Group Assignment

Information Search and Retrieval
Graz University of Technology
WS 2012/2013

New Trends in Automatic Question Answering

– Group 7 –

Christian Gailer
(christian.gailer@student.tugraz.at)

Stefan Kohl
(stefan.kohl@student.tugraz.at)

Stephan Oberauer
(stephan.oberauer@student.TUGraz.at)

Supervisor
Univ.-Doz. Dr.techn. Christian Gütl

Institute for Information Systems and Computer Media (IICM),
Graz University of Technology, Austria
cguetl@iicm.edu and cguetl@acm.org

Abstract

The following paper focuses on *Automatic Question Answering* (Automatic QA), a subfield of Information Retrieval. The first chapters focus on historical developments and the definition of the field. Thereafter, an overview of current research topics and key aspects will be given, as well as a classification of the most interesting approaches. The main part of this paper is an analysis of new trends regarding Automatic QA, primarily focusing on approaches in association with web and new media technologies. A discussion of available tools will follow up. Finally, a summary of the things learned during the research on the topic will conclude this work.

Kurzfassung

In dieser Arbeit wird das Thema automatische Fragenbeantwortung behandelt. Es stellt eine Unterkategorie von Information Retrieval dar. Die ersten Kapitel konzentrieren sich auf Definitionen und historische Entwicklungen in diesem Bereich. Danach wird ein Überblick über aktuelle Forschungsthemen und den wichtigsten Aspekten, sowie eine Klassifizierung der interessantesten Ansätze gegeben. Der Hauptteil dieser Arbeit beschäftigt sich mit der Analyse von neuen Trends im Bereich der automatischen Fragenbeantwortung, wobei der Schwerpunkt auf Ansätzen in Verbindung mit Web und neuen Medien liegt. Anschließend werden verfügbare Werkzeuge erörtert. Schlussendlich folgen eine Zusammenfassung und sich aus der Arbeit zu diesem Thema ergebende Schlussfolgerungen.

Contents

1. Introduction 5

1.1. Problem Definition and Motivation . . . 5

1.2. Structure of the Work . . . 6

2. Definition and Historical Development 7

2.1. Structure . . . 7

2.2. Information Retrieval . . . 7

2.3. Natural Language Processing . . . 8

2.4. ELIZA . . . 8

2.5. SHRDLU . . . 9

3. Current Research Work 10

3.1. Human Question Answering Performance Using an Interactive Document Retrieval System . . . 10

3.2. Towards Automatic Question Answering over SocialMedia by Learning Question Equivalence Patterns . . . 10

3.3. An Automatic Answering System with Template Matching for Natural Language Questions . . . 11

4. Computational Linguistics 13

4.1. Recent Methods . . . 13

4.1.1. Maximum Entropy Model . . . 13

4.1.2. Decision Tree Learning . . . 13

4.1.3. Artificial Neural Networks . . . 13

4.2. Question Answering (QA) . . . 14

4.3. Additional Applications . . . 15

4.3.1. Information Extraction . . . 15

4.3.2. Machine Translation . . . 16

4.3.3. Natural Language Generation . . . 16

4.4. Issues . . . 17

5. Existing Tools 18

5.1. Commercial Software . . . 18

5.1.1. Apple Siri . . . 18

5.1.2. Google Now / Knowledge Graph . . . 19

5.2. Watson . . . 20

5.2.1. Technology . . . 20

5.2.2. Application . . . 21

5.3. Wolfram Alpha 22
5.3.1. Technology 23
5.3.2. Evaluation 23
5.4. START Natural Language System 24
5.4.1. Technology 24
5.4.2. Evaluation 25

6. Lessons learned 26

7. Conclusion and Future Work 28

A. Appendix 29
A.1. List of Figures 29

References 30

1. Introduction

1.1. Problem Definition and Motivation

In the beginning of *Automatic QA* the studies where in a shape of creating an intelligent computer system, which can interact with a human being. It evolved from simple interaction systems without a knowledge database relying on a domain specific field to complex systems, which are web-scaling and able to answer elaborate questions in an interactive and context based way.

Normal key word based search engines rely on the replying of a set of relevant documents to a specific query. As the information offer is constantly growing in the internet it can be hard to find correct information, so it should be possible to get specific answers to a query. One reason for Automatic QA is to relieve the crawling through many documents for the correct answer. On the other hand also supporting documents can be useful to confirm the answers so users get convinced that the information is relevant.

In the age of the rise of mobile phones users want to get quick and accurate answers to their questions. Like in Apples Siri or Googles Search speech recognition gets more popular and with this system the Automatic QA is in great demand. What comes with these systems is the challenge of understanding the question and respond to them in a convenient way.

Automatic QA systems nowadays are more interactive and result in dialogues to get related information to complex questions as a set of queries. As in older approaches single questions could be answered with a simple response, these days Automatic QA systems need to interact with the user. Also spelling mistakes need to be covered, maybe if there are non native speaker questions. As many answers to specific questions are not available in databases, the information has to be extracted out of documents.

Automatic QA is a huge field of diverseness. There exist many different concepts of analysing the question with computational linguistics and natural language processes, retrieving the documents and extract relevant information for the answers, mapping the documents to the answers or to reply to the user and the interaction with the human.

The popularity of social media has grown in the last years, in these systems it is mostly necessary to make a search or a query. For some social media platforms the alternative search has built an effective counterpart to the usual web search. For Automatic QA social media platforms offer a good field of activity. With its closed system and the large datasets it is possible for Automatic QA systems to easily answer questions about e.g.:

user activities, hobbies or preferences.

1.2. Structure of the Work

In this paper we focus on new trends in Automatic QA. In chapter 2 we will focus on the definition and historical development, which includes the structure of a Automatic QA system and the basic approaches. Also 2 early systems will be discussed here. Chapter 3 contains the current research in this science field, which faces interactive document retrieval, social media approaches and template matching. In the fourth chapter the subject of computational linguistics will be discussed. Chapter 5 shows existing tools, which rely on the idea of Automatic QA. The lessons we learned about this work will then be covered in chapter 6. At the end chapter 7 draws a conclusion and shows future work.

2. Definition and Historical Development

2.1. Structure

The Automatic QA Process can be divided into 4 major parts:
Question Analysing, Preparing the Dataset, Text Processing and Data Mapping. At first the question needs to be analysed to identify the semantic of the possible answers. The preparation of the dataset includes the preprocessing of the document collection and the selection of the most valuable documents. The step of the text processing contains the filtering of the answers in the document collection. In the end the data mapping forms the list of answers by extracting the text from the documents. (Blooma, Chua, Goh & Keong, 2009)

In the paper of J. Cowie from 2000(Cowie, Ludovik, Molina-Salgado, Nirenburg & Scheremetyeva, 2000) it is mentioned that Question Answering is a pure information retrieval process. Nowadays it can be divided into three major approaches, which are a main part of the question answering system. These approaches are *Information Retrieval*, *Natural Language Processing* and the *template-based question answering model.* (Andrenucci, 2008)

2.2. Information Retrieval

The *Information Retrieval* part in the *Automatic QA* project takes care of finding documents, which contain useful information for the question answering. As it was common in the previous years that in *Automatic QA* only document retrieval was used, the information retrieval approach nowadays can also deliver text passages where valuable information can be extracted. The amount of documents can be a problem in retrieval systems. A possibility to avoid this is to work in closed document collections, therefore Indexing and querying can be done in a less time consuming task. An advantage of the information retrieval approach is that it is not so much justified by the context, because of the extraction of the words from a document, which are present in the users question.(Andrenucci, 2008)

To get out the relevant passages an information retrieval system like OKAPI can be used. In the paper (Plamondon & Kosseim, 2002) they use OKAPI for their QA system QUANTUM to retrieve the relevant paragraphs as variable length passages.

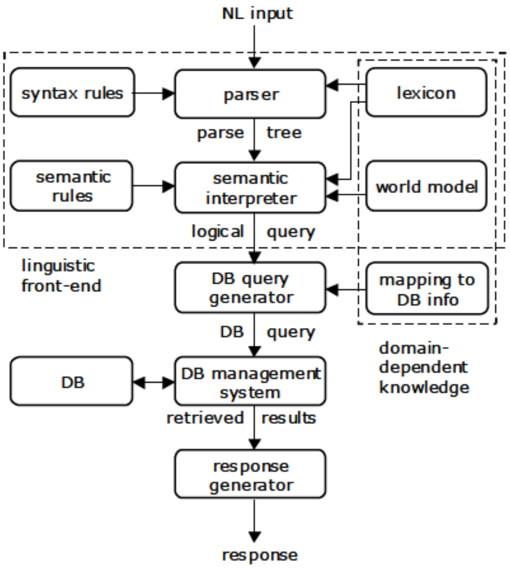


Figure 2.1.: Architecture of natural language processing system, Androutsopoulos et al., 1995

2.3. Natural Language Processing

The Natural Language Processing approach performs a semantic analysis of text. It therefore uses machine learning algorithms to learn rules for text analysis. It uses sets of theories and technologies. In the early years the most common algorithms were decision trees, the latest approaches base on statistical model and probalistic decisions.(Gunawardena, Lokuhetti, Pathirana, Ragel & Deegalla, 2010)

The figure 2.1 shows the process from Natural Language input till the generation of the response. In the figure it shows that there are a few steps to analyse the input of the user. At first the input is parsed with a lexicon and its syntax rules. The next step is about semantic interpretation, after this the information is mapped to the database management. The last step is to generate a response.

An advantage of the Natural Language Processing System are the probabilistic decisions. The main drawbacks of this approach rely on its computational heaviness, the long processing time and the difficulties of portability. (Andrenucci, 2008)

The subtopic Natural Language Understanding is the process of disassembling input and represent it in a more formal part. Therefore it is easier for machines to interpret the input. It is used in the first QA programs ELIZA and SHRDLU.

2.4. ELIZA

The science field *Automatic QA* evolved in the late 1960s. The first systems which where developed had a very specific domain to limit the possible answers and so make it easier for the system to answer most of the questions.

In 1966 Joseph Weizenbaum implemented one of the first kind of Automatic QA sys-

tem the so called ELIZA. The system relies on natural language processing and formed a question out of a statement by using simple pattern matching. It worked on a MAC time-sharing system at MIT. The program could run different scripts to simulate a human conversation partner.
One of the most famous scripts was the DOCTOR script, which simulated a psychotherapist. The main purpose of the program was to act as human as possible. Key words were filtered out of the input sentences and with simple rules answers were than created. If there was no key word found, then a content-free early transformation was used to reply. The keywords and transformation rules were stored in the scripts and not from the program itself.

Weizenbaum mentioned 5 fundamental problems in his paper (Weizenbaum, 1966) about the ELIZA systems:

1. Identification of key words
2. Discovery of minimal context
3. Choice of appropriate transformation
4. Generation of responses in the absence of key words
5. Provision of an editing capability for ELIZA scripts

As this program was only key pattern matching it had no real world knowledge, the system could only direct a user into a subsequent direction. The program so far could not answer a question, it was only remodel the sentences of the user and make a general statement.

2.5. SHRDLU

Another early natural language processing computer program was the so called SHRDLU. It was implemented from 1968-1970 at MIT from Terry Winograd. In the book *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language* Winograd describes the program, which is capable of answering questions, executing documents and simulating an English dialogue.

The main purpose of the computer system was to follow instructions from users, understand it and execute them. To make it easier for the system, the complexity was restricted to use basic block objects. In the simulation environment it could put simple objects like pyramids or blocks into a box and distinguish between the objects and its colors. It also had a small context related understanding and with its memory and rules it could calculate possible use cases for its world. In his book Winograd also mentioned that for understanding the input of a user, the computer needs to figure out the subject of the question.
At the end he concluded that if there was a better parsing also a semantic analysis could be included in his work what could involve to a search in a association net. This could improve the system by recognising meanings of words or to reduce the subjects to generate a more specific answer.(Winograd, 1971)

3. Current Research Work

3.1. Human Question Answering Performance Using an Interactive Document Retrieval System

In 1999, the TREC Question Answering (QA) was the first evaluation of QA systems. The goal was to return short answers to a particular question. In a survey of volunteers it turned out that the majority prefer a direct answer rather than a document in which they must search for the answer. (Voorhees, 2006)

In this paper a comparison of the performance of document retrieval systems and QA systems was employed. Here, the ability of the users answering the questions on their own using an interactive document retrieval system was analyzed. The result was compared and evaluated in terms of performance compared to QA systems.

It was observed that the users were able to find the information successfully using the document retrieval systems. However, it was found that the performance can be increased considerably using QA systems. (Smucker & Allan, 2012)

3.2. Towards Automatic Question Answering over SocialMedia by Learning Question Equivalence Patterns

In this paper, an approach is shown to answer new questions automatically in Collaborative Question Answering (CQA) systems. It is accessed on an existing archive, in which users answer each other questions. It is believed that many questions to be asked have already been asked and answered. This gives the possibility to find answers automatically for new questions.

The authors in this work assume that many questions are syntactically different but semantically equivalent. This raises the problem of dividing these questions into equal groups. The assignment to the different groups is based on the best answers that were marked by the askers as such. For each group equivalence patterns are generated for questions having syntactic similarities. These equivalence patterns are being compared with new questions. The best answer to a previously asked question is returned.

A filtering method is used to reduce accidental semantic similarities in questions with the same answer. Thereby results can be further improved.

200,000 questions from the Yahoo! Answers archive were used as test data. Regarding this approach a 66% precision can be achieved for new questions. (Hao, Liu & Agichtein, 2010)

3.3. An Automatic Answering System with Template Matching for Natural Language Questions

In QA systems there is a general distinction between two problems: open domain problems and closed domain problems. In open domain the answers to questions can be found in public information sources. Almost any question can be placed. In the closed domain, however, answers are stored by a domain expert in a database. The permissible questions are therefore limited to a specific topic. This work focuses on the closed domain problems.

To get a useful response of the natural language question, template matching is applied. It is aimed to provide a service for cell phones by SMS. Besides English, the system understands the typical SMS language. Since it is a closed system domain Frequently Asked Questions (FAQ) are used as sample data. The system could also be used for any other closed domain problems.

Figure 3.1 shows the system architecture.

The system is split up into three main modules:

1. pre-processing module
2. question template matching module
3. answering module

In the first module SMS abbreviations are converted to English words and stop words are removed.

In the second module the result from the previous module is matched with every template to find the best match. They use a special syntax to describe the templates. For each question the template has to be written manually. Furthermore the template matching approach has been improved by using synonym lists and adding robustness relating to spelling mistakes.

Every question has a defined answer. The third module returns the related answer of the question that matches best.

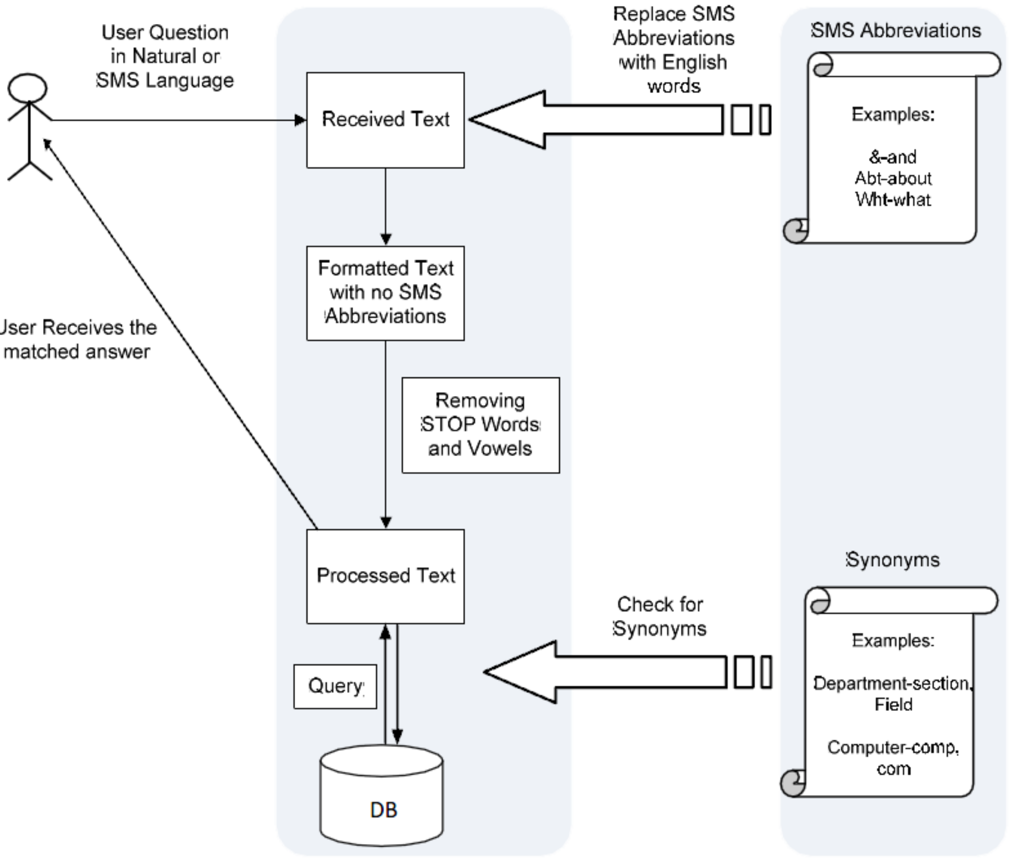

User Question in Natural or SMS Language
Received Text
Replace SMS Abbreviations with English words
SMS Abbreviations
Examples:
&-and
Abt-about
Wht-what
Formatted Text with no SMS Abbreviations
End User Receives the best matched answer
Removing STOP Words and Vowels
Processed Text
Synonyms
Check for Synonyms
Examples:
Department-section, Field
Computer-comp, com
Query
DB

Figure 3.1.: System architecture of the question answering system

The system was tested at an exhibition over seven days with thousands of users. The system worked as expected, even though it still had problems with handling SMS abbreviations and spelling mistakes. (Gunawardena et al., 2010)

4. Computational Linguistics

Computational linguistics (CL) and natural language processing (NLP) have always been important parts in the field of artificial intelligence. The processing of the language is a key element of human intelligence and offers a range of applications. (Patten & Jacobs, 1994)

In the last three decades, the field of CL has developed enormously. At that time it was a mixture of artificial intelligence and formal linguistics. Now it is an independent scientific discipline and already has a strong presence in the industrial development. (Clark, Fox & Lappin, 2010)

4.1. Recent Methods

This section gives a general overview of some recent methods in CL and NLP.

4.1.1. Maximum Entropy Model

The basic idea of the maximum entropy model is, that the probability distribution which best represents the current state of knowledge is the one with the largest entropy. It is used to build up models of many different sources with limited information. Maximum entropy models have been used for many applications, e.g. for tagging and parsing. (Mareček, Popel & Žabokrtský, 2010)

4.1.2. Decision Tree Learning

Decision tree learning is a method used in many domains of knowledge discovery, pattern recognition and data mining. Decision trees are hierarchical trees which try to predict an output variable for a given input. Each leave represents an attribute. Each path is a conjunction of the attributes which are on the path. The main advantages are that they are very intuitive for humans and have a good accuracy. (Barros, Cerri, Jaskowiak & de Carvalho, 2011)

4.1.3. Artificial Neural Networks

Artificial neural networks are mathematical models based on the biological neuron structure of the brain. The neurons, also called nodes, together with the weighted connections are the basic components. The advantages of artificial neural networks are the possibility of enhancing the processing speed by parallelization, adaptation of knowledge, robustness and implementation in low power applications. (Huang & Zhang, 1994)

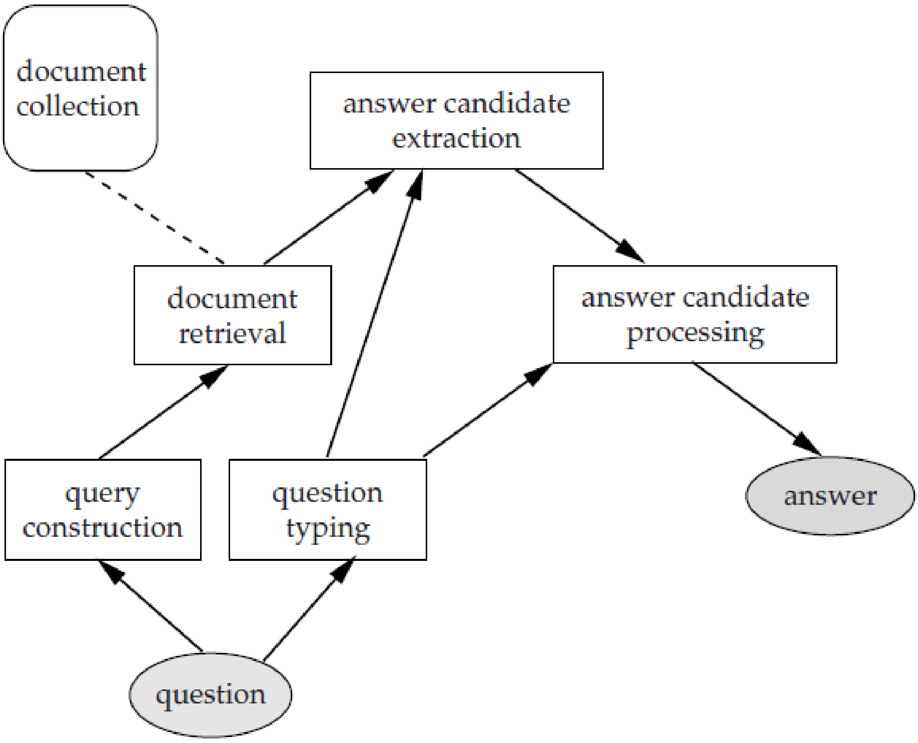


Figure 4.1.: Basic architecture of a QA system

4.2. Question Answering (QA)

The Internet provides an ever-growing amount of information. To gain access to the desired information, there are a variety of types of search methods. Conventional search engines usually provide a list of relevant documents, in which the requested information can be found. However, the goal of QA is to provide the information immediately to the user without having to search through all documents.

The following sections refer to (Clark et al., 2010).

In earlier QA systems, the information was gathered from databases with a fixed structure. Such historical systems were presented in Chapter 2. This procedure does not provide the desired results. For this reason, attempts are made, in order to find the desired information and extract it from a range of information sources instead of computing the information from a database. This kind of QA is called *open domain QA*. Its development is advanced by the Text Retrieval Conference (TREC).

In current QA systems an answer is dispensed as a result of several tasks on the input question. Figure 4.1 shows the basic architecture of a QA system. Each task can be described as follows.

- *Question typing*: What is the question type? What kind of information is sought? e.g., an abbreviation, definition of a specific word, specific information about a person

- *Query construction*: In which documents the answer could be found?
- *Text retrieval*: The way in which the query is created depends on the used text retrieval form. It can be distinguished in two types:
  - *Relevance-based retrieval*: Passages are retrieved that are relevant to the topic. The relevance is determined by, for example, weighted vectors (vector space model).
  - *Pattern-based retrieval*: Although languages usually have several different expressions of the same meaning, it is assumed that the answer is somewhere in the text formulated in the same way as in the question. In contrast to the relevance-based retrieval pattern-based retrieval uses these text blocks instead of just referencing it.
- *Answer candidate extraction*: Which passages or text modules are considered for building up the answer?
- *Answer candidate processing*: The different responses are evaluated. The best response is issued.

4.3. Additional Applications

In addition to QA, there are many other applications of CL and NLP which overlap partially. Some common applications are listed here.

4.3.1. Information Extraction

Information extraction is the application of automated procedures to extract information from an unstructured resource of a previously defined area. The extraction of information can be divided into three tasks. (Grishman, 2003)

Entity extraction: It identifies and classifies all phrases in a free text which refer to objects of semantic classes like names, nouns, pronouns etc. In addition all object mentions are linked together which refer to the same entity.

Relation extraction: The relations between entities are identified. A relation is always represented by two entities and can be described in many languages. A common framework to describe relationships for web data is Resource Description Framework which was orginally designed as a meta data model.
e.g.: relationship between a person and a location - "Peter lives in Vienna"

Event extraction: It is also a common application in order to derive specific knowledge from a text. Event extraction identifies events of particular types and the corresponding arguments. A type of an event would be for example "car crash" or "natural disaster".

Examples: Generation of meta data of unstructured documents, restructuring of large amounts of data, data mining with wrappers

4.3.2. Machine Translation

There are many different approaches to machine translation like hierarchical, rule-based, example-based, tree-based and hybrid approaches. The state of the art approach is phrase-based statistical machine translation which will be discussed in this section.

In contrast to the word-based translation, the aim of the phrase-based translation is to translate whole sequences of words rather than single words. Thereby the quality of the translation can be increased enormously. The length of the phrases and the word order can be different from language to language. A classification of the typical word order of subject, object and verb is therefor necessary. Phrases are limited to a number of three and are found by means of statistical methods.
A translation model is created which is based on the noisy channel model. Thus, Bayes' rule is used to compose the translation probabilities. (Koehn, Och & Marcu, 2003)

Examples: Automatic translation of text from one language into another

4.3.3. Natural Language Generation

Natural language generation is the process to construct a natural language text. The text should be syntactically and semantically correct and provide a formal presentation of the content. The challenge is to imitate human language ability by using computational effort.

Claude Shannon researcher has laid important foundations for natural language generation. Shannon introduced the ability to automatically generate text using Markov transition probabilities from one word to another in his paper "A Mathematical Theory of Communication" in 1948. He created the first theoretical model of a text generator. (Shannon, 1948)

For basic applications, the text generation process can be kept very simply by composing the text with standard text blocks and some link words. For more complex systems, however, more processing steps, as described in (Reiter & Dale, 2000), are needed.

The process is composed of several stages.

- *Content determination*: What information should be included in the output text
- *Document structuring*: Which content parts should be grouped
- *Lexicalisation*: What specific words should be used
- *Referring expression generation*: What expressions should be used to reference objects
- *Aggregation*: How the linguistic structure as sentences and paragraphs should be mapped
- *Linguistic and structural realisation*: Transformation of abstract representations of sentences, paragraphs and sections into real text

Examples: Weather forecast, automated reports, generation of jokes, summarisation of financial data or medical records, describing data for blind people

4.4. Issues

NLP is considered an AI-complete (artificial intelligence complete) problem in analogy to the complexity class NP-complete. This informal term was first used in Erik Mueller's Ph.D. dissertation. (Mueller, 1987) As an AI-complete problems are known, which can not be solved with a simple specific algorithm. It is believed that in addition to NLP computer vision or generally unexpected events in real world belong to this problem.

Artificial Intelligence systems can already do simple limited parts of such AI-complete problems. Once unexpected circumstances that do not move within the known problem arrise, there are problems. People can always fall back to the advantage of previously acquired skills. (Lenat & R. V. Guha, 1989)

5. Existing Tools

This chapter focuses on current existing tools implementing Automatic QA. At the beginning, two commercial applications will be analyzed before having a closer look on the three scientific solutions of a greater scale, namely IBM's *Watson*, *Wolfram Alpha* by Wolfram Research, and *START* by Boris Katz et al.

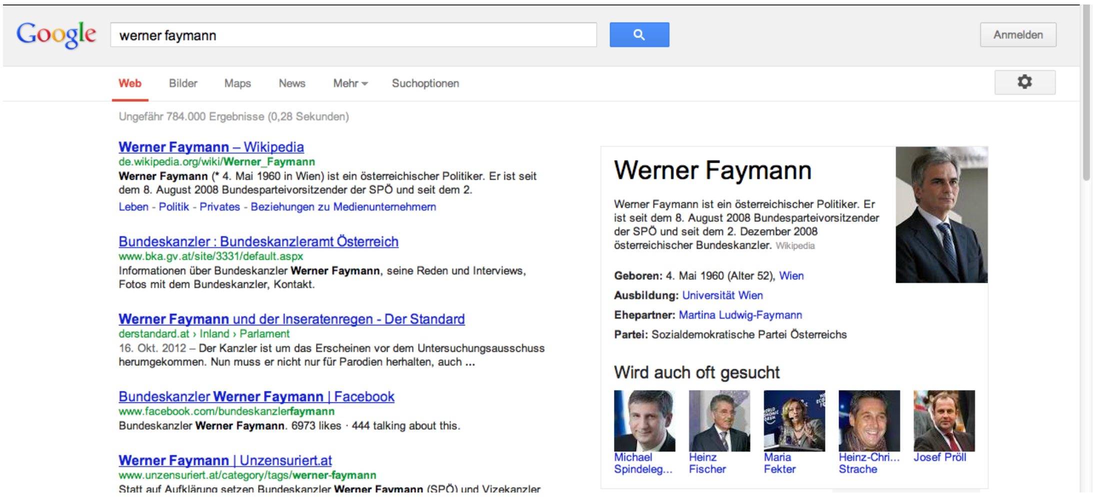
5.1. Commercial Software

5.1.1. Apple Siri

Apple's famous assistant *Siri* (Speech Interpretation and Recognition Interface) originates from the CALO project (Cognitive Assistant that Learns and Organizes), a project funded by DARPA and realized by SRI International. CALO focused neither on NLP nor on Automatic QA but on learning in an everyday's environment. Siri, Inc. was later founded by members of SRI who planned to develop an assistant which uses natural language processing. The company was acquired by Apple in April 2010. (SRI, 2011)

Siri consists of several components, forming a pipeline from the spoken word to the execution of a command or a query. In comparison to other systems, Siri uses speech recognition to acquire input from the user. This process precedes the actual question answering task and might be, along with the user, an additional source of inaccuracy. Different accents and language habits have always been the most challenging aspect of speech recognition. Siri is able to deal with these problems, some effort for calibration is needed, though. (Frädrich & Anastasiou, 2012) Other sources drew a different conclusion, criticizing Siri for only reaching 68 percent accuracy. (Scott, 2012) Nonetheless, Luc Barthelet from Wolfram Research states that "the technology [has] reached a threshold where people overall like it". Even if the field of research did not progress that much in the last 15 years, the popularity of the iPhone propagated better acceptance of intelligent software assistants and their technologies. (Geller, 2012)

As said before, Siri is capable of executing commands on the smartphone, regarding schedule, contacts, communication, etc. as well as searching for information. When asking a question (and thus using the latter capability), Siri relies on third-party web services to retrieve an answer. Which service to use is determined by the context of the question. Asking for restaurants causes Siri to query *Yelp*, information about stocks is retrieved from *Yahoo! Finance*. But the most interesting service is *Wolfram Alpha*, providing Siri with answers to factual questions. Wolfram Alpha will be analyzed later in this chapter in greater detail. (Wofford, 2011) Anyway, this web-service was the reason why Siri recommended the Nokia Lumia as the best smartphone, an issue that made Siri



Ungefähr 784.000 Ergebnisse (0,28 Sekunden)

Werner Faymann – Wikipedia
de.wikipedia.org/wiki/**Werner_Faymann**
Werner Faymann (* 4. Mai 1960 in Wien) ist ein österreichischer Politiker. Er ist seit dem 8. August 2008 Bundesparteivorsitzender der SPÖ und seit dem 2.
Leben - Politik - Privates - Beziehungen zu Medienunternehmern

Bundeskanzler : Bundeskanzleramt Österreich
www.bka.gv.at/site/3331/default.aspx
Informationen über Bundeskanzler **Werner Faymann**, seine Reden und Interviews, Fotos mit dem Bundeskanzler, Kontakt.

Werner Faymann und der Inseratenregen - Der Standard
derstandard.at › Inland › Parlament
16. Okt. 2012 – Der Kanzler ist um das Erscheinen vor dem Untersuchungsausschuss herumgekommen. Nun muss er nicht nur für Parodien herhalten, auch ...

Bundeskanzler **Werner Faymann** | Facebook
www.facebook.com/bundeskanzler**faymann**
Bundeskanzler **Werner Faymann**. 6973 likes · 444 talking about this.

Werner Faymann | Unzensuriert.at
www.unzensuriert.at/category/tags/**werner-faymann**
Statt auf Aufklärung setzen Bundeskanzler **Werner Faymann** (SPÖ) und Vizekanzler

Werner Faymann

Werner Faymann ist ein österreichischer Politiker. Er ist seit dem 8. August 2008 Bundesparteivorsitzender der SPÖ und seit dem 2. Dezember 2008 österreichischer Bundeskanzler. Wikipedia

Geboren: 4. Mai 1960 (Alter 52), Wien
Ausbildung: Universität Wien
Ehepartner: Martina Ludwig-Faymann
Partei: Sozialdemokratische Partei Österreichs

Wird auch oft gesucht

Figure 5.1.: Knowledge Graph presents facts related to the search term aside the search results.

to hit the headlines. This happened due the metrics Wolfram Alpha used to evaluate the quality of smartphones. (Sullivan, 2012) Finally, there is one restriction to the non-english versions of Siri: Wolfram Alpha is only available in English and therefore not used in other localizations. (Schwan, 2011)

5.1.2. Google Now / Knowledge Graph

Google Now is an intelligent software assistant developed by Google and used on Android smartphones. It is commonly seen as Android's counterpart of Apple Siri and thus handles similar tasks. Additionally, context-aware and location-aware data is used to display information that might be relevant to the user in his or her current situation. Question answering is done by using another service of Google, called *Knowledge Graph*, which was launched in 2012. KG goes the similar way Wolfram Alpha does, moving away from "being an information engine to being a knowledge engine", as Johanna Wright, product management director of Google, explained in a video on Google's YouTube channel.[1]. In Knowledge Graph, information is represented by graphs and knowledge is retrieved by querying the relations between entities. (Singhal, 2012) Wolfram Alpha, in contrast to Knowledge Graph, tends to handle knowledge in a more computational way (see below). The results are already available on Google Search when searching for well-described entities (figure 5.1).

As conclusion, despite the different way Google incorporates user data in its assistant, Now and Siri are using a similar overall architecture, doing speech recognition, natural language processing and command invocation on the mobile client, while question answering is performed on the remote service providers Wolfram Alpha and Knowledge Graph.

[1] Introducing the Knowledge Graph: http://www.youtube.com/watch?v=mmQl6VGvX-c

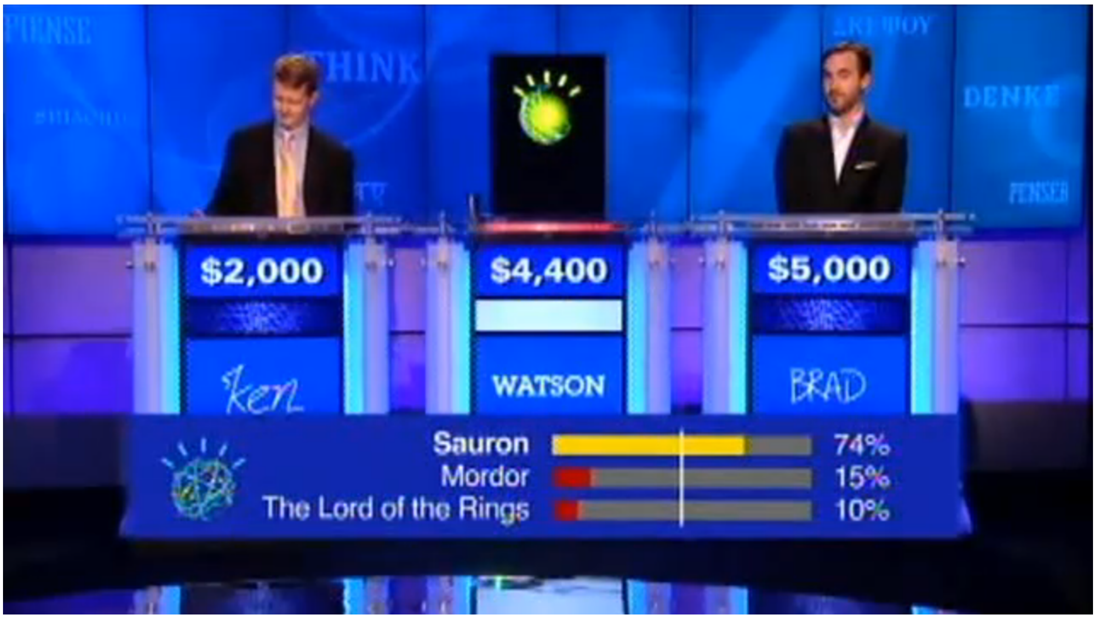


Figure 5.2.: Watson's evaluation results at *Jeopardy!*

5.2. Watson

IBM's *Watson* is known to a greater audience by his "performance" at *Jeopardy!* where the system beat its human competitors. Achieving this goal was one of the main challenges during research and development. Jeopardy, thus always focused by the Watson research team, is a famous, ongoing American quiz show with three participants competing against each other. The main requirements a participant must meet are *precision*, *confidence*, and *speed*, as formulated in (Ferrucci et al., 2010). Watson accepts natural language questions and uses a probabilistic approach, which differs from most decision-tree driven systems. IBM promotes three main capabilities, *Natural Language Processing*, *Hypothesis generation and evaluation* and *Evidence-based learning* on their website.[2]

5.2.1. Technology

The software behind Watson is called DeepQA, "a software architecture for deep content analysis and evidence-based reasoning". (DeepQA Research Team, 2011) According to the cited authors, Automatic QA cannot be solved by a single algorithm or program. Therefore, a system for deep content analysis and parallel hypothesis generation and evaluation will be necessary to accomplish this task. Figure 5.3 depicts the architecture of DeepQA, designed to fulfill these requirements. The architecture itself must be able to adapt for solving different kind of problems, like finding questions to given answers as it is done at Jeopardy. DeepQA processes the user input and generates different intermediate answers, which will then be evaluated for the best choice. This kind of reasoning – analyzing different hypotheses – is a key aspect of Watson and was also considered at Jeopardy by displaying the top three highest rated possibilities in an information panel

[2] What makes Watson different.Retrieved December 2012, from http://www-03.ibm.com/innovation/us/watson/what_makes_watson_different.shtml

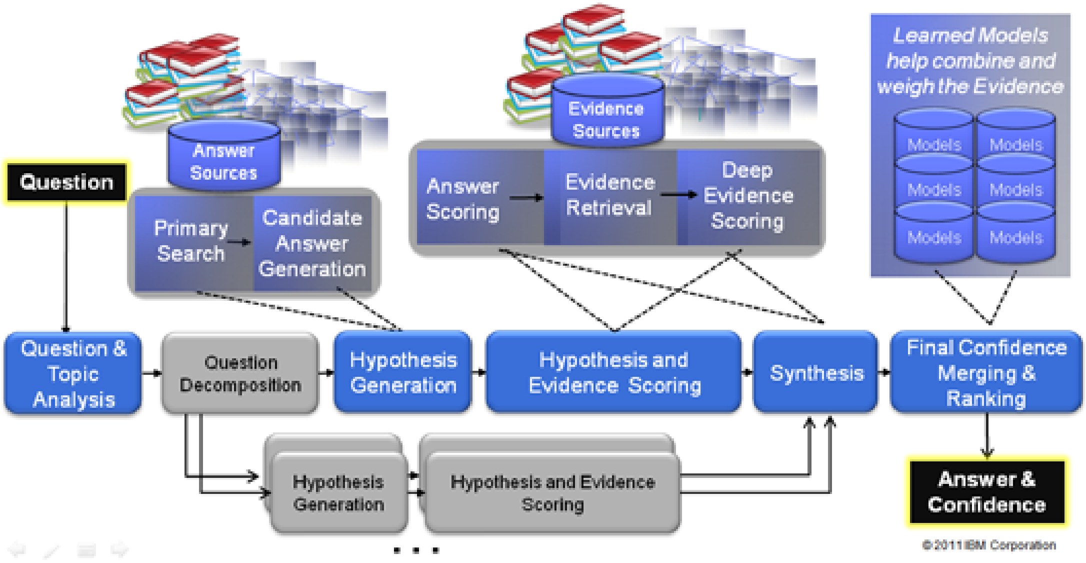


Figure 5.3.: DeepQA Software Architecture (Copyright: IBM Research)

(see figure 5.2). The DeepQA approach, as described by (Ferrucci et al., 2010), splits the entire process of Automatic QA into these three steps:

(1) **Content Acquisition:** The first step focuses on acquiring the content needed for answering questions and providing evidence. This mainly consists of manually analyzing example questions and automatic statistics on the application domain to determine the basic structure of the answer. Afterwards, this baseline is fed with information from appropriate sources during the *corpus expansion process*. This is done initially to prepare for the actual question answering.

(2) **Question Analysis:** Step 2 is processed during run-time when a query is performed. Watson analyzes the question by “mixture of experts” (i.e. analyzing the question on different levels, e.g. logically, semantically, etc.) to find out that the user wants to know and how to proceed to find an answer. This will be done within four sub-steps in the following order: question classification, focus and lexical answer type detection, relation detection, and decomposition.

(3) **Hypothesis Generation:** Finally, the results of the previous steps will be used as candidate answers and evaluated to find out which answer is most likely regarding the question.

5.2.2. Application

On their website[3], IBM Research proposes two main professional fields of application best suited for working with Watson:

Healthcare: In healthcare, one out of five decisions made are incorrect. However, lots of data and reports exist due publication in journals. The amount is doubling every five

[3] More than a quiz show champion. Watson goes to work. Retrieved December 2012, from IBM-2012:http://www.research.ibm.com/articles/watson.shtml

years, yet polls state that many doctors only spend five hours a month or less for reading journals. The information for making adequate decisions is available but hard to find in the vast amount of records. In addition, many reports imply further knowledge that has to been looked up. Watson can be used to retrieve information queried by users. The system takes data and notes about the patient and incorporates them in already available knowledge (e.g. from relatives, diagnosis, research, etc.) to form a hypothesis. Watson is already utilized in healthcare since 2011. In 2012, it was announced that IBM and Memorial Sloan-Kettering Cancer Center formed a partnership to use Watson diagnosis and treatment finding.

Finance: Another field of application is Finance because of its huge amount of data. Even within one single day many thousands of reports and millions of emails are created, as well as information about all financial transactions. IBM proposes to use Watson for recommendations on financial products by analyzing the market situation and the customers financial background.

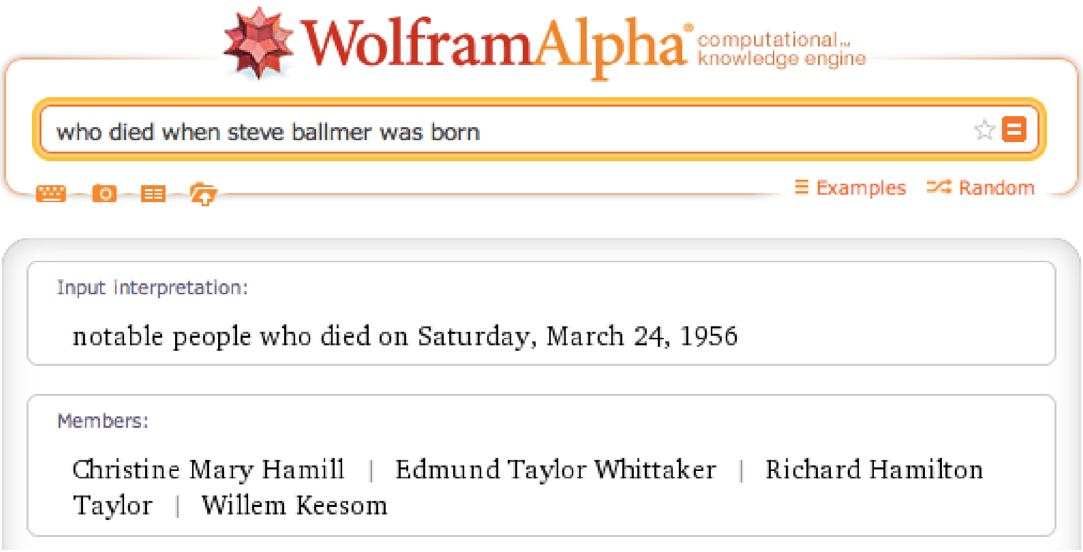
5.3. Wolfram Alpha

Wolfram Alpha calls itself a “Computational Knowledge Engine” and aims to “make all systematic knowledge immediately computable and accessible to everyone”. Systematic knowledge is objective, factual knowledge that can be expressed as a so-called *model*. Models are defined in nearly every field of science, as they systemize relations and properties of the entities of interest. Handling this source of data computationally (as Wolfram Alpha does) means that knowledge will be derived by a mathematical calculus. It is thereby self-evident that Wolfram Alpha best performs at mathematical questions.[4] Like the tools introduced before, Wolfram Alpha also uses NLP. Though, the official FAQs[5] proposes to use as less keywords as possible rather than formulate a full question because of ambiguities.

The main differences between Wolfram Alpha and search engines in general are the source of processed data and the type of result produced. Search engines crawl the *surface web* to gather information. A computational knowledge engine uses the *deep web*. The content of the surface web is retrievable via search engines and belongs to the sphere of the web which is considered to be the world wide web itself in a common sense. But the total amount of information available on the web is far more than most search engines have ever indexed. This information is part of the deep web and often stored in databases. These datasets will not be displayed until they are queried explicitly via a web-interface, somewhat a web-crawler is not able to do. There are no recent estimations of how big the deep web is. The last empirical analysis took place in 2001 and estimated the deep web to cover about 500 times more data than the surface web. (Dörner, S., 2010) Most

[4] Making the world’s knowledge computable. Retrieved December 2012, from http://www.wolframalpha.com/about.html

[5] Frequently Asked Questions. Retrieved December 2012, from http://www.wolframalpha.com/faqs.html



who died when steve ballmer was born

Input interpretation:

notable people who died on Saturday, March 24, 1956

Members:

Christine Mary Hamill | Edmund Taylor Whittaker | Richard Hamilton Taylor | Willem Keesom

Figure 5.4.: Wolfram Alpha web-interface after an answer has been retrieved (1).

information in the deep web is scientific, financial, business, medical, or of any other kind that is factual. It makes sense for Automatic QA to process these contents, since the knowledge engine does not need to refer to external websites which provide an answer to the questions (and probably do not even exist). So the question is answered by Wolfram Alpha itself by using the data of the deep web. The success of finding an answer entirely depends on the knowledge base of Wolfram Alpha, which receives updates consecutively. Therefore, results may improve over time. But the quality of the answers also depends on the question, or, to be more precise, on how Wolfram Alpha interprets the question.

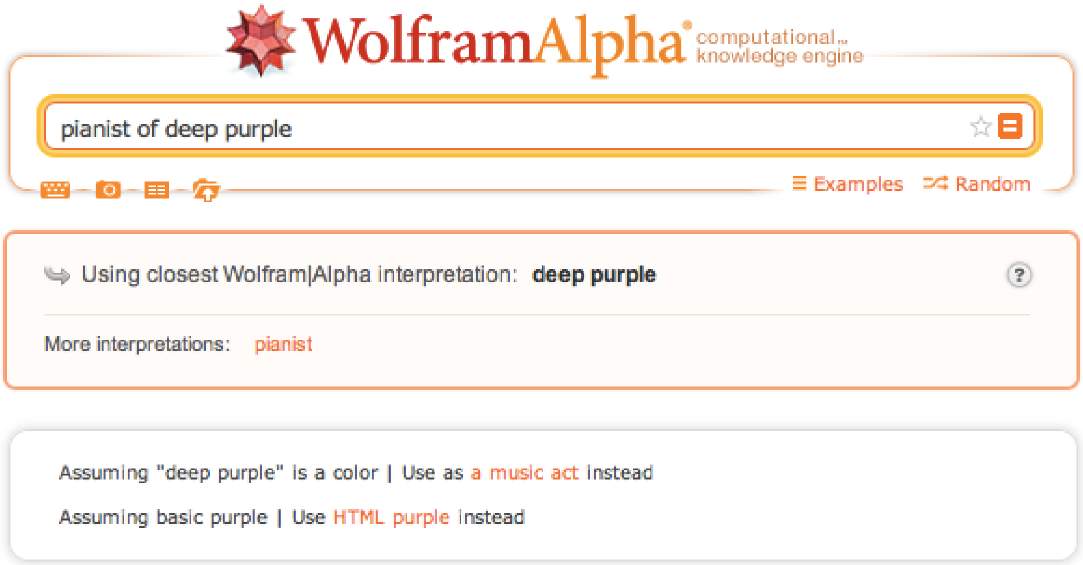
5.3.1. Technology

According to the Wolfram website, Wolfram Alpha was developed with Mathematica and Wolfram Workbench. It uses gridMathematica for distributed computing and webMathematica to present the results on the web interface.[6]

5.3.2. Evaluation

As said above, it is recommended to keep the wording of the question as easy as possible. Wolfram Alpha analyzes the question using NLP and generates a new calculus representation, which will then be used for knowledge computation (see figure 5.4). The drawbacks of the strict scientific approach based on discrete facts are obvious. As figure 5.5 shows, no answer was found to an informal question. Nevertheless, it seems to be a lack of information available, since there are questions that can be answered although they are not formal in any means, see example 5.6. One explanation might be that the latter one is a definition since the term “the day the music died” (first used by Don McLean in his song “American Pie”) has established in popular culture. Other critics noticed the high

[6] How Mathematic made Wolfram Alpha possible. Retrieved December 2012, from http://www.wolfram.com/mathematica/how-mathematica-made-wolframalpha-possible.html



pianist of deep purple

Using closest Wolfram|Alpha interpretation: **deep purple**

More interpretations: pianist

Assuming "deep purple" is a color | Use as a music act instead

Assuming basic purple | Use HTML purple instead

Figure 5.5.: Wolfram Alpha web-interface after failing to retrieve an answer.

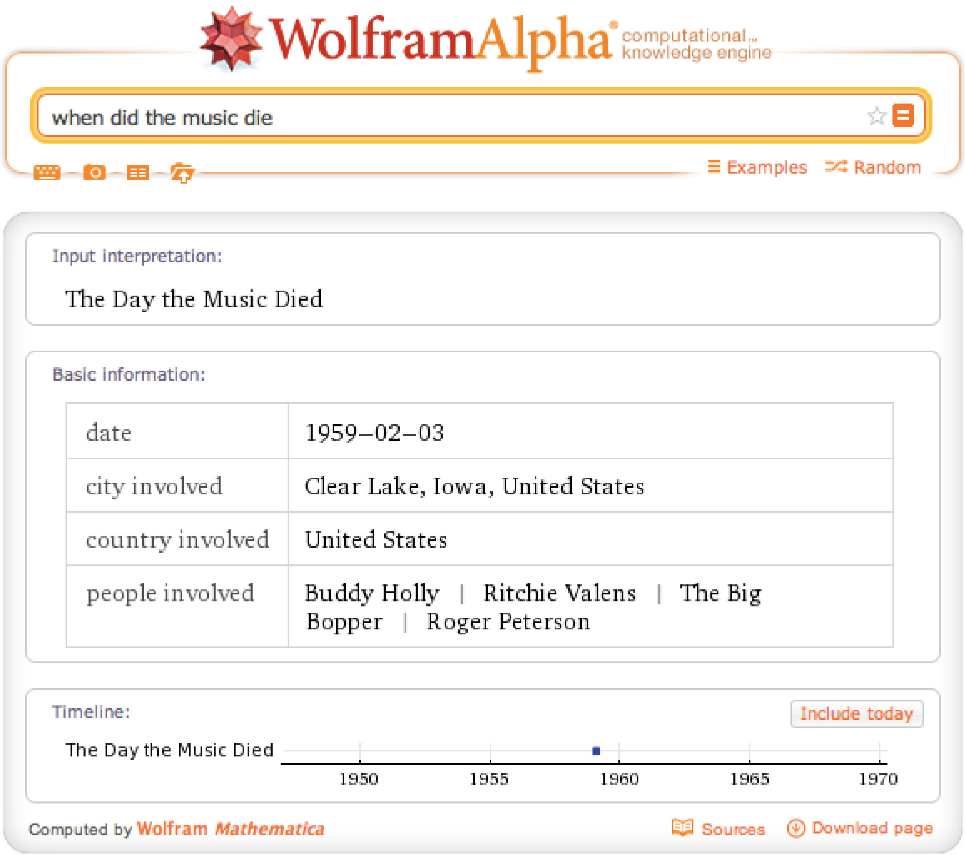
academic barrier of accessibility. Advanced skills are necessary to use Wolfram Alpha in an efficient and gainful way.

5.4. START Natural Language System

START (SynTactic Analysis using Reversible Transformations) is a web-based question answering system developed by Boris Katz and his staff at the MIT and online since December 1993, undergoing further improvements. Compared to Watson and Wolfram Alpha, START seems to be less known to a greater audience duo to its academic environment.

5.4.1. Technology

START uses constructs called *natural language annotations*. Katz defines them as "computer-analyzable collections of natural language sentences and phrases that describe the contents of various information segments." (Katz, Borchardt & Felshin, 2006) Natural language annotations are parsed from the content and *nested ternary expressions* are generated out of these. Nested ternary expressions are stored in the *knowledge base* and keep a reference to the original source. When the user sends a query to the system, START derives the nested ternary expression and searches in the knowledge base for matches. The reference to the previously analyzed content can then be retrieved. Nevertheless, an additional mechanism is needed when dealing with a large amount of data. *Parameterized annotations* make use of "parallel material" occurring in contents by identifying fixed language elements and combining them with variable parameters. An example is brought by (Katz, Borchardt & Felshin, 2005): The fixed language element "x people live in the metropolitan area of y" is stored once while the parameters for every entity are stored separately in the knowledge base. An additional database system called OMNIBASE supports START by providing object-based parameterized data.



when did the music die

Input interpretation:

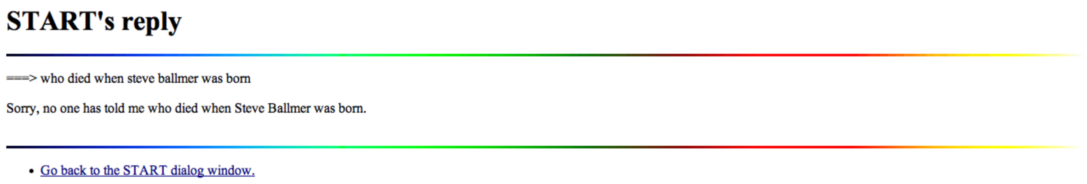
The Day the Music Died

Basic information:

date	1959–02–03
city involved	Clear Lake, Iowa, United States
country involved	United States
people involved	Buddy Holly | Ritchie Valens | The Big Bopper | Roger Peterson

Timeline:

Figure 5.6.: Wolfram Alpha web-interface after an answer has been retrieved (2).



START's reply

===> who died when steve ballmer was born

Sorry, no one has told me who died when Steve Ballmer was born.

- Go back to the START dialog window.

Figure 5.7.: START web-interface after failing to retrieve an answer.

5.4.2. Evaluation

Although START is online a significant longer timespan than Wolfram Alpha, it proves to be less flexible in parsing questions. The example questions provided on the website, however, are much more promising than those asked randomly by the user (figure 5.7). Again, demonstrating the concept needs to be done by preparing questions, which is problematic for an engine that wants to answer any factual question the user asks. Some further evaluations on which questions START knows an answer showed that a lack of information might be the reason for this issue.

6. Lessons learned

For the research of the paper of *New Trends in Automatic QA* we can draw the lesson, that there is much progress from the historical beginning of this science field until now. The diversity of the field in Automatic QA has emerged and the systems have become more complex. Automatic QA can now be interactive, they are used in new social media and are also very popular in smartphones like Siri in iOS.

In current research approaches, there is a strong tendency that social networks are more and more involved in the process of natural language processing. In this domain, there are some interesting possibilities that could be used for Automatic QA. Furthermore, the development still goes towards open domain question answering systems to stay abreast of changes of the ever-growing amount of information.

Some recent methods could be shown that are currently used in computational linguistic. Furthermore, the basic architechture for question answering systems was demonstrated. For computational linguistic there are a number of application areas that overlap and complement each other to a large extent. It is an independent scientific field, which is being pursued by many associations, such as military, industry and academia.

Although there have been some notable releases and demonstrations over the last two years, the current tools are far away from answering questions in an everyday usage. Apple Siri and Google Now benefit from mobile phone microphones, which are generally more sophisticated in recording speech. They serve their purpose in executing spoken commands, but they are no fully fledged Automatic QA systems. This is due to a lack of information in backend services and an insufficient interpretation of the question. Wolfram Alpha best performs on strict computational datasets. Other questions will be reduced to some single keywords and answered by presenting a definition. One main drawback, occurring on all systems, is the fact that the user needs to have some know-how on how to use the knowledge engines, and – even worse – in some cases it seems that already knowing the answer is mandatory for producing a question that can be answered. Finding example questions for this paper was indeed very time-consuming, since most answers were not suitable for demonstrating the advantages of (computational) knowledge engines. The most successive task is still retrieval of factual knowledge, even though most users would not be satisfied by knowing the weather of a day some decades ago. The most promising system is Watson, since it performs pretty reliable even on questions with high background knowledge requirements. Watson, however, is available for private use.

In spite of the weaknesses of currently existing tools, the field of Automatic QA experiences a significant upturn in offering practical implementations of algorithms and concepts that evolved over several decades. Research and evaluation of usage is still going on and it is likely that new insights will be considered in future implementations.

7. Conclusion and Future Work

In this paper the definition of the different approaches and the historical development are covered, the current research as well as the existing tools are presented. Some very powerful systems like Wolfram Alpha and IBM Watson have been shown, amongst others that were already introduced in the daily life, like Apple Siri and Google Now.

There is a huge possibility that complex Automatic QA systems can replace simple web search systems or at least become an competitive counterpart, but in the Automatic QA science field are still non-trivial research fields like the question analysis with the semantic and syntactic sciences. Document and Information Retrieval, the Natural Language Processing and Computational Linguistics are all huge sciences with many different approaches, which are still not all fully developed. And at the end there is the difficult part of the extraction of the relevant information to form an accurate answer.

In the book of Alexander Clark (Clark et al., 2010) are also three important directions mentioned in which the trend of Automatic QA heads:

1. extending the relation between question and corpus
2. broadening the range of answerable questions
3. relation between user and system

With the point extending the relation between question and corpus Clark means that answers can infer from other text, so that not a corpus as large as the world wide web is needed. In this topic the Recognising Textual Entailment (RTE) approach is evolving.

In the subitem broadening the range of answerable questions, Clark points out the future work for the answering of *how* and *why* questions. There is often no simple answer to these questions and the problem of evaluating the answers is still a complex process.

The relation between the user and the system is important for questions with different correct answers. So the system has to find out, which is the most likely correct answer for the specific user.

Finally we can conclude that in the future the speech recognition and interactive Automatic QA systems will be more and more popular, because of the still growing smartphone market, which gives new impulses, and the possibilities, which are offered in combination with new media.

A. Appendix

A.1. List of Figures

2.1. Architecture of natural language processing system, Androutsopoulos et al., 1995 8

3.1. System architecture of the question answering system 12

4.1. Basic architecture of a QA system 14

5.1. Knowledge Graph presents facts related to the search term aside the search results. 19
5.2. Watson's evaluation results at *Jeopardy!* 20
5.3. DeepQA Software Architecture (Copyright: IBM Research) 21
5.4. Wolfram Alpha web-interface after an answer has been retrieved (1). 23
5.5. Wolfram Alpha web-interface after failing to retrieve an answer. 24
5.6. Wolfram Alpha web-interface after an answer has been retrieved (2). 25
5.7. START web-interface after failing to retrieve an answer. 25

References

Andrenucci, A. (2008). Automated question-answering techniques and the medical domain. In *International conference on health informatics.* : HEALTHINF 2008.

Barros, R., Cerri, R., Jaskowiak, P. & de Carvalho, A. (2011). A bottom-up oblique decision tree induction algorithm. In *Intelligent systems design and applications (isda), 2011 11th international conference on.* : IEEE Computer Society.

Blooma, M., Chua, A., Goh, D.-L. & Keong, L. (2009). A trend analysis of the question answering domain. In *Sixth international conference on information technology: New generations.* : .

Clark, A., Fox, C. & Lappin, S. (Hrsg.). (2010). *The handbook of computational linguistics and natural language processing.* : Wiley-Blackwell.

Cowie, J., Ludovik, E., Molina-Salgado, H., Nirenburg, S. & Scheremetyeva, S. (2000). Automatic question answering. In *Proceedings of the ieee/wic/acm international conference on web intelligence.* Paris, F: .

DeepQA Research Team. (2011). *Deepqa.* Retrieved December 2012, from http://researcher.watson.ibm.com/researcher/view_project_subpage.php?id=2159.

Dörner, S. (2010). *Die dunkle seite des internets.* Retrieved December 2012, from http://www.handelsblatt.com/technologie/it-tk/it-internet/deep-web-die-dunkle-seite-des-internets/3543420.html.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., ... Welty, C. (2010). Building watson: An overview of the deepqa project. In *Ai magazine fall.* : Association for the Advancement of Artificial Intelligence.

Frädrich, L. & Anastasiou, D. (2012). *Siri vs. windows speech recognition.* Retrieved December 2012, from http://www.bokorlang.com/journal/61dictating.htm.

Geller, T. (2012). Talking to machines. In *Communications of the acm.* : Association for Computing Machinery.

Grishman, R. (2003). Information extraction. In *The oxford handbook of computational linguistics.* : Oxford University Press.

Gunawardena, T., Lokuhetti, M., Pathirana, N., Ragel, R. & Deegalla, S. (2010). Evaluating question answering system performance. In *Information and automation for sustainability (iciafs), 2010 5th international conference on.* : IEEE Computer Society.

Hao, T., Liu, W. & Agichtein, E. (2010). Towards automatic question answering over social media by learning question equivalence patterns. In *Proceedings of the naacl hlt 2010 workshop on computational linguistics in a world of social media.* Stroudsburg, PA, USA: Association for Computational Linguistics.

Huang, S. & Zhang, H. (1994). Artificial neural networks in manufacturing: concepts, applications, and perspectives. In *Components, packaging, and manufacturing tech-*

nology, part a, ieee transactions on. : IEEE Computer Society.

Katz, B., Borchardt, G. & Felshin, S. (2005). Syntactic and semantic decomposition strategies for question answering from multiple resources. In *Proceedings of the aaai 2005 workshop on inference for textual question answering.* Pittsburgh, PA, USA: Association for the Advancement of Artificial Intelligence.

Katz, B., Borchardt, G. & Felshin, S. (2006). Natural language annotations for question answering. In *Proceedings of the 19th international flairs conference.* Melbourne Beach, FL, USA: The Florida AI Research Society.

Koehn, P., Och, F. & Marcu, D. (2003). Statistical phrase-based translation. In *Naacl '03 proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology - volume 1.* : Association for Computational Linguistics.

Lenat, D. B. & R. V. Guha, R. V. (Hrsg.). (1989). *Building large knowledge-based systems.* : Addison-Wesley.

Mareček, D., Popel, M. & Žabokrtský, Z. (2010). Maximum entropy translation model in dependency-based mt framework. In *Wmt '10 proceedings of the joint fifth workshop on statistical machine translation and metricsmatr.* : Association for Computational Linguistics.

Mueller, E. T. (1987). *Daydreaming is but one more ai-complete problem: if we could solve any one artificial intelligence problem, we could solve all the others.* Unveröffentlichte Dissertation, University of California, Los Angeles.

Patten, T. & Jacobs, P. (1994). Natural-language processing. In *Ieee expert.* : IEEE Computer Society.

Plamondon, L. & Kosseim, L. (2002). Quantum: A function-based question answering system. In *In proceedings of the 15th conference of the canadian society for computational studies of intelligence (ai 2002.* : .

Reiter, E. & Dale, R. (Hrsg.). (2000). *Building natural language generation systems.* : Cambridge University Press.

Schwan, B. (2011). *Was siri versteht.* Retrieved December 2012, from http://www.heise.de/mac-and-i/meldung/Was-Siri-versteht-1355676.html.

Scott, P. (2012). *How accurate is siri, really?* Retrieved December 2012, from http://sociable.co/mobile/how-accurate-is-siri-really/.

Shannon, C. (1948). A mathematical theory of communication. In *The bell system technical journal.* : .

Singhal, A. (2012). *Introducing the knowledge graph: things, not strings.* Retrieved December 2012, from http://googleblog.blogspot.co.at/2012/05/introducing-knowledge-graph-things-not.html.

Smucker, M. D. & Allan, J. (2012). Human question answering performance using an interactive document retrieval system. In *Proceedings of the 4th information interaction in context symposium.* New York, NY, USA: ACM.

SRI. (2011). *Siri, the virtual personal assistant for the apple iphone.* Retrieved December 2012, from http://www.sri.com/work/timeline/siri.

Sullivan, D. (2012). *With fix in place, wolfram alpha explains how siri "recommended" the lumia by mistake.* Retrieved December 2012, from http://searchengineland.com/with-fix-in-place-wolfram-alpha-explains

-how-siri-recommended-the-lumia-by-mistake-121671.

Voorhees, E. M. (2006). Evaluating question answering system performance. In *Advances in open domain question answering.* : Springer.

Weizenbaum, J. (1966). Eliza a computer program for the study of natural language communication between man and machine. In *Computational linguistics.* Massachusetts Institute of Technology: .

Winograd, T. (1971). Procedures as a representation for data in a computer program for understanding natural language. In *Massachusetts institute of technology.* : .

Wofford, J. (2011). *How siri works.* Retrieved December 2012, from http://www.jeffwofford.com/?p=817.