Evaluating Long Contexts 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
{xmedved1, xsabol, hales}@fi.muni.cz

Abstract. In the search for the answer to an open-domain question, the size of the search window, or the answer context, can greatly influence the resulting determination of the answer. The presented paper offers a detailed evaluation of different sizes of the answer context in case of Czech question answering. We compare six different context types in four different lengths. The conclusion of the experiments is that prolonging the context can improve the precision for specific types but in general the best results are obtained with one-sentence contexts.

Keywords: Question answering · Answer selection · Answer context · Evaluation

1 Introduction

The longer the preceding answer context is, i.e. the more we know about the question subject in advance, the more precise and certain the sought answer is. At least, this is a common assumption for the way how people search for an answer. In the computer Question Answering (QA) task, the benefits of longer contexts has not yet been thoroughly evaluated.

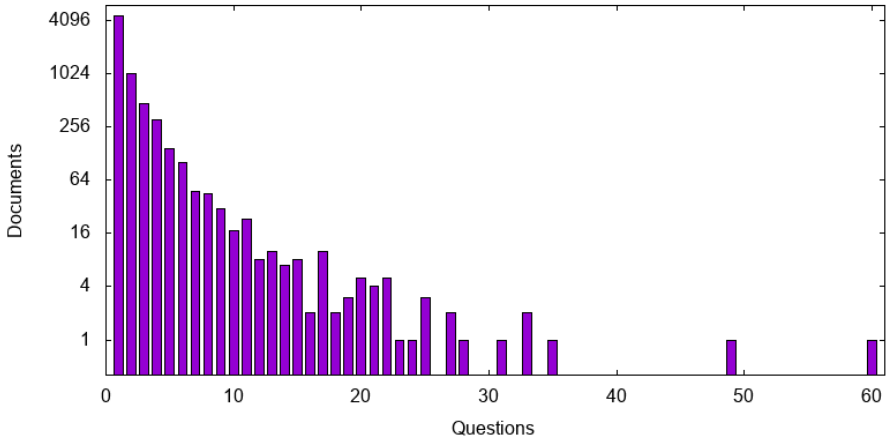
In this paper, we try to find the best answer context length experimentally. We evaluate and compare six different answer contexts setups each in four different lengths. The evaluation uses the Simple Question Answering Database (SQAD [4,6]) in version 3.1 and the results are compared with the answer selection task, i.e. the identification of the right document sentence which contains (or supports) the exact answer phrase.

To improve system performance, several related works examine context as a source of additional information. In [13], the authors used entities recognized in the question and a candidate concept and created an entity description based on Wiktionary definition. Afterwards, they employed this external entity descriptions to provide contextual information for knowledge understanding and achieved best results among non-generative models.

In [3], the authors modified BiDAF's [10] passage and question embedding processes to use the context information. According to their experiments, the context enhanced model outperformed the standard setup.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021, pp. 61–69, 2021.

Fig. 1: Histogram of the numbers of questions per document.



2 Contexts in the SQAD Dataset

The latest Simple Question Answering Database (SQAD) introduced in [9,7] consists of 13 473 records created from 6 898 different Czech Wikipedia articles. Figure 1 displays the actual frequencies of the numbers of questions per documents.

The detailed statistics concerning the average length of sentence and question in the SQAD database are introduced in Table 1.

In the latest update of SQAD v. 3.1 introduced in [7], the database is enriched by contextual data in two main forms. Recurrent network (RNN) word embeddings are used as the first group of contexts that are added to each sentence during learning. They are formed by a sequence of individual word vectors to be concatenated with the candidate answer sequence during the learning process. The first sentence uses the text title as a context, because in many cases the title carries important information.

The second group of contexts is based on BERT-based sentence embeddings that are added into the model as one vector obtained from BERT model. In the

Table 1: SQAD text and question length statistics

Type	In tokens
Average text sentence length	20.18
Max text sentence length	205
Min text sentence length	1
Average question sentence length	8.22
Max question sentence length	43
Min question sentence length	1

experiments, several BERT based pre-trained modes have been used to encode the content of previous sentences.
The available context types that are used in the training phase are:

- RNN context types:
  - list of previous sentences (SENT)
  - list of link named entities[1] extracted from previous sentences (NE)
  - list of noun phrases extracted from previous sentences (PHR)
- transformer contexts types (transformer encodes previous sentences):
  - Czert [11] (CZT)
  - RobeCzech [12] (RC)
  - Slavic BERT [2] (SLB)

Each context type can be used in different sizes. Table 2 shows average context length in terms of tokens and items (item can be phrase, named entity or sentence) for RNN contexts. The transformer based only uses N vectors. The length determines how far back in text the context is calculated. The context length can have different impact to the final system performance (as we can see in Section 4). Additional features learned from context can therefore improve or degrade the final answer selection module performance used in the AQA [8] system.

Table 2: Average context lengths (in tokens) and average numbers of context items (e.g. number of different phrases) per the variable context window

Context type	context window (sentences)	average length of context in tokens	average number of context items
NE	1	2.29	1.49
	2	4.48	2.97
	3	6.70	4.45
	4	8.93	5.93
	5	11.16	7.41
PHR	1	13.77	5.08
	2	27.71	10.22
	3	41.55	15.33
	4	55.42	20.45
	5	69.30	25.58
SENT	1	19.97	1.00
	2	40.12	2.00
	3	60.24	3.00
	4	80.41	4.00
	5	100.60	5.00

Table 3: Running times of experiments with respect to the context type and window

Context type	Time (h)			
Window Size	1	2	3	4
PHR	10.75	14.4	18.2	20.81
NER	9.82	10.32	10.81	12.18
SENT	11.32	13.8	18.09	20.48
Transformer	13.56	13.71	13.88	13.96

3 Experiments

The answer selection module performs a ranking task, where each sentence of a document obtains a score according to its semantic similarity to the question. The neural network input is a triplet of a *question*, a *candidate answer*, and its *context*. Both the question and the answer are represented as a sequence of 500-dimensional *word2vec* word embeddings, while the context representation depends on the current context type as described in Section 2.

The first step utilizes a *Bidirectional Gated Recurrent Unit* (*BiGRU*) network to re-encode both the question and answer sequences into a hidden representation where their position in the sequence enriches each token. For RNN contexts, the same BiGRU layer is used to transform them into their hidden representation. However, a separate BiGRU has to be used instead for the transformer contexts, as the sequences are derived from a different language model. In both cases, the resulting hidden context vectors are concatenated to the candidate answer.

The following process involves an attention layer that assigns an importance score to each question token according to its importance in the answer and vice versa. This process also applies to both transformer and RNN contexts at the tail of the answer sequence (for example, an importance score can be assigned for the entire previous sentence vector in the transformer context).

The created attention vectors are multiplied with their corresponding hidden sequence. They result in two equally sized vectors, where their cosine similarity is the final ranking for the input triplet.

The SQAD dataset is partitioned into train/validation/test sets in the ratio 60:10:30. The partitions are balanced with regards to the ratio of question and

[1] See [5] for details about the specific named entity recognition technique.

Table 4: The best hyperparameter values for various context types

Context Type	BiGRU Hidden Size	Learning Rate	Dropout
SENT RNN context	380	0.0004	0.4
PHR RNN context	380	0.0002	0.4
NER RNN context	320	0.0006	0.4
SENT Transformer ctx	480	0.0007	0.2

Table 5: Mean average precision for each context type and context window size

Context type	Mean Average Precision							
Window Size	1		2		3		4	
MAP	S	M	S	M	S	M	S	M
PHR	*82.24*	*84.92*	82.23	84.98	80.56	83.41	80.55	83.31
NER	82.58	85.3	82.16	84.94	*82.71*	*85.53*	82.4	85.04
SENT	*81.9*	*84.76*	80.9	83.39	79.31	82.2	78.54	81.56
CZERT	**83.39**	**85.79**	82.71	85.38	82.76	85.36	82.78	85.35
ROBECZECH	*82.75*	85.29	82.46	85.05	82.69	*85.44*	82.56	85.14
SLAVIC_BERT	83.05	85.59	*83.19*	*85.91*	82.74	85.49	82.88	85.55

answer types. The training partition contains 8,059 records and is used to optimize the weights of the model. The validation set has 1,401 records and is used for an unbiased evaluation and early stopping (models are trained on 25 epochs, but the epoch with the best validation accuracy will be chosen as the result). The test set contains 4,013 records and is used for the final evaluation of the model.

We will refer to the number of preceding sentences from which the context is derived as the context *window size*. The primary goal of the following experiments is to determine the most optimal context window for each context type, and compare their performances. For this purpose, a window size from 1 to 4 is used for each type of context presented in Section 2. Larger context windows (`PHR_5` or `SENT_5`) could not be realized due to the technical limitations of the GPU. Each of the setups is repeated three times where the resulting mean average precision (MAP) score is recorded as the result of all runs.

4 Results and Discussion

The experiments were performed on Metacentrum adan clusters and were accelerated using the NVIDIA Tesla T4 graphics cards. Table 3 shows differences

Table 6: Best models per question type with different context types

Question type	Non context MAP in %	best context	window	best MAP in %	worst context	window	worst MAP in %
VERB_PHRASE	82.64	NE	3	**83.63**	SENT	4	76.71
ENTITY	79.40	SLB	1	**81.62**	SENT	4	75.47
NUMERIC	78.50	NE	1	**79.79**	SENT	4	72.95
ADJ_PHRASE.	83.89	SLB	1	**84.19**	SENT	4	79.53
CLAUSE	74.82	SLB	2	**75.78**	SENT	4	66.19
DATETIME	84.52	CZT	1	**84.80**	SENT	3	79.93
LOCATION	83.13	CZT	1	**86.61**	SENT	4	81.83
PERSON	81.33	CZT	1	**85.17**	SENT	3	81.59
ABBREVIATION	91.75	NE	4	**94.16**	SENT	2	90.03

Question:
Kolik sportovců se zúčastnilo XXVIII. letních olympijských her 2004 v Aténách?
[*How many athletes participated in the XXVIII-th Summer Olympic games in 2004 in Athens?*]
Answer from non-context model:
Her se zúčastnilo 202 zemí.
[*202 countries took part in the games.*]
Answer from the NER context model (window size of one sentence)
Účastnilo se jich 10625 sportovců z 201 zemí světa.
[*10625 athletes from 201 countries took part in them.*]
1st context item
letní olympijské hry
[*Summer Olympic games*]
2nd context item
Athénách
[*Athens*]

Fig. 2: An example answer where the NER context improved the system performance (record 000252)

in running times for various types of context. We can observe that in RNN contexts, the running time increases substantially with the increasing context window. For the transformer context, the running times are overall longer due to the additional BiGRU layer, which brings more parameters to optimize for the model. However, the increase in running times w.r.t window size is minimal as these contexts have more compact representations than the RNN ones.

The hyperparameters of the model were optimized semi-automatically using the Optuna hyperparameter optimization framework [1]. The original hyperparameter values from [7] have been used with increased context sizes. The list of the parameter setups per context can be seen in Table 4.

Table 5 presents the results for each context type and window size. The MAP scores in the *S* columns refer to the version where each record assumes only one single correct answer in the document, while *M* refers to the version where any sentence containing the exact answer is a correct answer, i.e. multiple correct answer sentences are allowed. The best result of each row is in italic, while the best result globally is in bold font. For the `PHR` and `SENT` contexts, the performance gradually degrades with the increasing context window. The decrease is due to a large number of tokens in the context, making it more difficult for the model to capture the dependencies of the sequence items. The `NER` context is more compact and produces slightly better results for the window size 3.

For the transformer contexts, a slight improvement in accuracy with the `RobeCzech` model and window size of 2 are recorded. Otherwise, the window size of 1 results in the best performance. Overall, the best setup uses the `Czert` transformer context with window size 1 and achieves the MAP score of 83.39 % in the single answer setup and 85.79 % in the multiple answers setup.

Question:
Je Jeruzalém jedno z nejstarších měst na světě?
[*Is Jerusalem one of the oldest cities in the world?*]
Answer from the non-context model:
Historie města sahá až do 4. tisíciletí př. n. l. a činí tak z Jeruzaléma jedno z nejstarších měst na světě.
[*The history of the city dates back to the 4-th millennium BC and makes Jerusalem one of the oldest cities in the world.*]
Answer from the SENT context model (window size of 4 previous sentences)
Nachází se v něm však také množství významných starověkých křesťanských míst a je považováno za třetí nejsvětější místo islámu.
[*However, there us also located a number of important ancient Christian sites and is considered the third holiest site in Islam.*]
1st context item
Jeruzalém se nachází v Judských horách na hranici úmoří Středozemního a Mrtvého moře na okraji Judské pouště.
[*Jerusalem is located in the Judean Mountains on the border of the Mediterranean and the Dead Sea, on the edge of the Judean Desert.*]
2nd context item
Současný Jeruzalém se rozrůstá daleko za hranicemi Starého Města.
[*Today's Jerusalem is growing far beyond the Old City.*]
3rd context item
Historie města sahá až do 4. tisíciletí př. n. l. a činí tak z Jeruzaléma jedno z nejstarších měst na světě.
[*The history of the city dates back to the 4-th millennium BC and makes Jerusalem one of the oldest cities in the world.*]
4th context item
Jeruzalém je nejsvětějším místem judaismu a duchovním centrem židovského národa.
[*Jerusalem is the holiest site of Judaism and the spiritual center of the Jewish nation.*]

Fig. 3: An example answer where longer sentence context degraded the system performance (record 009720)

We have also evaluated the answer selection module performance (mean average precision – MAP) with the new context types in relation to different question types. Table 6 reveals a significant improvement in the module performance when supplying some context to the training phase. A comparison among the question type results shows that two transformer contexts and one RNN context outperform the other context types. While also here for most question types the shorter context windows win, the NE model achieves the best performance for *verb phrase* with window size 3 and for *abbreviations* with window size 4. Presumably, these question types are frequently explained in longer texts than the other types of questions. The SENT context with large window sizes significantly decreases the module performance.

Examination of the results shows why the named entities (NE) context improves the module performance. Figure 2 shows that named entities extracted from previous sentences provide the important additional information that

helps the system to choose the right sentence. The entities of *Summer Olympic games* and *Athens* resolve the anaphora appearing in the correct answer and finds the important antecedents contained in the question.

The Slavic BERT and Czert context types do not offer such explainable representation of the context. Overall, their dense sentence representation allows to encode the important aspects of the sentence even slightly better than the NE context even though they do not specifically point at the important pieces of information in the context.

On the other hand if we look on the performance of the SENT context model with window size of 4 previous sentences, we can see significant decrease in the final module performance. A specific example is presented in Figure 3, where the resulting sentence context is too long. This finally confuses the model with too much additional information. Also the context of the selected sentence contains the correct sentence which should have been selected as the correct answer.

5 Conclusions

In the paper, we have evaluated the assets of using several answer contexts in varying context lengths to solve the answer selection task. The results reveal that for specific question types, such as verb phrases or abbreviations, longer contexts in the form of important entities improve the performance. In all cases, the context representation is better than a model with no context information. However, in prevailing number of cases, the best context size uses just one preceding sentence as the source of context information and with widening the context window the benefits of using the context diminish and actually degrade the performance.

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

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2623–2631 (2019)
2. Arkhipov, M., Trofimova, M., Kuratov, Y., Sorokin, A.: Tuning multilingual transformers for language-specific named entity recognition. In: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing. pp. 89–93. Association for Computational Linguistics, Florence, Italy (Aug 2019).

https://doi.org/10.18653/v1/W19-3712, https://www.aclweb.org/anthology/W19-3712
3. Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W., Choi, Y., Liang, P., Zettlemoyer, L.: Quac : Question answering in context. CoRR **abs/1808.07036** (2018), http://arxiv.org/abs/1808.07036
4. Horák, A., Medveď, M.: SQAD: Simple question answering database. In: Eighth Workshop on Recent Advances in Slavonic Natural Language Processing. pp. 121–128. Tribun EU, Brno (2014)
5. Medved, M., Sabol, R., Horák, A.: Efficient Management and Optimization of Very Large Machine Learning Dataset for Question Answering. In: RASLAN 2020. pp. 23–34 (2020)
6. Medveď, M., Sabol, R., Horák, A.: Employing Sentence Context in Czech Answer Selection. In: International Conference on Text, Speech, and Dialogue, TSD 2020. pp. 112–121. Springer (2020)
7. Medveď, M., Sabol, R., Horák, A.: Comparing RNN and Transformer Context Representations in the Czech Answer Selection Task. In: Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022). SCITEPRESS, Setúbal, Portugal (2022), in print
8. Medveď, M., Horák, A.: Sentence and word embedding employed in open question-answering. In: Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018). pp. 486–492. SCITEPRESS - Science and Technology Publications, Setúbal, Portugal (2018)
9. Sabol, R., Medveď, M., Horák, A.: Czech question answering with extended SQAD v3.0 benchmark dataset. In: Proceedings of the 13th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2019. pp. 99–108. Tribun EU (2019)
10. Seo, M.J., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. CoRR **abs/1611.01603** (2016), http://arxiv.org/abs/1611.01603
11. Sido, J., Pražák, O., Přibáň, P., Pašek, J., Seják, M., Konopík, M.: Czert – Czech BERT-like Model for Language Representation. arXiv preprint arXiv:2103.13031 (2021)
12. Straka, M., Náplava, J., Straková, J., Samuel, D.: RobeCzech base (2021), http://hdl.handle.net/11234/1-3691, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
13. Xu, Y., Zhu, C., Xu, R., Liu, Y., Zeng, M., Huang, X.: Fusing context into knowledge graph for commonsense reasoning. CoRR **abs/2012.04808** (2020), https://arxiv.org/abs/2012.04808