Karlsruhe Institute of Technology

Development of a Domain-Independent Interactive Question Answering System

Bachelor Thesis of

Florian Kaiser

Interactive Systems Labs (ISL)
Institute for Anthropomatics
Karlsruhe Institute of Technology

Reviewer:	Prof. Dr. Alexander Waibel
Second reviewer:	Dr. Sebastian Stüker
Advisor:	M.A. Maria Schmidt

13 February 2015 – 12 June 2015

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Karlsruhe, den 12. Mai 2015

Zusammenfassung

Das Internet als vernetzte Plattform, die von Milliarden von Menschen genutzt und aktiv mitgestaltet wird, hält eine schier unendliche Menge an verschiedensten Informationen bereit. Jedoch ist der überwiegende Teil der Informationen im Internet unstrukturiert hinterlegt. Also braucht es geeignete Methoden, um diesen Datenschatz für Menschen auf effiziente und möglichst natürliche Art und Weise nutzbar zu machen.

Interaktive Frage-Antwort-Systeme zeigen in diese Richtung einen Weg auf. Sie bieten eine Schnittstelle um in natürlicher Sprache Informationen aus den Weiten des Internets zu erfragen. Dabei rücken sie ein entscheidendes Stück näher an die Bedürfnisse und Gewohnheiten der Menschen, indem sie den Kontext einer Frage und Referenzen innerhalb einer Frage interpretieren und versuchen aufzulösen. Falls dies nicht möglich ist, kann das System die Initiative ergreifen und aktiv mit dem Benutzer zusammenarbeiten, um den Mangel an Informationen aufseiten des Systems zu beseitigen.

Im Rahmen dieser Bachelorarbeit wurde ein domänenunabhängiges, interaktives Frage-Antwort-System entwickelt, das eine textbasierte Interaktion erlaubt. Dafür wurde ein quelloffenes Frage-Antwort-System benutzt und um interaktive Funktionalitäten erweitert. Damit ist es dem System möglich Kontextinformationen zu benutzen und bei Unsicherheit vom Benutzer klarstellen zu lassen. Aus̈erdem kann es sich auf so gewonnene Informationen stützen um dem Benutzer bessere und genauere Antworten zu liefern.

Contents

1. Introduction

This chapter gives an introduction to Interactive Question Answering (IQA). To understand IQA in its context and origins, it will be necessary to get an overview of Question Answering (QA) and dialogue systems first. This will lead to the motivation of extending pure Question Answering. The end of this chapter further outlines the scope of this thesis.

1.1. Question Answering

Question Answering (QA) is concerned with the retrieval of accurate answers to natural language questions. The interaction between a user and a QA system is characterised by queries from the user, i. e. the user has the initiative throughout the interaction. In particular, the QA system is not able to pose any questions if the user's utterance has not been understood correctly or the information given is not sufficient. Instead, the user is supposed to ask a single, independent question and the system will respond with an answer or a list of possible answers. In general, QA systems will not provide any further communication.

Most of the research in QA has been focused on answering factoid questions. Answers to such types of questions include named entities like persons, organisations or locations (Konstantinova and Orasan, 2013). Research has also been carried out, though, in answering definition questions(Blair-Goldensohn et al., 2004), *why* questions (Verberne et al., 2010), and complex questions(Bilotti and Nyberg, 2006).

QA systems can vary a lot in their design and their intended usage. The major decisions in designing a QA system is whether it should address open-domain or closed-domain questions, whether it utilises an unstructured knowledge base like the Web or a (semi-) structured knowledge base like Wikipedia or a database(Ferrucci et al., 2009), whether the system is multimodal or unimodal, text-based or speech-to-speech.

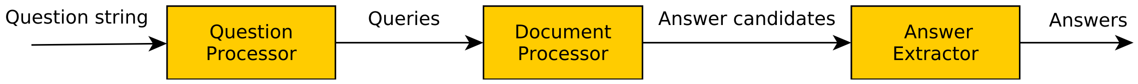


Figure 1.1.: The typical QA pipeline.

Nevertheless, the core components of a standard QA system consist of a *question processor*, *document processor*, and *answer extractor* (Harabagiu and Moldovan, 2003) as illustrated in figure 1.1. The question processor interprets the question, whereas the document processor collects and organises the relevant paragraphs that have been found. Finally, the answer extractor extracts possible answers from the paragraphs and evaluates them in order to generate a ranking of all possible answers.

1.2. Dialogue Systems

The term *dialogue system* is used in many different contexts today. Thus, the focus of its definition differs among them. A general definition is given by the *Journal of Dialogue Systems*:

A computational device or agent that (a) engages in interaction with other human and/or computer participant(s); (b) uses human language in some form such as speech, text, or sign; and (c) typically engages in such interaction across multiple turns or sentences. (Konstantinova and Orasan, 2013)

According to this definition, a dialogue system always has a user that interacts with it in human language. Moreover, it is important to note that the intention of such a system is not to reply to a single query but in leading a coherent conversation.

Depending on their intended area of application, dialogue systems vary significantly in how natural their interaction with the users are. While some offer a purely functional interface to navigating through a device for instance, others are deployed in customer service with an emphasis on its intuitive, human-like interaction. At the end of this spectrum are those dialogue systems that try to pass the Turing test by not being distinguishable from a human chat partner.

1.3. Interactive Question Answering

Interactive Question Answering (IQA) is a research field at the intersection of Question Answering (QA) and dialogue systems (Konstantinova and Orasan, 2013). More specifically, it incorporates a dialogue system into a QA system, allowing it to engage with the user in a deeper, more natural way.

Consequently, IQA systems can for instance take the initiative to clarify a question (e. g. resolving ambiguities). Furthermore, IQA systems allow the user to ask additional questions based on previous questions or answers, for instance to explore a topic further by making use of the current dialogue context or to refine a given answer. Thus, Webb and Webber (2009) define IQA as “a process where the user is a continual part of the information loop - as originator of the query, as arbitrator over information relevance, as consumer of the final product”.

1.4. Motivation

As De Boni and Manandhar (2005) states, to fulfil a user's need for information, it is often not enough for the user to ask a single question. Instead, the user might want to build upon the information given and elaborate a topic further. On the other hand, the system might require some more information from the user to satisfy the user's query, especially since natural language is too complex to be always interpreted correctly by a dialogue system. In this case the IQA can take the initiative and pose a question to the user in order to clarify the request.

The analysis of the characteristics of human dialogues by Jurafsky and Martin (2009) reveals that *grounding*, which is the mutual background of interlocutors, is essential for the success of a dialogue. Therefore, it is crucial for the correct interpretation of a question to have a common ground with the user, either by understanding the context of the question or by querying the missing information from the user.

The need of coherent interactions with a common ground has also been reflected by the TREC evaluation framework for Question Answering, which has started to introduce questions that require information from the context in 2001 (Voorhees and Harman, 2001).

Utilising its possibilities of context information and clarification, IQA systems cannot only improve the user experience by offering an intuitive, natural interaction but also improve its accuracy of answering questions. Consequently, IQA takes natural language-based information retrieval systems a step further.

1.5. Scope

In this project, a domain-independent interactive question answering system has been designed and developed based on an open-source question answering framework.

The IQA system provides an intuitive way of querying factoid information in a natural way. Through the ability to interpret questions in a given context, the interaction with the system is less repetitive and more efficient. However, the right context and references cannot always be interpreted correctly. In that case, the system tries to recover and asks the user for the missing information.

With the extra information received through clarification questions and the possibility to get entities disambiguated by the user for instance, the IQA system can actively contribute to more accuracy in the system's answers.

The developed system does not intend to have any small-talk with the user. Instead, users are supposed to be cooperative in their way of querying the system. Furthermore, the system does not focus on a highly sophisticated dialogue manager to keep track of all kinds of information and states. The dialogue manager is functional in its design so that it is able to deal with common dialogue situations. The mode of interaction offered is text-based.

1.6. Outline

Chapter 2 gives a detailed overview of the structure and essential features of the *Ephyra* QA framework that the developed IQA system has been based on.

In Chapter 3, it is described how the *Ephyra* framework has been integrated into the IQA system and with which components the QA framework has been extended to transform it into an interactive system.

Chapter 4 describes the setup and discusses the results from the conducted user study to evaluate the performance of the implemented IQA system.

Finally, Chapter 5 summarises the work done for this thesis and the developed IQA system.

2. The Open Source QA Framework *Ephyra*

The Interactive Question Answering system developed with this thesis is based on the open source Question Answering framework *Ephyra* (Schlaefer et al., 2007). In this chapter, its structure and essential features are described since chapter 3 will then refer to some concepts introduced in this chapter when elaborating the extensions of *Ephyra*.

2.1. Architecture

Ephyra also follows the typical pipeline structure of a QA system as depicted in figure 1.1. It has been developed with an emphasis on a highly modular structure. Hence, individual modules can be easily replaced. In the following sections, the design of the components *Query Formation*, *Information Retrieval*, and *Answer Selection* are presented.

2.1.1. Query Formation

According to the typical QA pipeline, *Ephyra*'s first component processes the user question. Figure 2.1 outlines the structure of this component. First, linguistic techniques are applied to the question string to normalise the question. For instance, the question normalisation replaces verbs with their corresponding lemmas and removes auxiliary verbs.

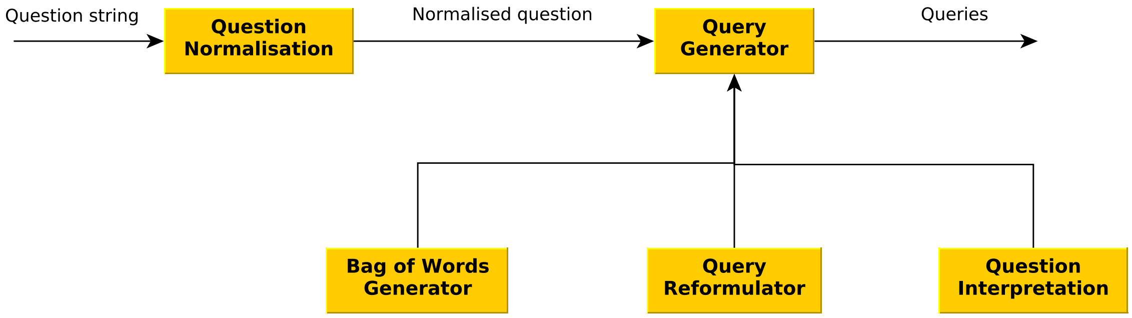


Figure 2.1.: Overview of the query formation component.

The normalised string is then used to generate appropriate queries. Three different methods are applied to generate queries: *bag of words*, *query reformulation*, and *question interpretation*. *Bag of words* generates a query based on a set of keywords extracted from the normalised string, whereas the *query reformulation* takes the question and rephrases it so that the words are in a form that is likely to occur in a text. For the reformulation there are fixed rules defined.

The *question interpretation* module is based on the observation that each question can be reduced to the three essential components *property*, *target*, and *context* and, hence, can be abstracted from its formulation. According to that, a question should ask for a *property* of a *target* in a *context*. As an example, the question "How many calories are there in a Big Mac?" could be interpreted as follows:

- Property: NUMBER
- Target: "calories"
- Context: "Big Mac"

To put it into other words, according to that interpretation the original question asks for a numeric value of the target object "calories" in the context of "Big Mac" (Schlaefer, 2005). The extraction of those three components are accomplished through question patterns that have been defined manually for about 70 different properties.

With the approach of using three different methods for query generation, a wide variety of queries are covered. Moreover, answer candidates retrieved from semantically and syntactically more complex queries are valued higher than basic queries that convey little semantic information. Accordingly, answer found in queries that have been generated from the components *query reformulation* or *question interpretation* are preferred over answers from a simple *bag of words* query.

2.1.2. Information Retrieval

This component receives queries generated by the *query formulation* component and uses different knowledge sources to fetch text snippets that contain possible answer candidates. There are two different types of searchers: *knowledge miners* that query unstructured knowledge sources like common search engines and *knowledge annotators* on that other hand that have (semi-)structured knowledge bases as source like *Wikipedia* (Schlaefer, 2005).

All searchers are merely required to offer a basic interface so that it is convenient to add a searcher for a new knowledge source. For computational reasons, knowledge miners only extract the text snippets returned by the search engine and does not follow the link to the source document as a whole.

2.1.3. Answer Selection

The *answer selection* component receives text snippets from unstructured and (semi-)structured knowledge sources and applies a set of various filters to those text snippets in order to extract all answer candidates. Additionally, some filters directly influence the score assigned to an answer candidate.

One of the core filters is the *sentence segmentation filter*, which parses each text snippet and splits it into its single sentences. Each sentence is then independently analysed further. Another essential filter is the *answer extraction filter*, which applies answer patterns to the candidates to receive exact answer phrases. The appropriate answer patterns are learned automatically from a tagged answer corpus.

The *answer type filter* determines the *property* type as defined in section 2.1.1 of the answer candidate and compares it to the property that has been assigned to the question when it was interpreted by the query formation component.

After all filters have been applied, each answer candidate has a score assigned to it according to its knowledge source, the query that it originates from, and the rating from different filters. At this point it is important to note that the score of an answer cannot be interpreted as an absolute measurement but as a relative evaluation among all answers. Finally, this component returns a ranked list of answers.

2.2. Interfaces

The QA framework *Ephyra* provides different user interfaces. Two of them are text-based: a command line tool and a web interface. As an alternative, *Ephyra* also offers the possibility of a spoken interaction, for which a speech recognition and speech synthesis have been integrated.

3. Design of Interactive Question Answering System

In this chapter, the design and features of the developed Interactive Question Answering system are described. The IQA system is based on the open-source QA framework *Ephyra* that has been introduced in chapter 2.

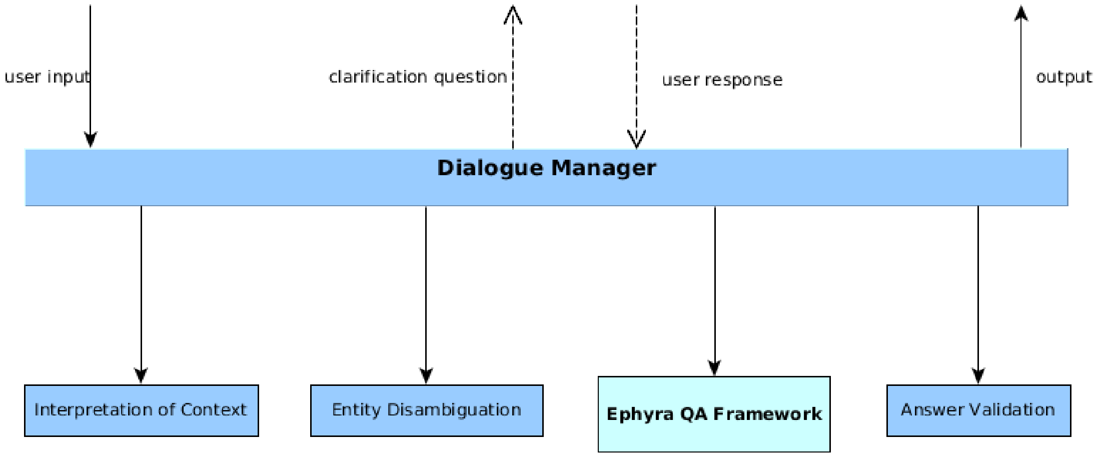


Figure 3.1.: Overview of IQA system.

3.1. Dialogue Management

The Dialogue Manager (DM) is the central component of the IQA system. As illustrated in figure 3.1, the DM handles the interface to the user and the execution of all other components. The user input is received, passed to the components for interpreting the context of the utterance and to the component for entity disambiguation. Moreover, the DM holds a representation of the state of the dialogue including the dialogue history, the interpretation of recent user questions, and entities that have been disambiguated by the user. This information about the dialogue is also used by the components to interpret the current question.

After the interpretation of the context and entity disambiguation has been finished and possible missing information has been clarified by the user, the interpreted question is passed to the Ephyra QA system. Finally, if an entity disambiguation was necessary, the DM invokes the component for answer validation and provides it with the list of answers from Ephyra and the entities to check. The final answer is then returned to the user and the system is ready for the next query.

3.2. Context-Sensitive Interpretation and Recovery

To analyse the occurrence of contextual phenomena in a question answering dialogue, Bertomeu et al. (2006) have conducted a Wizard-of-Oz experiment. Their results indicate that 36.16% of user utterances are context-dependent. Furthermore, the most prevailing phenomena they detected are fragments, bridging, and anaphoric pronouns.

```
Q1: Are there any projects on spell checking in Europe in the year 2006?
Q2: And in the year 2005?
Q3: How is the contact for that project?
Q4: Homepage?
```

Figure 3.2.: A dialogue sequence illustrating two fragments in utterances *Q2* and *Q4* (Bertomeu et al., 2006).

Fragments are utterances that are missing some parts of a full sentence structure. In many cases they have an analogous structure to previously uttered sentences that they are building upon. Figure 3.2 shows two examples of fragments.

```
Q1: The Speech TEK West 2006, when does it take place?
A: 2006-03-30 - 2006-04-01.
Q2: Until when can I hand in a paper []?
```

Figure 3.3.: A dialogue sequence containing bridging in *Q2* (Bertomeu et al., 2006).

Bridging describes the omission of an entity that is related to the entity of current focus. Question *Q2* in figure 3.3 contains a bridging. Anaphoric pronouns are words like *he* or *she* that refer to a specific entity.

Consequently, the designed IQA system contains an anaphora resolution component to deal with anaphoric pronouns and a component to address omitted context information like fragments and bridging.

3.2.1. Anaphora Resolution

This component checks whether the utterance of the user contains any anaphoric reference. If a reference in form of a personal or possessive pronoun has been detected, the coreference resolution module (Raghunathan et al., 2010) as part of the *Stanford CoreNLP Toolkit* is used to locate the actual entities that the pronouns are referring to. As discourse frame the two most recent dialogue turns are provided, taking into account the locality of references (Bertomeu et al., 2006). Figure 3.4 shows an example of the output of the coreference resolution module to the given dialogue context.

In case the coreference resolution module could not assign the referent to a pronoun in the user's utterance, the anaphora resolution component enters a subroutine in which it addresses a question back to the user asking for the correct referent. The clarification by the user is now used to recover from the missing reference. Figure 3.5 illustrates a subroutine for clarifying the referent.

```
Question: What is the name of Barack Obama's wife?
Answer: Michelle
Question: When did he meet her for the first time?

CHAIN1-["Barack Obama 's" in sentence 1, "he" in sentence 3]
CHAIN6-["Michelle" in sentence 2, "her" in sentence 3]
```

Figure 3.4.: Sample output of coreference chains from the coreference resolution module.

```
Question: When was Michelle Obama's first daughter born?
Answer: 1998
Question: Where was she born?
Clarification question: Who are you referring to by "she"?
Answer:
```

Figure 3.5.: Clarification subroutine after coreference resolution failed to find referent.

3.2.2. Recovery of Omitted Context Information

The detection of omitted context information that is relevant to interpret the user's question correctly is a non-trivial task. As described in section 2.1.1, the Ephyra QA system parses a user question to the three components *property*, *target*, and *context*. Therefore it is an indication for missing context information if one of the components remains empty after Ephyra has parsed the question.

A low confidence score (see section 2.1.3) for the highest ranked possible answers that have been retrieved is a second evidence for the need of including contextual information. In this case the IQA system fills the missing component(s) held by Ephyra with alternatives from the recent context. If the resulting answer set for one of those alternatives receives a high confidence score, it is assumed that the chosen interpretation of the question has been correct. If not, the user is asked to rephrase the question.

3.3. Named Entity Management

Mentions of named entities are in many cases ambiguous due to the similarity of names. For a human it is usually not a problem to infer the correct entity in a dialogue from the common ground and the similar world knowledge base of all interlocutors. For instance, speaking about *Michael Jordan* without giving any further context, most humans might immediately refer the name to the famous former basketball player. If Michael Jordan is mentioned in the context of basketball, this fact would even reassure the human interpreter that the named entity is actually the former basketball player.

However, the IQA system does not share a similar world knowledge base with the user. Thus, from the view of the system the name ”Michael Jordan” could possibly refer to any person with that name. Even if it is assumed that the mentioned person is eminent in some regard, a look at Wikipedia[1] reveals a whole list of eminent people with the name Michael Jordan as shown in figure 3.6.

In a Question Answering system, name ambiguities affect the system's performance negatively by leading to wrong answers and poor results. Hence, for the performance of a QA system it is essential to provide a mechanism for named entity disambiguation. As a

[1] www.wikipedia.org

```
Michael Jordan (born 1963) is an American basketball player
Michael Jordan (mycologist), English mycologist
Michael Jordan (footballer) (born 1986), English goalkeeper...
Michael Jordan (insolvency baron) (born 1931), English businessman
Michael Jordan (Irish politician), Irish Farmers Party TD from Wexford, 1927-1932
Michael B. Jordan (born 1987), American actor
Michael I. Jordan (born 1957), American researcher in machine learning
Michael H. Jordan (1936-2010), American executive for CBS, PepsiCo, Westinghouse
Michael-Hakim Jordan (born 1977), American professional basketball player
Michal Jordan (born 1990), Czech ice hockey player
```

Figure 3.6.: List of people with the name Michael Jordan returned by Wikipedia.

second step, the knowledge about the corresponding entity can be used to validate possible answers and therefore lead to a better overall performance and a higher user satisfaction. The following two subsections describe the design of the components to disambiguate named entities and validate answers based on the obtained information.

3.3.1. Named Entity Disambiguation

This component makes use of the semantic knowledge of the free online encyclopedia Wikipedia to detect the ambiguity of a named entity and to disambiguate it if necessary. Wikipedia has been created through decentralised, collective efforts of thousands of collaborators Remy (2002) and is permanently kept up-to-date (Bunescu and Pasca, 2006). As of 7 May 2015, the English version of Wikipedia has nearly 5 million content articles[2], which makes it the largest encyclopedia in the world. As a semi-structured knowledge base it makes comprehensive human knowledge available in an easily accessible way.

To use Wikipedia for disambiguating named entities, it needs to be understood how it organises entities and how names and content pages are linked. As described by Bunescu and Pasca (2006), there exists a many-to-many correspondence between names and entities in general, since an entity is often mapped to more than one name referring to it. For instance *Barack Obama* and *Barack Hussein Obama* are linked to the same entity. This relation is represented through *redirect* and *disambiguation* pages.

Redirect pages link all recorded alternative names for an entity to the that particular entity. In the above example, *Barack Obama* and *Barack Hussein Obama* are alternative names for the same entity. If there is a name that is linking to more than one entity, those entities are listed in a *disambiguation page*. However, Wikipedia has an important exception regarding the usage of disambiguation pages. However, if one of the entities of a disambiguation page is marked as *primary topic*, no disambiguation takes place and the user is directly referred to that particular entity. There are two major aspects for a topic to be primary:

- "A topic is primary for a term, with respect to usage, if it is highly likely-much more likely than any other topic, and more likely than all the other topics combined-to be the topic sought when a reader searches for that term.
- A topic is primary for a term, with respect to long-term significance, if it has substantially greater enduring notability and educational value than any other topic associated with that term."[3]

[2] `http://en.wikipedia.org/wiki/Wikipedia:About`

[3] `http://en.wikipedia.org/wiki/Wikipedia:Disambiguation`

If the user's question contains a named entity, the Wiki bot framework *Wiki-java*[4] is used to query Wikipedia with the name that has been parsed. The returned page to the query is then classified as either *redirect* (as a primary topic or due to its clear reference) or as a *disambiguation page*.

```
Question: Who is Bush?
Clarification question: Which of the following entities are you referring to?
    George H. W. Bush (born 1924), the 41st president of the United States of America
    George W. Bush (born 1946), the 43rd president of the United States of America
    Jeb Bush (born 1953), the former governor of Florida
    Bush family, the political family that includes both presidents
Answer: George W. Bush
```

Figure 3.7.: Clarification question of the IQA system to disambiguate the mentioned named entity.

In case Wikipedia classifies the query as a *redirect*, the IQA system links the mentioned name to the returned entity. Otherwise, the disambiguation list is parsed and a subroutine with a clarification question is invoked, in which the user is asked to choose the right entity out of the disambiguation list presented. Figure 3.7 shows a dialogue sequence with a clarification question to disambiguate the named entity.

3.3.2. Validation of Answer

This component validates the possible answers returned by the Ephyra QA system. The disambiguated named entities from subsection 3.3.1 are used to check whether the documents from which each answer has been retrieved are referring to the correct named entity.

Figure 3.8 illustrates the two highest ranked answers to the question "When was Bush born?". As noticed in figure 3.7, the name *Bush* is highly ambiguous. Consequently, the birthdays of *George H. W. Bush* and his son *George W. Bush* rank first and second among the answers and the confidence scores of both answers are close to each other.

From the component described in subsection 3.3.1, the system knows which entity is meant in the current discourse. Additionally, each answer provides a list of source documents from which the answer has been extracted. To decide which of the possible entities an answer refers to, the source documents of each answer are compared semantically with the Wikipedia article of each potential entity. After removing stopwords, the component calculates the *Term Frequency - Inverse Document Frequency (TF-IDF)* as a similarity measure of the source documents and the entities' Wikipedia articles. As a result, answers that are likely to refer to a wrong entity are penalised.

[4] http://github.com/mer-c/wiki-java

```
Question: When was Bush born?
Answers:

1) July 6th, 1946
            Confidence score: 1.0805564

            URLs of source documents:
            http://www.askhoo.org/elections/where-was-president-george-bush-born/
            http://www.conservapedia.com/Harold_Welch
            http://www.imdb.com/name/nm0124208/
            http://www.softschools.com/timelines/george_w_bush_timeline/206/
            http://www.evi.com/q/how_old_is_barbara_bush
            http://www.chacha.com/question/where-was-george-w-bush-born-in
            http://georgebushsuck.com/george-bush/Where-Was-George-W-Bush-Born.html

2) June 12, 1924
            Confidence Score: 0.64384484

            URLs of source documents:
            http://midtownblogger.blogspot.com/2014/06/born-today-george-h-w-bush...
            http://www.florencesbushangelfdn.org/#!about_us/csgz
            http://www.answerl.com/q/when-and-where-was-george-h-w-bush-born
```

Figure 3.8.: Sample output of highest ranked possible answers to the question “When was Bush born?”

4. Evaluation

To evaluate the developed IQA system presented in this thesis, a user study was conducted. In this chapter the setup of the user study is explained, followed by a discussion of its results.

4.1. Setup of User Study

Each participant of the user study was given the same list of 12 pieces of factual information that he should find out by querying the IQA system. As a benchmark, the participants were asked to retrieve the same 12 facts with the *Ephyra* QA system introduced in section 2. To ensure an objective evaluation, the participants did not know the differences between the two systems they were using. Furthermore, the order, in which the two systems were tested, was alternated for each participant to eliminate any bias in the way a system is queried or evaluated.

Figure 4.1 shows the list of 12 pieces of factual information that each participant was asked to find out using both systems. To encourage a natural dialogue structure, the 12 facts were grouped into 4 topics with 3 facts each, which is comparable to the context questions introduced in the *TREC 2001* task (Voorhees and Harman, 2001).

After finishing the dialogue with each system, the participants were asked to rate the performance of the system and naturalness of the dialogue on a scale from 1 to 5 (reaching from *not satisfied* to *satisfied* and *artificial* to *natural*, respectively). Having finished both dialogues, the participants were asked to choose the system between the two presented that they preferred, followed by some further questions about the participant. The survey has been conducted using LimeSurvey[1]. The full list of survey questions can be found in the appendix.

Nine people participated in the user study with 5 participants being male and 4 female. The age of the participants ranged from 23 to 30 years with an average age of 25 years. All dialogues of the user study have been logged.

4.2. Results of User Study

The analysis of the results from the user study is divided into the subjective evaluation by the participants and the objective measure of the number of correct facts returned by each system.

[1] `www.limesurvey.org`

```
Topic: Current Coach of German National Football Team
    Name
    Birthday
    year in which he became coach

Topic: Obama
    Profession
    Name of Wife
    Place of Birth

Topic: United Nations
    Number of Member States
    Year of Establishment
    Location of Headquarters

Topic: Marathon World Record
    Name of Record Holder
    Nationality
    World Record Time
```

Figure 4.1.: List with 12 pieces of factual information that had to be found out.

Regarding the subjective evaluation by the participants, figure 4.2 shows the satisfaction with the performance of each participant for both systems. The mean satisfaction ratings for the *Ephyra* QA system and the IQA system are 3.44 and 4, respectively. The standard deviation for the former is 0.85 and for the latter 0.67.

Figure 4.3 shows the answers to the question of how natural the dialogue was with each system. The *Ephyra* QA system scored an average of 2.22 with a standard deviation of 0.47 points. The mean score of the IQA system for the naturalness of the dialogue was 3.22 with a standard deviation of 1.09 points.

From the logs of each dialogue, the number of correct facts retrieved from each participant and system has been analysed. The results are displayed in figure 4.4. From the 12 facts that the participants should find out, the participants of the user study could obtain 4 correct facts on average with a standard deviation of 2.06 by using the *Ephyra* QA system. Querying the IQA system, participants could get 4.88 correct answers with a standard deviation of 0.78.

Even though the results cannot be interpreted as highly significant with 9 participants, there is a clear indication that the extension of the QA system to an interactive question answering system has led to a more natural interaction and especially to an improved performance and better, more precise answers. Furthermore, 8 out of 9 participants stated that they would prefer to use the IQA system over the *Ephyra* QA system.

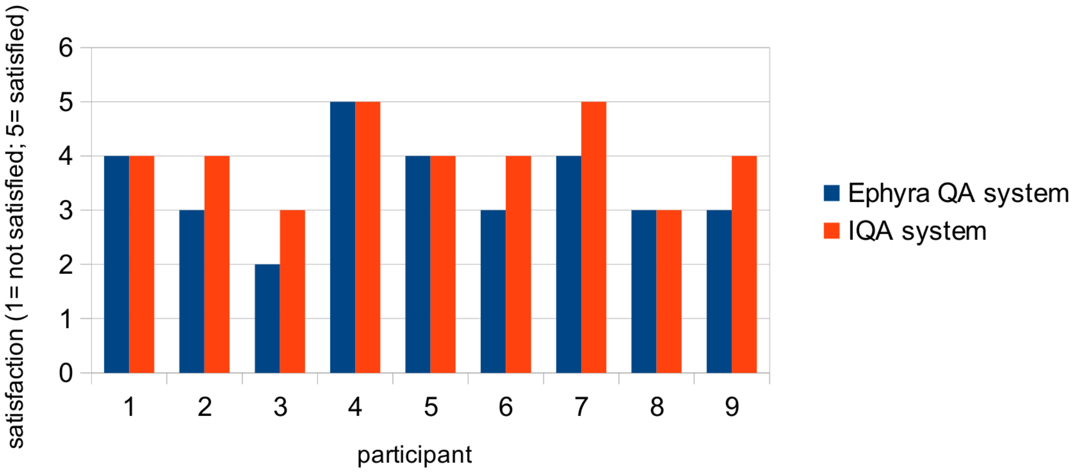


Figure 4.2.: Comparison of satisfaction with the performance of both systems for each participant.

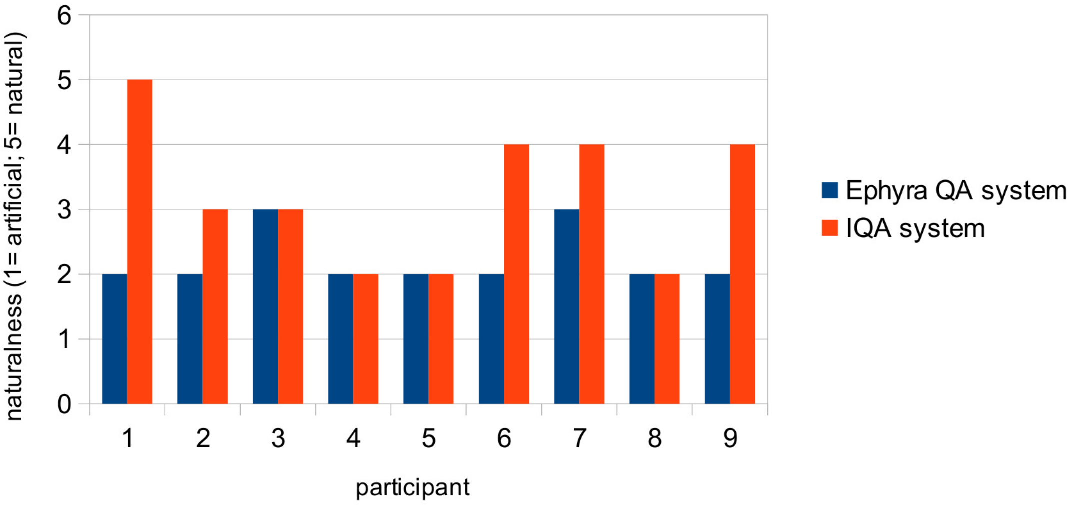


Figure 4.3.: Comparison of naturalness of the dialogues of both systems for each participant.

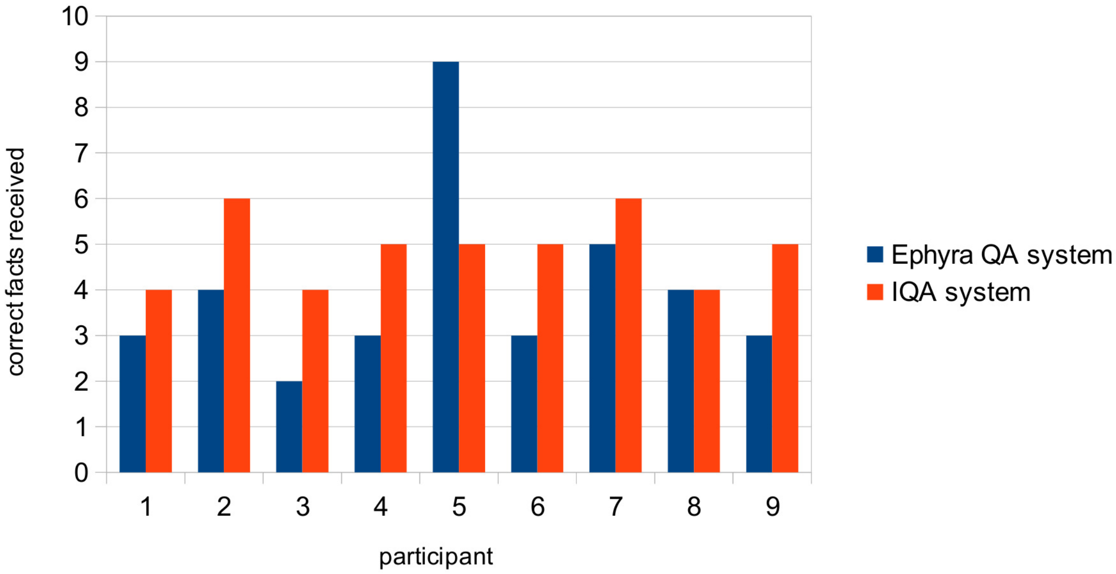


Figure 4.4.: The number of correct facts retrieved by each participant and for each system.

5. Conclusion

In this project, I designed and developed a domain-independent interactive question answering system based on the open-source *Ephyra* QA framework. The system is providing a text-based mode for interaction. When interpreting questions by the user, the system uses the current context of the dialogue and information about the entities present in the discourse to add those information to the question and to pass a fully defined question to the QA components.

If it is not possible for the system to make sense of the user input by itself, the IQA system is posing a clarification question back to the user. In that way, it can make sure that the input is interpreted correctly and, thus, reduce the probability of wrong answers.

Those clarification questions are also important for the user since he is then reassured that his input has been understood correctly. Moreover, the clarified information is used to improve the accuracy of the results returned by the system.

As a result, querying the IQA system is more intuitive and feels more naturals for human users as shown by the user study. Much more importantly, the analysis of the user study has revealed that users could retrieve more correct answers by the system when using the extended interactive question answering system.

As future work it should be analysed how the dialogues could be made even more natural and how the context-sensitive interpretation of questions could be improved further.

Bibliography

Bertomeu, N., Uszkoreit, H., Frank, A., Krieger, H.-U., and Jörg, B. (2006). Contextual phenomena and thematic relations in database qa dialogues: results from a wizard-of-oz experiment. In *Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006*, pages 1–8. Association for Computational Linguistics.

Bilotti, M. W. and Nyberg, E. (2006). Evaluation for scenario question answering systems. In *Proceedings of the International Conference on Language Resources and Evaluation.*

Blair-Goldensohn, S., McKeown, K., and Schlaikjer, A. H. (2004). Answering definitional questions: A hybrid approach. *New directions in question answering*, 4:47–58.

Bunescu, R. C. and Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *EACL*, volume 6, pages 9–16.

De Boni, M. and Manandhar, S. (2005). Implementing clarification dialogues in open domain question answering. *Natural Language Engineering*, 11(04):343–361.

Ferrucci, D., Nyberg, E., Allan, J., Barker, K., Brown, E., Chu-Carroll, J., Ciccolo, A., Duboue, P., Fan, J., Gondek, D., et al. (2009). Towards the open advancement of question answering systems. *IBM, Armonk, NY, IBM Res. Rep.*

Harabagiu, S. and Moldovan, D. (2003). Question answering.

Jurafsky, D. and Martin, J. H. (2009). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Konstantinova, N. and Orasan, C. (2013). Interactive question answering. *Emerging Applications of Natural Language Processing: Concepts and New Research*, page 149.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. (2010). A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics.

Remy, M. (2002). Wikipedia: The free encyclopedia. *Online Information Review*, 26(6):434–435.

Schlaefer, N. (2005). Pattern learning and knowledge annotation for question answering. Student Project Thesis.

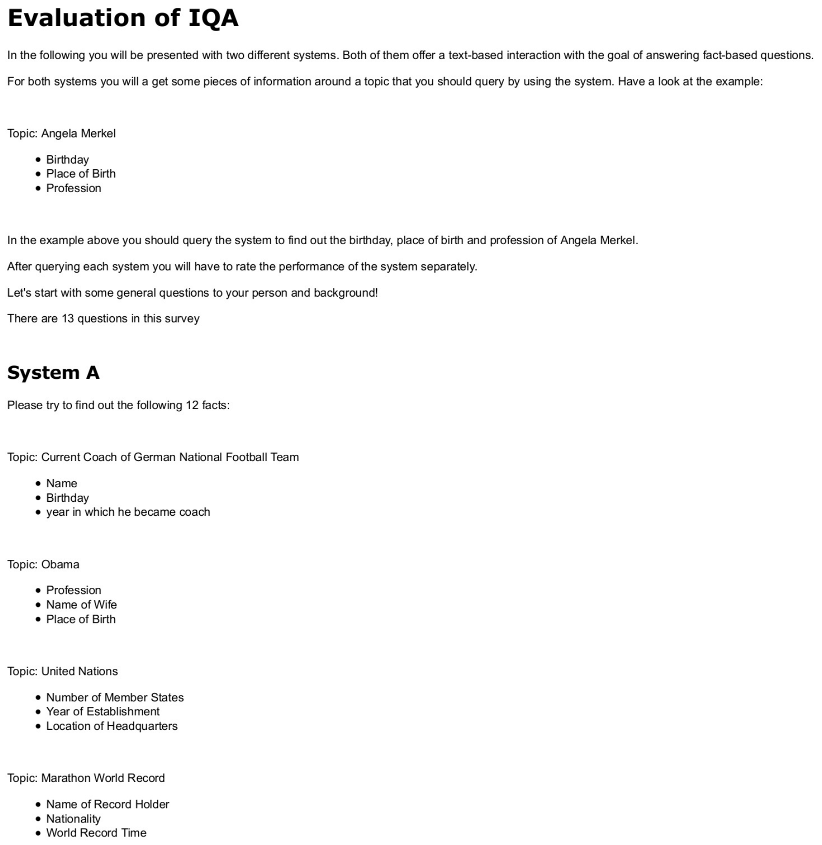
Schlaefer, N., Ko, J., Betteridge, J., Pathak, M. A., Nyberg, E., and Sautter, G. (2007). Semantic extensions of the ephyra qa system for trec 2007. In *TREC.*

Verberne, S., Boves, L., Oostdijk, N., and Coppen, P.-A. (2010). What is not in the bag of words for why-qa? *Computational Linguistics*, 36(2):229–245.

Voorhees, E. M. and Harman, D. (2001). Overview of trec 2001. In *Trec.*

Webb, N. and Webber, B. (2009). Special issue on interactive question answering: Introduction. *Natural Language Engineering*, 15(1):1–8.

A. Appendix



Evaluation of IQA

In the following you will be presented with two different systems. Both of them offer a text-based interaction with the goal of answering fact-based questions.

For both systems you will a get some pieces of information around a topic that you should query by using the system. Have a look at the example:

Topic: Angela Merkel

- Birthday
- Place of Birth
- Profession

In the example above you should query the system to find out the birthday, place of birth and profession of Angela Merkel.

After querying each system you will have to rate the performance of the system separately.

Let's start with some general questions to your person and background!

There are 13 questions in this survey

System A

Please try to find out the following 12 facts:

Topic: Current Coach of German National Football Team

- Name
- Birthday
- year in which he became coach

Topic: Obama

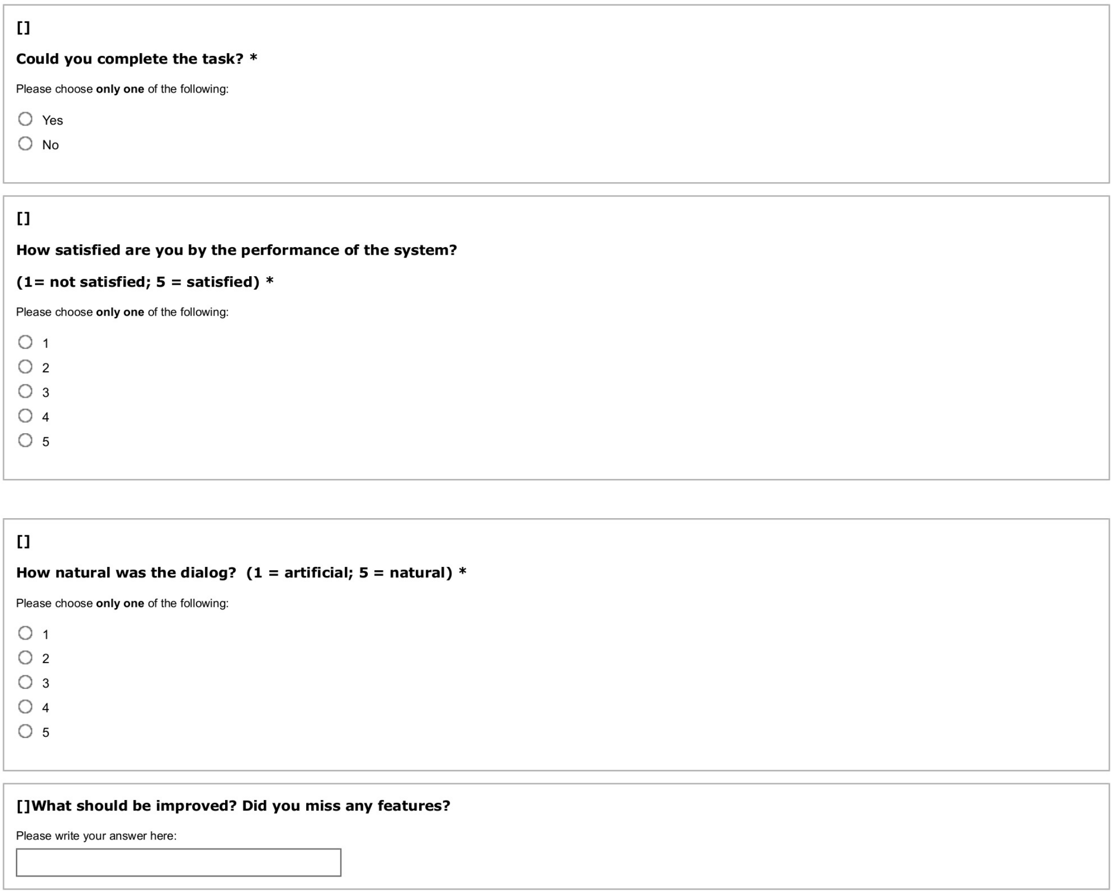
- Profession
- Name of Wife
- Place of Birth

Topic: United Nations

- Number of Member States
- Year of Establishment
- Location of Headquarters

Topic: Marathon World Record

- Name of Record Holder
- Nationality
- World Record Time



[]

Could you complete the task? *

Please choose **only one** of the following:

○ Yes
○ No

[]

How satisfied are you by the performance of the system?

(1= not satisfied; 5 = satisfied) *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

[]

How natural was the dialog? (1 = artificial; 5 = natural) *

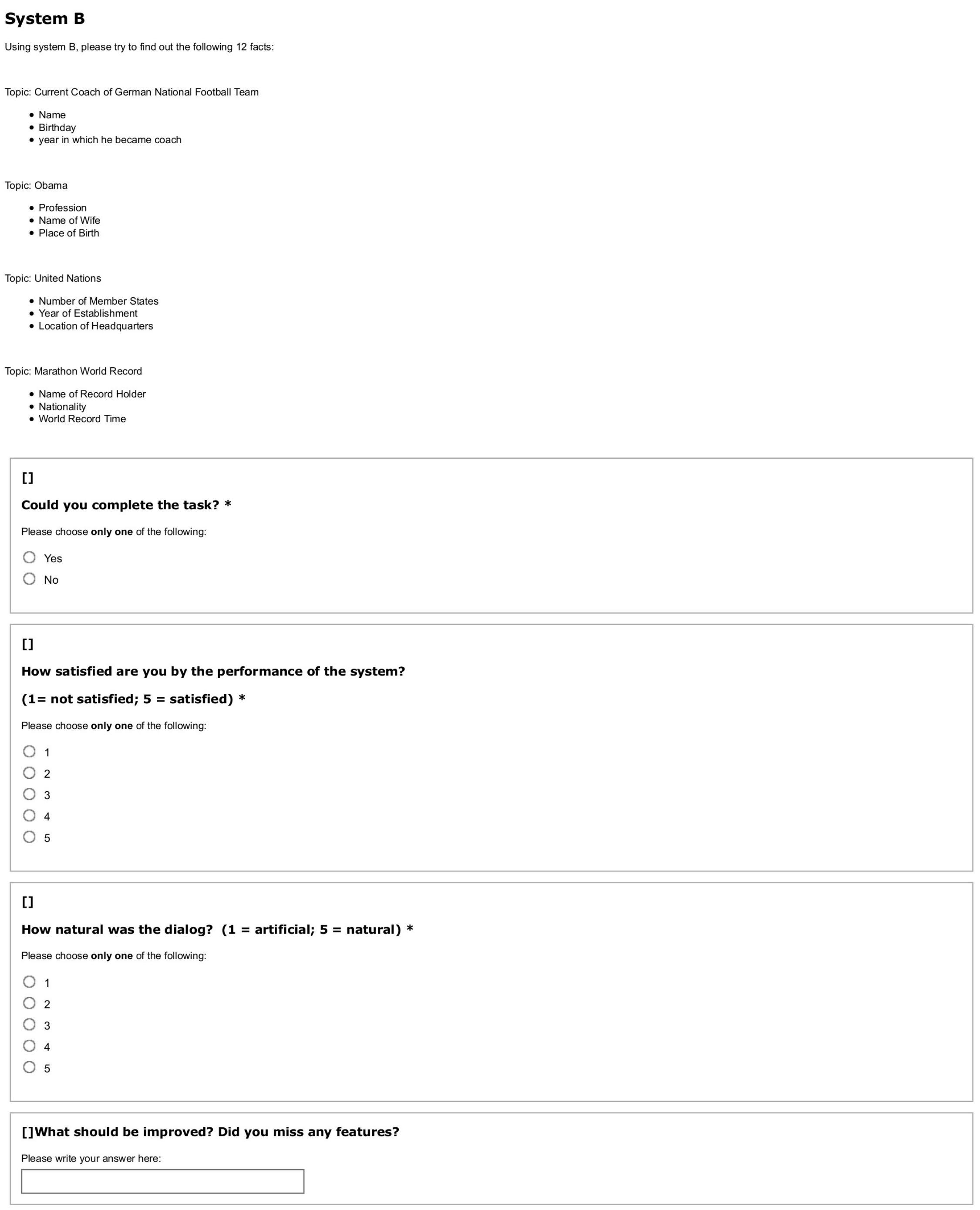
Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

[]What should be improved? Did you miss any features?

Please write your answer here:

□



System B

Using system B, please try to find out the following 12 facts:

Topic: Current Coach of German National Football Team

- Name
- Birthday
- year in which he became coach

Topic: Obama

- Profession
- Name of Wife
- Place of Birth

Topic: United Nations

- Number of Member States
- Year of Establishment
- Location of Headquarters

Topic: Marathon World Record

- Name of Record Holder
- Nationality
- World Record Time

[]

Could you complete the task? *

Please choose **only one** of the following:

○ Yes
○ No

[]

How satisfied are you by the performance of the system?

(1= not satisfied; 5 = satisfied) *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

[]

How natural was the dialog? (1 = artificial; 5 = natural) *

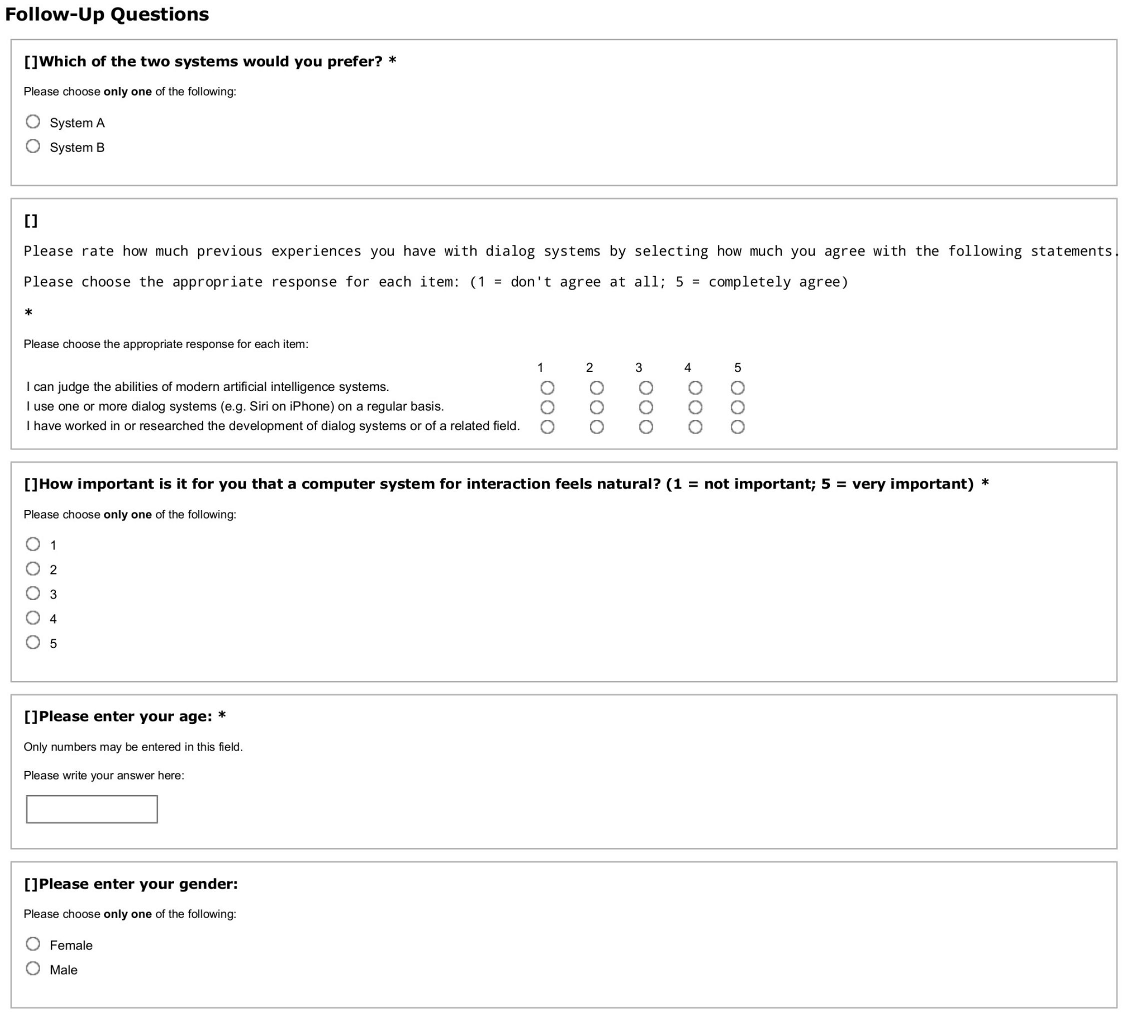
Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

[]What should be improved? Did you miss any features?

Please write your answer here:

□



Follow-Up Questions

[]Which of the two systems would you prefer? *

Please choose **only one** of the following:

○ System A
○ System B

[]

Please rate how much previous experiences you have with dialog systems by selecting how much you agree with the following statements.

Please choose the appropriate response for each item: (1 = don't agree at all; 5 = completely agree)

*****

Please choose the appropriate response for each item:

	1	2	3	4	5
I can judge the abilities of modern artificial intelligence systems.	○	○	○	○	○
I use one or more dialog systems (e.g. Siri on iPhone) on a regular basis.	○	○	○	○	○
I have worked in or researched the development of dialog systems or of a related field.	○	○	○	○	○

[]How important is it for you that a computer system for interaction feels natural? (1 = not important; 5 = very important) *

Please choose **only one** of the following:

○ 1
○ 2
○ 3
○ 4
○ 5

[]Please enter your age: *

Only numbers may be entered in this field.

Please write your answer here:

□

[]Please enter your gender:

Please choose **only one** of the following:

○ Female
○ Male