Interfacing Virtual Agents With Collaborative Knowledge: Open Domain Question Answering Using Wikipedia-Based Topic Models

Ulli Waltinger, Alexa Breuing and Ipke Wachsmuth
Artificial Intelligence Group
Bielefeld University, Germany
{uwalting,abreuing,ipke}@techfak.uni-bielefeld.de

Abstract

This paper is concerned with the use of conversational agents as an interaction paradigm for accessing open domain encyclopedic knowledge by means of *Wikipedia*. More precisely, we describe a dialog-based question answering system for German which utilizes *Wikipedia*-based topic models as a reference point for context detection and answer prediction. We investigate two different perspectives to the task of interfacing virtual agents with collaborative knowledge. First, we exploit the use of *Wikipedia* categories as a basis for identifying the broader topic of a spoken utterance. Second, we describe how to enhance the conversational behavior of the virtual agent by means of a *Wikipedia*-based question answering component which incorporates the question topic. At large, our approach identifies topic-related *focus terms* of a user's question, which are subsequently mapped onto a category taxonomy. Thus, we utilize the taxonomy as a reference point to derive topic labels for a user's question. The employed *topic model* is thereby based on explicitly given concepts as represented by the document *and* category structure of the *Wikipedia* knowledge base. Identified topic categories are subsequently combined with different linguistic filtering methods to improve answer candidate retrieval and reranking. Results show that the topic model approach contributes to an enhancement of the conversational behavior of virtual agents.

1 Introduction

In recent years, the field of Question Answering (QA) has evolved considerably in the scientific community [Giampiccolo *et al.*, 2007]. In general, QA is a task within the areas of Information Retrieval (IR) and Natural Language Processing (NLP) that aims to automatically answer a natural language question as asked by a user. The expected answer may thereby refer to a single word or expression (e.g. *Q: "Who invented Coca-Cola?"* $\mapsto$ *A: "John Stith Pemberton"*), or to an entire sentence (e.g. *Q: "Who is John Pemberton?"* $\mapsto$ *A: "John Stith Pemberton was an American druggist and the inventor of Coca-Cola."*). In this context, most QA systems are using a collection of natural language documents (e.g. local or web-based text corpus) for document retrieval, and apply selective methods in order to extract a single answer or a list of answer candidates. Using open domain encyclopedic information as a knowledge base, such as provided by the *Wikipedia* project, has captured the attention of QA researchers lately [Ahn *et al.*, 2004; Buscaldi and Rosso, 2006]. However, most of the proposed *Wikipedia*-based QA systems focus primarily on the *document collection* of *Wikipedia* for answer retrieval, thus disregard the complex hierarchical representation of knowledge by means of its *category taxonomy*, which can also be valuable in the context of QA systems.

In this paper, we approach the *Wikipedia* collection from a different point of view. We exploit the use of the *Wikipedia* category taxonomy as a reference point for identifying the broader topic of a user's question in order to deduce from the topic to a set of *expected* answer candidates. More precisely, we are heading towards *accessing* and *activating* only those *areas* of our knowledge base (e.g. sentences and phrases categorized by a certain set of categories) which are primarily *topically relevant* to the subject of the question. As an example, consider the following user-agent-based QA scenario:

User	Question	*Who invented Coca-Cola?*
Agent	Reasoning	Who $\mapsto$ People (Male,Female)
Agent	Reasoning	Coca-Cola $\mapsto$ Soft drinks, Company,...
Agent	Task	Access knowledge base by *topic*: 'People','Coca-Cola','Soft drinks',... and by *property*: 'invent','invented','inventor',...
Agent	Answer	*John Pemberton was ... ... and the inventor of Coca-Cola*
Agent	Topic	*Oh, we speak about Soft drinks*

After the user entered a natural language question, the agent activates his reasoning module by means of analyzing both, the question structure (e.g. *Who* $\mapsto$ *Person*), and the subject matter (e.g. *Coca-Cola* $\mapsto$ *Soft drink*). Building on that a query expansion (e.g. *invented* $\mapsto$ *inventor*) is combined with the identified subject labels to retrieve the answer candidate. Consequently, the subject context, as represented by the *Wikipedia* categories, will be memorized by the agent for the next dialog. Since our QA component is employed within an existing architecture of the virtual human *Max* [Kopp *et al.*, 2005], this approach contributes to an enhancement of the agent‘s conversational behavior for two reasons: First, *knowledge awareness* enables our virtual agent to access and explore the rich knowledge of the collaborative network in a more structured manner by means of utilizing the category taxonomy of *Wikipedia* as a reference point. Dialogue-based QA obviously plays an important role here[Sonntag, 2009]. Second, *subject awareness* enables the agent to identify and to label a user's utterance (question) by its topic during the dialogue [Breuing, 2010]. Overall, we thereby aim to realize a more *human-tailored access*, as argued by [Cimiano and Kopp, 2010], to and with the aid of the rich knowledge drawn from *Wikipedia*, and consequentially we aim to improve the interaction with human dialogue partners.

The rest of the paper is structured as follows: In Section 2 we review related work. Section 3 describes the method of the QA system using *Wikipedia*-based topic models and outlines the implementation within the architecture of our conversational agent. In Section 4 we present the results of an experiment. Finally, Section 5 summarizes and concludes the paper.

2 Related Work

QA has been a popular research topic in recent years. Though, most of the unrestricted accessible QA applications, such as the web-based QA systems *QuALiM* [Kaisser, 2008] or *LogAnswer* [Furbach *et al.*, 2008], focus on presenting online a weighted list of answer candidates rather than presenting only one (exact) answer. However, to frame only a single answer is a mandatory precondition within our system architecture. In general, we can identify three different branches of QA systems with reference to their comprised knowledge base. Most popular, several technologies are using web-based search engines [Kaisser, 2008; Adafre and van Genabith, 2009], such as *Google* or *Yahoo*, and/or interlink static knowledge bases with web crawlers for document retrieval and answer candidate ranking, such as the *START* system [Katz *et al.*, 2002] or *Answerbus* [Zheng, 2002]. Other systems [Tunstall-Pedoe, 2010; Lopez *et al.*, 2010] build on combining RDF resources , such as the *DBpedia* collection [Bizer *et al.*, 2009], with a reasoning component for answer prediction. The third branch of QA systems uses semi-structured resources, such as the *Wikipedia* collection [Ahn *et al.*, 2004; Buscaldi and Rosso, 2006; Fissaha Adafre *et al.*, 2007; Furbach *et al.*, 2008], as a knowledge base.

With respect to QA systems that are using German as the target language, only few are accessible: [Neumann and Sacaleanu, 2004] presented a cross–language QA system for German and English. Their approach uses an English system in combination with machine translation in order to build a so-called *bag-of-objects* representation. Subsequently, the subset of objects a query and an answer candidate have in common are used in order to assess the answer candidates (accuracy of up to 15%). Our system also uses an overlap measure for candidate ranking, though not involving machine translation but shallow parsing. The method of [Buscaldi and Rosso, 2006] uses *Wikipedia* category information in order to determine a set of question-related articles within the *Wikipedia* collection. The results show an improvement of 14.5% in recall. Their system is in parts similar to the system presented here in terms of using category information as a reference point to improve the answer retrieval. However, it differs in that we are not using string comparison for category selection, but employing a *Wikipedia*-based topic model involving taxonomy traversal.

[Koehler *et al.*, 2008] presented a QA system for German using a web search engine as a backend. In addition, a morphological linguistic resource is used in order to convert nouns into verbs and vice versa to increase the recall (precision: 20.9%; recall 86.0%). We adapted their linguistic method by incorporating a lexical resource for query extension. Most recently, [Furbach *et al.*, 2008] presented *LogAnswer*, an open domain question answering system, which uses the *Wikipedia* dataset as the knowledge base and employs an automated theorem prover to infer correct sentences to natural language questions (precision of 54.8% for support passages). Similar to their approach, our system also uses a sentence-based representation of the *Wikipedia* document collection as a knowledge base, but additionally regards the category taxonomy to infer answer sentences.

With reference to the topic labeling task of utterances, [Lagus and Kuusisto, 2002] presented an approach using neural networks in order to recognize the subject of a long dialogue. We adapt their approach in focusing on *topic and focus words* who occur in the individual utterances. However, our method differs in that we are not using these features as a semantic representation of a topic, but as a reduced representation of a question's subject, which is consequently mapped onto the category taxonomy. That is, our topic labels refer not necessarily to term features that occurred within the spoken dialogue. In this context, a related approach is the so-called *Explicit Semantic Analysis* as proposed by [Gabrilovich and Markovitch, 2007]. Their approach utilizes the articles of the document collection of *Wikipedia* as proxies for a concept-based representation of natural language texts. That is, they classify documents with respect to an explicitly given set of *Wikipedia* articles. Related to it are the method of [Schönhofen, 2009] and the *Open Topic Model* approach [Waltinger and Mehler, 2009], which both utilize the *Wikipedia* category taxonomy for the topic labeling task. The latter is in most parts similar to our method, however, it differs in that we are not using natural language text documents as an input representation but utilize *focus terms* from utterances only.

3 Question Answering using Wikipedia-based Topic Models

The overall method for the Question Answering using *Wikipedia*-based topic models can be subdivided into several phases within the processing pipeline: (1) question processing; (2) focus term detection; (3) topic identification; (4) query formulation; (5) sentence retrieval; (6) answer extraction. Figure 1 gives an overview of our approach and the

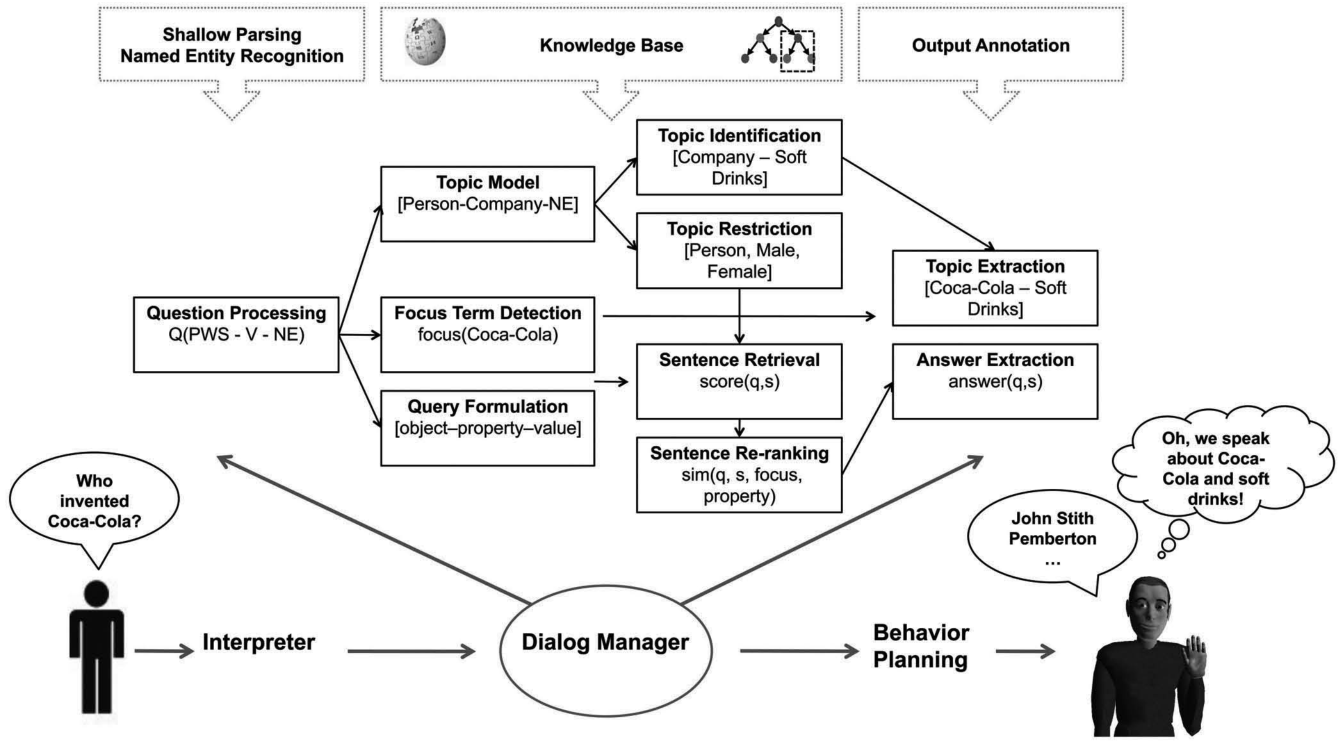

Figure 1: Overview of the QA architecture within the dialog system of our conversational agent *Max*.

corresponding modules. Since this system will be employed within an existing architecture of an embodied conversational agent, there are several specifications to meet. The first challenge is the runtime property of the entire QA component. The system must respond within a few seconds, even if the knowledge base consists of more than 30 million entries. The second challenge refers to the robustness in confidence of answer extraction. That is, to have a plausible conversation with a virtual agent, it is critical to present rather no answer and consequently to fall back on the existing dialog manager, than to output a (completely) wrong statement. Therefore, in terms of evaluation metrics, we focus on precision instead of recall. A third challenge for QA with conversational agents is answer presentation. Unlike other QA applications, the system needs to answer natural language questions via the virtual agent. This means that it is not adequate to present an answer or a list of answer candidates with supporting passages using a (hyper-) text representation, but we need to vocalize the answer using speech synthesis. In the following we describe the processing stages of the system.

3.1 Question Processing

At first, the question processing module is activated. All incoming natural language queries (questions) are linguistically analyzed using the shallow processing tool *TreeTagger* [Schmid, 1994]. It consists of several pre-processing components for tokenization, sentence boundary detection, Part-of-Speech (PoS) tagging and lemmatization. In addition, an embedded chunk parser defines the type of the syntactic chunks (e.g. NC, PC or VC) that occur in the question. Chunker and PoS-Tagger were trained on the German *Negra* treebank using the *STTS* tagset[1]. Named Entity Recognition is done using a rule-based approach as provided by the *ANNIE* module within the *GATE* framework [Cunningham *et al.*, 2002]. Therefore, each question is represented by the chunk structure and its corresponding wordform, lemma, PoS and named entity class information as determined by the shallow parser.

Thereupon, the question type of the input query is identified. In this context, the analyzed query representation is matched against several classification patterns, which have been pre-defined for a set of comprised question types (see Table 1) using the dataset of [Cramer *et al.*, 2006] for building the question classification rule set. More precisely, we have annotated for each question type a number of *Wikipedia* categories, where we expect to find the list of answer candidates by means of their taxonomy membership. For example, a question starting with the terms *'Who is ...'* is most likely to be a question about a specific person. In this case, our method activates, at first, only those knowledge base entries, which are annotated by the *Wikipedia* categories *'people','male' or 'female'*. At second, in the case of a mismatch, the entire knowledge base is activated for answer retrieval.

3.2 Focus Term Detection

Focus term detection builds upon the shallow parsing component. The goal of this module is to identify *topically relevant words in the utterances* [Lagus and Kuusisto, 2002, pp. 95].

[1] http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html

Question type:	Expected Answer Features:	Expected Position:	Expected Category:
Who - [Wer, Wie heisst]	NN, NE followed by VAFIN	first sentence / title	people, male, female,...
What - [Was, Womit]	PDS, PRELS followed by NN, NE	first sentence / title	topic model categories
Where - [Wo, Wohin]	APPR followed by NE, NN	first sentence / title	country, city,...
When - [Wann]	expression of dates, CARD	paragraph	topic model categories
How many - [Wie viele]	CARD, expression of length (e.g. meter, km,...)	paragraph	topic model categories

Table 1: Question types by *Wikipedia* article structure, expected answer features, and expected category membership.

This is done, in order to have a topic-based input representation of a user's question – which is needed for the *Wikipedia*-based topic model. The idea, from sentence to topic, follows thereby primarily the definition of [Schank, 1977, pp. 422], who argues that *a topic is any object, person, location, action, state, or time that is mentioned in the sentence to be responded to.* In our context, we see the set of topically relevant terms within an utterance, defined as *focus terms*, as a proxy of a question's topic. *Focus term* detection is processed by means of the analyzed syntactic chunks. More precisely, we utilize the concatenated noun and prepositional chunks (NC,PC) by their PoS-Tag (NE) as our topic representation. For example, the question *'Who invented Coca-Cola?'* is represented by the single focus term *'Coca-Cola'*. The extracted *focus terms* are further used as an input for the topic identification module.

3.3 Topic Identification

The purpose of the topic identification component is to equip our virtual agent with a topic-based reasoning module and consequently to assist the topic-based answer retrieval. In this context, we aim to map any given input question onto the category taxonomy of *Wikipeda*, where the graph structure of the taxonomy is used to derive the broader subject from the input query and the category nodes are used as distinctive labels of the question topic. A recent example of such a topic model is presented through the *Open Topic Model* [Waltinger and Mehler, 2009] in which natural language documents are classified with respect to *Wikipedia* categories. At large, their approach maps any given input stream q onto a high-dimensional real-valued concept space, C_{wiki}, using *Wikipedia* articles as proxies for concepts.

$$f : q \to C_{wiki} \quad (1)$$

The entries of the resultant vector $c_{art} \in C_{wiki}$ of q reflect thereby the strength of association between q and the respective *Wikipedia* articles. In a further processing step, they utilize the set of top-ranked articles from c_{art} to retrieve associated category nodes from C_{topic}:

$$f : c_{art} \to C_{topic} \quad (2)$$

At last, the graph structure of the category taxonomy is used to identify topic-related concepts within a certain scale of generalization. In our QA application, we adopted their approach, though not using natural language documents but the *focus term* representation (see Table 2), as described in the previous section.

3.4 Query Formulation

In order to enhance the recall of the sentence retrieval component, the input question is expanded to a set of search query variants. This is done by means of triple extraction, in terms of *object-property-value* detection, using the shallow parsed chunk representation. The *object* thereby refers to the *focus term* representation, the *property* to the *verbal chunk* of the question, and the *value* to the answer we are looking for. In addition, this module also takes the inflectional and derivational morphology of the terms into account. That is, verbs and nouns are replaced by their lemma and synonyms utilizing a manually annotated lexical dictionary using data from the *Wiktionary*[2] project. For example, the question *'Who invented Coca Cola?'* is translated into the following triple queries: *['Coca-Cola','invent','?']*, *['Coca-Cola','invented','?']*, *['Coca-Cola','inventor','?']*, and so on.

Question:	*Who invented Coca-Cola?*
Focus term:	*Coca-Cola*
Top-ranked articles:	1. The Coca-Cola Company
	2. Coca-Cola
	3. Coca
Top-ranked topics:	1. Soft drink
	2. Beverage company
	3. Company (Atlanta)

Table 2: Outline of the *Wikipedia*-based topic model applied to a natural language question.

Type:	Quantity:
articles	1.063.772
paragraphs	6.649.455
sentences	30.890.452
categories	88.749

Table 3: Quantity of utilized content items using the German *Wikipedia* collection (Version 10/2010).

3.5 Sentence Retrieval

The sentence retrieval component utilizes the German *Wikipedia* dump as a QA knowledge base. More precisely, we use *Apache Lucene* [Hatcher *et al.*, 2010] to index the document collection, utilizing 1.063.772 articles and 88.883 categories (see Table 3). The entire corpus was linguistically analyzed and subdivided into 30.890.452 sentences. That is, each sentence poses as a *Lucene document* that consists of seven *fields*: *Title*, which contains the title of the *Wikipedia* article in which the sentence occurred; *Text*, which stores the individual sentence; *Chunk*, which stores the shallow parsed repre-

[2]http://de.wiktionary.org/

sentation; *Position*, which lists the position of the sentence in the article; *Backlink*, which stores the number of hyperlinks pointing to the respective sentence (article); *Header*, which utilizes the headings of (sub-)sections within the article; sand finally *Category*, which stores the *Wikipedia* categories attached to the title page. For sentence retrieval, we apply the *MultiFieldQueryParser* using the *Lucene* search score:

$$score_{lucene}(q,s) = \sum_{t \in q}(tf(t \in s) \cdot idf(t)^2 \cdot t_b \cdot norm(t,s)) \quad (3)$$

where $tf(t \in s)$ correlates to the terms frequency in the currently scored sentence s; $idf(t)$ represents the inverse document frequency applied to the sentence representation. t_b is a search time boost of term t in the query q, applied to *focus terms* only. $norm(t,s)$ encapsulates a few (indexing time) boosts and length factors with reference to *Lucene's* document and field boost property [Hatcher *et al.*, 2010].

Note that we combine all query variants, obtained from the *query formulation*, the *question processing*, and the *Wikipedia*-based *topic model* component, to one query. In addition, category labels, as assigned from the latter module, are used as a *mandatory parameter* for the sentence retrieval task. This means that the *type* and the *topic* of a question influences significantly the query formulation process. Thus, answer candidates are filtered by their *article position*, *expected answer features*, and *taxonomy membership* (see Table 1). To give an example, for the question "*Who is John Pemberton?*", we 'activate/query' only those sentences in which the focus terms *John* and *Pemberton* occur in the first sentence of the respective article (*Position:1*) and which are additionally affiliated to one of the following topics: *Category:Male, Category:Female, Category:Human name disambiguation pages* (question type), *Category:American chemists, Category:Coca-Cola* (topic model).

3.6 Answer Extraction

In the current QA setup, we disregard the task of answer paraphrasing and extract the final answer by means of its sentence-based representation. Moreover, we only use the top-ranked sentence as an output for our conversational agent. To give an example answer[3]:

> ***John Pemberton*** *[a] (1831–1888) was an American druggist and the* ***inventor*** *[p] of* ***Coca-Cola*** *[o].*

Sentence re-ranking is performed by combining four different evidence scores. First, we normalize the *Lucene* retrieval similarity. Second, we score the lexical overlap between the question and the answer candidate using the *Jaccard* similarity index:

$$score_{jac}(q,s) = \frac{a_{q,s}}{a_{q,s} + b_q + c_s} \quad (4)$$

where the size of the intersection between the sentence s and the query q gets divided by the size of their union. Note that we use the *object* and *property* chunks only as a set-based representation of q. Third, we score the normalized *word index distance* of *object* (o) and *property* (p) in the answer candidate, defined as

$$score_t(o,p) = 1/|dis(o,p)| \quad (5)$$

Fourth, we apply the *word index distance* to the *object* (o) and the expected answer candidate (a) within the sentence. The rationale behind the latter heuristic scores is that we favor shorter sentences as answers by means of their syntactic structure (object-property and object-answer distance). Note that for re–ranking the respective lemma representation of s, q (o,p), and a is used. Subsequently, all evidence scores are summated and normalized. In addition, in the case of an ambiguous question, such as "*Who is John Pemberton?*[4]", we apply a backlink strategy for answer retrieval. That is, we use the hyperlink topology of *Wikipedia* as a proxy for commonsense knowledge (see Table 4). Consequently, we only use the top ranked sentence as the predicted answer candidate.

Title:	Frequency Backlinks:
John Pemberton (inventor)	198
John C. Pemberton (general)	113
John Pemberton (footballer)	53
John Pemberton (anthropologist)	17

Table 4: Frequency of backlinks for ambiguous input question: *"Who is John Pemberton?"*

4 Evaluation

For the evaluation of the system we utilized 200 questions from the *CLEF-2007* monolingual QA task, using German as the target language (best–in–class exact answer accuracy results by *DFKI* 39.29% and *Freie Uni Hagen* 28,57%) [Giampiccolo *et al.*, 2007]. However, the conducted evaluation setup differs slightly to the *CLEF* task. First, we manually performed the anaphora-resolution challenge within the evaluation dataset. Second, we evaluated the answers by means of their sentence representation only. That is, the exact answer has not been extracted, but had to be included in the answer sentence as determined by the system. Results grouped by question type are displayed in Table 5. The results show that the topic model approach allows to achieve an overall accuracy of 44% for the lenient task.

Question Type	Frequency	Accuracy
All	88/200	**44.0**
Factoid	65/164	39.6
Definition	21/28	**75.0**
List	2/8	25.0

Table 5: Results of the German QA task grouped by question type and accuracy.

Even though the evaluation indicates a mediocre performance for list-based and factoid-based question types with respect to definition-based question types, such as "*Who is ...*"

[3]http://en.wikipedia.org/wiki/John Pemberton (disambiguation)

[4]There exist six different *John Pemberton* in the Wikipedia dataset.

or "*What is* ...", the method presented in this paper performs very well (accuracy of 75%). Obviously, list-based question types are a hardly feasible task within this kind of evaluation setup, since the exact answer *list* has to occur within *one* sentence. With respect to the results for factoid-question types, we can identify with an accuracy of 39.6 that our QA system achieves only average results. The evaluation showed that there are two main reasons for this: First, the combination of an incorrectly deduced *Wikipedia*-based topic model and the question type classification has led to an inaccurate category selection within the sentence retrieval module. More precisely, in 72 cases of the 112 answers, which were classified as incorrect (64.0%), the QA system returns no answer at all. Second, the system rates an incorrect sentence misleadingly as *correct* , if the sentence has all 'ingredients' of a plausible answer. As for example, consider the following question from the *CLEF* task: *Who was the director of "Gone with the wind"?* The system returns: *Gone with the Wind: As a director 'George Cukor' started.*[5] However, the actual director of the movie was Victor Felming, who replaced Cukor after less than three weeks of shooting. This information is mentioned in the next sentences of the used *Wikipedia* article. This example clearly shows the drawbacks of the sentence-based QA approach, which disregards the sentence context (e.g., analyzing the entire section or paragraph) for answer prediction. Currently, we focus on integrating additionally the *RDF* dataset from the *DBpedia* project [Bizer *et al.*, 2009], in order to overcome the shortcomings of the factoid-question types. At large, one of the main effects of the *Wikipedia*-based topic model QA approach determined in the evaluation is that utilizing category filtering and re-ranking leads the system to rather return no answer instead of retrieving a wrong one, which has a positive effect in our framework. That is, in 160 cases of the 200 questions (80.0%) from the *CLEF* dataset, the QA system returns either the *correct* or *no* answer. A second effect is that the system favors shorter and more general answers (definition) due to the global topic categories and the overlap re–ranking. This contributes to the performance of definition-based question types. The satisfying results (accuracy of 75%) for the latter question types can be traced back to the good performance of the *Open Topic Model* and the underlying structure of our knowledge base representation (e.g., sentence positions within the articles). In this context, we can state that by the access to definitions of more than one million entities, our approach contributes to the exploration of collaborative knowledge via virtual agents.

5 Conclusion

In this paper we examined two different aspects for the task of interfacing virtual agents with collaborative knowledge. First, we explored the use of *Wikipedia* categories as a basis for identifying the broader topic within a dialog. The proposed approach identified topic-related focus terms of a user's question, which were subsequently mapped onto the category taxonomy of *Wikipedia* using a *Wikipedia*-based topic model. Second, we described a question answering framework for German which utilizes the category taxonomy as a reference point for context detection and answer prediction. Results showed, with an average accuracy of 44%, that *Wikipedia* is a useful resource to enhance the conversational behavior of our virtual agent. In the future, we envision to explore the usefulness of taking the topic context within longer dialogues of human-agent interaction into account. Moreover, we plan to enhance the answer re–ranking model by means of syntactic relation patterns, to integrate an additionally *RDF* query component, and to perform a comprehensive evaluation of the system using different QA reference datasets for the German language.

[5] The original German answer: "Als Regisseur begann 'George Cukor' mit der Arbeit."

Acknowledgments

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

References

[Adafre and van Genabith, 2009] Sisay Fissaha Adafre and Josef van Genabith. A hybrid filtering approach for question answering. In *Proceedings of the 14th International Lexical Functional Grammar Conference (LFG-09)*, Cambridge, UK, 2009.

[Ahn *et al.*, 2004] David Ahn, Valentin Jijkoun, Gilad Mishne, Karin Mller, Maarten de Rijke, and Stefan Schlobach. Using wikipedia at the trec qa track. In *Proceedings of TREC 2004*, 2004.

[Bizer *et al.*, 2009] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semant.*, 7:154–165, September 2009.

[Breuing, 2010] Alexa Breuing. Improving human-agent conversations by accessing contextual knowledge from wikipedia. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03*, WI-IAT '10, pages 428–431, Washington, DC, USA, 2010. IEEE Computer Society.

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

[Cimiano and Kopp, 2010] P. Cimiano and S. Kopp. Accessing the web of data through embodied virtual characters. *Semantic Web Journal*, 2010.

[Cramer *et al.*, 2006] Irene Cramer, Jochen L. Leidner, and Dietrich Klakow. Building an evaluation corpus for german question answering by harvesting wikipedia, 2006.

[Cunningham *et al.*, 2002] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary*

Meeting of the Association for Computational Linguistics, 2002.

[Fissaha Adafre *et al.*, 2007] Sisay Fissaha Adafre, Valentin Jijkoun, and Maarten de Rijke. Fact discovery in wikipedia. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 177–183, Washington, DC, USA, 2007. IEEE Computer Society.

[Furbach *et al.*, 2008] Ulrich Furbach, Ingo Glöckner, Hermann Helbig, and Björn Pelzer. 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, 2008. Springer-Verlag.

[Gabrilovich and Markovitch, 2007] Evgeniy Gabrilovich and Shaul Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12, 2007.

[Giampiccolo *et al.*, 2007] Danilo Giampiccolo, Pamela Forner, Jesús Herrera, Anselmo Peñas, Christelle Ayache, Corina Forascu, Valentin Jijkoun, Petya Osenova, Paulo Rocha, Bogdan Sacaleanu, and Richard F. E. Sutcliffe. Overview of the clef 2007 multilingual question answering track. In *CLEF*, pages 200–236, 2007.

[Hatcher *et al.*, 2010] Erik Hatcher, Otis Gospodnetic, and Mike McCandless. *Lucene in Action*. Manning, 2nd revised edition. edition, 2010.

[Kaisser, 2008] Michael Kaisser. The qualim question answering demo: supplementing answers with paragraphs drawn from wikipedia. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, HLT-Demonstrations '08, pages 32–35, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

[Katz *et al.*, 2002] Boris Katz, Sue Felshin, Deniz Yuret, Ali Ibrahim, Jimmy J. Lin, Gregory Marton, Alton Jerome McFarland, and Baris Temelkuran. Omnibase: Uniform access to heterogeneous data for question answering. In *Natural Language Processing and Information Systems, 6th International Conference on Applications of Natural Language to Information Systems, NLDB 2002, Stockholm, Sweden, June 27-28, 2002, Revised Papers*, volume 2553 of *Lecture Notes in Computer Science*, pages 230–234. Springer, 2002.

[Koehler *et al.*, 2008] Florian Koehler, Hinrich Schütze, and Michaela Atterer. 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*, 2008.

[Kopp *et al.*, 2005] S. Kopp, L. Gesellensetter, N. Krämer, and I. Wachsmuth. A conversational agent as museum guide – design and evaluation of a real-world application. In Panayiotopoulos and others (Eds.), editors, *Intelligent Virtual Agents*, LNAI 3661, pages 329–343, Berlin, 2005. Springer.

[Lagus and Kuusisto, 2002] Krista Lagus and Jukka Kuusisto. 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, 2002. Association for Computational Linguistics.

[Lopez *et al.*, 2010] Vanessa Lopez, Andriy Nikolov, Marta Sabou, Victoria S. Uren, Enrico Motta, and Mathieu d'Aquin. Scaling up question-answering to linked data. In *Proceedings of the EKAW - Knowledge Engineering and Management by the Masses - 17th International Conference, EKAW 2010, Lisbon, Portugal, October 11-15, 2010.*, pages 193–210, 2010.

[Neumann and Sacaleanu, 2004] Günter Neumann and Bogdan Sacaleanu. A Cross Language Question Answering System for German and English. In Carol Peters, Julio Gonzalo, Martin Braschler, and Michael Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems*, volume 3237 of *Lecture Notes in Computer Science*, pages 101–109. Springer Berlin / Heidelberg, 2004.

[Schank, 1977] Roger C. Schank. Rules and topics in conversation. *Cognitive Science*, 1(4):421–441, 1977.

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

[Schönhofen, 2009] Peter Schönhofen. Identifying document topics using the wikipedia category network. *Web Intelli. and Agent Sys.*, 7:195–207, April 2009.

[Sonntag, 2009] Daniel Sonntag. Introspection and adaptable model integration for dialogue-based question answering. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1549–1554, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[Tunstall-Pedoe, 2010] William Tunstall-Pedoe. True knowledge: Open-domain question answering using structured knowledge and inference. *AI Magazine*, 31(3):80–92, 2010.

[Waltinger and Mehler, 2009] Ulli Waltinger and Alexander Mehler. 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, 2009. IEEE Computer Society.

[Zheng, 2002] Zhiping Zheng. Answerbus question answering system. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 399–404, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.