Enlargement of the Czech Question-Answering Dataset to SQAD v2.0

Terézia Šulganová, Marek Medveď, and Aleš Horák

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00, Brno, Czech republic
{xsulgan1, xmedved1, hales}@fi.muni.cz

Abstract. In this paper, we present the second version of Czech question-answering dataset called SQAD v2.0 (Simple Question Answering Database). The new version represents a large extension of our original SQAD database. In the current release, the dataset contains nearly 9,000 question-answer pairs completed with manual annotation of question and answer types.
All texts in the dataset (the source documents, the question and the respective answer) are provided with complete morphological annotation in plain textual format. We offer detailed statistics of the SQAD v2.0 dataset based on the new QA annotation.

Key words: question answering, QA dataset, SQAD

1 Introduction

Question Answering (QA) is a rapidly evolving field of Natural Language Processing (NLP) and Informatics. We may regards QA as a basis for next generation search engines. If we set aside natural language interfaces to database queries, each QA system uses a large enough knowledge base that provides information for the answer, often in the form of large textual documents.

In this paper, we introduce a new version of a QA evaluation dataset created from a document collection coming from the Czech Wikipedia. SQAD v2.0 is an largely extended version of the previous SQAD database [1] used in evaluation of a syntax based question-answering system AQA [2]. The key assets of the new dataset version are the 3× increased dataset size with nearly 9,000 question-answer pairs (the previous SQAD contained 3,301 QA pairs). During the development of the dataset, the original short answer contexts have been expanded into full Wikipedia articles.

Besides the exact answer phrases, the new corpus contains the full answer sentence, which can be used in a separated evaluation of the answer selection process. All QA pairs have now been also annotated for the question type and answer type via manual annotation process by two annotators.

From the technical point of view, the database was also modified to avoid document duplicities by applying symbolic links between texts shared across multiple records.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2017, pp. 79–84, 2017.

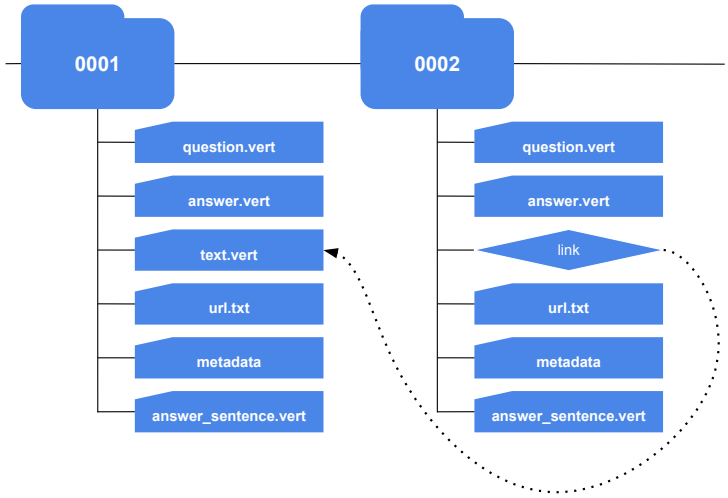

Fig. 1. SQAD v2.0 components structure

In the following text, we describe the whole process of developing the new SQAD v2.0. In section 2, we define the database structure and the data content. Section 3 describes the necessary automatic and manual adjustments of the SQAD v2.0 dataset including new metadata. In section 4, we provide statistics of token and sentence numbers, numbers of question and answer types.

2 Database Structure

Structure of the new SQAD database generally follows the previous one with some additional files. The previous version [1] consisted of files in plain text form denoting the original text, the question, the answer, Wikipedia URL and the question author. These input files were processed by morphological annotating pipeline formed by the Unitok [3] tokenizer and the Desamb [4] tagger.

The previous SQAD version did not allow easy sharing of manual modifications in either the underlying texts or the morphological annotation in case when there were multiple questions with the same source text. If the plain text form changed, the morphological annotation had to rebuilt and in case of manual corrections, all of them had to be reapplied again.

Therefore the new SQAD version has several changes in the files structure. The new version consists exclusively of vertical files, which represent a textual format of morphologically annotated text. A schema of two question-answer pairs is displayed in Figure 1. The standard plain text version of each component can be generated from this vertical files on request. `question.vert`, `answer.vert`, `answer_sentence.vert` and `text.vert` are sentences in the

Question:

word/ token	*lemma*	*POS tag*
<s>		
Z	z	k7c2
jakého	jaký	k3yQgInSc2
roku	rok	k1gInSc2
pochází	pocházet	k5eAaImIp3nS
školní	školní	k2eAgFnSc1d1
budova	budova	k1gFnSc1
v	v	k7c6
obci	obec	k1gFnSc6
Paršovice	Paršovice	k1gFnSc2
<g/>		
?	?	kIx.
</s>		

Answer sentence (part of):

word/ token	*lemma*	*POS tag*
<s>		
Další	další	k2eAgFnSc7d1
paměti-	paměti-	k1gFnSc7
hodností	hodnost	
v	v	k7c6
Paršovicích	Paršovice	k1gFnPc6
je	být	k5eAaImIp3nS
školní	školní	k2eAgFnSc1d1
budova	budova	k1gFnSc1
z	z	k7c2
roku	rok	k1gInSc2
1898	**#num#**	**k4**
<g/>		
,	,	kIx,
tehdy	tehdy	k6eAd1
nazvaná	nazvaný	k2eAgFnSc1d1
…		

Fig. 2. Vertical format of the question "*Z jakého roku pochází školní budova v obci Paršovice? (What year does the school building in Paršovice come from?)*" and (a part of) the answer sentence "*Další pamětihodností v Paršovicích je školní budova z roku 1898, tehdy nazvaná … (Another place of interest in Paršovice is the school building of the year 1898 at that time named as …)*" with the expected answer of "**1898**" marked with bold font.

vertical format (see Figure 2 for an example). The `url.txt` and `metadata` are in plain text form. The `text.vert` file can be also represented by a symbolic link that leads to file with the same content used in a different record. This provides consistency in changes and decrease redundancy in the whole dataset.

3 Database Adjustments

Before the final SQAD v2.0 release, we have perform multiple automatic and manual changes to correct tagger/tokenizer mistakes and supplement the corpus with additional metadata.

3.1 Automatic Adjustments

As in the previous version, we provide tokenization adjustments, fix out-of-vocabulary mistakes and a few regular morphological errors.

Apart form these changes, we have prepared a semi-automatic process to identify answer sentences. Such file is important for QA system evaluation on

the sentence selection level – whether the QA system is able to pick the correct sentence with the answer from the whole knowledge base.

3.2 Manual Adjustments

After the automatic changes, there were several manual changes performed on the data.

Very valuable information for QA development is represented by the question and answer type annotation of each record. This annotation was provided manually and the following question and answer types have been used.

The collection of annotated *Question types* takes an inspiration form The Stanford Question Answering Dataset [5]. Each question in SQAD v2.0 is tagged by one of the following question type:

(i) Date/Time
(ii) Numeric
(iii) Person
(iv) Location
(v) Other Entity
(vi) Adjective phrase
(vii) Verb phrase
(viii) Clause
(ix) Other

The annotated *Answer types* were taken form the first layer of Li and Roth's [6] two-layered taxonomy with a few adaptations. Each answer was thus assigned one of following types:

(i) Date/Time
(ii) Numeric
(iii) Person
(iv) Location
(v) Entity
(vi) Organization
(vii) YES/NO
(viii) Other

The remaining manual corrections of SQAD v2.0 are mostly technical and are related to the fields of Wikipedia URL, question and answer files to harmonize them with current Wikipedia content. First, the URL of Wikipedia articles changes quite often which makes problems when the article is moved to a new URL and the previous URL is redirected to another article with different content. Because of the live Wikipedia community, the articles changes frequently and that is why several questions and answers had to be adapted to the current text where the previous information was no longer present in the current data.

Table 1. SQAD v2.0 knowledge base statistics

Number of tokens	20,272,484
Number of sentences	911,014
Number of sentence selections	6,349
Number of source documents	3,149

4 Dataset Characteristics

The final SQAD v2.0 dataset consist of 8,566 question-answer pairs related to 3,149 documents obtained from Czech Wikipedia. The documents' texts are included as the underlying knowledge base with the corpus size of 20,272,484 tokens. The answer sentence selection list contains 6,349 sentences – see Table 1 for a summary of the SQAD knowledge base proportions.

Each question is annotated with of the selected questions types. The characteristics of the expected answers are categorized with the corresponding answer type. Overall statistics of the question and answer type distributions are presented in Tables 2 and 3.

Table 2. Question type statistics in SQAD v2.0

Date/Time	1,848
Numeric	900
Person	940
Location	1,436
Other Entity	1,440
Adjective phrase	253
Verb phrase	944
Clause	774
Other	31

Table 3. Answer type statistics in SQAD v2.0

Date/Time	1,847
Numeric	904
Person	943
Location	1,442
Entity	811
Organization	199
YES/NO	940
Other	1,480

5 Conclusions and Future Work

In this paper, we have introduced a new extended and manually annotated version of the SQAD dataset for question-answering evaluation. The second version contains nearly nine thousand records with manual annotation of question and answer types with each record.

Current planned steps of the work with the fresh new database concentrate on providing an evaluation of the syntax based question-answering system AQA based on this enhanced and enlarged evaluation dataset.

Acknowledgements. This work has been partly supported by the Czech Science Foundation under the project GA15-13277S and by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071.

References

1. Horák, A., Medveď, M.: SQAD: Simple question answering database. In: Eighth Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU (2014) 121–128
2. Horák, A., et al.: AQA: Automatic Question Answering System for Czech. In: International Conference on Text, Speech, and Dialogue, Springer (2016) 270–278
3. Michelfeit, J., Pomikálek, J., Suchomel, V.: Text tokenisation using unitok. In: 8th Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU. (2014) 71–75
4. Pavel Šmerk: Towards morphological disambiguation of Czech (2007)
5. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
6. Li, X., Roth, D.: Learning question classifiers. In: Proceedings of the 19th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics (2002) 1–7