Comparing RNN and Transformer Context Representations in the Czech Answer Selection Task

Marek Medveď, Radoslav Sabol and Aleš Horák
Natural Language Processing Centre, Faculty of Informatics, Masaryk University,
Botanická 68a, 602 00, Brno, Czech Republic

Keywords: Question Answering, Answer Context, Answer Selection, Czech, Sentece Embeddings, RNN, BERT.

Abstract: Open domain question answering now inevitably builds upon advanced neural models processing large unstructured textual sources serving as a kind of underlying knowledge base. In case of non-mainstream highly-inflected languages, the state-of-the-art approaches lack large training datasets emphasizing the need for other improvement techniques. In this paper, we present detailed evaluation of a new technique employing various context representations in the answer selection task where the best answer sentence from a candidate document is identified as the most relevant to the human entered question. The input data here consists not only of each sentence in isolation but also of its preceding sentence(s) as the context. We compare seven different context representations including direct recurrent network (RNN) embeddings and several BERT-model based sentence embedding vectors. All experiments are evaluated with a new version 3.1 of the Czech question answering benchmark dataset SQAD with possible multiple correct answers as a new feature. The comparison shows that the BERT-based sentence embeddings are able to offer the best context representations reaching the mean average precision results of 83.39% which is a new best score for this dataset.

1 INTRODUCTION

The main strategy of open domain question answering (QA) systems which exploit a large underlying unstructured textual knowledge base usually follows the same core schema. First, the documents which are most related to the answer are identified (*document selection*), then the "central point" (e.g. a sentence) containing or supporting the answer is searched for (*answer selection*), and finally the specific *exact answer* is extracted or deduced from the selected sentence (*answer extraction*). Such split approach is not always followed with big models e.g. ALBERT (Lan et al., 2019) or ELECTRA (Clark et al., 2020) fine-tuned on large QA datasets such as SQuAD (Rajpurkar et al., 2016), RACE (Lai et al., 2017) or GLUE (Warstadt et al., 2019). But for smaller datasets with non-mainstream languages this approach allows to fine tune the individual modules separately.

In this paper, we present the results of a new enhancement in the answer selection task for highly inflected languages. The proposed method is described and evaluated within the QA system AQA (Medveď and Horák, 2016; Medveď and Horák, 2018) and the Czech question answering dataset SQAD (Sabol et al., 2019) consisting of more than 13,000 QA pairs. The main idea of the new technique lies in employing various types of answer context representations to supplement the information contained in candidate answer sentences with directly preceding communication. For the purpose of context specification, two main approaches are evaluated and compared – word vector representations with recurrent neural networks and sentence vector representations with the transformer models.

Section 2 details the answer selection task within the Czech SQAD database and describes the core neural network architecture used in the answer selection module. The next section presents the seven context representations for the two model types that were implemented and evaluated. And the last section offers the results and discussion for each of the context types with selected output examples. The best representation was based on a Czech BERT model named Czert (Sido et al., 2021) reaching 83.39% of mean average precision as a new best result with the SQAD dataset.

Medved, M., Sabol, R. and Horák, A.
Comparing RNN and Transformer Context Representations in the Czech Answer Selection Task.
DOI: 10.5220/0010827000003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 3*, pages 388-394
ISBN: 978-989-758-547-0; ISSN: 2184-433X

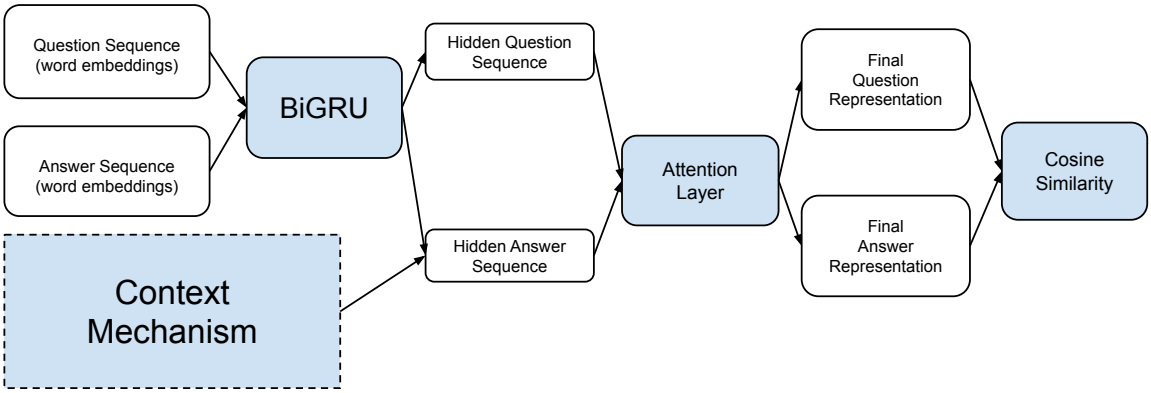

Figure 1: The overall Answer Selection module architecture.

2 THE ANSWER SELECTION TASK

The Simple Question Answering Database, or SQAD (Medved' et al., 2020), is a benchmark dataset designed for evaluating various Question Answering (QA) tasks for the Czech language. The latest version SQAD 3.1 consists of 13,473 question-answer records harvested with an underlying knowledge base of 6,571 different Czech Wikipedia articles. Each record contains several metadata files. The *Question* file stores the question text created by a human annotator. The *Text* file contains the full Wikipedia article. The *Answer Selection* file presents the sentence containing the exact answer (the shortest phrase fully answering the question), which is instantiated in the *Answer Extraction* file. The remaining metadata files describe the category (type) of the question and the answer.

In the final SQAD database, all records are preprocessed and enriched with automatic PoS-tagging and word embedding information. To achieve speed and space efficiency, the ZODB[1] Python library is employed to encode the database information without duplicities. The *Vocabulary* table stores all words with their base forms, part of speech tags, and the corresponding embedding vectors in different dimensions. FastText (Bojanowski et al., 2017) vectors are prepared with the dimensions of 100, 300 and 500 and BERT (pre-trained Slavic Bert model (Arkhipov et al., 2019a)) vectors with the size of 768. The *Knowledge Base* table contains all the involved Wikipedia articles where all words are represented by word IDs leading to the chosen vector representation. Moreover, each sentence is stored in several sentence vector embedding forms such as precomputed Sentence-BERT (Reimers and Gurevych, 2019) vectors or the CLS token vector provided by the Hugging Face BERT[2] transformer, both loading the Slavic BERT model.

2.1 The Neural Model and the Learning Process

The neural network architecture of the answer selection module (see Figure 1) is based on the Siamese network schema as introduced by Santo et al (Santos et al., 2016). The network input is thus formed by a pair (q,a), where q and a are matrices of word embeddings for the question and a candidate answer respectively.

The first layer is defined as a bi-directional Gated Recurrent Unit (BiGRU) with shared weights for both inputs q and a. The output of this layer forms matrices Q (question) and A (answer). The following layer uses an attention mechanism computing the matrix $G = Q^T WA$, where W is an attention matrix of learnable parameters. Row-wise and column-wise max pooling is used to create the final representations of q and a that are compared using the cosine similarity measure as the final score of the input pair.

In terms of learning, the objective to minimize is defined as the *Pairwise Ranking Loss*,

$$loss(p^+, p^-) = max\{0, m - p^+ + p^-\}$$

where p^+ and p^- represent positive and negative examples, and m is a constant margin (0i is used in this paper). For performance and memory reasons, the algorithm samples 50 negative examples along with the single correct example. As there are multiple negative examples during the training procedure, the one

[1]https://zodb.org/en/latest/

[2]https://huggingface.co/transformers/model_doc/bert.html

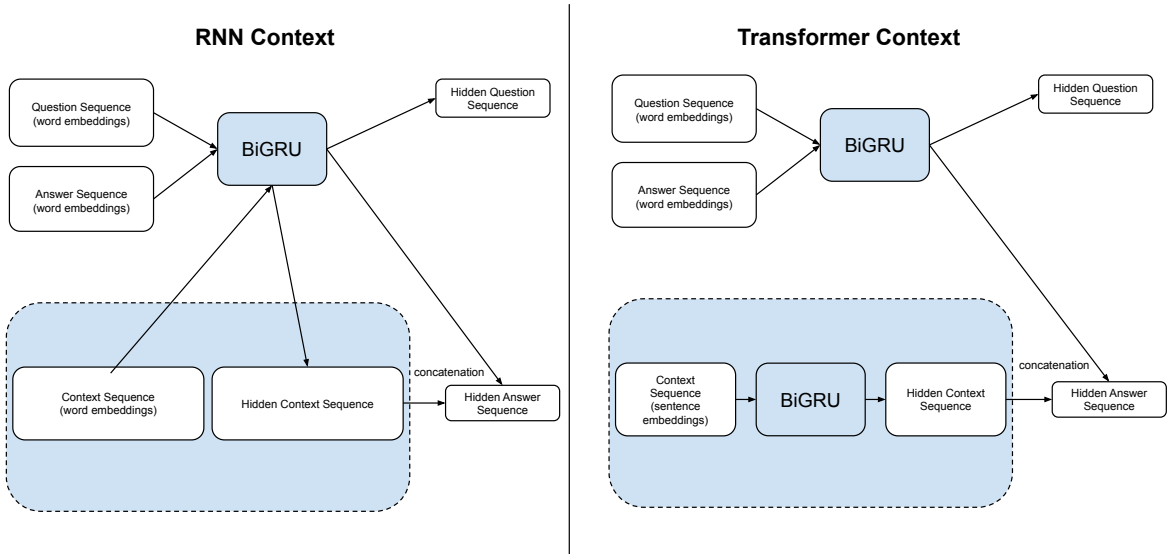

Figure 2: The specification of Answer Selecion context for both RNN and transformer-based approaches.

which produces the highest loss is used for the backpropagation.

The training algorithm iterates 25 times over the training data, and the model with the best validation accuracy is selected for the final evaluation. The forward pass and backpropagation is computed on a single *NVIDIA Tesla T4* GPU with 16GB VRAM, 320 Turing Tensor Cores, and 2560 CUDA cores available.

3 ANSWER CONTEXT

The key idea behind the exploitation of answer context in Question Answering is to find out if some features/knowledge obtained from answer context can help the final model to identify the correct answer more precisely. Very often the question directly specifies several important clues that navigate the answer selection process through the text. However, this clue information is often distributed not only in the final answer sentence but also in the preceding text.

For example, Figure 3 contains the phrase "band Apocalyptica" as the main clue of the question while the main focus of the question is "number of band members". Here the answer sentence does not even mention the "band Apocalyptica" clue. Without the previous context sentence, even a human reader is not able to confirm whether the answer sentence provides the required information about the Apocalyptica band or about some other band that can be mentioned in the article.

The final system therefore needs to benefit from the information represented by the context sentences if the database connects the final answer to its context.

3.1 Context Types in SQAD 3.1

The new SQAD v3.1 database is extended with several new context representations in two types. The first type offers a direct extension of the candidate answer sentence serving the context as input to the same recurrent network (RNN) layer as used for the answer. The second type uses various pre-trained transformer representations of the whole preceding sentence (or sentences) as the context. Each of these types needs a different inclusion technique into the module neural architecture as detailed in the following section.

In case of the RNN context type, three context variants have been evaluated. The most straightforward approach uses the full preceding sentence(s), word by word, as the context. The second variant extracts only noun phrases, as the most probable bearer of a clue information. In this case, the noun phrases are identified with the SET parser (Kovář et al., 2011). The third variant again extracts possible information clues which are now specified as so called *link named entities* (Medved' et al., 2020). The entities are searched in the text using a separate BERT-NER model trained to mark all phrases used as internal links in the Czech Wikipedia.

3.2 Encoding the Context

The answer selection module as specified in Section 2.1 can incorporate different *context mechanisms*. The RNN contexts come in the form of individual word vectors that represents the context sentences or

Table 1: The best hyperparameter values for various context types.

Context Type	BiGRU Hidden Size	Learning Rate	Dropout
SENT_1 RNN context	380	0.0004	0.4
PHR_2 RNN context	380	0.0002	0.4
NER_5 RNN context	320	0.0006	0.4
SENT_1 Transformer ctx	480	0.0007	0.2

Question:
Kolik členů má v současnosti finská hudební skupina Apocalyptica?
[How many members does the Finnish band Apocalyptica currently have?]
Correct answer:
Skupina je složena ze tří (původně čtyř) klasických violoncellistů.
[The band is composed of three (originally four) classical cellists.]
Answer context (previous sentence):
Apocalyptica je finská hudební skupina, jejíž zvláštností je interpretace původně heavy metalových skladeb osobitým způsobem aranžovaných pro violoncello.
[Apocalyptica is a Finnish band whose peculiarity is the interpretation of originally heavy metal compositions arranged in a special way for cello.]

Figure 3: Example of answer context.

phrases. The transformer contexts are in the form of sentence vectors encoding the whole sentences.

Figure 2 details two necessary architecture modifications for the two main context types. The first modification is composed of contexts represented by a sequence of word embeddings separated by a special `[SEP]` (separator) token. The separator introduces each full sentence, noun phrase of link named entity selected for the context.

The particular word embeddings are derived from the same embedding model[3] as used for the questions and candidate answers, and the internal sequence representation is provided by the same RNN layer. The resulting context word vectors are then concatenated to the answer sequence before the attention mechanism.

The second context form is constructed with sentence embeddings derived from a transformer-based model. Unlike in the first group, each sentence is represented by one vector. As this information is not directly comparable to the question/answer words, the internal context representation is thus obtained from a separate BiGRU layer trained with the sentence embeddings input. The new representation is concatenated in the same way as in the first group. Current implementation evaluates four sentence embedding models:

- the CLS vector provided by the Slavic BERT model (Arkhipov et al., 2019b)
- the CLS vector of the Czert model (Sido et al., 2021)
- the CLS vector of the RobeCzech model (Straka et al., 2021)
- the Sentence-BERT model representation (Reimers and Gurevych, 2019)

[3]The FastText word embedding vectors are used in the presented experiments.

4 EXPERIMENTS AND EVALUATION

In the following experiments, all the presented context variants have been evaluated with the SQADv3.1 dataset. The dataset comes divided into three balanced subsets used for training, test and validation with 8,059, 4,013 and 1,401 records.

4.1 Multiple Correct Answers in SQAD 3.1

Detailed error analyses of question answering techniques with previous SQAD versions revealed a possible misconception in identifying a single sentence as the only correct answer selection result. An example in Figure 4 demonstrates a case where the exact answer is present in *multiple* sentences. Any of these sentence then serves as a correct input to the final answer extraction module. Therefore the current SQAD version extends the metadata with a list of all sentences containing the exact answer as a possible answer selection result. Table 3 presents overall dataset statistics of the multiple answer lists where almost a half of the records contains more than one possible answer sentence. For comparison purposes, the mod-

Table 2: The answer selection results for different context type models. The *best RNN model* is in italics, the **overall best model** is bold.

	Single answers		Multiple answers	
Context Type	MAP	MRR	MAP	MRR
SENT_1 RNN context	81.94	88.03	84.76	90.09
PHR_2 RNN context	82.23	88.20	84.98	90.20
NER_5 RNN context	*82.42*	*88.40*	*85.14*	*90.35*
SENT_1 with Czert	**83.39**	**89.18**	**85.79**	**90.93**
SENT_1 with RobeCzech	82.75	88.68	85.29	90.55
SENT_1 with Slavic BERT	83.05	88.88	85.59	90.74
SENT_1 with S_BERT	82.65	88.58	85.14	90.41

Question:
Proč dostal Albert Einstein Nobelovu cenu za fyziku?
[Why Albert Einstein receive the Nobel Prize in Physics?]
Selected answer:
Za vysvětlení fotoelektrického jevu získal Einstein roku 1921 Nobelovu cenu za fyziku.
[For explaining the photoelectric effect, Einstein won the 1921 Nobel Prize in Physics.]
Correct answer:
V roce 1921 byl Einstein oceněn Nobelovou cenou za fyziku za " vysvětlení fotoefektu a za zásluhy o teoretickou fyziku " .
[In 1921, Einstein was awarded the Nobel Prize in Physics for "explaining the photo effect and for his contributions to theoretical physics."]

Figure 4: Answer in two sentences example (record 011880).

ule offers the evaluation results fort both the single answer and the multiple answer forms.

4.2 Results and Comparison

The following experiments use a fixed context window depending on the type of context. A single previous sentence (`SENT_1`) was used for both the RNN full sentence context and for the transformer contexts. The noun phrases RNN context was evaluated with all noun phrases from preceding two sentences (`PHR_2`) and the link named entity context used five preceding named entities (`NER_5`).

The answer selection model was trained using the RMSprop optimization algorithm, where the learning rate, dropout, batch size, and the BiGRU hidden size were the optimized hyperparameters. The best hyperparameter settings were found using the Optuna hyperparameter optimization framework, as shown in Table 1.

Table 3: SQAD 3.1 statistics of multiple answer sentences. Number of sentences displays how many sentences of the underlying document contain the exact answer. In case of 0 sentences, the exact answer is not directly represented in the text, e.g. with a yes/no question.

0 sentences	1387 records
1 sentence	5964 records
2 sentences	1751 records
3 sentences	869 records
4 sentences	514 records
5 sentences	383 records
$\geq$ 6 sentences	2605 records

Table 2 presents the final results of the mean average precision (MAP) and the mean reciprocal rank (MRR) using both the single and multiple answer setups. We may see that the multiple answer setup brings the results in a narrower range eliminating possible correct exact answers that are evaluated as incorrect in the single answer setup. All values are averages from three runs of the best optimized model.

For a comparison, the most recent published best result for the SQADv3 test set was 82.91% MAP in (Medved' et al., 2020). The same model setup (RNN SENT_1) achieved 81.94% MAP score in the current SQADv3.1 version which thus serves as a baseline here. The best-performing context type was the transformer context using Czert (Sido et al., 2021) sentence embeddings, which reached MAP score of **83.39%**, a 1.45% improvement over the baseline. For the new evaluation with multiple correct answers, the same best model achieved **85.79%** of the mean average precision with less differences between the models. The NER_5 model is the winner of the RNN word models, proving the informative value of the link named entities selected as the clue information.

Figure 5 offers an example where the winning Czert sentence representation provided more informative

Question:
Kdy začala vysílat TV Nova?
[When did the Nova TV start broadcasting?]
Czert model result (correct):
TV Nova je česká komerční televizní stanice.
[Nova TV is a Czech commercial TV station.]
Licenci na vysílání získala v roce 1993, první vysílání proběhlo až 4. února 1994 v Praze.
[It obtained a broadcasting license in 1993, the first broadcast however started on the 4th of February 1994 in Prague.]
RNN SENT_1 result (incorrect):
Na Novu přichází i nové zahraniční seriály jako Námořní vyšetřovací služba, Kriminálka Miami, Kriminálka New York, Ztraceni, Dr. House a mnoho dalších.
[Nova TV starts broadcasting new foreign series such as NCIS, C.S.I. Miami, C.S.I. New York, Lost, House M.D. and more.]
Od poloviny října 2007 začala TV Nova vysílat zpravodajství (ranní, odpolední i večerní) v obrazovém formátu 16:9 a v rozlišení HDTV.
[From the half of October 2007, the Nova TV started to broadcast news (morning, afternoon and evening) in 16:9 format and in HD resolution.]

Figure 5: An example record (000267) where the Czert transformer model improved the RNN SENT_1 incorrect result. The **answer selection sentence** is in bold, while the *preceding context* is in italics.

context representation allowing to choose the correct answer sentence.

5 CONCLUSIONS AND FUTURE DIRECTIONS

We have presented the details and evaluation of a new technique of employing **various forms of preceding contexts** to improve the answer selection task. The designed model is able to incorporate both **word-based context** representations as well as **sentence-based context** representations. The effectiveness of this approach was evaluated with seven variants of both context types. The best results were achieved using transformer-based sentence context representation of the Czech BERT model named Czert. This model reached 1.45% improvement over the baseline and the resulting mean average precision of **83.39%** is a new best result for the SQAD benchmark dataset with the single answer setup and **85.79%** with the new multiple correct answer evaluation.

In the future work, we plan to test model setups with varying context lengths and to decide whether longer contexts can improve the performance. We will also plug the new answer selection module into the complete AQA pipeline for open domain question answering and evaluate the whole system with the latest version of the SQAD database.

ACKNOWLEDGEMENTS

This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101.

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

REFERENCES

Arkhipov, M., Trofimova, M., Kuratov, Y., and Sorokin, A. (2019a). Tuning Multilingual Transformers for Language-Specific Named Entity Recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Arkhipov, M., Trofimova, M., Kuratov, Y., and Sorokin, A. (2019b). Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders

as Discriminators Rather Than Generators. *arXiv preprint arXiv:2003.10555*.

Kovář, V., Horák, A., and Jakubíček, M. (2011). Syntactic Analysis Using Finite Patterns: A New Parsing System for Czech. In *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 161–171, Berlin/Heidelberg. Springer.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. (2017). RACE: Large-scale ReAding Comprehension Dataset From Examinations. *arXiv preprint arXiv:1704.04683*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv preprint arXiv:1909.11942*.

Medveď, M., Sabol, R., and Horák, A. (2020). Efficient management and optimization of very large machine learning dataset for question answering. In Horák, A., editor, *Proceedings of the Fourteenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2020*, pages 23–34, Brno. Tribun EU.

Medveď, M., Sabol, R., and Horák, A. (2020). Employing Sentence Context in Czech Answer Selection. In *International Conference on Text, Speech, and Dialogue, TSD 2020*, pages pp. 112–121. Springer.

Medveď, M. and Horák, A. (2016). AQA: Automatic Question Answering System for Czech. In *Text, Speech, and Dialogue, TSD 2016*, pages 270–278, Switzerland. Springer.

Medveď, M. and Horák, A. (2018). Sentence and word embedding employed in open question-answering. In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018)*, pages 486–492, Setúbal, Portugal. SCITEPRESS - Science and Technology Publications.

Medveď, M., Sabol, R., and Horák, A. (2020). Improving RNN-based Answer Selection for Morphologically Rich Languages. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020)*, pages 644–651, Valleta, Malta. SCITEPRESS - Science and Technology Publications.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100, 000+ Questions for Machine Comprehension of Text. *CoRR*, abs/1606.05250.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sabol, R., Medveď, M., and Horák, A. (2019). Czech Question Answering with Extended SQAD v3.0 Benchmark Dataset. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019*, pages 99–108.

Santos, C. d., Tan, M., Xiang, B., and Zhou, B. (2016). Attentive pooling networks. *arXiv preprint arXiv:1602.03609*.

Sido, J., Pražák, O., Přibáň, P., Pašek, J., Seják, M., and Konopík, M. (2021). Czert – Czech BERT-like Model for Language Representation. *arXiv preprint arXiv:2103.13031*.

Straka, M., Náplava, J., Straková, J., and Samuel, D. (2021). RobeCzech base. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Warstadt, A., Singh, A., and Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.