Czech Question Answering with Extended SQAD v3.0 Benchmark Dataset

Radoslav Sabol, Marek Medveď, and Aleš Horák

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00, Brno, Czech republic
{xsabol, xmedved1, hales}@fi.muni.cz

Abstract. In this paper, we introduce a new version of the Simple Question Answering Databases (SQAD). The main asset of the new version lies in increasing the number of records to a total of 13,473 records. Besides the database enlargement, the new version incorporates new restrictions of specifying different formats of the expected answer for a given question. These new restrictions are connected with automatic database consistency checks where new sub-processes safeguard the database correctness and consistency.
We also introduce a new on-line annotation tool used which offered a unified environment for extending the SQAD data in a crowdsourcing experiment.

Keywords: question answering, QA benchmark dataset, SQAD, Czech

1 Introduction

The evaluation of question answering (QA) results has substantially improved in recent years. Detailed comparison of new state-of-the-art QA tools is now possible thanks to large shared benchmark datasets, especially for English. These datasets consists of thousands of records, which makes them an important resource in the QA field for both evaluation and training of unsupervised approaches.

The Stanford Question Answering Dataset (SQuAD [6]) is one of the best-known QA benchmark dataset. SQuAD consists of more than 100,000 questions with several correct answers for each question. The state-of-the-art tools reach more than 92% F1 score with this dataset.

The ReAding Comprehension from Examinations (RACE [3]) dataset is another benchmark dataset for the QA task which also consists of nearly 100,000 questions where each question has 4 candidate answers. Current evaluation result using this dataset reach nearly 90% F1 score.

Majority of current state-of-the-art QA tools use unsupervised machine learning approaches which make them dependent on large dataset such are those mentioned above. Without such large datasets, the unsupervised models are not able to provide good generalizations of the problem. A problem arises

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019, pp. 99–108, 2019.

when the successful approaches are to be applied to a non-mainstream language for which a corresponding dataset is unavailable.

In this paper, a new version of the Czech benchmark QA dataset called SQAD (Simple Question Answering Database) is introduced. The latest version is numbered 3.0 and consists of almost 13,500 question-answer pairs. Besides the enlargement of the dataset, the format of the record items is enhanced to satisfy two main requirements. The answer selection item[1] must consist of a single sentence from the text and this sentence must contain the expected answer. The answer extraction item (the expected exact answer) has to be a proper sub-phrase of the answer selection sentence. This new restrictions ensure database consistency and allow for automatic database correctness and consistency checks.

During the preparation of the new SQAD version, a new online annotation tool, named AddQA, for unifying the process of creating new SQAD records was developed.[2] This tool implements multiple sub-processes. Firstly, the indicated article content is downloaded and preprocessed from the Czech Wikipedia dump. Secondly, the selected text parts (questions, answer selection and exact answer) and the whole text are automatically annotated with morphological information and part-of-speech tags.

In the evaluation section, the final statistics of the new SQAD v3.0 database and the current results of the AQA [4] system with this new SQAD version are presented.

2 SQAD Version 3.0

The new version 3.0 of the Czech QA benchmark dataset SQAD features a substantial enlargement of SQAD v2.1 introduced in [8]. All SQAD versions comprise both the carefully processed manual annotations of questions and answers by human annotators and the database accompanying tools which automatically preprocess, store and check the data throughout the whole development process. From version to version, the tools are adapted to improve the uniqueness and consistency of the annotated data. In version 3.0, new web interface for annotators called AddQA was implemented and several new consistency checks were added to make the data preparation process more streamlined and less time consuming.

In the first part of this section, we describe new web based interface for annotators and in the second part we introduce rules (restrictions) according to which the new records have to be created.

2.1 New AddQA Online Annotation Tool

The process of adding new records to the SQAD dataset is quite complicated and multilevel. For a substantial increase in the number of the dataset records

[1] The sentence which contains the expected answer to the given question.

[2] All new records in SQAD v3.0 have been created by this new tool.

Remaining annotations: 198 of 200

Question type	Missing QA	
ABBREVIATION	**10** of 10	Add new ABBREVIATION QA
ADJ_PHRASE	**50** of 50	Add new ADJ_PHRASE QA
CLAUSE	**10** of 10	Add new CLAUSE QA
DATETIME	**9** of 10	Add new DATETIME QA
ENTITY	**28** of 28	Add new ENTITY QA
LOCATION	**9** of 10	Add new LOCATION QA
NUMERIC	**4** of 4	Add new NUMERIC QA
PERSON	**28** of 28	Add new PERSON QA
VERB_PHRASE	**50** of 50	Add new VERB_PHRASE QA

Already annotated: 2 of 200

Question ID	Q type	A type	Question	Answer	Sentence	URL
19	LOCATION	LOCATION	Kde se nachází Kuba?	v severním Karibiku	Kuba se nachází v severním Karibiku a její břehy omývají Karibské moře, Mexický záliv a Atlantský oceán.	https://cs.wikipedia.org/wiki/Kuba
21	DATETIME	DATETIME	Kdy se narodil Jeremy Clarkson?	11. dubna 1960	Jeremy Clarkson (* 11. dubna 1960), celým jménem Jeremy Charles Robert Clarkson, je anglický hlasatel, žurnalista a spisovatel, který se specializuje na motorismus.	https://cs.wikipedia.org/wiki/Jeremy_Clarkson

Fig. 1: AddQA overview page: display of the question types to create and a listing of previously created records.

(question-answer pairs), a new online interface, denoted as AddQA, was developed where the annotators do not deal with routine work (split text to sentences, annotate words, etc.) and concentrate only on the expert annotations that can not be automated. The AddQA process consists of multiple steps that lead to a final new SQAD record. As it was introduced in [8,2], each SQAD record consists of several files:

- the *question*
- the *full article* text
- the *answer selection* sentence which contains the expected answer as a sub-phrase
- the *answer extraction* result, i.e. the exact expected answer
- the *URL* of the original article in the Czech Wikipedia
- the QA *metadata* – the question and answer types

When entering all this information, each annotator has to go through the following steps:

- Create a new record or to edit a previously created one. In order to create a well balanced dataset, each annotator was assigned a predefined composition of the requested question type classes to add. The annotator thus knows in advance what kind of questions is needed and he or she can pick up a suitable article from Wikipedia. This is implemented in the first annotation phase displayed in Figure 1. The top part of the image informs the user how much records of each question type has to be added and the bottom part lists all records created by the user until now. Besides creating a new record, the possibility to adjust previously created question or answer is accessible from this page.

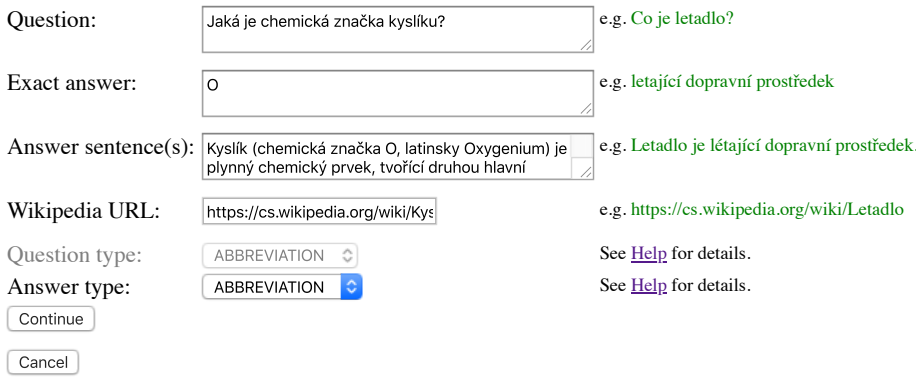

Fig. 2: The new record form for the question "Jaká je chemická značka kyslíku (What is the chemical symbol of oxygen)?"

- Fill the (new) record form. After selecting a predefined question type to add, the annotator is navigated to editing form. Here, the corresponding question information is connected to a text (the full document, the answer sentence and the exact answer) from a chosen Wikipedia article identified by its URL. The user also enters the type of the expected answer. An example for an ABBREVIATION question is presented in Figure 2.
- Final check of record correctness. When the filled form is submitted, two main sub-processes are triggered. The first one automatically uses the Wikipedia API to download the raw representation of the article. The second process takes care about an automatic annotation of sentence boundaries, lemmata, morphological categories and part-of-speech tags by the Unitok [5], Majka [7] and Desamb [9] tools. The final record is displayed in Figure 3.

The last but very important part of web annotation tool is help page where annotators can find description of question and answer types and see demo examples for each type to get main idea how to form a new one (see Figure 4).

3 SQAD Format Update

To be able to automatically check the dataset consistency, a set of rules for each new record has been defined. In the current version, these rules have been enforced by a semi-automatic batch processing. In a new AddQA version, they are planned to be incorporated in a form of an automatic sub-process.

The main goal of these restrictions is to create a coherent and consistent database. The first two restrictions are focusing on the answer selection part, the third rule handles the format of the expected answer.

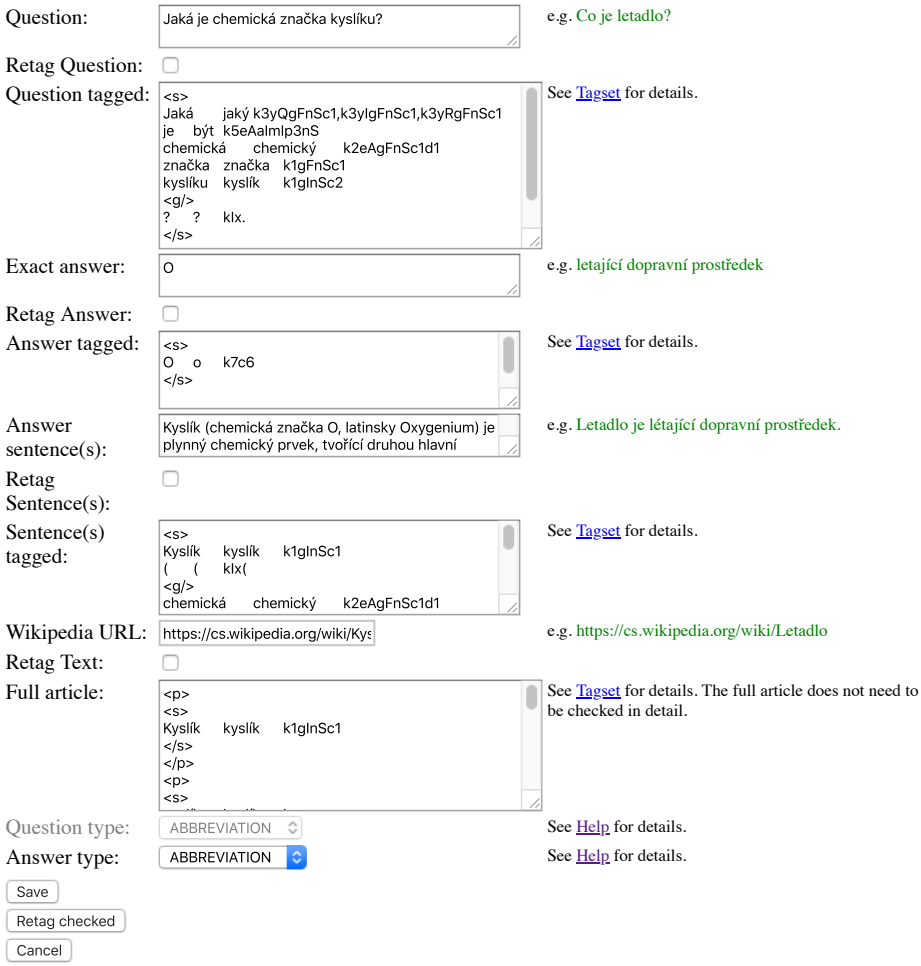

Fig. 3: Final check of record correctness.

3.1 The Answer Selection Format

A Wikipedia article is a conventional text therefore there are naturally connecting anaphoric references such as "*Peter was a famous singer. He was also a famous English song writer.*" Here "*Peter*" is an antecedent[3] of the pronoun "*He*". In case an annotator creates a question "*What is the name of a famous English song writer?*", an issue of choosing the of correct answer selection sentence arises. In previous versions of the SQAD database, such case could be annotated in three possible ways. The first option was to mark just the sentence containing the exact

[3] the target of an anaphoric reference

Question type	Description	Example question	Example Answer
ABBREVIATION	The question asks for abbreviation of some name.	Jakou chemickou značku má vápník?	Ca
ADJ_PHRASE	Question asks about some specific group of things, that is usually specified by adjective.	Jaká je tradiční barva Oxfordské univerzity?	tmavě modrá
CLAUSE	Question is general and the answer can be any general clause.	Proč se chtěla Marie Terezie spojit s Francií?	Protože by Prusko ztratilo svého důležitého spojence
DATETIME	Main goal of question is to determine certain point in time.	Kdy se narodila Petr Kvitová?	8. března 1990
ENTITY	Main goal of question is to name a thing (not a person) that meets all conditions of question.	Jak se jmenuje největší planeta sluneční soustavy?	Jupiter
LOCATION	Main goal of question is to determine certain place.	Kde zemřel Josef Kajetán Tyl?	Plzeň
NUMERIC	Main goal of question is to determine certain number.	Do kolika větví je rozdělena Armáda České republiky?	Do tří
PERSON	Main goal of question is to name a person that meets all conditions of question.	Kdo byl 33. prezidentem Spojených států amerických?	Harry S. Truman
VERB_PHRASE	Main goal of question is to find out if something happened. The answer can be general verb phrase or confirmation in form of YES/NO.	Je Brno sídlem fotbalového týmu Bohemians 1905?	ne

Answer type	Description	Example question	Example Answer
ABBREVIATION	The answer is abbreviation of some name.	Jakou chemickou značku má vápník?	Ca
DATETIME	Answer is a certain point in time.	Kdy se narodila Petr Kvitová?	8. března 1990
ENTITY	Answer is a name of some thing (not a person).	Jak se jmenuje největší planeta sluneční soustavy?	Jupiter
LOCATION	Answer denotes a certain place.	Kde zemřel Josef Kajetán Tyl?	Plzeň
NUMERIC	Answer is a number.	Do kolika větví je rozdělena Armáda České republiky?	Do tří
ORGANIZATION	Answer is name of some organisation, band, company ...	Frontman jaké kapely je Jarda Svoboda?	Traband
OTHER	Answer is a general phrase that is not belong to other answer type	Co je hard rock?	hudební styl
PERSON	Answer is a name a person that meets all conditions of question.	Kdo byl 33. prezidentem Spojených států amerických?	Harry S. Truman
DENOTATION	Answer is a general name of a field, area or an approach.	Jak se nazývá obor, který se zabývá studiem chování živočichů?	Etologie
YES/NO	The answer is YES or NO.	Je Brno sídlem fotbalového týmu Bohemians 1905?	ne

Fig. 4: Question and answer type description for annotators

answer[4] (in this case "*Peter was a famous singer*"). The second possible answer selection contained the sentence which corresponds directly to the question but the answer is just referred by the anaphora (for the example "*He was also a famous English song writer*"). And the third possibility annotated both these sentences as the answer selection. That is why this process has been unified in SQAD 3.0 with the following three rules:

1. The answer selection must contain exactly one sentence.
2. The answer selection sentence must contain the expected answer (as a sub-phrase).
3. If multiple sentences (with anaphora) are needed to uniquely identify the answer to the question, the answer sentence and the respective antecedent sentence(s) are stored in a new "question_context" file.

3.2 The Expected Answer Format

After choosing the answer selection sentence, the annotator is to demarcate the expected answer (answer extraction in the SQAD terminology) as a part of the annotated sentence. The answer should be as short as possible but still contain enough information to answer the question. In previous SQAD versions, the

[4] The shortest answer for the given question. In this example, the exact answer is "*Peter*".

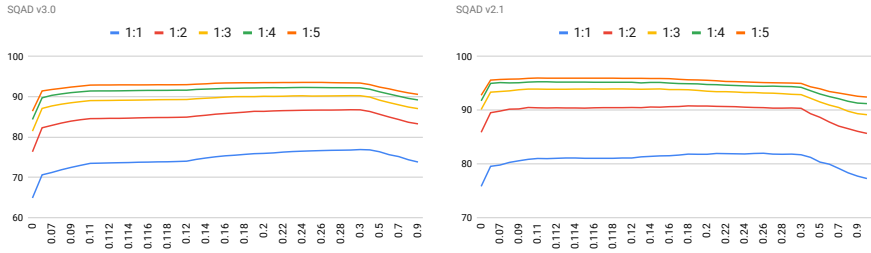

Fig. 5: Comparison of SQAD v2.1 and SQAD v3.0 on the document selection level (combined score on top 5 documents)

expected answer sometimes just "followed" from the answer sentence, but it was not a concrete sub-phrase of it. For the same reasons of consistency as mentioned in the previous section, the exact answer is now expressed in two forms. The original (expected) *answer* continues to be in the form as being formulated by a human after reading the given question and the answer selection (plus possible context) sentence(s). A new "answer_extraction" file now contains the real exact sub-phrase of the answer selection sentence (in the same form as stated in the text) and the *answer* should be its reformulation (or just a copy). This allows for extra automatic consistency checks of the answer extraction annotation.

Table 1: SQAD v3 statistics

	SQAD v3.0	SQAD v2.1
No. of records	13,473	8,566
No. of different articles	6,571	3,930
No. of tokens (words)	28,825,824	20,288,297
No. of answer contexts	378	0

Q-Type statistics:	SQAD v3.0	SQAD v2.1
DATETIME	14.7 %	21.6 %
PERSON	13.1 %	11.9 %
VERB_PHRASE	16.8 %	10.97 %
ADJ_PHRASE	11.2 %	2.7 %
ENTITY	18.4 %	20.4 %
CLAUSE	3.5 %	2.8 %
NUMERIC	7.3 %	10.7 %
LOCATION	12.4 %	17.8 %
ABBREVIATION	2.5 %	0.95 %
OTHER	0.1 %	0.18 %

A-Type statistics:	SQAD v3.0	SQAD v2.1
DATETIME	14.6 %	21.5 %
PERSON	13.2 %	12.3 %
YES_NO	16.8 %	10.95 %
OTHER	16.7 %	9.6 %
ENTITY	13.1 %	12.7 %
NUMERIC	7.4 %	10.7 %
LOCATION	12.3 %	17.6 %
ABBREVIATION	2.4 %	0.96 %
ORGANIZATION	2.1 %	2.5 %
DENOTATION	1.4 %	1.2 %

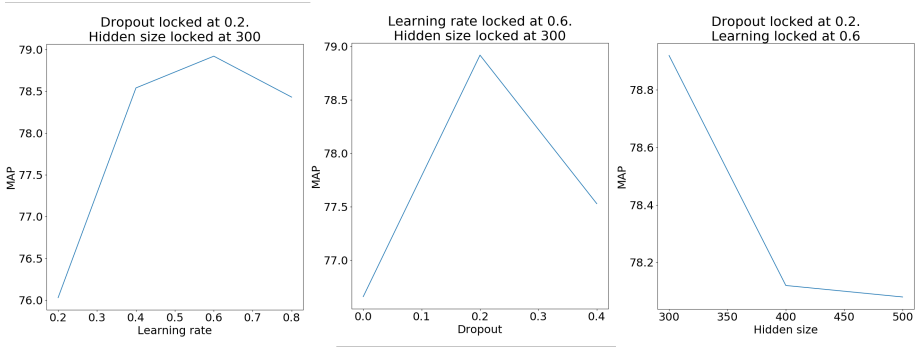

Fig. 6: The sensitivity graphs of the hidden size, dropout and learning rate hyperparameters in the SQADv3 evaluation.

4 Evaluation

The new SQAD version 3.0 is larger than all previous versions. It contains 13,473 records (question-answer pairs) which is almost 5,000 more than SQAD v2.1. For fine-grained statistics about the new version see Table 1.

The first tests of the Automatic Question Answering (AQA [4]) tool evaluated separately the document selection and the answer selection modules. The results of the document selection module are graphically illustrated and compared to the previous version in Figure 5.

The parameters of the AQA answer selection were adjusted according to their performance with SQADv2. The experiments included hidden representation vector dimensions of 300, 400, and 500 (400 was the most successful with the previous version of the dataset). The dropout values ranged for 0 (no dropout), 0.2 and 0.4 (0.6 was omitted for previous low performance). The learning rate range was extended to cover the values of 0.6 and 0.8, due to fact that 0.4 was the

Table 2: The SQADv3 answer selection accuracy per question and answer types

Question type	Count	MAP (%)	Answer type	Count	MAP (%)
VERB_PHRASE	546	80.06	YES_NO	539	79.73
NUMERIC	212	74.13	NUMERIC	215	73.63
ADJ_PHRASE	363	79.08	OTHER	526	76.79
CLASUE	99	67.81	DATETIME	470	81.88
DATETIME	473	82.12	LOCATION	415	84.87
ABBREVIATION	71	78.89	ENTITY	403	76.47
LOCATION	417	84.76	PERSON	421	77.53
ENTITY	571	77.16	ABBREVIATION	66	78.57
PERSON	418	77.41	ORGANIZATION	65	75.58
OTHER	1	33.33	DENOTATION	51	87.93

Table 3: Answer selection Precision at k ($P@k$)

Position k	Count	$P@k$	Position k	Count	$P@k$
1.	3,171	79.00%	6.	31	0.77%
2.	377	9.39%	7.	20	0.50%
3.	125	3.11%	8.	16	0.40%
4.	66	1.64%	9.	13	0.32%
5.	59	1.47%	$\geq$ 10.	136	3.40%

best and also the maximum value for previous runs, so the current experiments combined the learning rates of 0.2, 0.4, 0.6, and 0.8.

The training and evaluation procedure remained the same as in SQADv2. For this purpose, SQADv3 was partitioned into the train, validation, and test sets with the ratios of 60:10:30. All these sets comprise balanced proportions of the question and answer types. The answer selection models were trained for 25 epochs using Stochastic Gradient Descend [1] process, where the weights with the best performance with the validation set were applied to the test set. All 36 possible combinations of parameters were evaluated three times, which means that overall 108 models were trained. The best parameter combination for SQADv3 had the hidden size value of 300, the dropout value of 0.2, and the initial learning rate of 0.6. The Mean Average Precision (MAP) of this combination with the test set was **78.92%**, and the Mean Reciprocal Rank (MRR) of **85.95**. The sensitivity of various model parameters can be found in Figure 6. Table 2 shows the accuracy results per question/answer types. Precision at position k ($P@k$) of one of the three best performing models can be seen in Table 3.

To evaluate the dataset independently of the predefined train-validate-test split, 5-fold cross validation test were run with both SQADv2 and SQADv3 with the results presented in Table 4. The technique involves splitting the dataset to n (in this case n=5) equally sized partitions. For each of n runs, one of the partitions is labeled as test set with its own validation set (300 questions) while all other serve as the training set. The overall MAP is then computed as the average value of all n trained models reaching **77.68** which is a 0.51% improvement when compared to SQADv2. Th 5-fold cross validation was performed only with the best parameter combination for both versions of the dataset (using Adadelta optimizer).

Table 4: Results of 5-fold cross validation for both versions of the SQAD dataset

Test partition no.	SQADv3.0 MAP	SQADv2.1 MAP
1	82.76%	87.16%
2	74.26%	81.28%
3	71.98%	76.20%
4	80.85%	73.85%
5	78.55%	67.35%
overall	**77.68%**	**77.17%**

5 Conclusions and Future Work

In this paper, we have introduced a new version of the Czech benchmark dataset called Simple Question Answering database (SQAD). The new version 3.0 contains almost 13,500 records. In addition to new content, a new AddQA online tool for annotations of new SQAD records was presented.

In the future work, the AddQA online annotation tool is planned to incorporate all the consistency restrictions described in Section 3.

We have also presented the first results of the document selection and answer selection modules with the SQADv3 dataset reaching the best results of 78.92% mean access precision (MAP) and 85.95 mean reciprocal rank (MRR).

Acknowledgements This work has been partly supported by the Czech Science Foundation under the project GA18-23891S.

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. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010, pp. 177–186. Springer (2010)
2. Horák, A., Medved', M.: SQAD: Simple Question Answering Database. In: Eighth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2014. pp. 121–128. Tribun EU, Brno (2014)
3. Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: RACE: Large-scale ReAding Comprehension Dataset From Examinations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 785–794 (2017)
4. Medved', 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)
5. Michelfeit, J., Pomikálek, J., Suchomel, V.: Text tokenisation using unitok. In: RASLAN 2014. pp. 71–75. Tribun EU, Brno, Czech Republic (2014)
6. Rajpurkar, P., Jia, R., Liang, P.: Know What You Don't Know: Unanswerable Questions for SQuAD. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 784–789. Associationfor Computational Linguistics, Melbourne, Australia (2018)
7. Šmerk, P.: Fast Morphological Analysis of Czech. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009. pp. 13–16 (2009)
8. Šulganová, T., Medved', M., Horák, A.: Enlargement of the Czech Question-Answering Dataset to SQAD v2.0. In: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2017. pp. 79–84 (2017)
9. Šmerk, P.: K počítačové morfologické analýze češtiny (in Czech, Towards Computational Morphological Analysis of Czech). Ph.D. thesis, Faculty of Informatics, Masaryk University (2010)