sciendo
DOI 10.2478/jazcas-2023-0054

SLOVAK QUESTION ANSWERING DATASET BASED ON THE MACHINE TRANSLATION OF THE SQUAD V2.0

JÁN STAŠ[1] – DANIEL HLÁDEK[1] – TOMÁŠ KOCTÚR[2]
[1] Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
[2] Deutsche Telekom IT & Telecommunications Slovakia s.r.o., Košice, Slovakia

STAŠ, Ján – HLÁDEK, Daniel – KOCTÚR, Tomáš: Slovak Question Answering Dataset Based on the Machine Translation of the SQuAD v2.0. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 381 – 390.

Abstract: This paper describes the process of building the first large-scale machine-translated question answering dataset SQuAD-sk for the Slovak language. The dataset was automatically translated from the original English SQuAD v2.0 using the Marian neural machine translation together with the Helsinki-NLP Opus English-Slovak model. Moreover, we proposed an effective approach for the approximate search of the translated answer in the translated paragraph based on measuring their similarity using their word vectors. In this way, we obtained more than 92% of the translated questions and answers from the original English dataset. We then used this machine-translated dataset to train the Slovak question answering system by fine-tuning monolingual and multilingual BERT-based language models. The scores achieved by EM = 69.48% and F1 = 78.87% for the fine-tuned mBERT model show comparable results of question answering with recently published machine-translated SQuAD datasets for other European languages.

Keywords: language modeling, machine reading comprehension, machine translation, natural language processing, question answering

1 INTRODUCTION

In the field of natural language processing (NLP), understanding text and answering questions about this text is an important and challenging task for machines. Question answering (QA) is a conversational task in NLP that retrieves correct answers to questions asked by humans in natural language, which is useful for searching for an answer in a text document. Machine reading comprehension (MRC) is a type of QA in which a machine can read any text, understand it, and answer questions about that text. Therefore, the relationship between QA and MRC is very close and is often altered.

Recent methods for textual QA rely on large-scale human-annotated datasets. Existing QA datasets are primarily built in English. The most popular benchmark datasets for evaluating QA models is the Stanford Question Answering Dataset (SQuAD). It contains more than 100k pairs of questions and answers posed by crowd workers on a set of Wikipedia articles. The answers to each question are a text fragment from the corresponding reading paragraph of the article (Rajpurkar et al. 2016). The SQuAD v1.1 dataset was later

updated to the SQuAD v2.0, which includes additional 50k questions, labeled as unanswerable concerning the corresponding target text, and forces QA models to understand when a question cannot be answered given the context (Rajpurkar et al. 2018).

Building large-scale resources is labor-intensive and expensive for others, especially low-resourced languages. However, the unavailability of such datasets makes it challenging to train multilingual QA systems with performance comparable to the English one.

One solution is to apply machine translation (MT) to an existing QA dataset. The advantage is that it is fast and does not require human labor and finances. That is why we decided to create a QA dataset for the Slovak language based on machine translation of the original English SQuAD v2.0 dataset. The dataset will be publicly available at.[1]

The paper is organized as follows. In Section 2 we briefly describe current approaches to creating QA datasets using machine translation and summarize 25 existing non-English QA datasets that have been created in recent years. In Section 3, we describe in detail the proposed approach to create a machine-translated QA dataset for the Slovak language. Next, we fine-tuned several available monolingual and multilingual BERT (Bidirectional Encoder Representations from Transformers) language models on a machine-translated SQuAD-sk dataset and evaluate them. Finally, Section 5 summarizes the contribution of our work and concludes the paper with future directions.

2 STATE-OF-THE-ART

For the task of question answering, many SQuAD-like datasets have been created in the last ten years. Most of them were created only for English. By the end of 2022, we would have identified 25 machine-translated SQuAD datasets in 21 other languages. The primary data source was the English SQuAD in version 1.1, after 2019 also in version 2.0, containing a translation of unanswerable questions. Tab. 1 summarizes the basic information about them.

When creating MT datasets, authors encounter the following two problems:

- the errors caused by the machine translation;
- correct alignment of the translated answers in the translated paragraph context.

For accurate machine translation, authors often use Google Translate. In rare cases, other MT systems, such as DeepL for Finnish and Italian, Moses for Spanish and Danish, LINDAT for Czech, Yandex for Russian, and Marian for German were used.

In this case, each paragraph is thus automatically translated with all relevant questions and marked answers. Some of these triplets cannot be translated due to unknown characters or low lexical coverage of the MT system. Also, the MT system can some-times be unable to recognize named entities without context and transliterate them. Furthermore, translations of the questions can result in multiple translation variants (Lee et al. 2020).

[1] https://huggingface.co/TUKE-DeutscheTelekom

Due to the differences in nature and complexity between languages, the precision of machine translation varies greatly. Some languages produce better translation results than others. Machine translation works fine for most Latin and Western languages, but the accuracy of some under-resourced African or Asian languages is often poor.

Language	Title	Translation	v2.0	Size	Reference
Arabic	Arabic-SQuAD	Google Translate	N	60k	Mozannar et al. 2019
Brazilian Portuguese	BraSQuAD	Google Cloud API	Y	-	Esposito 2020[2]
Bengali	Bengali-SQuAD	Google Cloud API	Y	91k	Mayeesha et al. 2021
Czech	Czech-SQuAD	LINDAT Translator	Y	117k	Macková – Straka 2020
Danish	SQuAD-da	TAR Method/Moses	Y	156k	Carrino et al. 2020[3]
Finnish	SQuAD_v2_fi	DeepL	Y	-	Kylliäinen 2022[4]
French	SQuAD-fr	Google Translate	N	100k	Cattan et al. 2021
French	French-SQuAD	Google Translate	N	-	Kabbadj 2019[5]
German	German SQuAD	Marian NMT	N	-	Giagoulas 2020[6]
Hindi	Hindi SQuAD	Google Translate	N	18.5k	Gupta et al. 2019
Indonesian	SQuAD-id	Google Translate	-	-	Wikidepia 2021[7]
Italian	SQuAD-it	DeepL	N	60k	Croce et al. 2018
Korean	K-QuAD	Google Translate	N	77k	Lee et al. 2018
Latvian	SQuAD-lv	-	Y	-	Tilde 2019[8]
Persian	ParSQuAD	Google Translate	Y	95k	Abadani et al. 2021
Polish	PolAQ	Google Cloud API	N	14k	Jurkiewicz 2020[9]
Polish	SQuAD-pl	Google Translate	Y	-	Brodzik 2022[10]
Portuguese	SQuAD_v1.1_pt	Google Cloud API	N	100k	Carvalho 2019[11]
Portuguese	SQuAD_v2.0_pt	Google Cloud API	Y	-	Janiake 2020[12]
Russian	SQuAD-ru	Yandex Translate	N	-	Semyonov 2017[13]
Spanish	SQuAD-es-v1.1	TAR Method/Moses	N	87.5k	Carrino et al. 2020
Spanish	SQuAD-es-v2.0	TAR Method/Moses	Y	46k	Carrino et al. 2020
Swedish	SQuAD-v2-sv	Google Translate	Y	125k	Okazawa 2021[14]
Turkish	SQuAD (Tr)	Google Translate	N	-	Gemirter – Goularas 2021
Ukrainian	SQuAD-uk	Google Cloud API	N	30k	Tiutiunnyk – Dzomkin 2019

Tab. 1. Summary of available machine-translated SQuAD datasets

[2] https://github.com/piEsposito/br-quad-2.0
[3] https://github.com/ccasimiro88/TranslateAlignRetrieve/tree/multilingual/squads-tar/da
[4] https://github.com/ilmarikyl/SQuAD_v2_fi
[5] https://github.com/Alikabbadj/French-SQuAD
[6] https://github.com/agiagoulas/german_squad
[7] https://github.com/Wikidepia/SQuAD-id
[8] https://github.com/tilde-nlp/SQuAD-LV
[9] https://github.com/tjur/polaq_master_thesis
[10] https://github.com/brodzik/SQuAD-PL
[11] https://github.com/nunorc/squad-v1.1-pt
[12] https://github.com/cjaniake/squad_v2.0_pt
[13] https://github.com/lightforever/SQUAD_RU
[14] https://github.com/susumu2357/SQuAD_v2_sv

On the other hand, to verify whether the translation between the languages has been done correctly, it is approached either by laborious manual correction or by using back translation, which is the conversion of a translated text back to its original language. Using the back translation and then comparing the original document with the back translation, we can gauge how closely the translation mirrors the meaning of the original text. This also gives us feedback on how accurate the translation between languages is.

Back translation can also be used when creating MT datasets without any additional grammar correction of the translated questions and answers. An example of this approach is the Ukrainian SQuAD-uk or Polish SQuAD-pl dataset.

The second issue is the problem with alignment after translation. It occurs when the start and end points of the answers to the relevant questions are not correctly marked in a paragraph or if the links between words in the adjacent context are broken due to translation. Therefore, it is necessary to look for the correct translated answer in the paragraph context. This is usually done by:

- using special tags (for example quotation marks) that are inserted before the machine translation (may break the context);
- using attention of the MT system and choosing the words aligned to the source answer (very difficult);
- using an approximate search for the translated answer and measuring the minimal distance between translated answer and translated context, for example using Levenshtein distance or by cosine similarity between these two word vectors.

3 SLOVAK TRANSLATION OF THE SQUAD V2.0

3.1 Machine translation from English to Slovak

To translate the original English SQuAD v2.0 into Slovak, we used an efficient and freely available neural MT framework Marian (Junczys-Dowmunt et al. 2018) together with the Helsinki-NLP Opus English-Slovak model[15] (Tiedemann – Thottingal 2020). The BLEU score for this model was evaluated at 36.8.

The process of machine translation can be described as follows (Hládek et al. 2023):

- we loaded the original English SQuAD v2.0 in JSON file format and transformed it into a set of triplets: paragraphs, questions, and answers;
- for each paragraph, question, and answer in English, we obtained their translations into Slovak;
- we searched for the translated answer in the translated paragraph; the result is part of the paragraph that contains the translated answer;

[15] https://huggingface.co/Helsinki-NLP/opus-mt-en-sk

- in validation, we eliminated the triplets where the answer could not be found;
- we compiled the translated paragraphs, questions, and answer areas into a new Slovak SQuAD-sk dataset.

3.2 Proposed approach for the approximate search of the translated answer

The SQuAD dataset contains the exact position of the answer in the context of the paragraph. Marks for the start and end of the answer are not suitable for being inserted into the translated context because unknown and unexpected tags can disrupt the context and reduce the quality of the translation. They can also get lost in the translation process. Therefore, it is necessary to translate the answer separately and look for it again in the translated paragraph.

The main disadvantage is that the independent translation of the answer can be affected by the missing context since the answer that was translated on its own may differ from the answer in the translated context.

When searching for an answer in the paragraph context, we follow this procedure:

- exact search for the original answer;
- exact search for the translated answer;
- approximate search for the translated answer.

First, we searched for the exact occurrence of the original, untranslated answer in the translated paragraph. We started from the assumption that the most well-known named entities for persons, organizations and geographical locations will not be changed by machine translation. If we did not find the original answer, we tried to find a translated answer. This option handles cases where the translation of the answer is unambiguous and does not depend on the context. If we did not find the answer in this case either, then its translation was influenced by the context, and the independently translated form differed from the form occurring in the paragraph context. Since Slovak is a morphologically rich language, the answer in the translated context will likely contain word forms different from those of the independently translated answer.

Therefore, we proposed an approximate search that would identify the part of the answer that is semantically similar. We assume that a translation in context will be similar in meaning to a translation without context but may differ by using a different synonym or morphological form. We also assume that the number of words in the answer in context will be the same as in the case of the answer without context.

We evaluated the semantic similarity of two texts a_O, and a_P using word vectors. If the answer is one word, the distance between the word vectors will be smaller than the threshold T as follows:

$$D\left(V\left(a_o\right), V\left(a_p\right)\right) \leq T,$$

where V is the vector of texts a_O and ap, respectively, and D is the distance metric between the two word vectors. If the answer is multi-word, we choose the arithmetic mean of the word vectors of all words as the semantic representation.

As a distance metric between two word vectors, we chose cosine distance, which is the negative value of cosine similarity. Cosine similarity is defined as the cosine angle of two word vectors and is often used in determining semantic similarity.

We used the spaCy library (Honnibal et al. 2020) because it allows easy use of the word vector model and the method for approximate search using cosine similarity. We created a model of word vectors using fastText (Bojanowski et al. 2016) with the Floret[16] subword units. The word vector model was trained on a web corpus of Slovak written texts and created as part of the Slovak spaCy model available at.[17]

We searched in the paragraph context using a floating window. We set the size of the floating window to the same as the number of words in the translated answer. For the floating window, we calculated its average word vector. Next, we calculated the cosine distance between the mean word vectors of the translated answer and the floating window.

We selected the window that had the minimum distance from the translated answer. In this way, we found the translated answer in the context. We ignored distances greater than the constant T to limit spurious results. Only in a few cases could no region with a similarity less than T be found. We discarded these answers.

A visual inspection of the translated database revealed that most errors were caused by the translation itself, sometimes not all words were found, and at other times the answer did not grammatically match the question. These types of errors are usually caused by differences between the Slovak and English languages. The easiest way is to identify such grammatically incorrect answers and eliminate them. However, a deeper analysis would be needed in this area, which may be the subject of further research.

However, in most cases, the result was understandable and grammatically correct. The number of answers that could not be found by an approximate search in the paragraph context is very low (see Tab. 2).

Number of	**SQuAD v2.0**	**SQuAD-sk**	**% of original**
Documents	442	442	100
Paragraphs	19,035	18,931	99.45
Questions	130,319	120,239	92.27
Answers	86,821	79,978	92.12
Unanswerable questions	43,498	40,261	92.56

Tab. 2. Statistics on the English SQuAD v2.0 and the MT Slovak SQuAD-sk datasets

[16] https://github.com/explosion/floret

[17] https://github.com/hladek/spacy-skmodel

Note that the proposed approach is applicable to another language that has a trained word vector model and for which an MT model exists. The approach is independent of the MT system; any other translation model can be used.

4 EXPERIMENTAL RESULTS

The experiment was focused on the evaluation of the performance of several available monolingual and multilingual pretrained BERT models after fine-tuning on a machine-translated SQuAD-sk dataset in the task of automatically answering questions in Slovak. For this purpose, we created two evaluation datasets. The first consisted of a machine-translated English SQuAD development set into Slovak. The second consisted of 9,583 hand-annotated questions for 7,822 answers marked in 2,568 paragraphs, which were obtained from 940 Slovak documents published on Wikipedia. The annotation was performed by crowd workers, in a similar way as it was in the original English SQuAD.

We used the following types of pre-trained BERT models that support Slovak:

- our own monolingual Slovak RoBERTa-base cased model;
- monolingual SlovakBERT-base cased model (Pikuliak et al. 2022);
- multilingual mBERT-base cased model (Devlin et al. 2019);
- multilingual XLM-RoBERTa-base cased model (Conneau et al. 2020).

To evaluate the fine-tuned BERT models, we used the exact match (EM) and F1 scores. EM measures the percentage of predictions that match any reference answers at the token level. The F1 score is the harmonic mean of precision and recall, where precision measures the ratio of correct tokens in the prediction and recall rates the ratio of the correct tokens in the prediction to the correct response.

Training set	**Evaluation set**	**SK RoBERTa**		**SlovakBERT**		**mBERT**		**XLM-R**	
		EM	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**
MT SQuAD-sk	machine-translated	52.50	58.73	56.02	62.78	56.02	63.02	56.56	62.91
MT SQuAD-sk	hand-annotated	48.16	57.75	62.97	71.14	69.48	**78.87**	**69.76**	78.03

Tab. 3. Results of the Slovak question answering

The experimental results of the Slovak question answering for selected BERT models fine-tuned on the SQuAD-sk dataset are summarized in Tab. 3. It should be noted that the number of parameters of pre-trained models was approximately the same in all cases.

As we can see from these results, the huge multilingual models outperform the monolingual models. This is because the multilingual models have a vocabulary several times larger than monolingual models and can therefore better capture the context of answers in paragraphs. To improve monolingual models, it would be necessary to significantly increase the size of the training data or the number of model parameters.

5 CONCLUSION

SQuAD-sk is the first machine-translated QA dataset in Slovak. The experimental results show the usefulness of the dataset for monolingual and multilingual QA. It will mainly contribute to the creation of new systems to generate answers to questions in natural language. The dataset will be published free of charge for scientific use.

In the future, we would like to focus more on errors and biases caused by machine translation to improve the SQuAD-sk dataset. Further research should focus on machine translation of other types of QA datasets that do not only contain factoid questions but also multiple-choice or causal questions.

ACKNOWLEDGEMENTS

This research was supported by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic, and the Slovak Academy of Sciences under the project VEGA 2/0165/21 funded by the Ministry of Education, Science, Research and Sport of the Slovak Republic; and by the Slovak Research and Development Agency under the grants APVV-SK-TW-21-0002, APVV-22-0261, and APVV-22-0414.

Our special thanks go to Deutsche Telekom IT & Telecommunications Slovakia s.r.o. for fruitful cooperation and personal and financial support.

References

Abadani, N., Mozafani, J., Fatemi, A., Nematbakhsh, M., and Kazemi, A. (2021). ParSQuAD: Persian question answering dataset based on machine translation of SQuAD 2.0. International Journal of Web Research, 4(1), pages 34–46.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. Trans. of the ACL, Vol. 5, Cambridge, MA, pages 135–146. Accessible at: https://aclanthology.org/Q17-1010.pdf.

Carrino, C. P., Costa-Jussa, M. R., and Fonollosa, J. A. R. (2020). Automatic Spanish translation of the SQuAD dataset for multilingual question answering. In Proc. of LREC, Marseille, France, pages 5515–5523. Accessible at: https://arxiv.org/abs/1912.05200.

Cattan, O., Servan, C., and Rosset, S. (2021). On the usability of transformers-based models for a French question-answering task. In Proc. of RANLP, Varna, Bulgaria, pages 244–255. Accessible at: https://hal.archives-ouvertes.fr/hal-03336060/.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proc. of ACL, Online, pages 8440–8451. Accessible at: https://aclanthology.org/2020.acl-main.747.pdf.

Croce, D., Zelenanska, A, and Basili, R. (2018). Neural learning for question answering in Italian. In C. Ghidini – B. Magnini – A. Passerini – P. Traverso (eds): Advances in Artificial Intelligence, LNAI vol. 11298, Springer, Cham, pages 389–402. Accessible at: https://doi.org/10.1007/978-3-030-03840-3_29.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. of NAACL, Minneapolis, Minnesota, pages 4171–4186. Accessible at: https://aclanthology.org/N19-1423/.

Germirter, C. B., and Goularas, D. (2021). A Turkish question answering system based on deep learning neural networks. Journal of Intelligent Systems: Theory and Applications, 4(2), pages 65–75. Accessible at: https://dergipark.org.tr/tr/download/article-file/1361881.

Gupta, D., Ekbal, A., and Bhattacharyya, P. (2019). A deep neural network framework for English Hindi question answering. ACM TALLIP, 19(2), Article No. 25, pages 1–22.

Hládek, D., Staš, J., Juhár, J., and Koctúr, T. (2023). Slovak dataset for multilingual question answering. IEEE Access, Vol. 11, pages 32869–32881. Accessible at: https://ieeexplore.ieee.org/document/10082887.

Honnibal, M., Montani, I., Landeghem van, S., and Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. Accessible at: doi 10.5281/zenodo.1212303.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In Proc. of ACL, Melbourne, Australia, pages 116–121. Accessible at: https://aclanthology.org/P18-4020.pdf.

Lee, K., Yoon, K., Park, S., and Hwang, S. W. (2018). Semi-supervised training data generation for multilingual question answering. In Proc. of LREC, Miyazaki, Japan, pages 2758–2762. Accessible at: https://aclanthology.org/L18-1437.

Macková, K., and Straka, M. (2020). Reading comprehension in Czech via machine translation and cross-lingual transfer. In Proc. of TSD, Brno, Czech Republic, pages 171–179. Accessible at: https://arxiv.org/abs/2007.01667.

Mayeesha, T. T., Sarwar, A. Md., and Rahman, R. M. (2021). Deep learning based question answering in Bengali. Journal of Information and Telecommunication, 5(2), pages 145–178. Accessible at: https://doi.org/10.1080/24751839.2020.1833136.

Mozannar, H., El Hajal, K., Maamary, E., and Hajj, H. M. (2019). Neural Arabic question answering. In Proc. of WANLP, Florence, Italy, pages 108–118. Accessible at: https://arxiv.org/abs/1906.05394v1.

Pikuliak, M., Grivalský, Š., Konôpka, M., Blšták, M., Tamajka, M., Bachratý, V., Šimko, M., Balážik, P., Trnka, M., and Uhlárik, F. (2022). SlovakBERT: Slovak masked language model. In Proc. of EMNLP, Abu Dhabi, United Arab Emirates, pages 7156–7168. Accessible at: https://aclanthology.org/2022.findings-emnlp.530.pdf.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In Proc. of EMNLP, Austin, Texas, pages 2383–2392. Accessible at: https://aclanthology.org/2021.emnlp-main.530.pdf.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In Proc. of ACL, Melbourne, Australia, pages 784–789. Accessible at: https://aclanthology.org/P18-2124.pdf.

Tiedemann, J., and Thottingal, S. (2020). OPUS-MT – Building open translation services for the Worlds. In Proc. of EAMT, Lisboa, Portugal, pages 479–4810. Accessible at: https://aclanthology.org/2020.eamt-1.61.pdf.

Tiutiunnyk, S., and Dyomkin, V. (2019). Context-based question-answering system for the Ukrainian language. In Proc. of MS-AMLV, Lviv, Ukraine, pages 81–88. Accessible at: https://ceur-ws.org/Vol-2566/MS-AMLV-2019-paper17-p081.pdf.