

# Question and Answer Classification in Czech Question Answering Benchmark Dataset

Daša Kušniráková, Marek Medved' and Aleš Horák

*Natural Language Processing Centre, Faculty of Informatics, Masaryk University*

**Keywords:** Question Answering, Question Classification, Answer Classification, Czech, Simple Question Answering Database, SQuAD.

**Abstract:** In this paper, we introduce a new updated version of the Czech Question Answering database SQuAD v2.1 (Simple Question Answering Database) with the update being devoted to improved question and answer classification. The SQuAD v2.1 database contains more than 8,500 question-answer pairs with all appropriate metadata for QA training and evaluation. We present the details and changes in the database structure as well as a new algorithm for detecting the question type and the actual answer type from the text of the question. The algorithm is evaluated with more than 4,000 question answer pairs reaching the F1-measure of 88% for question typed and 85% for answer type detection.

## 1 INTRODUCTION

Open domain question answering (QA) systems have seen a jump in the accuracy recently, mostly with employing recurrent neural networks (Wang et al., 2018; Hu et al., 2018) and large benchmarking datasets, e.g. SQuAD (Rajpurkar et al., 2016). The availability of these large datasets containing hundreds of thousands QA pairs allows to learn the necessary features for all QA stages based on word-level or character-level neural network processing of the question and answer sentences. In case of non-mainstream languages, preparation of such large QA datasets is not feasible. The current state-of-the art algorithms cannot be easily transferred to a less-resourced language without the accuracy drop of 5-40%, see e.g. (Pamela et al., 2010).

With morphologically rich languages, a QA system may exploit linguistic syntax-based ques and search for a syntactic-semantic match between the question and answer phrases. With an example dataset for the Czech language, as a representative of such morphologically rich languages, the syntax based approach is being developed and evaluated providing a correct or partially correct answer in 46% with the Czech SQuAD v1.1 (Medved' and Horák, 2018).

In the following text, we present a new version v2.1 of the Czech SQuAD (Simple Question Answer-

ing Database). The database was developed as a benchmarking dataset of question-answer pairs based on the Czech Wikipedia texts with rich structured annotations related to all QA subtasks (question classification, answer selection, answer extraction). The new version SQuAD v2.1 consists of 8,566 questions and offers a new updated system of question type and answer type specification.

In the next section, we describe the current structure of the SQuAD database. Section 3 introduces the presented question/answer type extraction tool, which is evaluated in detail in Section 4.

## 2 THE SQuAD DATABASE

The Czech Simple Question Answering Database, or SQuAD (Horák and Medved', 2014; Šulganová et al., 2017), is a question-answering (QA) benchmarking dataset resource consisting of human made question-answer pairs based on the content of Czech Wikipedia articles. The intended application of this data source is to provide a consistent and representative data source for model training and tool evaluation in Czech, as a morphologically rich language that allows to use syntax-based clues in the QA process.

The SQuAD v2.1 database currently contains 8,566 question-answer pairs, which are related to the content of 3,149 Czech Wikipedia articles. The underly-

ing texts of all these articles form a corpus of more than 20 million words. The SQuAD database is organized in structured records (one QA pair corresponds to one record) consisting of 6 items:

- the *question*,
- the *correct answer* (as can be extracted from the document),
- *answer selection* – the context of the correct answer, one or two sentences,
- the full *article text*
- the *source URL* in Wikipedia
- *question-answer metadata*, which contain the *types of the question and of the correct answer*.

All texts are provided in both the plain text and the annotated data in the vertical format (see Figure 1), which supplements a word with lemma and Part-Of-Speech (POS) tag. The POS annotation was obtained by the Czech automatic POS tagger Desamb (Šmerk, Pavel, 2010; Šmerk, 2009) with subsequent manual checking and corrections. A schema of the SQuAD database structure is presented in Figure 2.

<i>word/ token</i>	<i>lemma</i>	<i>POS tag</i>
<s>		
V	v	k7c6
jakém	jaký	k3yRgInSc6
roce	rok	k1gInSc6
vznikla	vzniknout	k5eAaPmAgFnS
kapela	kapela	k1gFnSc1
Rammstein	Rammstein	k1gMnSc1
<g/>		
?	?	kIx.
</s>		

Figure 1: Vertical format of the question “V jakém roce vznikla kapela Rammstein? (In what year was the Rammstein band formed?)”.

The correct answer contains a smallest multiword or numeric expression, which denotes the full expected answer as specified by the question and as present in the document text. This item is the expected result of the final QA module of the answer extraction subtask. The answer selection item represents a sentence in the article text, which contains the correct answer. In case the sentence needs larger context to disambiguate pronouns, there may be usually one extra adjoining sentence in this item. Altogether there are 9,289 sentences in the answer selection items in the whole database leading to an average of 1.08 sentences per each QA pair. The correct answer is always a sub-phrase of the answer selection text.

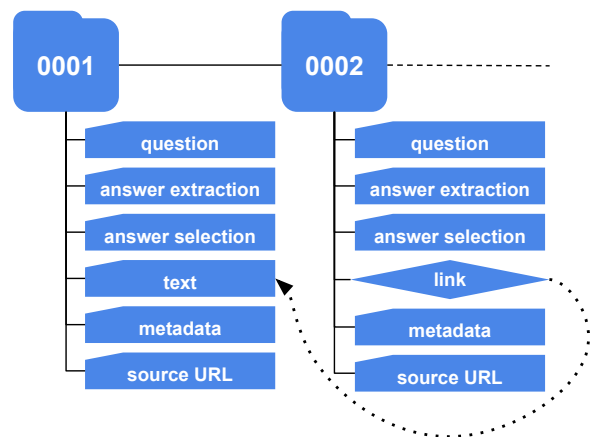


Figure 2: A SQuAD record schema visualization.

## 2.1 Question and Answer Types in SQuAD

Since the introduction of the extended version SQuAD v2.0 containing more than 8,000 QA pairs, the dataset was annotated with classification of each record into categories for the question type and the actual correct answer type. The sets of possible types (Šulganová et al., 2017) took inspiration from the large benchmark dataset for English, the Stanford Question Answering Dataset (Rajpurkar et al., 2016).

Further checks of the manual annotation of the QA types from the previous version have shown the need for a distinction of a substantial proportion of the general types denoted as OTHER and to reclassify questions starting with a relative clause. As presented in Table 1 and Table 2, the main changes are with the questions of type "CLAUSE", which has been distributed among more specific classes, and the answer type "OTHER" which has been newly divided into more specific classes like "ABBREVIATION" and "DENOTATION".

Table 1: SQuAD v2.1 question type statistics.

Database	SQuAD v2.0	SQuAD v2.1
PERSON	940	1,023
ENTITY	1,436	1,745
ADJ_PHRASE	253	233
DATE/TIME	1,848	1,851
LOCATION	1,436	1,524
NUMERIC	900	913
ABBREVIATION	-	81
CLAUSE	774	241
VERB_PHRASE	944	940
OTHER	31	15

Table 2: SQAD v2.1 answer type statistics.

Database	SQAD v2.0	SQAD v2.1
PERSON	943	1,050
DENOTATION	-	102
ENTITY	811	1,085
OTHER	1,480	819
ORGANIZATION	199	216
DATE/TIME	1,847	1,845
LOCATION	1,442	1,511
NUMERIC	904	918
ABBREVIATION	-	82
YES/NO	940	938

The distribution of the question classes over the answer classes is displayed in Table 3, which also shows that (DATE/TIME, DATE/TIME), (LOCATION, LOCATION), (NUMBER, NUMBER), (PERSON, PERSON) and (VERB\_PHRASE, YES/NO) classes are very consistent in the question-answer type pairs.

### 3 QUESTION AND ANSWER TYPE DETECTION

The first tool that has been evaluated with the SQAD database is the AQA system (Medved' and Horák, 2016; Medved' and Horák, 2018). The current development version employs a new module for the question type and answer type detection, which forms an important part of the Question processor and aims for improvement of the Answer extraction module (see Figure 3 for the AQA system schema).

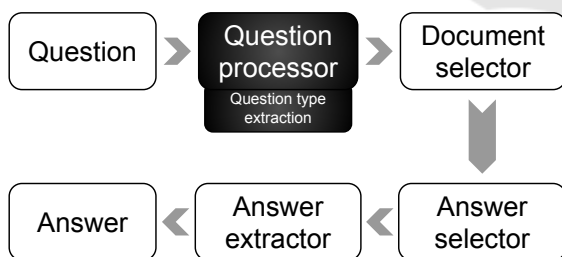


Figure 3: AQA system visualization.

The QA type detection algorithm is based on induced rules exploiting POS tagging and lexical features. Before the rules are applied to the input question, the system extracts the *question keyword*, which is represented by the main (head) question meaning noun.

The question keyword is recognized by the following three rules:

- If the relative pronoun "*který*" (which) or "*jaký*" (what) is present in the question and such word is not part of a relative sentence, then the candidate

question: 'Na jakém ostrově se nachází newyorský městský obvod Manhattan?' (On which island the Manhattan district is located?)

keyword: 'ostrov' (island)

hypernyms: ['ostrov', 'teritorium', 'obvod', 'region', 'poloha', 'lokace', 'entita'] (island, territory, district, region, position, location, entity)

rule: (LOCATION; LOCATION) -> "ostrov" in keyword.hypernym OR "teritorium" in keyword.hypernym OR ...

Figure 4: A question/answer type rule example: *if the question has "ostrov" (island) in keyword hypernyms or "teritorium" (territory) in keyword hypernyms or ..., then the question type is LOCATION and the answer type is also LOCATION.*

for the question keyword is the first noun after this word.

- Otherwise the first noun after the first verb in question is returned as the question keyword candidate.
- If the keyword candidate is one of words "*název*" (title), "*pojem*" (concept), "*termín*" (term), "*typ*" (type), "*část*" (part), or "*větev*" (branch), then the first following noun is returned as the final keyword, otherwise the candidate becomes the keyword.

After the question keyword extraction, the type detection rules are applied to the question. They combine three main features that are recognized from the question text and POS structures:

- question structure:  
Example: "k5" in words.tag\_at\_index(0) -> first word in the sentence is verb<sup>1</sup>
- important word recognition:  
Example: "<word>" in sentence
- keyword hypernym match:  
Example: "<word>" in keyword.hypernym

Keyword hypernyms are obtained by means of the Czech Wordnet API (Rambousek et al., 2017). Before the system is able to use the keyword hypernyms, a two-step process has to be executed. In the first step the system queries the Czech Wordnet to find all possible senses of the extracted keyword. For three most

<sup>1</sup>see (Šmerk, 2009) for more information about the POS tagset.

Table 3: SQUAD v2.1 distribution matrix of question and answer types.

Q type /A type	PER.	DENOT.	ENTITY	OTHER	ORG.	D./T.	LOC.	NUM.	ABB.	YES/NO
PERSON	1,016	0	2	0	3	0	2	0	0	0
ENTITY	20	101	1,031	378	204	1	7	1	2	0
ADJ_P.	7	0	8	216	0	0	0	2	0	0
D./T.	0	0	1	2	0	1,844	0	4	0	0
LOC.	1	0	14	5	3	0	1,501	0	0	0
NUM.	1	0	0	2	0	0	0	910	0	0
ABB.	0	1	0	0	0	0	0	0	80	0
CLAUSE	1	0	27	205	6	0	1	1	0	0
VERB_P.	2	0	1	0	0	0	0	0	0	937
OTHER	2	0	1	11	0	0	0	0	0	1

common word senses<sup>2</sup> the Wordnet API is queried and a list of available hypernyms for each meaning is created (see Figure 4 for an example of such rule).

The answer type detection then follows in two main steps:

- The input plain text question sentence is preprocessed and enriched by three important pieces of information:
  - automatic POS tagging
  - question keyword extraction
  - list of Wordnet hypernyms of the extracted keyword
- After the preprocessing part, the rule based analysis is performed. The algorithm traverses the predefined rules from the specific ones to the more general.

A schematic description of the QA type detection process is presented in Figure 5.

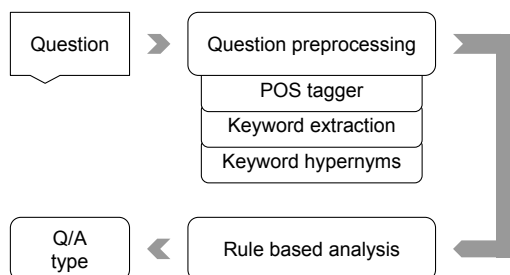


Figure 5: Question/Answer type detection schema.

The detection rules are in human readable and easy to edit form. They are applied step by step during the classification process. If a question meets the rule's conditions, then the appropriate labels are returned as the question and answer types.

<sup>2</sup>According to our tests, less than three senses give a too narrow list of hypernyms and, on the contrary, more than three senses give a too broad list of hypernyms, so both of them have a bad impact of system performance.

## 4 EVALUATION

In this section, we offer a thorough evaluation of the QA types detection with the new version of the SQUAD database. The database has been split into two equal parts – the development set and the testing set. The set splitting is properly balanced to maintain a similar representation of each question type present in both sets. The rules for QA types detection were developed using the development set consisting of 4,279 records and evaluated with the testing set.

The final evaluation is present in Table 4. The overall precision for both types combined is 82% with the answer type precision going up to 85%. The question type detection reaches both high precision and recall with the F1 measure of 88%. A detailed confusion matrix of all the expected and predicted question types is presented in Figure 5. We may see that the prediction of the ENTITY class is among the most complex ones as entities can be expressed in several ways. The detailed evaluation of the answer type detection is displayed in Figure 6, where the most misclassified classes are also ENTITY, OTHER and PERSON. This may call for further specification of the members of the OTHER class.

Table 4: QA types detection evaluation with SQUAD v2.1.

	precision	recall	F1
question t.	88.77%	87.79%	88.28%
answer t.	85.05%	84.52%	84.78%
both types	82.43%	82.93%	82.68%

### 4.1 Error Analysis

We have thoroughly analyzed the misclassified cases of the QA detection module and we have identified the following main error sources:

- The DATE/TIME answer type values are in some cases expressed as a single number representing just the year of the questioned event. Such

Table 5: Question type confusion matrix.

predicted	expected									
	AB	AJ_P	CLS	D/T	ENT	LOC	NUM	OTH	PER	V_P
ABBR	37	1	1	0	19	3	1	0	0	0
ADJP	1	52	4	0	49	6	6	0	4	0
CLAUSE	1	0	35	0	14	4	0	0	5	0
D/T	0	0	1	916	16	0	2	0	1	1
ENTITY	0	44	71	3	685	41	13	2	40	8
LOC	0	6	1	0	22	695	3	0	3	1
NUM	1	4	1	4	8	0	422	0	0	0
OTHER	0	1	3	2	25	7	7	5	3	6
PERSON	0	8	3	0	33	6	2	0	455	0
VERBP	0	0	0	0	0	0	0	0	0	454

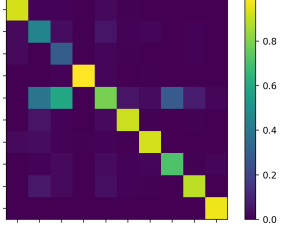
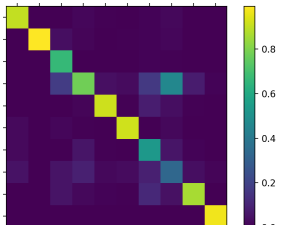


Table 6: Answer type confusion matrix.

predicted	expected									
	AB	D/T	DEN	ENT	LOC	NUM	ORG	OTH	PER	Y/N
ABBR	37	0	0	9	3	1	1	9	2	0
D/T	0	915	2	7	0	2	1	8	1	1
DEN	0	0	38	1	1	1	1	3	0	0
ENT	0	2	10	405	32	14	19	191	40	5
LOC	0	0	0	7	693	3	9	15	3	1
NUM	1	3	1	3	0	423	0	9	0	0
ORG	1	0	0	30	5	0	61	24	6	0
OTH	2	2	3	46	16	14	10	138	19	7
PER	0	0	3	12	7	2	13	18	452	0
Y/N	0	0	0	0	0	0	0	0	0	454



cases cannot be easily distinguished from the NUMERIC class. A possible solution to such situation would be to allow more than one correct answer type as a result of the detection serving as a hint to the Answer extraction module to look for all possible answer forms.

Example:

Q: Od kdy byla Ermesinda Lucemburská lucemburskou hraběnkou? (Since when was Ermesind Luxemburk the Luxemburg's countess?)

A: 1196

- Named entity recognition is currently not part of this module, which causes errors in questions that ask about PERSON or LOCATION but the detection recognizes it as an ENTITY. Since the base of the QA system already works with named entities, their processing in the QA type detection is a planned future step.

Example:

Q: Jak se původně jmenoval Alice Cooper? (What is the original name of Alice Cooper?)

A: Vincent Furnier

- The Czech Wordnet contains about 40,000 synonymical sets, but still about 20% of nouns that appear in common texts are not covered. In such cases, even when the question keyword is correctly determined, the keyword hypernym list is

empty so no appropriate rule can be applied to the question and consequently, wrong answer type is recognized. Such cases can be improved by employing broad coverage word embedding model to propose potential hypernym/synonym candidates to be used in the induced rules. Here again, the AQA system already employs word embeddings in phrase similarity detection, so this enhancement of unknown keyword rule induction is a logical future work.

## 5 CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced a new version of the Czech Question Answering database called SQA. The new version 2.1 incorporates improved question and answer type labeling that details several overly broad classes which have been divided into more fine grained classes reflecting the expected structure of the correct answer.

In the second part of the paper, we have described the implementation of the question and answer type detection used in the Question processor and Answer extraction modules of the question answering system AQA. The detection is based on a set of induced rules that drive the decision from lexical and structured in-



formation obtained by the question processing. The module was trained on the development set of 50% of the SQuAD questions and evaluated with the testing set of the same size. The resulting precision was 89% for question types and 85% for answer types with the respective recall of 88% and 85%. The combined overall F1 measure was 83%. The error analysis of the detection module directs the future work to employment of named entity recognition and word embedding similarity score for question keywords missing in the Czech Wordnet.

The open domain question answering system AQA was evaluated with the previous version of the SQuAD database, where it was able to point at the correct answer in 46%. The newly implemented question and answer type detection module aims at improving this result in the AQA evaluation. Apart from the answer type extraction module, a new module for AQA answer selection is currently in development and it is also planned for evaluation with the new SQuAD v2.1 benchmark dataset.

## ACKNOWLEDGEMENTS

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

## REFERENCES

- Horák, A. and Medved', M. (2014). SQuAD: Simple Question Answering Database. In *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2014*, pages 121–128, Brno. Tribun EU.
- Hu, M., Peng, Y., Huang, Z., Yang, N., Zhou, M., et al. (2018). Read + Verify: Machine reading comprehension with unanswerable questions. *arXiv preprint arXiv:1808.05759*.
- Medved', M. and Horák, A. (2016). AQA: Automatic Question Answering System for Czech. In Sojka, P. et al., editors, *Text, Speech, and Dialogue 19th International Conference, TSD 2016 Brno, Czech Republic, September 12–16, 2016 Proceedings*, pages 270–278, Switzerland. Springer International Publishing.
- Medved', 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.
- Pamela, F., Danilo, G., Bernardo, M., Anselmo, P., Rodrigo, Á., and Sutcliffe, R. (2010). Evaluating multilingual question answering systems at clef. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.
- Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, pages 2383–2392. Association for Computational Linguistics.
- Rambousek, A., Pala, K., and Tukačová, S. (2017). Overview and Future of Czech Wordnet. In McCrae, J. P., Bond, F., Buitelaar, P., Cimiano, P., 4, T. D., Gracia, J., Kernerman, I., Ponsoda, E. M., Ordan, N., and Piasecki, M., editors, *LDK Workshops: OntoLex, TIAD and Challenges for Wordnets*, pages 146–151, Galway, Ireland. CEUR-WS.org.
- Šmerk, P. (2009). Fast Morphological Analysis of Czech. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009*, pages 13–16.
- Šmerk, Pavel (2010). *K počítačové morfologické analýze češtiny (in Czech, Towards Computational Morphological Analysis of Czech)*. PhD thesis, Faculty of Informatics, Masaryk University.
- Šulganová, T., Medved', M., and Horák, A. (2017). Enlargement of the Czech Question-Answering Dataset to SQuAD v2.0. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2017*, pages 79–84.
- Wang, W., Yan, M., and Wu, C. (2018). Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1705–1714.