SQAD: Simple Question Answering Database

Aleš Horák, Marek Medveď

Natural Language Processing Centre Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic

6.12.2014

Supported by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

Definitions

Question answering:

- computer science discipline, which is concerned with building systems that automatically answer questions posed by humans in a natural language

Question answering system:

- systems that process the input question, go through a knowledge base and provide a reasonable answer to the input question

SQAD: Simple Question Answering Database

- developed for accuracy evaluation of question answering systems
- created from Czech Wikipedia
- created by students of computation linguistic course
- contains 3301 records

SQAD record

- SQAD record consists of:
  - the original sentence(s) from Wikipedia
  - a question that is directly answered in the text
  - the expected answer to the question as it appears in the original text
  - the URL of the Wikipedia web page from which the original text was extracted
  - name of the author of this SQAD record

Example of SQAD record

Example

Original text:

Létající jaguár je novela spisovatele Josefa Formánka z roku 2004.

[Létající jaguár is a novel of writer Josef Formánek form the 2004.]

Question:

Kdo je autorem novely Létající jaguár?

[Who is the author of the novel of Flying jaguar?]

Answer:

Josef Formánek

URL:

`http://cs.wikipedia.org/wiki/L%C3%A9taj%C3%ADc%C3%AD_jagu%C3%A1r`

Author:

chalupnikova

SQAD structure

Example

```
sqad/1877/:
    01question.txt
    01question.vert
    02answer.txt
    02answer.vert
    03text.txt
    03text.vert
    04url.txt
    05author.txt
```

SQAD: Automatic morphological annotation

- texts are processed by `Unitok` and `Desamb` tool
- to obtain high-quality data, the tagged texts were checked and corrected by semi-automatic and manual adjustments

SQAD: Tokenization adjustments

- wrong tokenization for large numbers

Example

a)

<s>
1
200
300
</s>

⟶

b)

<s>
1 200 300
</s>

Unitok: a) wrong and b) adjusted tokenization of number *"1 200 300"*.

SQAD: Out-of-vocabulary words

- the system Desamb is used for morphological tags disambiguation according to the word context and working over the attributive Czech tagset of the Majka system
- the Desamb tool cannot determine correct tag in this two main cases:
  - the context is too narrow
  - Majka system does not contain word form

SQAD: Out-of-vocabulary words

Desamb: Narrow context

- SQAD answers that contains only number
- we change "k?" (unknown tag) tag to "k4" (tag for numerals) if the word is number

SQAD: Out-of-vocabulary words

Desamb: Narrow context

Example

<s>
120 120 k? ⟶ <s>
120 120 k4
</s> </s>

Desamb: unrecognized number

SQAD: Out-of-vocabulary words

Desamb: Unrecognized word form

- Majka does not recognize all existing words, especially proper names and abbreviations
- for unrecognized words Desamb system returns “k?” (unknown tag) as a resulting tag
- for proper names and abbreviations we changed the unknown tag to:
  - “k1” (nouns) for all words that start with an upper case letter
  - “kA” (abbreviations) for words that contains only upper case letters, words ending with dot or words containing dots between upper case letters

SQAD: Out-of-vocabulary words

Desamb: Unrecognized proper names and abbreviations

Example

<s>				<s>		
Los	Los	k?		Los	Los	k1
Angeles	Angeles	k?		Angeles	Angeles	k1
</s>			⟶	</s>		
<s>				<s>		
LA	LA	k?		LA	LA	kA
</s>				</s>		

Desamb: unrecognized proper names and abbreviations

SQAD: Out-of-vocabulary words

Desamb: Unrecognized word form

- SQAD database is extracted from Czech Wikipedia and contain original forms of proper names

Example

Original form of word *"Tokio"* is "東京"

- we extracted remaining unknown words into one file keeping the original file name, word position and unknown word with its lemma and tag from Desamb
- the file was than manually annotated and programmatically applied back to the original annotated file

SQAD: Out-of-vocabulary words

Manually annotated file

Original Desamb output stored in file 03text.txt:

Example

```
Tokio   Tokio   k1gInSc1
(       (       kIx(
jap.    jap.    kA
東京    東京    k?
```

Record of unknown word extracted from 03text.txt file:

Example

```
./0000/03tex.txt|3|東京    東京    k?
```

SQAD: Out-of-vocabulary words

Manually annotated file example

Record from `03text.txt` with manual changes:

Example

./0000/03tex.txt|3|東京 東京 k1

File `03text.txt` with changes:

Example

Tokio	Tokio	k1gInSc1
(	(	kIx(
jap.	jap.	kA
東京	東京	k1

SQAD: Mistakes in morphological analysis

- wrong lemma for foreign words

Example

For word "Las" (from proper name "Las Vegas") the output of Desamb is “Las laso k1gInSc1”

- we checked all the SQAD database records and extracted a file with morphological analysis mistakes
- the file was than manually annotated and programmatically applied back to the original annotated file

SQAD

Table : SQAD mistakes

mistake type	number of found mistakes
out-of-vocabulary words	618
morphological analysis	160

SBQA: Syntax-based question answering system

- we used the SQAD database to evaluate the accuracy of a first version of SBQA
- the SBQA system was developed by M. Pavla at Faculty of Informatics, Masaryk University
- input of SBQA system is a plain text question which is then preprocessed by Unitok and Desamb system and passed to SET parser to identify dependencies and phrase relations within the question
- SBQA finds the answer in its knowledge base based on a match on corresponding syntactic structures
- SBQA knowledge base is made from plain text documents, which are automatically processed with Unitok, Desamb and SET
- to evaluate SBQA we use SQAD database as a knowledge base

Evaluation

Table : Evaluation of SBQA system

total questions	correct	partially correct	incorrect	not found
3,301 100%	758 23%	60 1%	2,003 61%	480 15%

Classification of SBQA errors

- we manually checked 200 questions and find:
  - errors in implementation of SBQA system
  - errors in tokenization or syntactic analysis
  - phenomena not covered by the current implementation of SBQA system

SBQA: Errors in implementation

- we have identified the following error types that are caused by SBQA implementation:
    - answer in brackets
    - part of speech requirement
    - comparison of dates or numbers
    - wrong question type

Answer in brackets

Example

Text: Ing. Miloš Zeman (* 28. září 1944 Kolín) je český politik.

[Ing. Miloš Zeman (* 28. September 1944 Kolín) is a Czech politician.]

Question: Kde se narodil Miloš Zeman?

[Where was Miloš Zeman born?]

Original answer: Kolín

SBQA answer: který na Novém Zélandu

[which in New Zealand]

Part of speech requirement

Example

Text: Hlaholice je nejstarší, dnes již neužívané slovanské písm.
[Glagolitsa is the oldest, not being used today, Slavic writing system.]

Question: Co je nejstarší slovanské písmo?
[What is the oldest Slavonic writing system?]

Original answer: Hlaholika
[Glagolitsa]

SBQA answer: písmo staroevropské civilizace
[writing system of the Middle-European civilization]

Comparison of dates or numbers

Example

Text: George Walker Bush je bývalý 43. prezident Spojených států amerických.
[George Walker Bush was 43. president of United States of America.]

Question: Byl George W. Bush 40. prezidentem Spojených států amerických?
[Was George W. Bush 40. president of United States of America?]

Original answer: Nie
[No]

`SBQA` answer: Ano
[Yes]

Wrong question type

Example

Text: Angličtina patří do skupiny západogermánských jazyků.
[English language belongs to group of West Germanic languages.]

Question: Do skupiny jakých jazyků patří Angličtina?
[To which group of languages the English language belongs to?]

Original answer: západogermánských
[West Germanic]

`SBQA` answer: Ano
[Yes]

SBQA: Errors in tokenization, tagging or syntactic analysis

- there are three types of such errors that appear in the current SQAD database:
  - Unitok incorrectly detects sentence boundaries and splits one sentence into two or more sentences
  - Desamb incorrectly tagged a word thus the syntactic analysis is incorrect and SBQA system cannot derive the required answer
  - SET incorrectly parses a sentence and creates an incorrect syntactic tree. This usually leads to incorrect answer.

Error in tokenization

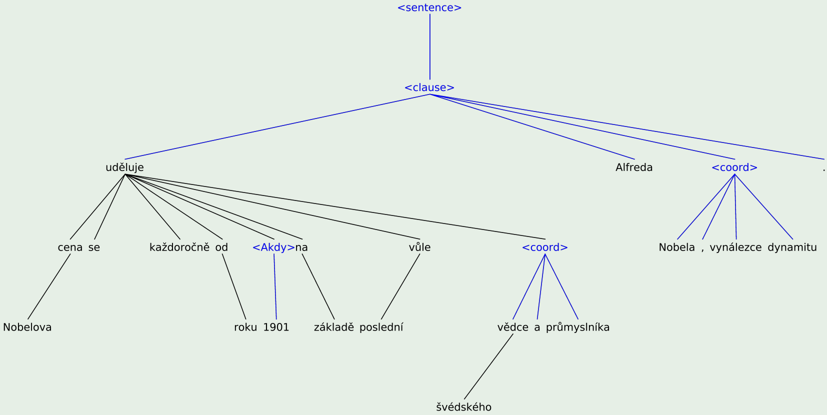
Example

Text: Lilongwe je hlavní město afrického státu Malawi.
[Lilongwe is a capital city of African state Malawi.]

Lilongwe	Lilongwe	k6eAd1
je	být	k5eAaImIp3nS
hlavní	hlavní	k2eAgNnSc1d1
město	město	k1gNnSc1
...		

Error in syntactic analysis

Example



SBQA: Uncovered phenomena

- the SBQA system has not yet implemented advanced NLP techniques such as anaphora resolution

Evaluation

Table : Classification of SBQA errors (on 200 examples)

total questions	error in SBQA system	error in tokenization or syntax analysis	uncovered phenomena
200 100%	119 59.5%	43 24.5%	38 19%

Conclusions

- we have presented new Czech question answering database called SQAD
- SQAD record consists of an annotated question, the annotated answer, the annotated sentence containing the full answer, Wikipedia URL as a source of the statement and the author name of this question-answer pair
- morphological annotation of SQAD was obtained automatically and manually corrected

Future Directions

- improve SQAD database
- according to SQAD database refine SBQA system
- add new phenomena into SBQA system

Thank you for your attention.