Noname manuscript No.
(will be inserted by the editor)

Connecting Question Answering and Conversational Agents

Contextualizing German Questions for Interactive Question Answering Systems

Ulli Waltinger · Alexa Breuing · Ipke Wachsmuth

Received: date / Accepted: date

Abstract Research results in the field of Question Answering (QA) have shown that the classification of natural language questions significantly contributes to the accuracy of the generated answers. In this paper we present an approach which extends the prevalent question classification techniques by additionally considering further contextual information provided by the questions. Thereby we focus on improving the conversational abilities of existing interactive interfaces by enhancing their underlying QA systems in terms of response time and correctness. As a result, we are able to introduce a method based on a tripartite contextualization. First, we present a comprehensive question classification experiment based on machine learning using two different datasets and various feature sets for the German language. Second, we propose a method for detecting the focus chunk of a given question, that is, for identifying which part of the question is fundamentally relevant to the answer and which part refers to a specification of it. Third, we investigate how to identify and label the topic of a given question by means of a human-judgement experiment. We show that the resulting contextualization method contributes to an improvement of existing question answering systems and enhances their application within interactive scenarios.

Keywords Interactive Question Answering · Question Classification · Topic Spotting · Machine Learning

Ulli Waltinger[†] · Alexa Breuing[π] · Ipke Wachsmuth[π]
Siemens AG[†], Corporate Technology,
Otto-Hahn-Ring 6, 81739 Munich, Germany
Artificial Intelligence Group[π], Bielefeld University,
P.O. Box 100131, 33501 Bielefeld, Germany
E-mail: ulli.waltinger@siemens.com[†],
{abreuing,ipke}@techfak.uni-bielefeld.de[π]

1 Introduction

Question Answering (QA) has become an important research topic in the field of Information Retrieval (IR) and Artificial Intelligence [Giampiccolo et al., 2007, Ferrucci et al., 2010]. Different to traditional IR approaches (e.g. search & browse) in which users need to wade through a large set of query-related documents, the domain of QA allows for the delivery of succinct answers to natural language questions as posed by a user. This is of relevance for all types of intelligent user interfaces as QA abilities significantly help to improve human-computer interaction. Question and answer-type classification, representing the tasks of identifying the expected question and answer categories of a user's query, can be regarded as the most fundamental tasks of most existing QA systems [Li and Roth, 2002, Ferrucci et al., 2010]. Generally speaking, these tasks aim at classifying any given input question with reference to a given set of output categories, that is, identifying the kind of answer formation, entity, or concept being asked. Examples are determining the question types as *factoid*, *list*, and *definition* (e.g. *How tall is ...*, *In which movies played ...*, *What is a ...*), or, in the context of answer types, the appropriate named entity class for the answer (e.g. *numeral, person, company, date, height, currency*).

Previous research [Suzuki et al., 2003, Zhang and Lee, 2003, Quarteroni et al., 2007, Blooma et al., 2009] has shown that question contextualization [Lin et al., 2003, Bradesko et al., 2010] contributes significantly to the accuracy of QA systems. With reference to the QA track at the TREC conferences [Voorhees, 2007, Peñas et al., 2010], current state-of-the-art QA systems show a reasonable performance (accuracy up to 0.80) when focusing

Table 1 Example question entry from the *CLEF-2007* monolingual QA task [Giampiccolo et al., 2007] with enhanced contextualization (*German: Wann wurde Pearl Harbor von den Japanern angegriffen?*)

Q: [*When was*]^q [*Pearl Harbor*]^s [*attacked*]^p *by the* [*Japanese*]^{spec}*?*
Question Type: *Factoid* **Answer Type:** *Num:Date*
Subject: *Pearl Harbor* **Predicate:** *attacked* **Object:** ⟨ *Num:Date* ⟩
Focus Node: *Pearl Harbor* **Focus Specification:** *Japanese*
Topic Hints: *Attack on Pearl Harbor* *Battle of the Pacific War* *World War II*

on factoid question types (e.g. *What is the height of Mount Everest?*). However, with reference to all question types, only mediocre results can be achieved (average accuracy of QA systems at *ResPubliQA* 2009 [Peñas et al., 2010]: German: 0.44; English: 0.61; Spanish: 0.44; French: 0.45;). Hence, despite the significant improvements of current QA systems, the field of QA still remains challenging [Giampiccolo et al., 2007, Voorhees, 2007,Peñas et al., 2010].

In this paper, we investigate an extended question contextualization method to improve the interpretation of German questions and, eventually, to enhance the performance of an *interactive* QA system. More precisely, we focus on a tripartite contextualization (see Table 1) to tackle the following questions:

1. **Question and Answer Type Classification:** What is the expected question type (e.g. list, fact, or definition) and what is the expected answer type (e.g. numeral, person name) being asked?
2. **Focus Detection:** Which part of the query is at the centre of attention (fundamentally relevant to the answer) and which part of the question refers to a specification of it?
3. **Topic Spotting:** What is the primary topic of the question? To which topic may the answer belong to? What is the question (or conversation) about?

Question type classification obviously refers to the most traditional task of contextualization. Focus phrase detection is to support the decoding of natural language questions into a triple representation (e.g. *subject, predicate, object*) as demanded and applied in most existing question answering systems that use RDF resources as a knowledge base. The purpose of the topic spotting component is again tripartite: First, it allows us to deduce a set of expected answer candidates from the identified topic by means of their thematical membership [Waltinger et al., 2011]. It enables confining the used knowledge base by topic. Second, it allows the incorporation of the (*interactive*) context information of the entire conversation within a certain timeframe. That is, the identified topic hints can be incorporated as an answer context for the next question (e.g. in the context of the Pearl Harbor example, the next question of the user might be: *What was the first ship to be sunk?*). Third, since this application is embedded within an existing conversational agent architecture, it allows us to summarize the entire conversation by means of its dialog topic (e.g. *We talked about World War II and the Attack on Pearl Harbor!*). Consequently, the proposed approach aims to enhance the conversational behaviour of a conversational agent by means of *knowledge awareness*, in terms of connecting question answering and conversational agents, and *subject awareness*, in terms of connecting topic detection and interactive user dialogues [Waltinger et al., 2011].

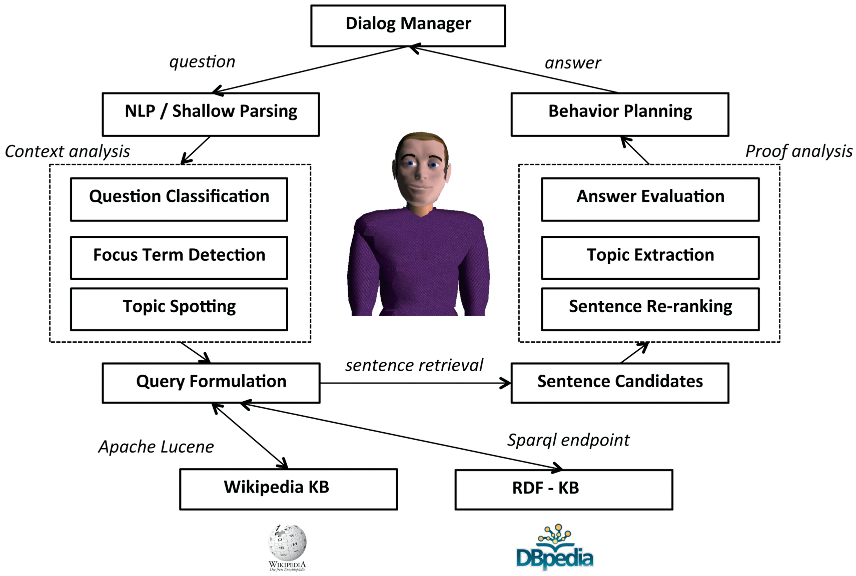
2 An Overview of WikiQA

In the *KnowCIT*[1] project, we extend the conversational abilities of the conversational agent *Max* [Kopp et al., 2005] by making the agent more context and topic aware in natural language interactions with humans. In this project, we connect two research areas which have moved closer to each other in recent years: Question Answering, here utilizing the Wikipedia-based question answering system *WikiQA*[2], and conversational agents, here: *Max*. Using information and answers drawn from the online encyclopaedia *Wikipedia*, *Max* is able to answer questions posed by his human dialogue partner as well as being able to identify the topic of an on-going dialogue [Breuing et al., 2011, Waltinger et al., 2011]. The overall architecture of the QA system employed in *Max* (see Fig. 1) can be subdivided into the classical processing pipelines for QA systems such as *context analysis* (e.g. question processing, shallow parsing, query formulation), *knowledge base retrieval* (e.g. semi-structured and/or RDF-based resources), and *proof analysis* (e.g. sentence candidate selection, re-ranking and answer evaluation). This specific system setup, however, also implicates several challenges: First and foremost, the response time. It is a mandatory precondition of our project setup that the QA system returns, out of millions of sentences, only

[1] Knowledge Enhanced Embodied Cognitive Interaction Technology (`www.cit-ec.de/research/knowcit`)

[2] The system is available at `www.wikiqa.de`

Fig. 1 Overview of the QA architecture within the dialog system of the conversational agent *Max*.



one single answer within a few seconds, in order to sustain the on-going conversation. Second, robustness is, in this context, of high priority. That is, the QA system must have a 'sense of confidence' about the answer, otherwise the interlocutor may not take the conversational agent's answers seriously in subsequent dialogues. In this regard, a contextualization of the users' questions, as proposed in this paper, additionally contributes to the accuracy, speed, and adequacy of the returned answers which in turn enables a more flexible, fluent, and coherent interaction between the artificial and the human interlocutors. Thus, we integrated our question contextualization approach into the agent's existing system architecture and evaluated a wide range of feature types and learning methods to exploit the applicability of our tripartite question contextualization in the context of interactive question answering.

The rest of this paper is structured as follows: In Section 3 we review related work. Section 4 describes the methods for the tripartite question contextualization with reference to question classification, focus detection, and topic spotting. We present the results of the classification experiments and the unsupervised topic detection method, which is evaluated through a human-judgement experiment. Finally, Section 5 summarizes and concludes the paper.

3 Related Work

Question classification is an important step to narrow down the search space of question answering and dialog systems. In recent years, many approaches to this problem have been proposed. Most notably with reference to the primarily comprised category structure for question classification, [Li and Roth, 2002] presented a two-layered taxonomy (see Table 2), which consists of six coarse categories and a total of 50 finer categories. The authors used a hierarchical classifier (accuracy: 0.91) combining lexical and syntactic features (e.g. Part-of-Speech (PoS), Named Entity (NE), and head chunks) targeting the English language.

Table 2 Subset of the two-layered question classification taxonomy for typical answers in the *TREC* task by [Li and Roth, 2002] (six coarse and 50 fine named-entity types)

ABBR	abr.	exp.		
HUMAN	group	individual	title	...
NUM	date	money	distance	...
LOC	city	country	state	...
ENTITY	animal	body	term	...
DESC	definition	manner	reason	...

[Solorio et al., 2004] also used the two-layered question classification taxonomy within their language independent classification method. By combining word features and machine learning-based Support Vector Machines (*SVM*), they obtained an accuracy of 0.82 on English, 0.88 on Italian and 0.80 on Spanish. [Zhang and Lee, 2003] presented a comprehensive question classification evaluation using *SVM*, k-Nearest Neighbor (*k-NN*), *Naive Bayes*, the *Sparse Network of Winnows*, and *Decision Trees*. They found that the syntactic structures of questions support the question classification task. The proposed syntactic tree kernel *SVM* exhibits the best performance (accuracy 0.90). Similar results could be achieved by [Suzuki et al., 2003] using *HDAG Kernel* on Japanese questions and by [Blooma et al., 2009] employing the *Yahoo! Answers Dataset* (accuracy 0.75). Using head words and their hypernyms as features for an *SVM*-based question classification, [Huang et al., 2008] report an accuracy of 0.89. With reference to the German language, as targeted in this paper, [Davidescu et al., 2007] presented an extensive evaluation using various machine learning algorithms applied to the 50 hierarchically organized classes of the *SmartWeb* ontology [Sonntag and Romanelli, 2006]. Davidescu and collegues used shallow and syntactical features for the classification task and report an accuracy of about 0.45. Their comprehensive evaluation has clearly shown the complexity of the task of question classification for the German language. The German *LogAnswer* system [Furbach et al., 2008, Glöckner and Pelzer, 2010] uses 240 classification rules for the question classification task.

The system proposed by [Koehler et al., 2008] recognized the question type by primarily focusing on the identification of (fourteen different) question words (e.g. *Who, Where, What*). [Neumann and Sacaleanu, 2004] presented a cross–language QA system for German and English using the lexicalized tree substitution grammar (LTSG) for question classification and query construction.

In this paper, we utilize part of the question dataset and classification taxonomy as provided by [Davidescu et al., 2007] and [Li and Roth, 2002] for our experiments (as a baseline - accuracy of about 0.45 for the German language), although we additionally analyse different feature types (e.g. lexical word net, class labels of the hierarchical structure, syntactic chunks, bag-of-word) for the classification task using an SVM-based approach.

In the context of focus detection, [Damljanovic et al., 2010] presented an approach of identifying the question focus by combining syntactic analysis and an ontology-based lookup technique based on user interaction. In this regard, their approach is similar to the approach proposed in this paper, though, instead of predicting the answer type by combining the head of the focus with ontology-based lookup, we combine syntactic analysis with a topic model technique applied to the Wikipedia dataset.

With respect to the domain of topic spotting for question contextualization [Lin et al., 2003, Bradesko et al., 2010] in dialogue systems, [Gerber and Chai, 2006] presented a regression model to identify topic terms. [Myers et al., 2000] proposed an approach for topic spotting in conversational speech (ten topics of the Switchboard corpus [Godfrey et al., 1992]) using the machine-learning program *BoosTexter* [Schapire and Singer, 2000] (accuracy of up to 88.3%). [Gupta and Ratinov, 2007] also comprised ten categories of the Switchboard corpus using a feature-generation approach to knowledge transfer. Their Naive Bayes classification approach has shown an error reduction of 17% using external knowledge (e.g. *Yahoo Answer dataset*, 500 *Wikipedia* clusters, and *Google* 5-grams collection). [Liu and Chua, 2001] proposed a semantic perceptron net approach for topic spotting using the *Reuters* corpus. [Lagus and Kuusisto, 2002] presented an approach using neural networks for subject recognition of dialogues.

Different to the approaches above, the method proposed in the present paper uses the *Wikipedia* dataset as the primary knowledge base for both answer extraction and topic detection. That is, we are not focusing on term or phrase extraction from a given input text, but utilize an external knowledge base to derive topic labels for a given question-answer pair. More precisely, we make use of the topic identification systems proposed by [Breuing and Wachsmuth, 2012]. This approach is mainly based on the five main tasks determined in the context of the Topic Detection and Tracking (TDT) research program [Allan, 2002]. Moreover, the Wikipedia category system is accessed to realize a dynamic online topic identification enabling the topical specification of previously unknown dialog contributions. Further examples for interactive systems identifying conversational topics are *Conversation Clusters* which visually highlight topics discussed in conversations using Explicit Semantic Analysis (ESA [Gabrilovich and Markovitch, 2007]) [Bergstrom and Karahalios, 2009], an emergency interface tool displaying relevant information sources according to the described emergency, and an embodied conversational agent identifying out-of-domain topics on the basis of the Google's directory structure [Mehta and Corradini, 2008]. In general, the *Wikipedia* dataset has received much attention in the field of information retrieval [Gabrilovich and Markovitch, 2007] and topic detection [Schönhofen, 2009, Waltinger and Mehler, 2009, Breuing et al., 2011], but also most recently to the domain of question answering [Buscaldi and Rosso, 2006, Fissaha Adafre et al., 2007, Furbach et al., 2008, Waltinger et al., 2011]. In our approach, we use the *Wikipedia* dataset as our resource to structure the knowledge base and to derive article and category information for the topic labeling task. Our method is using the dataset also for the question classification and focus detection task.

4 Contextualizing German Questions

In this section, we present the methods applied to the task of question contextualization. In general, we make use of two different methods for the experiments. First, we utilize Support Vector Machines (SVM) [Vapnik, 1995] as the classical apparatus in the context of (text) classification. The current implementation of the classifier module is based on SVM^{light} [Joachims, 2002], where linear and radial basis kernel functions are evaluated in leave-one-out cross-validation. Second, we apply the *Open Topic Model* approach as proposed by [Waltinger and Mehler, 2009]. More precisely, we utilize the German *Wikipedia* dataset in combination with the *Apache Lucene* framework [Hatcher et al., 2010] to rank *Wikipedia* article and category entries according to their strength of association to a given natural language question [Waltinger et al., 2011]. See Table 3 for an example ranking for a given input question.

Table 3 Excerpt from the Wikipedia-ranking for input question: *'When was Pearl Harbor attacked by the Japanese?'*.

Rank	Wiki Article Set	Wiki Category Set
1	Attack on Pearl Harbor	Battle of the Pacific War
2	Pearl Harbor	1941
3	Pearl Harbor (movie)	Hawaii
4	USS Pearl Harbor (LSD-52)	Sea Battle (World War II)

4.1 Dataset

For the experiments, we used two different question collections. First, we utilized 200 questions (Definition: 28; Factoid: 164; List: 8) from the *CLEF-2007* (8th Workshop of the Cross-Language Evaluation Forum) monolingual QA task using German as the target language [Giampiccolo et al., 2007]. Additionally, we used a subset of 200 questions (Definition: 59; Factoid: 138; List 3;) from the *SmartWeb* corpus [Cramer et al., 2006]. Note that we annotated each question by means of its category and a subset of the two-level-based question taxonomy as provided by [Li and Roth, 2002] (see Table 2). That is, the first level refers to the three most coarse-grained *question type* (Q-Type) categories (e.g. Definition, Factoid, and List). At the second layer, we differentiate between six different *question index* (Q-Index) categories (Abbr, Human, Num, Loc, Entity and Description). The third layer, which we refer to as the *answer type* (A-Type), comprises ten named entity classes (e.g. Title, Date, State, Distance, ...). Consequently, we used both datasets in combination, resulting in 400 sentences and an *answer type* category set of 14 different classes. All natural language questions were linguistically analysed using the shallow processing tool *TreeTagger* [Schmid, 1994]. That is, we applied tokenization, Part-of-Speech tagging, and lemmatization on the evaluation dataset. In addition, we utilized the embedded chunk parser to determine the syntactic chunks (e.g. NC: noun chunks, PC: prepositional phrase chunks, VC: verb chunks) of each question. See Figure 2 for an example question representation used for the experiments.

4.2 Question Classification

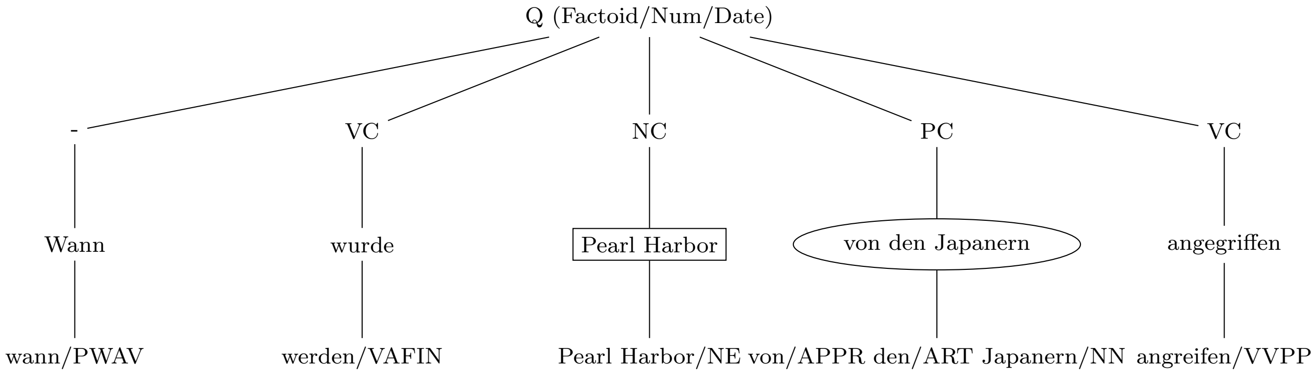
For the question classification task we employ an SVM-based approach. Previous research [Li and Roth, 2002, Davidescu et al., 2007] has already evaluated a wide range of machine learning classifiers, although, in the context of question classification, SVMs have not been extensively evaluated targeting the German language. In this work, we evaluated the following features to classify German questions by their question type:

- *Head words* refer to question words (e.g. *when was*, *what*, or *who is*) as the most obvious feature for question-type determination. Good performance for this feature has already been evaluated for the English language [Huang et al., 2008]. In our experiments, we used a quadgram-based approach to build the head word representation. More precisely, we focused on the first four words of each question by its lemmata and Part-of-Speech-tag (PoS) representation (e.g. feature set$f = \{wann, wann-werden, wann-werden-PearlHarbor, PWAV, PWAV-VAFIN, ...\}$).
- *Bag-of-words*, as the most traditional representation model in IR, represents each question as a set of words together with their frequency of occurrence, abstracting from its syntactic structure [Davidescu et al., 2007]. In our experiments, we built this representation by means of lemmata, PoS, and named entity class information using a trigram approach. That is, we allowed the incorporation of the syntactic structure to some extent, however, merging different feature categories (e.g. PoS, lemmata).

Table 4 German example questions from the *CLEF-2007* monolingual QA task [Giampiccolo et al., 2007] that are processed by the question contextualization pipeline.

Question	**Classification**	**Subject**	**Predicate**	**Focus Specification**
Wie hoch ist der Mount Everest?	Fac-Num-Dis	Mount Everest	hoch	
Wo lebt heute der Sohn von Audrey Hepburn?	Fac-Loc-Not	Audrey Hepburn	lebt	Sohn
Was ist Madame Tussaud?	Def-Def-Def	Madame Tussauds	ist	
Wo in Italien wurde die Villa Medici erbaut?	Fac-Loc-Not	Villa Medici	erbaut	Italien
Wie heißen die drei großen Wasserfälle im Canyon?	List-Ent-Pro	Canyon	heißen	drei großen Wasserfälle
Wie heißt das höchste Bergmassiv Afrikas?	Fac-Ent-Sub	Afrika	heißt	höchste Bergmassiv
Wie groß ist die Grundfläche des Pentagon?	Fac-Num-Size	Pentagon	groß	Grundfläche

Fig. 2 German example question from the *CLEF-2007* monolingual QA task [Giampiccolo et al., 2007] utilizing chunk, PoS-Tag and lemma representation after preprocessing. The expression marked with the rectangle highlights the focus term/phrase, the ellipse marks the specification of the question (English: *When was Pearl Harbor attacked by the Japanese?*)



– *Chunk* refers to the syntactic chunk representation of the respective question (see Figure 2). That is, we added a *bag-of-chunks* to the *bag-of-words* representation [Zhang and Lee, 2003] (e.g. feature set $f = \{VC, VC - NC, VC - NC - PC, ..\}$).
– *GermaNet* terms refer to hypernyms and hyponyms of the German lexical semantic wordnet *GermaNet* [Lemnitzer and Kunze, 2002]. That is, we enhanced any given input question by means of semantic relation information as identified by its labeled synset structure (e.g. angreifen $\mapsto$ beschädigen)
– *Wikipedia* articles and categories are used to enhance the question representation by its topical context [Waltinger and Mehler, 2009].
– *Taxonomy structure* features refer to categories of the used question taxonomy [Li and Roth, 2002]. More precisely, we enhance the question representation (*bag-of-words*) by its superordinate category label (e.g. factoid or definition) (e.g. Number $\mapsto$ Factoid). The rationale behind this approach is that since we are not aiming at a multi-label classification, the hierarchical structure allows us to narrow down the set of possible target categories.

As the *SVM* classifier expects questions represented as data vectors, each input was transformed to a weighted feature vector. Here we made use of the well-known TF-IDF [Salton and Buckley, 1988] weighting scheme.

The results of the question classification experiments on the *CLEF* corpus are shown in Table 5. The results of the *SmartWeb* and the combined dataset are shown in Table 6 and Table 7. Our baseline consists of the performance published by [Davidescu et al., 2007], who used 500 questions from the SmartWeb corpus. Their best results have shown an accuracy of 0.65 when using the Naive Bayes approach. In this context, our results (average 0.74 over all three layers of the SmartWeb; 0.77 using the combined dataset) can be regarded as a good performance. Interestingly, the feature enhancement approaches (in terms of *GermaNet* and *Wikipedia*

Table 5 Results of the question classification using *CLEF* dataset. We report F1-Measure by means of the *SVM*-based leave-one-out cross-validation using bag-of-words (bow), headwords (head), *Wikipedia* categories (wiki) and *GermaNet* terms (germ) representation. +Q-Type and +Q-Index refer to the enhancement by means of their taxonomy features. We leave out the chunk feature as it did not improve the classification performance.

Label	bow	head	wiki	germ
Q-Type	0.916	**0.930**	0.734	0.848
Q-Index	0.824	0.849	0.413	0.697
Q-Index (+Q-Type)	0.830	**0.856**	0.413	0.707
A-Type	0.682	0.681	0.374	0.586
A-Type (+Q-Index)	0.724	**0.781**	0.383	0.586

Table 6 Results of the question classification using SmartWeb corpus. We report F1-Measure by means of the *SVM*-based leave-one-out cross-validation using bag-of-words (bow) and headwords (head) representation.

Label	bow	head
Q-Type	0.850	**0.860**
Q-Index	0.742	0.732
Q-Index (+Q-Type)	0.776	**0.788**
A-Type	0.634	0.637
A-Type (+Q-Index)	0.721	**0.766**

Table 7 Results of the question classification on combined (co) dataset using *CLEF* and *SmartWeb* corpus. We report F1-Measure by means of the *SVM*-based leave-one-out cross-validation using bag-of-words (bow) and headwords (head) representation.

Label	bow	head
Q-Type	0.813	**0.823**
Q-Index	0.812	0.790
Q-Index (+Q-Type)	0.838	**0.820**
A-Type	0.695	0.724
A-Type (+Q-Index)	0.757	**0.762**

categories) did not improve the classification accuracy. In addition, utilizing the syntactic chunk information of a question also did not improve the performance (average accuracy of 0.72 on combined dataset). Moreover, it can be identified that using head words only (represented as quad-gram) exhibits the best performance on all three datasets (up to 0.82 at the *CLEF* dataset). That is, the first words opening a natural language question already indicate the type of the question as well as the type of the expected answer. As only a small amount of different wh-words exist, their particular range in terms of correct answers are clearly separated which significantly helps to narrow down the search space. Moreover, the enhancement of the hierarchical structure by means of superordinated categories obviously contributes, in addition, significantly to classification accuracy (up to 0.85 at the *CLEF*).

4.3 Focus Chunk Detection

The focus detection task locates the actual question object in front. That is, we need to identify which part (e.g. person name, company) of the question is at the centre of attention, and which part of the question refers to a specification of it, in order to identify the sequence of words which defines and disambiguates a given question [Moldovan et al., 1999]. Consider the following example: *'Name the 8 districts of Hiroshima'*. As shown in Figure 3, the PC chunk element (*Hiroshima*) can be identified as the question focus. The NC chunk, *the 8 districts*, serves as the specification of the object. Since *Hiroshima* is the only named entity in the question, this example is obviously of very basic nature. But what about the following questions: *'When did Audrey Hepburn marry Mel Ferrer?'* (focus on *Audrey Hepburn* or *Mel Ferrer*?), *'Who was the first African American who played for the Brooklyn Dodgers?'*, or to use the running example *'When was Pearl Harbor attacked by the Japanese?'* (NC is at the focus). Therefore, the task of focus chunk detection refers to identifying that part of the syntactic chunks which serves as the primary object of the question. In the current QA system, we need this information as a hint in which *Wikipedia* article we might find the desired information. To use the running example, we most probably find the information on the Pearl Harbor site instead of the *Wikipedia* Japan website. In addition, we need to extract the main object to query the RDF-based *DBpedia* dataset.

For the experiments, we manually annotated 200 questions of the *CLEF* question collection by means of their syntactic chunk representation and individually marked the focus and specification part in each question. Finally, we used the *bag-of-words* and the *bag-of-chunks* representations (as described in the previous section) for an SVM-based classification, using again leave-one-out cross-validation. In a second experiment, we applied the *Wikipedia*-based *Topic Model* approach. That is, for each input question we extracted a set of article and category titles from the *Wikipedia* dataset and ranked the respective chunk parts of the question with reference to the ranked *Wikipedia* results. More precisely, we focused on term overlap between each *Wikipedia* entry and each chunk part of the question. If the *Wikipedia* entry contains parts of the observed syntactic chunk, the latter will be labeled by the rank number of the respective *Wikipedia* rank score. Finally, we chose that part of the chunk set as our focus chunk which has the highest rank number. As shown in Table 8, the *Pearl Harbor* chunk is ranked higher than the *Japanese* chunk part (at rank 12). Therefore, the algorithm would detect *Pearl Harbor* as the focus chunk for the given input question.

The results of both experiments are shown in Table 9. The *SVM* results show that the incorporation of the syntactic chunk representation of questions let the performance only be mediocre (F1-measure[3] of 0.53). Interestingly, using only chunk and part-of-speech (including named entities) information, and therefore abstracting from the word (lemma) information, let us slightly increase the performance for this task. With respect to the (simple) un-supervised ranking approach

[3] F1-measure refers to the weighted harmonic mean of precision and recall.

Fig. 3 Parsed question representation of *"Name the 8 districts of Hiroshima"* used for focus chunk detection.

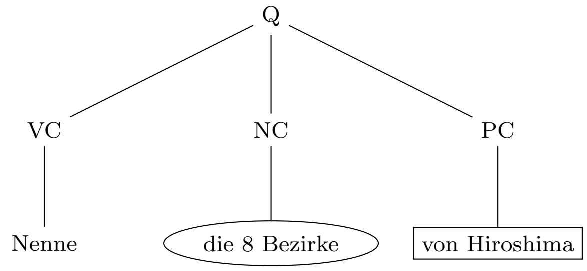


Table 8 Example *Wikipedia*-ranking of input question: *'When was **Pearl Harbor** (NC^{rank1}) attacked by the **Japan**ese (PC^{rank12})?'* for focus chunk detection task.

Rank	Wikipedia Topic Model
1	Attack on **Pearl Harbor**
2	Pearl Harbor
3	Pearl Harbor (movie)
4	USS Pearl Harbor (LDS-52)
5	1941
..	..
12	Naval battles involving **Japan**

Table 9 Results of the focus chunk detection experiment on *CLEF* corpus using *SVM* and leave-one-out cross-validation (F1-measure) and results of the *Wikipedia* topic model (accuracy).

Method	F1/Acc
SVM - with lemmata	0.529
SVM - without lemmata	0.533
Wikipedia - exact match	0.695
Wikipedia - substring match	**0.934**

by means of the *Wikipedia* topic model, we can identify that by using an exact match strategy for focus term detection, we clearly outperform the *SVM* approach (accuracy of 0.69). When allowing the incorporation of the substring strategy for the ranking, we achieve a quite reasonable accuracy of 0.93 for this task. That is, the proposed approach of ranking *Wikipedia* articles enables us to extract the main object of a user question (see Table 4). This information can be incorporated in a question answering system (unstructured or RDF-based) as a hint in which *Wikipedia* article the desired answer information might be located.

4.4 Topic Spotting

In the last experiment, we focus on the task of topic spotting in natural language questions. That is, we aim at labeling any given question by its thematic affiliation. Since the *question answering* system is embedded within a conversational agent architecture, which aims to detect and track the topic during the user-agent interaction, it is possible to not only return the answer but also the conversational topic. Thus, the agent is able to demonstrate his topic awareness by additionally presenting utterances such as "Hey, we speak about Pearl Harbor!" or "Hey, we speak about World War II!". Different to existing approaches, we thereby do not focus on term extraction or a clustering of a given dialogue dataset, but on labeling the questions as posed by the user individually (online) and by means of external topic labels using the *Wikipedia* dataset. More precisely, for each question we determine three different topic labels: First, we use the title of the best ranked article of the *Wikipedia* topic model as a topic label (e.g. Pearl Harbor). Second, we use the title of the most highly ranked category (e.g. 1941). Third, we utilize the title of the most highly ranked category, that is at least one link distant from the best ranked article in its category taxonomy structure (e.g. World War II). The rationale behind this approach is that an article title consists prevalently of terms that also occur in the question (e.g. as a substring), and can thereby be regarded as very 'close' to the input question representation (similar to the related clustering approaches). Predicted categories mainly refer to an abstract concept representation of the question and do not necessarily share features with the input question. Using categories which are at least one link distant from the respective article aims to label the input question by its broader topic (e.g. Pearl Harbor $\mapsto$ World War II; Helmut Kohl $\mapsto$ politics).

In the experimental setup, we were interested in which kind of label should be used for the interactive question answering system and which is regarded as an appropriate (human-like) response to the current topic by the user. Focusing on this aspect, we conducted a human-judgement experiment. We asked five volunteers to rate the different predicted topic labels by means of their thematical appropriateness for the given question embedded within the considered dialogue. Overall we used 200 questions from the *CLEF* task, each having three different topic labels given, where each label could be rated by three categories (a: fits well; b: mediocre; c: not appropriate). Finally, we calculated the inter-annotator agreement using Fleiss' Kappa [Fleiss, 1973] and the average pairwise percent agreement. The re-

Table 10 Results of the topic spotting experiment reporting Fleiss Kappa for inter-annotator agreement and average pairwise percent agreement on CLEF dataset.

Label	Fleiss' Kappa	Average pairwise
Article	0.43	**78,4%**
Category	0.33	60.6%
General	0.24	50.3%

sults are shown in Table 10.

We can identify that when using articles as a topic reference, a moderate agreement within all annotators can be achieved (average pairwise agreement of 78%). Using category information to label the given question, only a fair agreement can be achieved. What does that mean? Obviously, apart from the fact that some of the examples were not appropriate at all, the volunteers tend to rather expect terms that already occur in the question than labels within a certain scale of generalization. As, for example, to the question: "*Who is the singer of U2?*", they rather expect the label "We talk about *U2*" than "*Alternative rock-band*". Since we only used one single question as an input for the topic labeling task instead of providing an entire conversational sequence, the predicted topic labels rather refer to so-called *sentence topics* than *discourse topics* [Bublitz, 1989]. Embedded within an interactive, conversational system, the results may differ according to the topical context given by the present conversation. Evaluation on this aspect will be part of future work.

5 Conclusions

This work described a question contextualization approach for a German interactive question answering system employed within the architecture of the conversational agent *Max*. We thereby focused on a tripartite contextualization. *First*, as the most basic task, question type classification: We could identify that using headwords only as an input representation with *SVM*, a good performance with an overall accuracy of up to 0.85 can be achieved. *Second*, we proposed a method for detecting the focus chunk part of a given question. With an accuracy of 0.93, the results show a strong performance in this task. *Third*, we investigated how to label and directly display the topic of a given question using a large set of community-generated article and category labels from the *Wikipedia* dataset. The conducted user-judgement experiment suggested, with an average pairwise agreement of 78%, to use the *Wikipedia* article rather than category information to label single questions topically. Altogether, we have described three different approaches to the task of contextualizing German questions. The proposed approaches can be applied to improve existing unstructured and RDF-based question answering systems.

Acknowledgements We gratefully acknowledge financial support of the German Research Foundation (DFG) through EXC *277 Cognitive Interaction Technology (CITEC)* at Bielefeld University.

References

[Allan, 2002] Allan, J. (2002). *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic Publishers.

[Bergstrom and Karahalios, 2009] Bergstrom, T. and Karahalios, K. (2009). Conversation clusters: grouping conversation topics through human-computer dialog. In Jr., D. R. O., Arthur, R. B., Hinckley, K., Morris, M. R., Hudson, S. E., and Greenberg, S., editors, *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4-9, 2009*, pages 2349–2352. ACM.

[Blooma et al., 2009] Blooma, M. J., Goh, D. H.-L., and Chua, A. Y. K. (2009). Question classification in social media. *International Journal of Information Studies*, 1(2):101–109.

[Bradesko et al., 2010] Bradesko, L., Dali, L., Fortuna, B., Grobelnik, M., Mladenic, D., and Novalija, I. (2010). Contextualized question answering. *Journal of Computing and Information Technology*, 18(4).

[Breuing and Wachsmuth, 2012] Breuing, A. and Wachsmuth, I. (2012). Let's talk topically with artificial agents! providing agents with humanlike topic awareness in everyday dialog situations. In *Proceedings of the 4th International Conference on Agents and Artificial Intelligence (ICAART)*. to appear.

[Breuing et al., 2011] Breuing, A., Waltinger, U., and Wachsmuth, I. (2011). Harvesting wikipedia knowledge to identify topics in ongoing natural language dialogs. In *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2011). Lyon, France.*

[Bublitz, 1989] Bublitz, W. (1989). Topical coherence in spoken discourse. *Studia Anglica Posnaniensia*, 22:31–51.

[Buscaldi and Rosso, 2006] Buscaldi, D. and Rosso, P. (2006). Mining knowledge from Wikipedia for the question answering task. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy.

[Cramer et al., 2006] Cramer, I., Leidner, J., and Klakow, D. (2006). Building an evaluation corpus for german question answering by harvesting Wikipedia. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1514–1519.

[Damljanovic et al., 2010] Damljanovic, D., Agatonovic, M., and Cunningham, H. (2010). Identification of the question focus: Combining syntactic analysis and ontology-based lookup through the user interaction. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*.

[Davidescu et al., 2007] Davidescu, A., Heyl, A., Kazalski, S., Cramer, I. M., and Klakow, D. (2007). Classifying german questions according to ontology-based answer types. In *Advances in Data Analysis, Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V., Freie Universität Berlin, March 8-10, 2006*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 603–610. Springer.

[Ferrucci et al., 2010] Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefer, N., and Welty, C. (2010). Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3).

[Fissaha Adafre et al., 2007] Fissaha Adafre, S., Jijkoun, V., and de Rijke, M. (2007). Fact discovery in Wikipedia. In

Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI '07, pages 177–183, Washington, DC, USA. IEEE Computer Society.

[Fleiss, 1973] Fleiss, J. (1973). *Statistical Methods for Rates and Proportions.* Wiley.

[Furbach et al., 2008] Furbach, U., Glöckner, I., Helbig, H., and Pelzer, B. (2008). LogAnswer - a deduction-based question answering system (system description). In *Proceedings of the 4th International Joint Conference on Automated Reasoning*, IJCAR '08, pages 139–146, Berlin, Heidelberg. Springer-Verlag.

[Gabrilovich and Markovitch, 2007] Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12.

[Gerber and Chai, 2006] Gerber, M. and Chai, J. (2006). Topic term identification for context question answering. In *Proceedings of the Third Midwest Computational Linguistics Colloquium (MCLC)*, Urbana-Champaign, IL.

[Giampiccolo et al., 2007] Giampiccolo, D., Forner, P., Herrera, J., Peñas, A., Ayache, C., Forascu, C., Jijkoun, V., Osenova, P., Rocha, P., Sacaleanu, B., and Sutcliffe, R. F. E. (2007). Overview of the CLEF 2007 multilingual question answering track. In *CLEF*, pages 200–236.

[Glöckner and Pelzer, 2010] Glöckner, I. and Pelzer, B. (2010). The LogAnswer project at ResPubliQA 2010. In Braschler, M., Harman, D., and Pianta, E., editors, *CLEF (Notebook Papers/LABs/Workshops)*.

[Godfrey et al., 1992] Godfrey, J. J., Holliman, E. C., and Mcdaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1 of *ICASSP-92*, pages 517–520.

[Gupta and Ratinov, 2007] Gupta, R. and Ratinov, L. (2007). Topic spotting in dialogues using knowledge transfer. In *Proc. of the NIPS Workshop on Learning Problem Design*.

[Hatcher et al., 2010] Hatcher, E., Gospodnetic, O., and McCandless, M. (2010). *Lucene in Action.* Manning, 2nd revised edition.

[Huang et al., 2008] Huang, Z., Thint, M., and Qin, Z. (2008). Question classification using head words and their hypernyms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 927–936, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Joachims, 2002] Joachims, T. (2002). SVM light, http://svmlight.joachims.org.

[Koehler et al., 2008] Koehler, F., Schütze, H., and Atterer, M. (2008). A question answering system for German. Experiments with morphological linguistic resources. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco.*

[Kopp et al., 2005] Kopp, S., Gesellensetter, L., Krämer, N., and Wachsmuth, I. (2005). A conversational agent as museum guide – design and evaluation of a real-world application. In *Proceedings of Intelligent Virtual Agents (IVA 2005)*, pages 329–343. Springer.

[Lagus and Kuusisto, 2002] Lagus, K. and Kuusisto, J. (2002). Topic identification in natural language dialogues using neural networks. In *Proceedings of the 3rd SIGdial workshop on Discourse and dialogue - Volume 2*, SIGDIAL '02, pages 95–102, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Lemnitzer and Kunze, 2002] Lemnitzer, L. and Kunze, C. (2002). GermaNet - representation, visualization, application. In *Proceedings of the 4th Language Resources and Evaluation Conference*, pages 1485–1491.

[Li and Roth, 2002] Li, X. and Roth, D. (2002). Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics*, volume 1 of *COLING '02*, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Lin et al., 2003] Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., and Karger, D. (2003). What makes a good answer? The role of context in question answering. In *Proceedings of the 9th International Conference on Human-Computer Interaction.*

[Liu and Chua, 2001] Liu, J. and Chua, T.-S. (2001). Building semantic perceptron net for topic spotting. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, pages 378–385, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Mehta and Corradini, 2008] Mehta, M. and Corradini, A. (2008). Handling out of domain topics by a conversational characteristics. In *Proceedings of the 3rd International Conference on Digital Interactive Media in Enterteinment and Arts (DIMEA '08)*, pages 273 – 280.

[Moldovan et al., 1999] Moldovan, D. I., Harabagiu, S. M., Pasca, M., Mihalcea, R., Goodrum, R., Girju, R., and Rus, V. (1999). Lasso: A tool for surfing the answer net. In *TREC.*

[Myers et al., 2000] Myers, K., Kearns, M. J., Singh, S. P., and Walker, M. A. (2000). A boosting approach to topic spotting on subdialogues. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 655–662, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Neumann and Sacaleanu, 2004] Neumann, G. and Sacaleanu, B. (2004). A crosslanguage question/answeringsystem for German and English. In Peters, C., Gonzalo, J., Braschler, M., and Kluck, M., editors, *Comparative Evaluation of Multilingual Information Access Systems*, volume 3237 of *Lecture Notes in Computer Science*, pages 101–109. Springer, Berlin – Heidelberg, Germany.

[Peñas et al., 2010] Peñas, A., Forner, P., Sutcliffe, R., Rodrigo, A., Forăscu, C., Alegria, I., Giampiccolo, D., Moreau, N., and Osenova, P. (2010). Overview of ResPubliQA 2009: Question answering evaluation over european legislation. In Peters, C., Nunzio, G. M. D., Kurimo, M., Mandl, T., Mostefa, D., Peñnas, A., and Roda, G., editors, *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, volume 6241 of *Lecture Notes in Computer Science*, chapter 21, pages 174–196. Springer, Berlin – Heidelberg, Germany.

[Quarteroni et al., 2007] Quarteroni, S., Moschitti, A., Manandhar, S., and Basili, R. (2007). Advanced structural representations for question classification and answer reranking. In Amati, G., Carpineto, C., and Romano, G., editors, *Advances in Information Retrieval — Proceedings of the 29th European Conference on Information Retrieval (ECIR 2007), 2-5 April 2007, Rome, Italy*, volume 4425 of *Lecture Notes in Computer Science*, pages 234–245, Berlin–Heidelberg. Springer.

[Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

[Schapire and Singer, 2000] Schapire, R. E. and Singer, Y. (2000). BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168.

[Schmid, 1994] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

[Schönhofen, 2009] Schönhofen, P. (2009). Identifying document topics using the Wikipedia category network. *Web Intelligence and Agent Systems*, 7:195–207.

[Solorio et al., 2004] Solorio, T., Pérez-Coutiño, M., Montes-y Gémez, M., Villaseñor Pineda, L., and López-López, A. (2004). A language independent method for question classification. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Sonntag and Romanelli, 2006] Sonntag, D. and Romanelli, M. (2006). A multimodal result ontology for integrated semantic web dialogue applications. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italy.

[Suzuki et al., 2003] Suzuki, J., Taira, H., Sasaki, Y., and Maeda, E. (2003). Question classification using HDAG kernel. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12*, MultiSumQA '03, pages 61–68, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Vapnik, 1995] Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.

[Voorhees, 2007] Voorhees, E. M. (2007). Overview of TREC 2007. In *Proceedings of The Sixteenth Text REtrieval Conference, TREC 2007, Gaithersburg, Maryland, USA, November 5-9, 2007*. National Institute of Standards and Technology (NIST).

[Waltinger et al., 2011] Waltinger, U., Breuing, A., and Wachsmuth, I. (2011). Interfacing virtual agents with collaborative knowledge: Open domain question answering using Wikipedia-based topic models. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence - IJCAI 2011, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 1896–1902.

[Waltinger and Mehler, 2009] Waltinger, U. and Mehler, A. (2009). Social semantics and its evaluation by means of semantic relatedness and open topic models. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '09, pages 42–49, Washington, DC, USA. IEEE Computer Society.

[Zhang and Lee, 2003] Zhang, D. and Lee, W. S. (2003). Question classification using support vector machines. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 26–32, New York, NY, USA. ACM.

Ulli Waltinger is a research engineer at Siemens Corporate Technology Munich. He was a post-doctoral researcher at the Center of Excellence Cognitive Interaction Technology (CITEC) at Bielefeld University, where he also completed his PhD studies in the field of social semantics in information retrieval.

Alexa Breuing is a Ph.D. student and research assistent at the Center of Excellence Cognitive Interaction Technology (CITEC) in the KnowCIT project. Her main research interests are human-agent interaction and dialog topic detection.

Ipke Wachsmuth is chair of Artificial Intelligence at Bielefeld University. He is co-initiator and current coordinator of the Collaborative Research Center 'Alignment in Communication' (CRC 673) and a principal investigator in the Center of Excellence Cognitive Interaction Technology (CITEC).