

Context Tracking for Question Answering with Aqqu

Bachelor Thesis Presentation

Anushe Glushik

Examiner: Prof.Dr.Hannah Bast

Adviser: Niklas Schnelle

Chair of Algorithms and Data Structures

03.09.2019

1. Motivation

- 1.1 Conversational Question Answering
- 1.2 Aqqu Question Answering System

2. Approach

- 2.1 Aqqu Chatbot
- 2.2 Context Tracking

3. Evaluation

- 3.1 Datasets
- 3.2 Experiments

4. Conclusion

5. Future Work

Context: the immediate history of previous questions and answers.

Context simplifies a human question-answer dialog and makes it more convenient.

1.1 Conversational Question Answering

An example of conversational question answering:

Question 1: *Who is Johann Sebastian Bach?*

Answer 1: **Composer and Musician**

Question 2: *Who was he married to?*

Answer 2: **Anna Magdalena Bach**

Question 3: *Where was she born?*

Answer 3: **Zeitz**

In Question 2 and Question 3 no entity was specified, yet it is clear out of the context, that *Johann Sebastian Bach* is meant by *he* and *Anna Magdalena Bach* by *she*.

1.2 Aqqu Question Answering System

An example of NON-conversational question answering:

User: *Who is Johann Sebastian Bach?*

Aqqu: **Composer and Musician**

User: *Who was he married to?*

Aqqu: **World Health Organization, founders: Brock Chisholm**

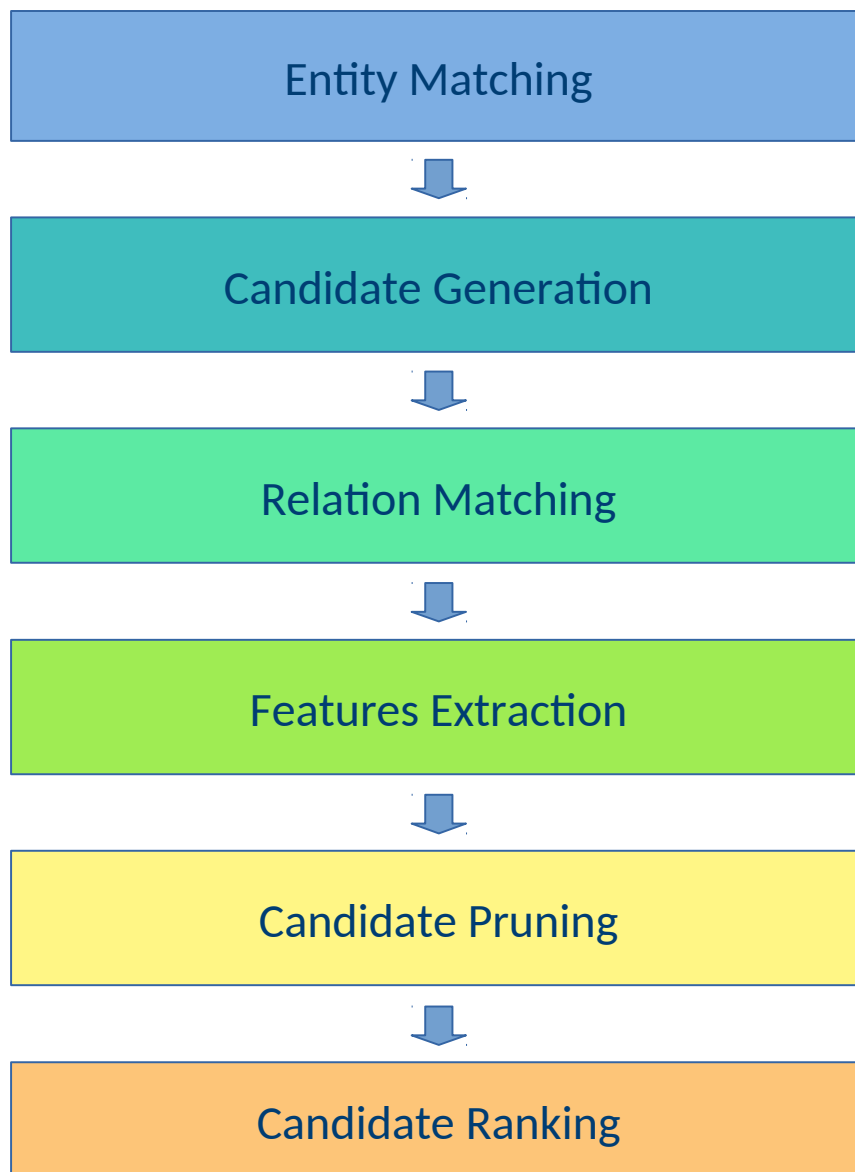
Aqqu treats the questions separately, isolated from each other.

In the example above Aqqu identifies **who** as the best suitable entity and gives an answer according to it.

1.2 Aqqu Question Answering System

- Aqqu gets answers from the knowledge base (Freebase).
- Aqqu extracts the possible entities and relations from natural-language query and creates possible matching SPARQL queries.
- With the SPARQL queries the answer candidates are extracted from the knowledge base.
- After that the candidates are pruned and sorted.

1.2 Aqqu Question Answering System



2. Approach

To implement the context tracking we:

- **Created a conversational web user interface in the form of a chatbot (Aqqu Chatbot).**
- **Added context tracking functionality in Aqqu backend.**

2.2 Aqqu Chatbot

The **Aqqu Chatbot** was built using **Flask**.

You can try it right now: <http://galera.informatik.privat:5000/>

2.1 Aqqu Chatbot – Example 1

Aqqu Chatbot

Hi, I'm Aqqu. I know a lot of stuff!

Where was Albert Einstein born?



Albert Einstein, place of birth: Ulm



Albert Einstein, location: Germany, Princeton, Munich, Bern



Albert Einstein, place of death: Princeton



Albert Einstein, location: Germany, Princeton, Munich, Bern

Question

Ask

Next answer

2.1 Aqqu Chatbot – Example 2

Aqqu Chatbot

Hi, I'm Aqqu. I know a lot of stuff!

Where was Albert Einstein born?

Albert Einstein, place of birth: Ulm

What did he do?

Albert Einstein, profession: Physicist, Scientist, Writer, Author, Mathematician, Teacher, Philosopher, Theoretical Physicist

Who was he married to?

Albert Einstein, spouse: Elsa Einstein, Mileva Marić

Question

Ask

Next answer

2.1 Aqqu Chatbot – Example 3

Aqqu Chatbot

Hi, I'm Aqqu. I know a lot of stuff!

Where was Albert Einstein born?

✓ Albert Einstein, place of birth: Ulm

Who was Isaac Newton and what did he do?

✓ Isaac Newton, profession: Chemist, Physicist, Scientist, Mathematician, Astronomer, Philosopher

✓ Albert Einstein, profession: Physicist, Scientist, Writer, Author, Mathematician, Teacher, Philosopher, Theoretical Physicist

Question

Ask

Next answer

2.1 Aqqu Chatbot – Example 4

Aqqu Chatbot

Hi, I'm Aqqu. I know a lot of stuff!

Is there a pattern behind prime numbers?

Sorry, I don't know the answer to: "Is there a pattern behind prime numbers?". Please check your question for typos.

Question

Ask

Next answer

2.1 Aqqu Chatbot – Example 5

Hi, I'm Aqqu. I know a lot of stuff!

What Is the answer to the ultimate question of life the universe and everything?

✓ Life, the Universe and Everything, original language: English Language

Question

Ask

Next answer

2.1 Aqqu Chatbot – Data Augmentation Functionality

Aqqu Chatbot

Hi, I'm Aqqu. I know a lot of stuff!

Who was Albert Einstein?



Albert Einstein, profession: Physicist, Scientist, Writer, Author, Mathematician, Teacher, Philosopher, Theoretical Physicist

Mark this answer as the most accurate.



Albert Einstein, Inventions: Einstein refrigerator



Albert Einstein, original Ideas: General relativity



Albert Einstein, Influenced by: Baruch Spinoza, George Bernard Shaw, Arthur Schopenhauer, David Hume, Fyodor Dostoyevsky, Isaac Newton, Mahatma Gandhi, Paul Valéry, Hermann Minkowski, Bernhard Riemann, Thomas Young, Moritz Schlick, Ernst Mach, Hendrik Lorentz, Karl Pearson, James Clerk Maxwell, Riazuddin

Question

Ask

Next answer

2.1 Aqqu Chatbot – Data Augmentation Functionality

Aqqu Chatbot

Hi, I'm Aqqu. I know a lot of stuff!

Who was Albert Einstein?



Albert Einstein, profession: Physicist, Scientist, Writer, Author, Mathematician, Teacher, Philosopher, Theoretical Physicist

Albert Einstein, Inventions: Einstein refrigerator

Albert Einstein, original Ideas: General relativity

Albert Einstein, Influenced by: Baruch Spinoza, George Bernard Shaw, Arthur Schopenhauer, David Hume, Fyodor Dostoyevsky, Isaac Newton, Mahatma Gandhi, Paul Valéry, Hermann Minkowski, Bernhard Riemann, Thomas Young, Moritz Schlick, Ernst Mach, Hendrik Lorentz, Karl Pearson, James Clerk Maxwell, Riazuddin



Albert Einstein, area of activism: Humanitarian, Nonviolence, Peace, Zionism, Anti-fascism

Question

Ask

Next answer

2.2 Context Tracking – Main Approach

The approach for the conversation tracking consists of the following steps:

Aqqu Chatbot side:

1. Store the identified entities (ID and name) after the system gets a result.

2. Look for pronouns in further queries.

3. If the processed query contains a pronoun – add the previous entities to the list of matched entities; if it does not – treat the query as usual.

Aqqu backend side:

4. Process the list of identified entities and identify the gender for each entity.

2.2 Context Tracking – Gender Identification

To identify the gender of an entity the system uses one of two methods:

1. Find a genus in the **gender.csv** file.
2. If the entity is not found in the gender.csv file, the system tries to guess the gender.
To guess the gender we used the **gender-guesser** python module.

2.2 Context Tracking – Mistakes in Gender Identification

Example:

The Bahamas should be identified as **Neutral**. However the name **Bahamas** is stored as **Male** in **gender.csv**. This may lead to:

User: What language is spoken in the Bahamas?

Chatbot: Bahamas, languages spoken: English Language, Bahamas Creole English Language

User: Where is it?

Chatbot: Bahamas Creole English Language, main country: Bahamas

The system does not know that **it** should be related to **the Bahamas**.

3. Evaluation

Influence of context tracking on the overall performance of the Aqqu System.

3.1 Datasets - Original Dataset

For the evaluation of the system performance the setup from original Aqqu and the WebQSP dataset with original split (70% for training and 30% for testing) were used.

The dataset consists of a list of questions.

Some important parameters of a question:

- **ProcessedQuestion** - *processed query text, lowercase and without a question mark.*
- **Sparql** - *the executed SPARQL query.*
- **TopicEntityName** - *the name of the entity.*
- **TopicEntityMid** - *the knowledge base ID of the entity.*
- **Answers** - *the answers for the question. A question can have many correct answers. For example, for a question **What language did ancient Romans write in?** there are two answers: **Greek Language** and **Latin Language***

3.1 Datasets – Conversational Dataset

The script concatenates the questions into groups according to their entities.

An example of a conversation from the conversational WebQSP:

Leading query: what did **Galileo** do to become famous?

Second query: what was **he** famous for?

Third query: what discovery did **he** make?

3.1 Datasets – Basic Conversational Dataset

- One conversation for each entity.
- One-question conversation with no additional pronoun-questions.

Dataset	Conversations	Longest Conversation	Average conversation	Questions
Training	1720	27	2	~3000
Evaluation	1072	14	1.7	1815

3.1 Datasets – Extended Conversational Dataset

- If an entity has N questions we create N conversations. *For example:*

First conversation:

- **Where was Albert Einstein born?**
- **Who was he married to?**

- Second conversation:

- **Who was Albert Einstein married to?**
- **Where was he born?**

- For entities with only one question – the first query is the original query, and the second one is the same query but with the entity replaced with a corresponding pronoun.
- *For example:*

First question: **Where was Albert Einstein born?**

Second question: **Where was he born?**

Dataset	Conversations	Longest Conversation	Average conversation	Questions
Training	3453	27	2	~3000

3.2 Experiments

We have conducted experiments with the Aqqu system trained on:

1. *Original Dataset*

2. *Basic Conversational Dataset*

3. *Extended Conversational Dataset*

3.2 Experiments – Trained on Original Dataset

Evaluation Dataset	Original	Conversational without Gender Identification	Conversational with Gender Identification
Average Precision	0.67	0.50	0.60
Average Recall	0.72	0.55	0.65
Average F1	0.657	0.495	0.586
Accuracy	0.478	0.355	0.417
Parse Accuracy	0.510	0.362	0.440

3.2 Experiments – Trained on Conversational Data

Training Dataset	Basic Conversational Dataset	Extended Conversational Dataset
Average Precision	0.59	0.60
Average Recall	0.63	0.645
Average F1	0.575	0.584
Accuracy	0.417	0.420
Parse Accuracy	0.440	0.425

3.2 Experiments – Comparison

Experiment	Original	Conversational
Average Precision	0.67	0.60
Average Recall	0.72	0.645
Average F1	0.657	0.584
Accuracy	0.478	0.420
Parse Accuracy	0.510	0.425

4. Conclusion

- The Conversational Aqqu is more convenient and intuitive to use.
- It has a slight decrease in accuracy of 7.1% compared to the original Aqqu system.
- Possible reasons for the errors:
 - Mistakes in gender identification during the creation of conversational data and during the question answering process.
 - More entities to process. The more entities – the more difficult it is to prune the incorrect answers and rank the correct ones.
- Aqqu Chatbot data augmentation functionality can help to improve the datasets for further training and evaluation.

5. Future Work

- Add a new feature, that indicates whether a candidate is generated from context or from a current question.
- Data augmentation functionality can be developed to additionally save the context of a question-answer pair in the WebQSP format.
- Collect more data with the Aquu Chatbot data augmentation functionality and train the system on a bigger dataset.
- Process additional single interrogative word questions: "Where?", "When?", etc.

Backup Slides: Evaluation Metrics 1

- $q_1, \dots, q_n$: questions
- $c_1, \dots, c_i$: the answer candidates
- $g_1, \dots, g_n$: multiple answers of the gold answer
- $a_1, \dots, a_n$: the answers from the system for the first candidate
- GA-Size: Gold answer size is the number of ground truth answers ($|g_1, g_2, \dots, g_n|$).
- BCA-Size: Best candidate answer size is the number of answers of the first candidate ($|a_1, a_2, \dots, a_n|$).
- Candidates: The number of all predicted candidates ($|c_1, c_2, \dots, c_i|$).

Backup Slides: Evaluation Metrics 2

- Precision: The precision shows what percentage of the answers from the best candidate are correct.

$$\text{Precision} = \frac{TP}{TP + FP},$$

where TP is a true positive, i.e. $TP = |a_k, \dots, a_m|$, where $a_k, \dots, a_m$ are correct answers and FP is a false positive, i.e. $FP = |a_l, \dots, a_p|$, where $a_l, \dots, a_p$ are false answers.

For example:

Utterance	GA-size	GA	BCA-Size	BCA
who does ronaldinho play for now 2011?	2	"Brazil national football team" "Clube de Regatas do Flamengo"	2	"Clube Atlético Mineiro" "Clube de Regatas do Flamengo"

Only one answer out of two is correct and only one correct answer is found.

Therefore $TP = 1$, $FP = 1$, $\text{Precision} = 1/(1+1) = 0.5$

Backup Slides: Evaluation Metrics 3

- Recall. The recall measures how well the system finds correct answers, i.e. what percentage of correct answers are found.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where FN is false negative, i.e. $\text{FN} = |g_l, \dots, g_p|$, where $g_l, \dots, g_p$ are correct answers that were not found by the system. For example:

Utterance	GA-size	GA	BCA-Size	BCA
what state does romney live in?	1	"Massachusetts"	2	"Massachusetts" "Bloomfield Hills"

Only one answer out of two is correct and only one correct answer is found.

Therefore $\text{TP} = 1$, $\text{FP} = 1$, $\text{Precision} = 1/(1+1) = 0.5$

- F1: It is the harmonic average of the precision and recall. The best value is 1 (Precision $\rightarrow 1$ and Recall $\rightarrow 1$) and the worst is 0 (Precision $\rightarrow 0$ and Recall $\rightarrow 0$).

$$\text{F1} = 2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Parse Match: This parameters shows if the candidate relation that gives an answer with the best F1 score is matched perfectly to the ground truth (>0.99 matching).

Backup Slides: Averaged Evaluation Metrics

- Questions: Total number of questions in the evaluation dataset.
- Average Precision: The average precision of all questions.
- Average Recall: The average recall of all questions.
- Average F1: The average F1 across all questions.

$$\text{average F1} = \frac{1}{n} \sum_{i=1}^n F1(g_i, a_i)$$

- Accuracy: The percentage of queries answered with the exact gold answer.

$$\text{accuracy} = \frac{1}{n} \sum_{i=1}^n I(g_i = a_i)$$

- Parse Accuracy: Average parse match across all questions.

Backup Slides: Evaluation Metrics 4

ID	Utterance	GA-Size	BCA-Size	Candidates	Precision	Recall	F1	Parse Match
0	what time zone is chicago in right now?	1	1	3	1	1	1	True
1	where to stay there tourist?	1	1	21	1	1	1	True
2	who does ronaldinho play for now 2011?	2	2	21	0.5	0.5	0.5	False
3	what is ella fitzgerald name?	2	2	9	1	1	1	True
4	what state does romney live in?	1	2	10	0.5	1	0.67	False
5	where did his parents come from?	1	1	31	1	1	1	True
6	what university did he graduated from?	1	6	21	0	0	0	False
7	where did he graduated college?	1	6	19	0	0	0	False
8	what colleges did he attend?	5	6	1	0	0	0	False
9	when did he become governor?	1	1	6	0	0	0	False
10	where is his family from?	1	1	8	0.5	1	0.67	False
11	what degrees does he have?	3	6	1	0	0	0	False
12	who does jeremy shockey play for in 2012?	1	1	19	1	1	1	True
13	what does bolivia border?	5	5	10	1	1	1	False