# 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

December 1, 2017



# Outline

- Introduction
- Database Structure
- Database Adjustment
- Dataset Characteristics
- Conclusion and Future Work

# Question Answering

# SQAD v1.0 vs. SQAD v2.0

- SQAD v1.0
  - 3,301 QA pairs
  - short answer context
  - redundant data
- SQAD v2.0
  - 8,566 QA pairs
  - full Wikipedia articles
  - shared texts

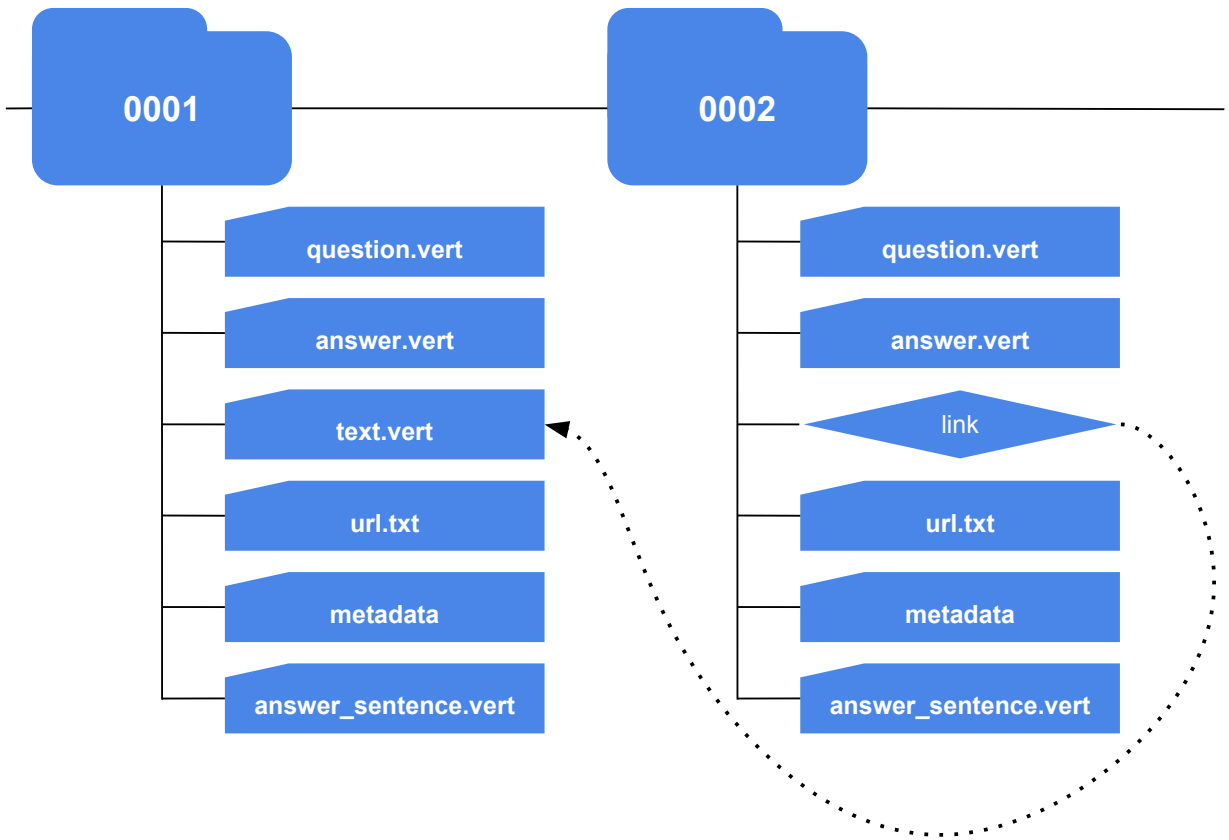
# Database Structure

# SQAD v1.0

- files in plain text form
  - original text
  - question
  - answer
  - Wikipedia URL
  - author
- does not support sharing of manual modifications

# SQAD v2.0

- generally the same structure
- texts in **vertical form**
- applying **symbolic links** to avoid document duplicities




Source: Marek Medveď

# SQAD v2.0 (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> | | |

# SQAD v2.0 (Answer)

| *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 |
| ... | | |

# SQAD v2.0

- metadata

**05metadata.txt**

```
<q_type>ENTITY</q_type>
<a_type>ENTITY</a_type>
```

- URL

**04url.txt**

```
http://cs.wikipedia.org/wiki/Wolfram
```

# Database Adjustments

# Automatic adjustments

- tokenization
- out-of-vocabulary mistakes
- morphological errors
- semi-automatic selection of answer sentences

# Manual adjustments

- question and answer type annotation
- technical problems
  - fix URL
  - fix answer sentence

# Manual adjustments

| Question types | Answer types |
|---|---|
| Date/Time | Date/Time |
| Numeric | Numeric |
| Person | Person |
| Location | Location |
| Other Entity | Entity |
| Adjective phrase | Organization |
| Verb phrase | YES/NO |
| Clause Other | Other |

# Dataset Characteristics

# SQAD database

- dataset build form Czech Wikipedia
- number of Q&A pairs: 8,566
- number of related documents: 3,149

# Knowledge base statistics

Table: 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 |

# Knowledge base statistics

Table: SQAD v2.0 question type statistics

| 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 |

# Knowledge base statistics

Table: SQAD v2.0 answer type statistics

| 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 |

# Conclusion and Future Work

# Conclusion and Future Work

- Conclusion
  - nearly 9000 question-answer pairs
  - annotated question/answer types
- Fututre work
  - evaluation of the AQA

# Questions and Answers