEU between MT and RPE on the test set, which sets quite high standards for learning a new APE model.

The train/dev/test partition was 80/10/10. We used the OpenNMT-tf framework (Klein et al., 2017)[4] to train a bidirectional encoder-decoder model with attention (Luong et al., 2015). A single-layer LSTM (Hochreiter and Schmidhuber, 1997) is used for both encoder and decoder. We trained using full vocabulary and the maximal length in the source and target; all the hyperparameters were tuned on the development set. The APE model was trained with a mini-batch size of 32, a word embedding size of 512, and a hidden unit size of 512. It was optimised with Adam with a starting learning rate of 0.0005. We used early stopping based on BLEU on the development set, as a result of that, the model was trained for 23 epochs. Decoding was done using beam search with a beam size of 5. As an evaluation metric, we used BLEU-4 (Papineni et al., 2002) calculated between our model predictions and RPE. BLEU and statistical significance were calculated on tokenised texts using COMPARE-MT tool (Neubig et al., 2019), which, in turn, uses the NLTK implementation of BLEU. Results are shown in Table 5.

The APE model performance reached parity with the baseline on dev and test data. The difference between scores was not statistically significant via the bootstrap resampling (1000 samples, $p < 0.05$). On the training data, the model yielded 94 BLEU, which indicates a possible overfitting.

System	Train	Dev	Test
Baseline	81.11	81.25	82.85
Our APE model	94.45	83.00	83.65

Table 5: BLEU-4 scores.

Our results are in line with the last findings of WMT18 APE shared task that correcting NMT-based translations is a challenging task: gains were only up to 0.8 BLEU points in the NMT track (Chatterjee et al., 2018).

6 Evaluation of Rule-Based Post-Editing

Evaluation was carried out only on the rule-based method output, since it is more robust than the neural approach, and since the APE model did not yield better results.

We analysed a sample of total 66 lexicalisations in 4 categories: Astronaut, University, Monument (2-7 triples) and ComicsCharacter (2-5 triples). Around two thirds of analysed named entities were

[4] version 1.22.0, https://github.com/OpenNMT/OpenNMT-tf

replaced correctly. Below we analyse common sources of errors for the erroneous NEs.

The most frequent case is unrecognised named entities. In 62% of the cases the replacement was not performed, which includes 28% of Latin transcriptions kept, 27% of kept Cyrillic translations, and 7% of acronyms. For the majority of these NEs, the original translations include unaccounted elements (not covered by the extracted rules) such as missing or wrongly inserted prepositions or punctuation marks.

Another common error is lack of grammatical adaptation of the NE. Wrong case marking occurred in 23% of all NEs (cf. example 2), and gender and number agreement make about 6.5%. The less frequent but important error categories are spelling errors, such as missing capitalisation, insertions of quotation marks, and gender or number agreement with anaphors, especially in texts verbalising 5-7 triples.

(2) En: 'The dean of Accademia di Architettura'
MT: 'Декан Accademia di Projecttura'
RPE: 'Декан Академия$_{nomn}$ архитектуры'
Correct: 'Декан Академии$_{gen}$ архитектуры'

To conclude, many errors are caused by irregularities in the translations (which, in turn, are often caused by misspelled input) and can be eliminated by introducing more variation to the replacement algorithm. Grammatical adaptation of NEs, however, requires more careful further investigation.

7 Conclusion

In this study, we reported an ongoing effort to translate the data-to-text WebNLG corpus in Russian. A detailed error analysis showed that roughly 80% of the neural translations contained an error and that 53% of these errors were due to incorrectly translated named entities. We provided a rule-based method which permits correcting these errors and trained a neural post-editing model.

In future work, we plan to extend the approach to other error types and to investigate whether the neural model can be improved to help generalise post-editing to errors not captured by the rule-based method.

Another possible direction for future research will be to identify named entities before the translation phase, perform translation on the texts stripped of named entities (cf. WebNLG delexicalised version of Castro Ferreira et al. (2018)), and then insert named entities, which were translated and verified separately.

References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Thiago Castro Ferreira, Diego Moussallem, Emiel Krahmer, and Sander Wubben. 2018. Enriching the webnlg corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176, Tilburg University, The Netherlands. Association for Computational Linguistics.

Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018. Findings of the wmt 2018 shared task on automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 710–725, Belgium, Brussels. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for nlg micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. Ms-uedin submission to the wmt2018 ape shared task: Dual-source transformer for automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 822–826, Belgium, Brussels. Association for Computational Linguistics.

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. Association for Computational Linguistics.

L. R. Komalova. 2017. Oshibki i netochnosti perevoda (svodnyj referat). *Social'nye i gumanitarnye nauki. Otechestvennaja i zarubezhnaja literatura. Ser. 6, Jazykoznanie: Referativnyj zhurnal*, (4):32–44.

M. A. Kunilovskaya. 2013. Klassifikacija perevodcheskih oshibok dlja sozdanija razmetki v uchebnom parallel'nom korpuse russian learner translator corpus. *Lingua mobilis*, (1 (40)):141–158.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Ondřej Bojar, and Pavel Pecina. 2016. Cuni system for wmt16 automatic post-editing and multimodal translation tasks. In *Proceedings of the First Conference on Machine Translation*, pages 646–654, Berlin, Germany. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics.

Shashi Narayan, Claire Gardent, Shay B. Cohen, and Anastasia Shimorina. 2017. Split and rephrase. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 606–616, Copenhagen, Denmark. Association for Computational Linguistics.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. ESCAPE: a large-scale synthetic corpus for automatic post-editing. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016. A neural network based approach to automatic post-editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286, Berlin, Germany. 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*.

Maja Popović. 2018. *Error Classification and Analysis for Machine Translation Quality Assessment*, pages 129–158. Springer International Publishing, Cham.

Ekaterina Rakhilina, Anastasia Vyrenkova, Elmira Mustakimova, Alina Ladygina, and Ivan Smirnov. 2016. Building a learner corpus for Russian. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 66–75, Umeå, Sweden. LiU Electronic Press.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, Copenhagen, Denmark.