SAARLAND
UNIVERSITY

Automatic Context-Relevant Off-Activity Talk Suggestion for Dialogue Contribution

By:

Omid Moradiannasab

Supervisors:

Ivana Kruijf-Korbayova
Günter Neumann
Johan Bos

Master of Science Thesis

Faculty of Science
Computational Linguistics and Phonetics Department

October 2015

Automatic Context-Relevant Off-Activity Talk Suggestion for Dialogue Contribution

By:

Omid Moradiannasab

Supervisors:

Ivana Kruijf-Korbayova

Günter Neumann

Johan Bos

Master of Art Thesis

Faculty of Art
Research Master Linguistics

October 2015

Declaration of Authorship

I hereby confirm that the thesis presented here is my own work, with all assistance acknowledged.

Saarbrücken, October 2015

Omid Moradiannasab

“Thinking is a human feature. Will AI someday really think? That’s like asking if submarines swim. If you call it swimming then robots will think, yes.”

Noam Chomsky

Saarland University

Abstract

Faculty of Science
Computational Linguistics and Phonetics Department

Master of Science

by Omid Moradiannasab

This thesis investigates the possibility of utilizing online resources for off-activity dialogue contribution. It is a first attempt to propose a tool which automatically suggests *off-activity talk*s in form of some sentences relevant to the dialogue context. We propose four approaches and comparatively evaluate them over two test-sets of open domain and health-related queries in a conversational quiz-like setting. The evaluation results show satisfying performance for some of the proposed approaches. The results suggest that the modular architecture implemented throughout this work provides an applicable and effective system to process dialogue context and suggest relevant *off-activity talks.*

Keywords: conversation system, dialogue contribution, human-computer interaction, small talk, off-activity talk, topic modeling

Acknowledgements

I would like to thank my supervisors in Saarbrücken, Dr. Ivana Kruijf-Korbayova and Professor Günter Neumann, as well as my supervisor in Groningen, Professor Johan Bos for their guidance and assistance throughout my thesis.

My sincere appreciation is extended to all my trusted friends and colleagues, near and far, who provided emotional support, and shared some of their brilliance with me in the form of feedback on chapters or presentations.

Last but not least, I would like to thank my parents for supporting me spiritually throughout writing this thesis and my life in general.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Conversational agents (e.g. virtual characters, chat-bots, etc) are software programs that interact with users employing natural language processing capabilities. This ability in companion with non-verbal functionalities of an agent can empower it to provide services to human on several applications in the fields of education, training, entertainment, help desks, or personalized services. *Traditional conversational systems* are designed for specific purposes, such as a banking service or navigating through a website. Dialogue in such systems is limited to *task-bound talks* (also called *activity talks*). These talks have a specific purpose and follow a particular structure. *Traditional conversational systems* are developed to attempt to engage the user in a natural, robust conversation in well-defined domains. However, empirical investigations reveal that the effect of these systems on engagement of the users with the system and their perception of the agent's intelligence is debatable [Dehn and Van Mulken, 2000]. *Relational agents*, on the other hand, are defined in literature as "computational artifacts designed to establish and maintain long-term social-emotional relationships with their users" [Bickmore and Picard, 2005]. In order to achieve such relationships, certain amount of user's trust and engagement is required (see 3.1). Various conversational strategies are employed in *relational agents* that comprise models of social dialogues with the aim of raising user's trust.

As an instance of a conversational strategy employment, *small talk* (also *social talk*) is discussed in literature. It is introduced as a kind of talk which executes conversational strategies. While interleaved between *task-bound talks*, *social talk* indirectly builds trust through the natural progression of a conversation. Bickmore and Cassell define *small talk* as "any talk in which interpersonal goals are emphasized and task goals are either non-existent or de-emphasized"[Bickmore and Cassell, 2001]. Kluwer defines it as a talk "often perceived as unsophisticated chit-chat in which content exchange is irrelevant and negligible"[Klüwer, 2015]. Following this definition, *small talk* (e.g. about the weather, events and objects around, stories about the environment or the virtual character itself) represents the opposite of *task-bound talk* which aids the execution of a particular task. Thus, the range of topics and contents is definitely much more unrestricted in *small talk* than in *task-bound talk*. *Small talk* is useful to develop the conversation and to avoid pauses. It can be used to ease the situation and to make a user feel more comfortable in a conversation with an agent [Cassell and Bickmore, 2000]. It is also introduced as a way of assuring certain amount of *closeness* to the user (e.g. before asking personal questions)[Bickmore and Cassell, 2001]. *Small talk* is also helpful to avoid repetitiveness of conversations which is counted in literature as a negative impact factor to the users' motivation in interaction with the agent. (see 3.2)

Similar to *small talk*, *off-activity talk* (also called *non-activity talk*) is another technique to employ conversational strategies. Table 1.1 briefly lists differences and similarities between the features of *small talk* and OAT. Both *small talk* and *off-activity talk* enrich the task-oriented dialogue via opening the structure of the conversation. However, OAT can be differentiated from *small talk* by the topic and the purpose of the talk. OAT has a specific purpose (e.g. knowledge exchange) and is about a specific topic, while a *small talk* is an independent talk without any functional topics. Several studies have found that the purpose of *small talk* is not to negotiate knowledge but to aid in management of social situation. On the other hand, *off-activity talk*, as we define it here, is to disclose some information relevant to the dialogue context. Therefore, whenever no divergence from the subject matter of the *task-bound talk* is required, OAT is preferred to *small talk*. Even though OAT is a diversion from the structure of the task-dialogue, it maintains the dialogue topic.

Table 1.1: Comparison of *small talk* and OAT

	Changes dialogue's structure	**Has a specific topic**	**Keeps dialogue's topic**
Small talk	✓	X	X
OAT	✓	✓	✓

OAT was first defined by [Kruijff-Korbayova et al., 2014] in resemblance to *small talk*. In their work, the purpose of *activity talk* is knowledge exchange or knowledge probing while *off-activity talk* is used to break out of the fixed structure of task-bound dialogues. They use OAT in form of prerecorded questions around a set of predefined topics (e.g. hobbies, diabetes, eating habits, friends, diary, etc) to encourage the user to talk about those topics in order to elicit information from them. In this thesis, we focus on *off-activity talk* with a similar definition but a different employment. An OAT in the current work is a follow-up added information relevant to the context of the previous interaction with the aim of encouraging users' engagement and possibly promoting their trust in knowledgeability and intelligence of the agent, thus no deliberate direct information elicitation is targeted in this thesis.

The initial motivation for this thesis originates from an EU-funded project called Aliz-E (The Adaptive Strategies for Sustainable Long-Term Social Interaction) which includes development of a *relational agent*. The aim of that project is to sustain a long-term interaction between a user (a diabetic child) and a robot through encouraging the user to follow the conversation[Kruijff-Korbayová et al., 2012]. Aliz-E is described in more details in 3.5.

The idea of context determination and relevant off-activity talk suggestion for dialogue contribution, in a broad sense, can be used in any task-oriented dialogue setting. However, due to its predominantly verbal character and naturally constrained interaction structure, a conversational quiz-game setting is chosen as a good test bed for the current thesis. This setting is similar to

Aliz-E Quiz activity scenario. In their scenario, the agent asks the user a multiple-choice question from an open domain. After the user selects one of the choices (which can be correct or not), the agent should give a verbal reaction (*off-activity talk*). Our setting reflects the same format with this simple difference that the follow-up can be uttered by any of the partners in the previous dialogue or even a third person rather than the questioner agent. In this setting, we need the follow-up to be related to the content of the previous interaction. It should be a piece of information on the main subject matter extracted from the available online resources and possibly, but not necessarily, confirm or give the correct answer. It can include a provision of some added information as a follow-up to the previous content. Some examples of dialogues in this setting, in addition to some sample OATs to be uttered by the agent right after this dialogue, are presented in table 1.2. It is worth noting that all of these examples are outputs of the tool developed in the course of this thesis from the test set of table 5.3 as input.

Table 1.2: Some samples of a suitable OAT

Previous Sub-dialogue	**Follow-up OAT**
A: What is the capital of Chechnya? B: Grozny	C: Not long ago, Grozny, the capital of Chechnya, was called ”the most devastated city on Earth.”
A: What is the capital of Chechnya? B: Grozny	C: Chechnya Republic is a federal subject of the Russian Federation, part of North Caucasian District.
A: When was the Berlin Wall built? B: in 1961	C: Originally a barbed wire fence, the first concrete sections were built in 1965.
A: When was Elvis Presley’s first record recorded? B: on July 5, 1954	C: On July 5 1954, Elvis Presley changed music world forever.
A: How many portions of fruit and vegetables should we try to eat? B: At least five a day	C: Healthy diet means 10 portions of fruit and vegetables per day, not five.

A dialogue manager component in a conversation system (see 3.4), which takes the major responsibility of controlling the conversation flow, is the consumer of the OAT suggestions. The dialogue manager decides when and how to come up with an OAT. When it decides so, it provides the context to the OAT suggestion component. Having the results returned form the OAT suggestion component as a ranked list, the dialogue manager will decide whether and which OAT to use. It can choose one of the suggested OATs from the ranked list right after the question-answer dialogue.

Overall, the current thesis is an exploration toward verifying the hypothesis that providing automatically generated talks by the agent out of the main activity (i.e. *off-activity talk*), analogously to the effects of *small talk*, leads to building up user’s engagement. Within this wider vision, the concrete objective of this thesis is to propose a tool which automatically suggests *off-activity talk*s in form of some sentences relevant to the dialogue context to be used in a conversational quiz-game setting.

1.1 Research Objectives

The goal of this project is to develop a tool which takes advantage of the Web as an online resource to seek for relevant sentence(s) to a given dialogue context. This tool is supposed to return a ranked list of appropriate candidates for dialogue contribution in order to suggest to the dialogue manager of a conversation system. A quiz-game setting is used as the test bed for this project. The main benefit of such a tool will be to break out of the fixed structure of task-bound dialogues between an artificial agent and a human user, probably leading to an increment in users' engagement.

As listed below, the overall task of context-relevant dialogue contribution using online resources can be divided into a number of sub-problems within a processing pipeline. The tool we implement in the course of this thesis is supposed to provide a solution to each of these sub-problems:

1. focus term detection
2. topic identification
3. query formulation/expansion
4. relevant content retrieval
5. ranking

At large, the procedure is as follows. The first step is to identify topic-related *focus terms* of a dialogue. Using these *focus terms*, the second step is to determine the major topic to be followed up in the conversation. The topic labels after this step do not necessarily need to match the terms that occurred within the previous dialogue. In closed-domain solutions, this can be done by mapping the *focus terms* onto a category taxonomy. In such a case, the taxonomy is utilized as a reference point to derive topic labels for a follow-up conversation. On the other hand, for an open-domain problem the task is not as straightforward. The reason to that is the lack of predefined taxonomy. Identified topic categories are subsequently combined with information retrieval and different linguistic filtering methods to improve candidate content retrieval. In the end, the retrieved content will be ranked based on their relevance to the context and properness for dialogue contribution.

In the process of this thesis, following research questions are to be answered:

1. What can be the architecture of a feasible tool able to suggest off-activity talks? (see 4.1)
2. Which approaches can be applied to achieve a high quality sentence selection? (see 4.3)
3. Which of the proposed approaches is more effective? (see 6)
4. How far can we reach by employing topic modeling techniques in the proposed tool? (see 4.1.7 and 6)

1.2 Thesis Structure

This thesis is organized in the following way. In chapter 2, we present the most related literature. In chapter 3, background information about the involved resources and concepts is provided. This chapter also presents tools to overcome the sub-tasks. Aliz-E project as the test bed for the knowledge and computational tools developed in this thesis will be discussed in this chapter as well. Chapter 4 describes the architecture of our system and any other work related to the research goals achieved in this thesis. Chapter 5 covers the experimental setup. Chapter 6 represents the results obtained by our baseline system and the proposed methods. The thesis concludes in chapter 7 with a summarization. Finally, some suggestions for future work are presented in chapter **??**.

Chapter 2

Related Works

There is a large body of literature on the techniques that conversational agents can employ to establish and maintain long-term social-emotional relationships with their users. However, in this chapter we present the most relevant works to the content of this thesis which is automated retrieval of relevant content from online sources for dialogue contribution to improve the interaction with human dialogue partners.

Our approach is unique in the sense that it combines current research in the field of Human Computer Interaction (HCI), especially the field of interaction design, with recently developed technology from Language Technology like information retrieval, natural language processing and topic modeling. As explained in 1.1, the target task of this thesis consists of a number of sub-problems each of which is studied in different works. For this reason we separately explain related works to each sub-problem in section 2.1. Section 2.2 gives an overview of topics related to evaluation. Some existing work regarding definition of *relevance* are presented in section 2.3.

2.1 Sub-problems

The most similar work to the current thesis is the one conducted by [Waltinger et al., 2011]. They describe a dialogue-based question answering system for German which utilizes Wikipedia-based topic models as a reference point for context detection and answer prediction. Even though they provide a system for a question answering task, analogously their approach can be cited as a relevant work for dialogue contribution. This is on the ground that their approach provides a solution to each of the sub-tasks of this thesis. A major difference between that work and this thesis is that they take use of Wikipedia categories as a basis for identifying the broader topic of a spoken utterance. Second, they 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. They utilize the taxonomy as a reference point to derive topic labels for a user's question. Results show that their topic model approach contributes to an enhancement in the conversational behavior of virtual agents. Following subsections will cover the related works to deal with each of the sub-problems of the current work.

2.1.1 Focus Term Detection

Focus term detection aims at identifying topically relevant words in the utterances. The goal is building a topic-based representation of a (sub-)dialogue which is needed for topic identification. The solutions to this task are usually built upon shallow parsing. The idea of extracting the topic from the uttered sentences primarily follows the definition of [Schank, 1977], 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 dialogue‘s topic.

Waltinger and colleagues use syntactic chunks to detect *focus terms* [Waltinger et al., 2011]. To be more precise, they utilize the concatenated noun and prepositional chunks (NC,PC) by their PoS-Tag (NE) as a proxy to the topic. For example, the question “Who invented the typewriter?” is represented by the single *focus term* “typewriter”. The extracted *focus terms* are further used as an input for the topic identification module. In [Figueroa et al., 2009], the authors propose a method which takes advantage of surface patterns in combination with corpus-based semantic analysis and sense disambiguation strategies to extract words from the question which represent the target concepts.

In order to guide answer search, it is common to include natural language question analysis in QA systems. The main aim of the question analysis in the context of a QA system is to identify the set of relevant keywords, the set of recognized named-entities and also the expected answer type. In [Figueroa and Neumann, 2006], the authors propose a method for ’answer context prediction’ by analyzing a natural language question. Beside local lexical and syntactical criterion, Neumann and colleagues confirm that in many situations larger syntactic units together with information extracted from external knowledge sources need to be considered in question analysis component of a QA system [Neumann and Sacaleanu, 2005].

2.1.2 Topic Identification

In human-computer interaction systems using natural language, the recognition of the topic from user’s utterances is an important task. In order to build up a successful dialogue, topic analysis must be carried out. As mentioned in [Breuing, 2010], *subject awareness* enables the agent to identify and to label a user’s utterance by its topic during the dialogue. The purpose of the topic identification task is to equip a conversational system with a topic-based inference and as a result to assist the topic-based content retrieval.

Waltinger and colleagues employ a *topic model* which is based on explicitly given concepts as represented by the document and category structure of the Wikipedia articles[Waltinger et al., 2011]. They also include some reasoning process by analyzing the subject matter. The authors report that this approach contributes to an enhancement of the agent’s conversational behavior in closed domain dialogues. This technique equips the agent with a level of so-called *knowledge awareness* which enables it to explore the source of content in a more structured manner by means of utilizing the category taxonomy of Wikipedia as a reference point.

Another related approach is the so-called *Explicit Semantic Analysis* as proposed by [Gabrilovich and Markovitch, 2007]. In their approach, they use the articles in 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 this method are the approach of [Schönhofen, 2009] and the Open Topic Model approach [Höppner et al., 2009], both utilizing Wikipedia category taxonomy for the topic labeling task. The latter is the most suitable one for our architecture. However, it differs in that we are not using natural language text documents as an input representation but utilize *focus terms* from utterances only.

[Lagus and Kuusisto, 2002] use neural networks to recognize the subject of a long dialogue, focusing on topic and *focus terms* of individual utterances. They describe two approaches in the analysis of topical information: (1) the use of *topically ordered document maps* for analyzing the overall topic of dialogue segments, (2) identification of topic and *focus terms* in an utterance for sentence-level analysis and identification of topically relevant specific information in short contexts.

Latent Dirichlet Allocation [Blei et al., 2003] is a generative probabilistic model which can be used to model text corpora. LDA and its derived models treat a document as a bag-of-words where word order and other important linguistic structures are neglected. Each document is modeled as a finite mixture over an underlying set of various topics and each topic is modeled as a probability distribution over an underlying set of topic probabilities. A topic in LDA has probabilities of generating various words with a distribution assumed to have a Dirichlet prior. A lexical word may occur in several topics with a different probability. The features of this model make it a good candidate to be used for topic modeling in an open domain problem. Some recent models also incorporate more local word dependencies such as Bigrams to capture sequential consecutive dependencies between words or grammatical regularities to detect the syntactically relevant topics. What is required in our project is sentence-level topic modeling. Two approaches to model topics at the level of sentence are using dependency relations and hidden markov model. These two approaches treat sentences as entities. While one uses dependency relations as features to model topic coherence, the other uses Hidden Markov Model and EM-like Viterbi algorithm to model the topic drifts.

[Ponte and Croft, 1997] investigate the problem of text segmentation by topic. What relates this problem to the current work is the focus of their study which is on data with relatively small segment sizes and for which within-segment sentences have relatively few words in common. The authors present a segmentation method where a query expansion technique is used to find common features for the topic segments. In order to score the segmentation generated by the algorithm, they first perform a least squares alignment with the correct segmentation. Then the distance between the two is measured in terms of insertions, deletions and moves.

2.1.3 Query Formulation/expansion

In order to enhance the recall of the sentence retrieval component, the input set of keywords is usually expanded by taking the inflectional and derivational morphology of the terms into account. In short, the query is expanded by the lemma and synonyms of verbs and nouns which it contains.

[Yi and Allan, 2009] discuss different ways of using topic modeling techniques in information retrieval. In order to include topic modeling they calculate a query related topic by employing topic models and use it for query expansion. They propose their approach analogously to the work by [Lavrenko and Croft, 2001]. Having a topic model trained (e.g. using LDA), they propose a probability formula to investigate whether word w can be used for query expansion:

$$P_{TM}(w|q) = \sum_{t_i} P_{TM}(w|t_i) * P_{TM}(t_i|q) \tag{2.1}$$

[Zhai and Lafferty, 2001] propose a feedback strategy based on language models and they use it as an extension of the language modeling approach. They use the feedback documents to update the language model based on the extra evidence carried by the feedback documents.

2.1.4 Relevant Content Retrieval

New technologies to search and retrieve information from the Web efficiently and effectively have enabled us to realize the potential of fetching relevant content for dialogue contribution. In general, we can identify different branches of QA systems depending on the knowledge base they comprise. However there are several well-known approaches using web-based search engines [Kaisser, 2008] [Adafre and Van Genabith, 2009], such as Google or Yahoo or interlink static knowledge bases with web crawlers for document retrieval and answer candidate ranking, such as the START system [Katz et al., 2002]. In [Figueroa and Neumann, 2008], the authors employ Genetic algorithms to extract exact answers from high-ranked web snippets by calculating the similarity of substrings in the snippets and contexts of already known answers.

The method proposed by [Buscaldi and Rosso, 2006] uses Wikipedia category information in order to determine a set of question-related articles within the Wikipedia collection. Their system is partly similar to the one presented by [Waltinger et al., 2011] in terms of using category information as a reference point to improve the answer retrieval. However, it differs in that the first uses string comparison for category selection, but the latter employs a Wikipedia-based topic model involving taxonomy traversal.

In [Sethy et al.] an iterative web crawling approach is described which utilizes a competitive set of adaptive language models comprised of a generic topic independent background language model. Given an initial set of example utterances, this system builds language models for a specific domain in a rather fast manner. The authors address the case where an initial representative set of documents for the domain of interest is available. According to them when the initial set of documents is too small to build a robust language model (e.g. a few utterances from the previous dialogue), then the web data plays a more important role. In addition to the initial topic model, they assume the existence of a generic topic independent language model and the corresponding documents on which it was built. The two language models, one topic dependent and the other topic independent (referred to as the background model) are used to generate search queries using a relative entropy measure.

[Magarreiro et al.] use movie subtitles as a source of dialogue content. The authors explore the possibility of using interactions between humans to obtain appropriate responses to Out-of-Domain (OOD) interactions, taking into consideration several measures, including lexical similarities between the given interaction and the responses. They consider movie subtitles as sequences of turns uttered between humans and use a corpus made of interactions obtained from movie subtitles to be used in OOD interactions.

[Banchs and Li, 2012] is a system demonstration paper which presents IRIS (Informal Response Interactive System), a chat-oriented dialogue system based on the vector space model framework. IRIS is an example-based dialogue system created to provide a means for participating in a game, or just for chitchat or entertainment. Its dialogue strategy is supported by a large database of dialogues that is used to provide candidate responses to a given user input. The dialogue database is first turned into a vector space model representation, in which each utterance is represented by a vector. The search for candidate responses is then performed by computing the cosine similarity metric into the vector space model representation.

2.1.5 Ranking

[Sahami and Heilman, 2006] introduce a method for measuring the similarity between short text snippets (even those without any overlapping terms) by leveraging web search results. In this method, the web search results provide a broader context for the short texts. The similarity function they present is based on query expansion techniques. Even though the traditional goal of query expansion has been to improve recall, they focus on using such expansions to provide a richer representation for a short text in order to make it potentially comparable to other short texts.

In Latent Dirichlet Allocation the topic probabilities provide an explicit representation of a document[Blei et al., 2003]. In the current work, we also implement a ranking method based on cosine similarity of these representations as a measurement method for the similarity of any arbitrary pair of texts.

2.2 Evaluation Measures

A dialogue system can be evaluated in various styles. The evaluation approach can be either subjective or objective. Evaluation metrics can be derived from questionnaires or log files [Paek, 2001]. The scale of the metrics can vary from the utterance level to the whole dialogue. The dialogue system can be treated as a “black box” or as a “transparent box”. This variety of styles beside lack of any agreed-upon standards in the research community and incompatibility of evaluation methods make evaluation of dialogue systems a challenge.

In case of objective evaluation, metrics like resources used (e.g. time, turns, user attention, etc) or the number of errors the system makes or inappropriate utterances made by the system can be mentioned. Specified definitions of task success in some cases is used as an objective metric. However, it is not always easy to define task success in an objective way.

The subjective measurement of the acceptance of an application or technology belongs to the group of usability evaluation. Usability evaluation focuses on users and the users' needs. Usability evaluation wants to know if a system can be used for the specific purpose from the user's point of view, and if it satisfies their expectations. The most important criterion for measuring usability is the user satisfaction. In contrast to objective evaluation techniques which are fairly well-established, subjective measurements are not as structured and straightforward. A difficulty which arises here is that dialogue systems and their users have different and sometimes inconsistent needs. The goals of the system designers do not necessarily always match with the ones of the users. Moreover, users can have multiple goals and their goals can change during a dialogue.[Danieli and Gerbino, 1995]

Information about user satisfaction is gathered through interviews and questionnaires. The type of application determines the aspects that are important for a usability evaluation. Test parameters usually cover aspects such as the efficiency in reaching a goal, and the effectiveness of single system characteristics. This is the case in the ISO standard 9241/10, for example. This standard is intended to calculate the usability summing-up values for effectiveness (percentage chance of achieving a goal), efficiency, and user satisfaction. Efficiency parameters include time required to reach a goal, the error rate and the amount of effort needed to achieve a goal. Another commonly used usability test is SUMI (Software Usability Measurement Inventory), the industry usability evaluation standard for analyzing users' opinions towards software products. SUMI covers most of the principles described in the ISO standard but focuses on the dimensions of efficiency, effect, helpfulness, control, and learnability. [Klüwer, 2015]

[Hone and Graham, 2000] introduce a methodology for the generation and analysis of questionnaires called SASSI. They state that the previously reported subjective techniques are unproven. They argue that their content and structure are, for the most part, arbitrary and the items chosen for a questionnaire or rating scales are based neither on theory nor on well-conducted empirical research. The reasons for choosing a particular structure in the previous studies (e.g. questions, statements or numerical scales) and sub-structure (presentation, number of points on a scale, etc.) are not reported. Therefore, they use factor analysis to determine the main components of user's attitude and also define suitable rating scales for each of these components. Resultant factors after labelling are:

1. **Response accuracy:** the system is accurate/reliable/it makes few errors, the interaction is predictable/efficient.

2. **Likeability:** the system is useful/pleasant/friendly/the user enjoys using the system/it is clear how to speak to the system.

3. **Cognitive demand:** The user feels confident/tense/calm using the system/a high level of concentration is required

4. **Annoyance:** the interaction with the system is repetitive/boring/irritating/frustrating/the system is too inflexible.

5. **Habitability:** The user always knows what to say/The user is not always sure what the system was doing.

6. **Speed:** the interaction with the system is fast/the system responds too slowly

A popular framework for evaluating dialogue systems usability is PARADISE. The basis for system usability in PARADISE is task success and task effort. [Walker et al., 1997]. PARADISE determines the contribution of various factors to a system's performance. Following the idea that one metric does not suffice, the authors form a linear combination of objective metrics such as efficiency measures and task success reflecting both task success and dialogue costs. PARADISE framework provides a representation that separates the domain task from the dialogue which to some extent facilitates comparison of dialogue systems in different domain tasks. However, [Kamm et al., 1999] questions the generalizability of PARADISE and reports noticeable problems with regard to comparability of different models when using PARADISE for evaluation.

Kluwer employs a subjective evaluation method based on user satisfaction. She embedded the developed component in a test bed application and measured the difference in the usability of that application.[Klüwer, 2015]

In the context of the current work, one of the measures to evaluate the quality of the output is by using a seemingly objective measurement method. An off-activity talk is useful to the degree of its *relevance* to the context. However, few attempts have been made to quantify this *relevance* (it will be discussed in section 2.3 that *relevance* is a subjective concept, even though at first it might seem not). [Iyer, 1998] used three measures of relevance (although she seems to prefer using the notion of similarity) , one for content relevance (taking into account word distributions in the compared corpora), one for style relevance (based on the posterior probability of POS n-grams), and another one, which is a modified version of the latter (POS n-grams are replaced by word n-grams),that attempts to account for both content and style.

TREC uses the following working definition of relevance: "If you were writing a report on the subject of the topic and would use the information contained in the document in the report, then the document is relevant". Only binary judgments (”relevant” or ”not relevant”) are made, and a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document). [Dang et al., 2007]

2.3 Relevance Definition

The notion of relevance plays an extremely important role in Information Retrieval. The high number of publications trying to define the concept of relevance, is evidence to the fact that how difficult it is to define and formally model this seemingly intuitive notion.

Defining *relevance* in a dialogue is a very difficult task for many reasons: *relevance* is a subjective and time-varying concept; the dialogue is heterogeneous and highly dynamic; dialogue partners have different expectations and goals. Grice's maxim of relation [Grice, 1975] implicitly defines *relevance* as a relation between a set of propositions and a discourse-topic but he mentions that formulation of this maxim brings about questions the treatment of which he finds "exceedingly difficult". Berg views *relevance* as "usefulness with regard to the conversational goals"[Berg, 1991]. Van Dijk defines *relevant* information as that information which is worth hearer's attention[Van Dijk, 1979].

There is a major difference between *relevance* and *appropriateness*. In [Leuski et al., 2009] the authors define *appropriateness* of an answer. In their definition "an evasive, misleading, or an honestly-wrong answer" from a character can also be counted as an acceptable output. This definition covers a broader set of possible answers which includes answers that might not be *relevant*. What is targeted in this thesis is the *relevance* and not the *appropriateness* of the OAT.

Relevance assessments are of critical importance to the evaluation of information retrieval systems. The Text REtrieval Conference (TREC) has established an evaluation practice where a binary relevance scale is combined with liberal relevance criteria. However, the low threshold for relevance criteria followed in TREC has been criticized in that it affects the ability to identify and develop IR methods capable of retrieving highly relevant documents. In the study presented in [Al-Maskari et al., 2008] , the authors examine the overlap in agreement between official TREC assessments and the relevance judgments created by new users within an Interactive IR experiment. Results show that 63% of documents judged relevant by their users matched official TREC judgments. One explanation for this agreement could be that TREC topics used in their experiment (and their associated relevance) were clear and lacked ambiguity. This high agreement might indicate that when a retrieval system believes documents are relevant, human subjects are also likely to agree on relevance.

Chapter 3

Background

This chapter describes the most important background knowledge necessary for understanding this thesis. It begins with description of some basic ideas and concepts as the ingredients for understanding this thesis. This chapter also briefly describes Aliz-E project which is used as the test bed for the knowledge and computational tools developed in this thesis. Topic modeling as one of the main ideas behind the developed methods is also introduced in this chapter. The technology stack (tools and techniques) employed in the system architecture will be also presented. Finally, some candidate online resources to be considered for finding relevant content are discussed in more detail.

3.1 Relational Agents

Although some level of trust is important in all human-computer and human-human interactions, trust and engagement are especially crucial in applications in which a change in the user is desired and which require significant cognitive, emotional or motivational effort on the user side [Cassell and Bickmore, 2000]. More trust and engagement of the user can help with achieving a long-term interaction. The long-term concern is of special significance in many applications because in many cases (e.g., health-care, education, entertainment, computer games), it is a necessity that the user keeps using the system voluntarily for a long-term period in order to achieve the main goals of the system. Long-term interaction in these systems is a series of several (as many as possible) non-repetitive encounters between the agent and a given user.

Bickmore is the first in the literature who describes the concept of a *Relational agent*. He defines it as "computational artifacts designed to establish and maintain long-term social-emotional relationships with their users" [Bickmore and Picard, 2005]. That is in opposition to the definition of *traditional conversation systems*. Dehn and colleagues through empirical investigations conclude that the effect of *traditional conversation systems* on engagement of the users with the system and their perception of the agent's intelligence is debatable[Dehn and Van Mulken, 2000]. An example of a *Relational agent* is those to interview patients or clients about their experiences and conditions and provide information and counseling using natural language dialogue. Recent work includes the employment of robots to motivate users to study [Kanda et al., 2004], to manage physical activity [Fasola and Mataric, 2012] or diet [Kidd and Breazeal, 2007]. *Relational agents* are intended to produce a relational response in their users, such as increased liking for or trust in the agent. Although people know these machines are not establishing social relationships, they pleasantly tend to use social talk with them. Several studies have found that human users tend to communicate with such agents in a social way, especially if the agents possess human features such as human voice.

Bickmore and Picard report that after a month of experiments over a *relational agent* (named *Laura*) in comparison to an equivalent task-oriented agent without relation-building features, users became significantly more eager to keep on working with the *relational agent*[Bickmore and Picard, 2005]. In [Bickmore et al., 2010], the same author investigates various aspects of the behavior of an agent in its interaction with humans and describes the effect of each aspect on long-term engagement. The assessment of their work is based on analyzing system usage on a daily basis (e.g. whether the users had a conversation with the agent every day; number of steps the users went through with the agent toward the defined goal;) and also through self-reports by the participants about their engagement and enjoyment with the system. The aim of their assessment process was to get an estimation of the users' interest. Some of the questions the participants answered in [Bickmore et al., 2010] are as follows: "How much do you like Laura?", "How would you characterize your relationship with Laura?" (ranging from "Complete Stranger" to "Close Friend"), and "How much would you like to continue working with Laura?". Bickmore compares *Laura* to a normal task-oriented agent lacking social and relationship-building features in a period of one month. Considering different aspects of the relationship (e.g. social psychology, sociolinguistics, etc), he reports that "the *relational agent* was respected more, liked more, and trusted more, even after four weeks of interaction".[Bickmore and Picard, 2005]

3.2 Small Talk

Various conversational strategies are employed in *relational agents*. In [Bickmore and Cassell, 2001], the authors discuss some conversational strategies that comprise a model of social dialogues with the aim to raise user's trust and solidarity. They discuss *small talk* as a kind of talk which executes their conversational strategies. According to them, *small talk*, while interleaved between *task-bound talks*, indirectly builds trust through the natural progression of a conversation.

Bickmore and Cassell define *small talk* as "any talk in which interpersonal goals are emphasized and task goals are either non-existent or de-emphasized"[Bickmore and Cassell, 2001]. kluewer defines it as a talk "often perceived as unsophisticated chit-chat in which content exchange is irrelevant and negligible"[Klüwer, 2015]. Following this definition, *small talk* (e.g. about the weather, events and objects around, stories about the environment or the virtual character itself) represents the opposite of *task-bound talk* which serves the execution of a particular task. Thus, the range of topics and content is definitely much more unrestricted in *small talk* than in *task-bound talk*.

Small talk is one of the fundamental instruments between humans to form a favorable and appropriate social relationship during their conversations, especially in situations in which people meet for the first time. It is useful to develop the conversation and to avoid pauses. Several authors emphasize the essential role of *small talk* for conversational agents. *Small talk* can be used to ease the situation and to make a user feel more comfortable in a conversation with an agent [Cassell and Bickmore, 2000].

Small talk in literature is also introduced as a way of changing the dimensions of trust in conversation. For example [Bickmore and Cassell, 2001] implements a discourse planner that interleaves task-bound talk and *small talk* during the conversation. The agent decides to use *small talk* whenever certain amount of *closeness* to the user is required (e.g. before asking personal questions). They define different functions for a *small talk* (e.g. transitional function, exploratory function, mutual acknowledgment function, expertise establishment, etc) and explain that each of these functions serve as a conversational strategy and indirectly build trust through the natural progression of a conversation.

Small talk also serves as a provider of the sense of uniqueness. According to [Bickmore and Picard, 2005], sense of uniqueness is also crucial for human-agent relationships. The authors count repetitiveness of conversations as a negative impact factor to motivation of the users in interaction with the agent. As one of the participants of the experiments declared in [Bickmore et al., 2010] puts it, “In the beginning I was extremely motivated to do whatever Laura [the relational agent] asked of me, because I thought that every response was a new response.”

3.3 Off-Activity-Talk

Similar to *small talk*, *off-activity talk* (also called *non-activity talk*) is another technique to employ conversational strategies. Both *small talk* and *off-activity talk* enrich the task-oriented dialogue via opening the structure of the conversation. However, OAT can be differentiated from *small talk* by the topic and the purpose of the talk. OAT has a specific purpose (e.g. knowledge exchange) and is about a specific topic, while a *small talk* is an independent talk without any functional topics. Several studies have found that the purpose of *small talk* is not to negotiate knowledge, but in the management of social situation. On the other hand, *off-activity talk*, as we define it here, is to disclose some information relevant to the dialogue context. Therefore, whenever no divergence from the subject matter of the *task-bound talk* is required, OAT is preferred to *small talk*. Even though OAT is a diversion from the structure of the task-dialogue, it maintains the dialogue topic.

OAT was first defined by [Kruijff-Korbayova et al., 2014] in resemblance to *small talk*s. In their project (specifically in Quiz game part), the purpose of the *activity talk* is knowledge exchange or knowledge probing, while *off-activity talk* is used to break out of the fixed structure of the task-bound dialogue. They use OAT in the form of prerecorded questions around a set of predefined topics (e.g. hobbies, diabetes, eating habits, friends, diary) to encourage the user to talk about those topics in order to elicit information from them (see 3.5). In this thesis, we focus on *off-activity talk* with a similar definition but different application. An OAT in the current work, is a follow-up added information relevant to the context of the previous interaction with the aim to encourage engagement of the user and possibly to promote their trust in knowledgeability and intelligence of the agent, thus no deliberate direct information elicitation is targeted in this thesis.

A suitable OAT, as defined in this thesis, is a verbal reaction which is required to be contextually *relevant* to the content of the previous interaction. It preferably includes a provision of some added-information. This verbal reaction can consist of one or more sentences and is extracted from the freely available online resources. The suggested OAT is not intended to be a social dialogue (small talk) and no information exploitation from the human user is expected, but it is meant to be an informative talk including one or more sentences to be uttered in a single turn by an artificial agent (i.e. a virtual character or a robot). An OAT can include some social features (e.g. personal information sharing, exchange of opinions, consultation and negotiation) but it is not meant to.

3.4 Dialogue Manager

A dialogue manager component in a conversation system takes the major responsibility of controlling the conversation flow and the verbal behavior of the robot during the interaction. It keeps track of the dialogue state and utilizes that to interpret user inputs. Selecting the most suitable next action is also the main responsibility of this component. Dialogue manager takes advantage of the service of the OAT suggestion component we develop in this thesis work. In the first place, dialogue manager decides whether to provide an off-activity talk as the next action and if so, it can choose one of the suggested OATs from the output of OAT suggestion component (i.e. the ranked list of suggestions). Kruijff-Korbayova illustrates the role of dialogue manager(DM) as the central component of the architecture of Aliz-E system[Kruijff-Korbayová et al., 2012].

3.5 Aliz-E Project

The initial motivation for this thesis originates from an EU-funded project called Aliz-E (The Adaptive Strategies for Sustainable Long-Term Social Interaction) which includes development of a *relational agent*. This section is a brief overview of the technology, data and methods used in Aliz-E. More information about Aliz-E can be found on the project website[1].

3.5.1 Introduction to Aliz-E: Aims, Approaches, Structure

Some parts of the research presented in this thesis were investigated alognside the research project Aliz-E, carried out at the Language Technology Department of the German Research Center for Artificial Intelligence (DFKI). Aliz-E (The Adaptive Strategies for Sustainable Long-Term Social

[1] http://aliz-e.org

Interaction) is an EU-funded project which aims at sustaining a long-term interaction between a user (a diabetic child) and a robot.

Kruijf-Korbayova and colleagues present Aliz-E as a project that aims at building a conversational system to enable an interactive robot (Nao robot platform as the infrastructure) to communicate with humans in a long-term, adaptive course in real-world settings[Kruijff-Korbayová et al., 2011]. Belpaeme reports on the approach taken to make progress toward this goal in [Belpaeme et al., 2012]. In [Kruijff-Korbayová et al., 2012], the authors give more details about the embedded conversational system, its architecture, and components.

The project includes three game-like activity scenarios. The setting defined for this thesis is in structure similar to the framework defined for quiz-game activity of ALIZ-E system. In [Kruijff-Korbayova et al., 2014], the author focuses on quiz-game scenario as a knowledge-exchange activity around specific domains. This activity includes questions asked by each side (both the robot and the user) from different domains in addition to provision of evaluation feedback. Beside activity-talk (i.e. conversation pertaining to the activity at hand), the interactions of the latest implementation of Aliz-E system also includes social components like greetings, personal introductions, performance feedbacks and also some indications of familiarity between the robot and the user through references to the previous experiences.

3.5.2 Definition of OAT in Aliz-E

The definition of off-activity-talk in [Kruijff-Korbayova et al., 2014] is in compliance with the one in this thesis. However in contrast to the similarity in definitions, there is a fundamental difference between the purpose of an OAT in this thesis and the one in that work. In the latter work, an OAT is used to encourage the user to talk about specific topics (e.g. hobbies, diabetes, eating habits, friends, diary) in order to elicit information from him/her, while in this thesis an OAT is a follow-up added-information relevant to the context of the previously asked question with the aim to encourage user's engagement and to promote their trust in knowledgeability and intelligence of the agent, thus no deliberate information elicitation is targeted in this thesis.

In [Kruijff-Korbayova et al., 2014], the authors specify a set of predefined topics for OATs. They codify some prerecorded questions around each topic of interest, by means of which the robot tries to prompt the user to disclose his/her ideas and experiences on that topic. Here stands another difference from that work with the current thesis; OATs in [Kruijff-Korbayova et al., 2014] are some hard-coded questions which are supposed to actively motivate the human to disclose information, but current thesis uses online resources and selectively provides relevant added-information. Even though their method raises the user's engagement, offline mechanism of that system causes repetitiveness of dialogues after few sessions or in long interactions. This is in contrast to Bickmore's advice of uniqueness [Bickmore and Picard, 2005]. As a further study, the difference of the characteristics of OATs in this thesis and the recent work can be investigated in order to study the correlation of these characteristics with overall effects of OAT.

In [Kruijff-Korbayova et al., 2014], as well as this thesis, *relevance* of a given OAT of a certain topic to the context ("such as a question with semantically related content") needs to be examined before deciding to express it. However, the authors did not explain how this *relevance* assurance is done in their work.

3.5.3 The outcome of including OATs in Aliz-E

The authors of [Kruijff-Korbayova et al., 2014] report their recent trial on adding the so-called non-activity talk (i.e. off-activity-talk in this thesis) to this quiz-game activity as a novel technique and study its impacts on human-robot relationship. In [Kruijff-Korbayov et al., 2014], the same authors focus on OAT design in their work and some experimental studies on human responses to system-initiated OATs. In their study, OATs are triggered in different points during the activity, each of which stands between question-answer sequences. As they describe, they employ OATs through which they aim to raise user's engagement in the dialog in order either to elicit information from the user about his/her habits and experiences or to teach them health-related concepts implicitly. The authors report the effects of including off-activity-talks in the long-term social interactions of a *relational agent* with the users (children of age 11-14). They investigate the attitudes of the user from three aspects:

1. They measure perception of the users through two questionnaires. First, a questionnaire for self-assessment of the users' engagement, their relationship to the robot, and their opinions about the interaction with the robot. Second, questionnaire is a multiple-adjective choice to describe their perception of the characteristics of the agent (e.g., beautiful, funny, tender, intrusive, boring, ...). In their analysis, results do not show any difference in the users' perception of the robot and the relationship. However they note that addition of system-initiated OATs did not cause the agent to be perceived as intrusive or too curious even though the OATs where questions about their personal habits and experiences.
2. They also measured user's interest in having other sessions of interaction through their intention expressions. This study confirms a *significant increase* in the likelihood of interested users.
3. Adherence of the user to writing a diary on a related topic is also measured. According to their study, addition of OAT feature did not have a significant effect on this measure.

Overall, they mention that inclusion of OATs resulted in an increase in users' interest to have further interactions with the conversational system and conclude that it is worth to include off-activity talk in *relational agents*. Beside engagement, they explain that when the robot asks more personal questions focused on the child, he/she becomes surprised and receive more "humanization" perception which proceeds to an emotionally warmer interaction.

It is worth noting that current thesis does not take account of specific considerations of Aliz-E on the range of the users' age and the result is not supposed to be bound to health-related topics. This makes the output of the current work applicable in a broader range of applications. That is why, we have a test-set for evaluation of our work in two parts: both free-domain queries and health-related queries.

3.6 Topic Modeling

Topic Modeling, as a branch of text mining, is the process of identifying patterns in the text in order to classify words into groups called ”topics”. A topic is defined as a probability distribution over the terms in the vocabulary. In this process, topics are assigned to documents and terms are assigned to topics each with specific probability distributions. There are different methods to achieve this goal: LDA, HDP, NNMF, etc. In this thesis, we take advantage of LDA method for topic modeling which will be explained in the coming subsection.

3.6.1 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation [Blei et al., 2003] is a generative probabilistic model which can be used to model text corpora. LDA and its derived models treat a document as a bag-of-words where word order and other important linguistic structures are neglected. Each document is modeled as a finite mixture over an underlying set of various topics and each topic is modeled as a probability distribution over an underlying set of topic probabilities. A topic in LDA has probabilities of generating various words with a distribution assumed to have a Dirichlet prior. A lexical word may occur in several topics with a different probability. The features of this model make it a good candidate to be used for topic modeling in an open domain dialogue contribution problem as the one discussed in the current thesis.

The two hidden variables to be estimated in the generative process of LDA are:

- $P_{TM}(w|t_i)$: The probability of term w being generated by topic t_i.
- $P_{TM}(t_i|q)$: The probability of topic t_i generating document q.

In the context of this thesis, training a topic model over a corpus is done by calculating $P_{TM}(w|t_i)$ for each given term in the vocabulary and each topic. Note that a topic is defined as a probability distribution over all the terms of the vocabulary. Usually the most likely terms of this distribution can be used as a representation of a topic. As an example, the following list is a sample output of training a topic model with 6 topics:

1. data, analysis, information, software, analyses, based, time, experiments, interface, experimental
2. patients, hospital, acute, bleeding, surgical, emergency, hours, perforation, failure, variceal
3. company, service, aol, time, warner, internet, media, companies, cable, advertising
4. teen, young, age, high, years, girls, younger, music, older, ago
5. dance, ballet, dancers, company, dancing, american, performance, choreography, titles, dancer
6. technology, computer, wireless, phone, computers, mail, apple, data, electronics, design

Note that each topic is a probability distribution, so to each term in a topic there should be a probability value assigned ($P_{TM}(w|t_i)$). This distribution is defined over all the terms of the vocabulary, but to keep it short, only the ten highest likely terms of each topic is listed in the example above.

Topic inference is the calculation of $P_{TM}(t_i|q)$ for each pre-trained topic and the given test document. Therefore, the inference process results in a probability distribution over all the predefined topics.

Some recent models also incorporate more local word dependencies such as bigrams to capture sequential consecutive dependencies between words or grammatical regularities to detect the syntactically relevant topics. What is required in our project is sentence-level topic modeling, in contrast to document-level topic modeling. Two approaches to model topics at the level of sentence are using dependency relations and hidden markov model. These two approaches treat sentences as entities. While one uses dependency relations as features to model topic coherence, the other uses Hidden Markov Model and EM-like Viterbi algorithm to model the topic drifts.

3.6.2 Measuring the confidence of a topic inference process

The result of a topic inference process is not always of the same quality. Given an inference of a text over a specific topic model, the amount of information that the inference provides is not always identical.

Let's consider the values of an exemplary topic inference in table 3.1. The results of an inference over three dialogues are computed for a topic model consisting of 5 topics. The probability distribution for dialogue 1 is roughly uniform (i.e. the probability values are almost equiprobable). This means that the inference does not provide much information regarding which topic is more probable for the given dialogue. This happens when the topic model is trained over a corpus that does not include texts around similar topics to the ones of the dialogue. For this extreme case, a good measure should return a value almost equal to zero. Another extreme case is the inference of dialogue 2. In this case, the process results in a decision on the topic of the dialogue with a high level of confidence. A good measurement of the confidence for this inference should return a value almost equal to 1.

Table 3.1: An example for topic inference quality

$P(t_i\|d)$	**topic 1**	**topic 2**	**topic 3**	**topic 4**	**topic 5**	**Sum**	**formula 3.1**	**formula 3.2**
dialogue 1	0.21	0.20	0.20	0.20	0.19	1.0	0.0042	0
dialogue 2	0.9999	0.000025	0.000025	0.000025	0.000025	1.0	0.9997	1
dialogue 3	0.55	0.25	0.10	0.05	0.05	1.0	0.27	0.25
dialogue 4	0.90	0.049	0.030	0.020	0.001	1.0	0.80	0.72

For this measurement the author suggests two formulas. The first one which only cares about the maximum and minimum probability values and neglects median ones is as below:

$$(\max_i P_i - \min_i P_i) * \max_i P_i \tag{3.1}$$

where P_i is a short form for $P(t_i|d)$. This formula does not vary by the probability of secondary topics. This makes it simple but neglects quality of the inference of those topics. So, if all the probability distribution is taken into account in a process, formula 3.1 cannot be a sound representation for its confidence.

The second formula which is defined with regard to the definition of *entropy* in *information theory* is as below:

$$1 - H/\log_2 T \tag{3.2}$$

where T is the number of topics in the topic model (e.g. 5 in the example) and H is *entropy* as defined in *information theory* $(-\Sigma_i P_i \log_2 P_i)$. *Entropy* is a measure for unpredictability of information content. It can also be interpreted as the number of bits required to store a variable. This variable in our problem is the topic to which the dialogue belongs. One needs $\log_2 T$ bits to store variable T. If the probability values are equiprobable, the entropy is equal to the number of bits. So, $\log_2 T$ is the maximum value of *entropy* for a probability distribution with T possible values. Dividing *entropy* by its maximum possible value, we assure that it varies between 0 to 1. Since *entropy* represents uncertainty and we need to measure confidence, the result is subtracted from 1.

Both measuring methods have relatively similar results. However, formula 3.1 is independent of median values while formula 3.2 takes them into account and slightly varies as the median values vary. Therefore, depending on how the result of inference is used, one or another formula might be more suitable. If only the representation of the first topic is taken into account (only first topic strategy; see 4.2.3), then formula 3.1 is more robust and straightforward. However, if all topics are included in the process (multiplication strategy; see 4.2.3), formula 3.2 is more precise and reasonable to consider.

3.7 Tools and External Libraries

In this section we present a brief introduction to the APIs, tools and online resources employed in our system and we describe the role of each in the system.

3.7.1 Stanford CoreNLP Natural Language Processing Toolkit

Stanford CoreNLP is a widely-used Java framework, which provides some of the common core natural language processing(NLP) tools, from tokenization through to coreference resolution[Manning et al., 2014]. The structure of these tools is formed as a pipeline of different annotators such as tokenizer, xml cleaner, sentence splitter, true case determiner, POS tagger, lemmatizer, named entity recognizer, syntactic parser, semantic analyzer, and coreference resolution, etc.

We use this tool as an online processing pipeline over the query strings. We also apply the functions this tool provides in different occasions in order to find the most suitable sentences among the ones fetched from the online resources for dialogue contribution.

Depending on the task domain, questions, imperative sentences and sentences including co-referent mentions (i.e. it,this,...) may not be appropriate for dialogue contribution. Therefore, we should be able to detect and prune them in case. For this purpose, we used the Stanford NLP tools to detect questions, imperative sentences and unresolved mentions. Since this is an intensive process *Parallel programming* paradigms are applied in the implementation of our system in order to boost this process as much as possible. (see 4.2.5)

3.7.2 Apache Lucene

Lucene is an open-source, scalable, high performance text search engine written entirely in Java programming language. It is an open source project freely available under the Apache license. It allows users to easily integrate search abilities to their application. It is suitable to tackle many common search problems. In Lucene, the documents are indexed using the Vector-space model and the search can be performed using several query types such as phrase, wildcard, proximity and range matching queries. Other features include scalable and high-performance indexing, accurate and efficient search algorithms, cross-platform solutions, the possibility to sort the documents based on the fields, to search in multiple indexes simultaneously, to apply stemmers in the document processing phase, to create custom rank functions, among others[Goetz, 2000][McCandless et al., 2010].

Our choice of this tool is guided by following benefits in its employment: wide variety of its features, ease of use, rapid implementation, flexibility of Lucene, and the fact that it has been widely used in the literature. Lucene is employed in the ranking component of the architecture of our system. We use it to rank candidate OATs by calculating the similarity of each OAT to the dialogue context.

3.7.3 MALLET

In order to employ topic modeling techniques in this research, among many topic modeling programs and libraries available, we use MALLET natural language processing toolkit which is a library in Java programming language. MALLET is a package for statistical natural language processing and some other machine learning applications to text. The MALLET topic modeling toolkit contains an implementation of Latent Dirichlet Allocation [Blei et al., 2003]. MALLET is an efficient and threaded piece of software which supports multicore processors. More details on topic modeling and LDA is explained in 3.6. Section 4.1.7 also illustrates how we take advantage of topic modeling techniques and how these techniques are embedded in the architecture of our system.

As an alternative tool for topic modeling, *Gensim* [2] can be employed. *Gensim* is developed in Python and includes implementations of variational methods, such as the online variational Bayes inference[Hoffman et al., 2010]. Online variational Bayes (VB) algorithm for Latent Dirichlet Allocation (LDA) is based on ”online stochastic optimization with a natural gradient step” guaranteed to converge. Roughly speaking, it is an approximate solution which is less accurate than sampling. Sampling implemented in MALLET is a more accurate fitting method.

3.7.4 WordNet

WordNet is a large lexical-semantic database for English developed at the Laboratory of Cognitive Science in Princeton [Miller, 1995]. Words are grouped according to their semantics each of which is called a synset. Every synset is a representation of a concept. Several semantic relations are provided between synsets: holonymy-meronymy, hypernymy, hyponymy and synonyms-antonyms. “Formally, WordNet is a semantic network, an acyclic graph”[Fellbaum, 1998]. In version 3.1 of WordNet there are 117,659 synsets. In our system WordNet is utilized to perform query expansion.

3.8 Candidate Online Resources for Dialogue Contribution

One of the very first questions to be answered in this thesis is where to find relevant content that can be used for dialogue contribution. What can be a good source for relevant sentences? Which online resources are available for this purpose? This section is to explain candidate sources and some discussion about pros and cons of each.

3.8.1 Snippets from web search engines

In order to retrieve relevant content, web search engines are commonly employed. The purpose of a web search engine is to retrieve, from documents available on the web (Blogs, News, Books sources, etc), the documents assumed to be *relevant* to a user query.

In order to investigate appropriateness of these tools to our needs, we tried to employ some of them. However, at the time of writing this thesis, it was not easy to find an accessible service. The Google Web Search API is officially deprecated as of November 1, 2010 . Similar obstacles are there for some other options: Google Web Search API (deprecated), Yahoo Boss (commercial), Blekko API (discontinued). A candidate alternative could be a free web search API called FAROO[3] but according to comparative observations with identical queries on several search engines, the number of results FAROO returns is way fewer than aforementioned search engines, almost all of which are not suitable to the requirements in this thesis. It seems that

[2] http://radimrehurek.com/gensim/

[3] http://www.faroo.com/hp/api/api.html

the number of web-pages indexed in this engine is to a great extent fewer than some of the aforementioned well-known engines.

There is an exception to what mentioned before about web search services. Even though Bing Web Search engine is a commercial service, Microsoft provides an interface for *.Net* programmers facilitating 5000 queries per month for free. The author managed to find an open source library[4] in order to integrate their service in Java framework. Therefore, Bing is the engine we chose for our information retrieval needs.

Although search engines have now become very accurate with respect to navigational queries, for informational queries the situation is less clear, especially for ambiguous queries. New generations of search engines try to address this issue by offering various *post-search* tools that help the user with handling large sets of retrieved results [Ferragina and Gulli, 2008]. One of these tools is called Query-biased Web-snippet and usually contains the URL, the title, links to live and cached versions of the document and sometimes an indication of file size and type. The fragment of the result page that summarizes the context of the searched keywords in that page is also presented as part of the query results in the form of snippets. Snippets give the searcher a sneak preview of the document contents. Provision of accurate snippets as part of the search results may considerably increase the value of the result to searchers by allowing them to make good decisions about which results are worth accessing and which can be skipped. [Cucerzan and Richardson, 2009] is a patent by Google on systems and methods that enable search engines to present relevant snippets.

As mentioned in chapter 1, for many reasons defining *relevance* is not an easy task. Beside that, the notion of *relevance* as defined in information retrieval tasks does not necessarily fully match to the notion of *relevance* and *appropriateness* in dialogue contribution tasks. Moreover, snippets from web search engines are to a great extent biased to the query. Despite these facts, using web-snippets as an online source of OATs was the first idea toward the goal of this thesis.

The first observation was done by querying the search engine with the whole question and different combinations of the answers (i.e. all choices, just correct answer, only the choice of the user). The top-ten snippets drawn on-the-fly from the results returned by Google, Blekko and Bing search engines is considered as the source of OATs. According to the observations, occasionally some sentences can be found in this set which have the potential to be used for our need. That is why we took this idea to define a baseline for the goal of this thesis. However, these sentences are not evenly distributed across the search results and without syntactic and/or semantic filtering, one cannot expect high levels of precision in the relevance or appropriateness of the results for dialogue contribution.

Another minor observation with snippets is as follows: as the number of keywords increases, the search engine tries to concatenate chunks of several sentences each including different keywords to form a single snippet. Each chunk is a sequence of terms extracted from the document with no guarantee to be a complete sentence. This leads to less chance of finding grammatical sentences in a snippet and lower number of items in the set of useful sentences. For this reason, we include

[4] https://code.google.com/p/azure-bing-search-java/downloads/detail?name=azure-bing-search-java-0.12.0.jar

the process of OAT-selection in our developed baseline (see 4.4). The OAT selection process is explained in 4.2.4.

3.8.2 A minor observation with Yahoo! Answers

There is also a basic idea of taking advantage of Yahoo! Answers service as a source of relevant and appropriate OATs for dialogue contribution. We tried to ask some of the questions in Yahoo! Answers and provided also the multiple choices in our entry. We asked the users of Yahoo! Answers service to give the right answer in addition to an explanation or comment. According to what we observed, with a correct categorization of the question, there is a chance to find a suitable answer and since they are generated by real human users, there is some levels of confidence that the output is rational and understandable.

However, to assure certain levels of accuracy we need the ranks by the users, which takes time. This makes the solution not adequate for real-time applications. That is why we preferred the idea of using snippets as a fair and comparable baseline for our task. However, one might try to use the highly ranked answers for each question in an offline mode with no real-time requirements to annotate dialogue knowledge-bases.

Another strategy could be using the original item in addition to the topics found and looking up Yahoo! Answers website. If a relevant question with an answer marked best by the website users is found, then the original question and the answer can be added to the agent database. This way we can be hopeful to find a question-answer pair which is relevant to the last question asked in the quiz. However in this thesis we are not expected to add new question to the database.

3.8.3 iGNSSMM

Third option investigated as a candidate online source of OATs was the output of a system developed for topic graph extraction from web content by Neumann and Schmeier[Neumann and Schmeier, 2011]. iGNSSMM is a system developed for topic graph extraction from web content with a web search engine as the back-end by Neumann and Schmeier[Neumann and Schmeier, 2011]. Basically, iGNSSMM performs topic-driven search/exploration through online resources.

The author tried two different versions of this system with different search engines as the back-end. (i.e. BLEKKO, Microsoft Bing). Basically, iGNSSMM performs topic-driven search/exploration. So, a query should semantically be a topic, e.g. entities. A topic in their work is textually a nominal phrase, and iGNSSMM tries to find its collocations with other topics. To query this system, we needed a mechanism to extract such topic keywords from the question-answer pairs in the dialogue interactions. For this purpose, in the latest version Stanford POS tagger was employed to extract words with specific part of speech tags (Query Processing Component; see 4.1.1).

Given a limited list of keywords, the available online interface of iGNSSMM[5] could return a list of extracted topics considered to be the topics that the user wants to know and learn about. A set

[5] http://www.dfki.de/gnssmm/GNSSMM/indexplus.php?query=SAMPLE

of snippets collected from a standard web search engine is also included in the returned output. For a given query as the input, iGNSSMM collects these snippets from the search engine and tries to compute a topic graph. We decided against taking this system as the baseline; firstly because the actual goal of this tool is to suggest relevant topics (nominal phrases) and not sentences; and secondly because it does not satisfy our need for a simple and reproducible strategy.

Chapter 4

System Architecture

This chapter describes the architecture of the implemented system. Section 4.1 introduces the architecture of our system and also briefly describes how each of the modules constructing the system works. Section 4.2 gives an overview of the building-blocks which are employed to construct the implementation of the proposed approaches. In section 4.3, we describe the processing pipeline of the four approaches we implemented through our architecture. These four approaches are intended to improve the baseline method defined in section 4.4.

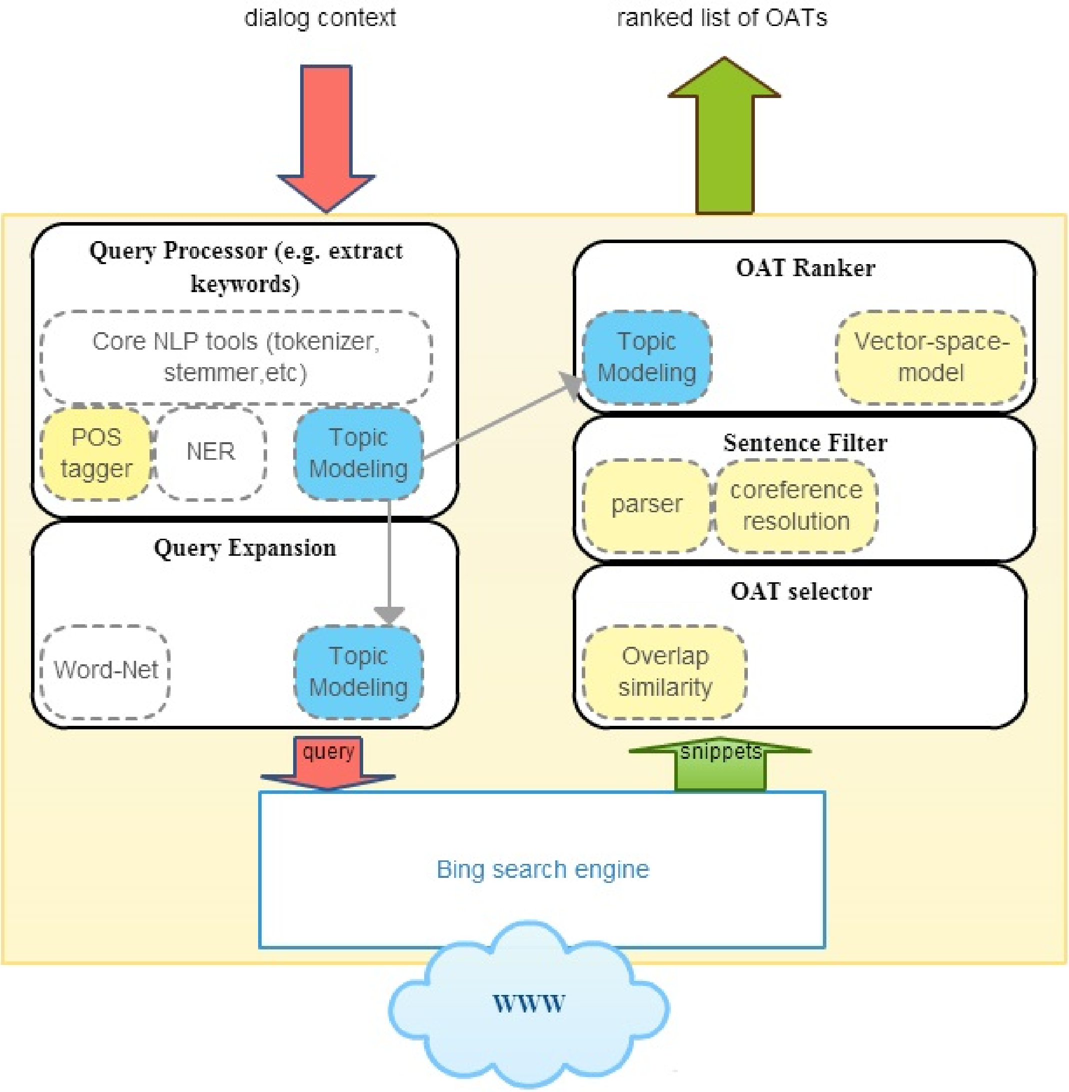
4.1 System Modules

Our tool is formed as a modular architecture that allows exploring different solutions to solve the tasks associated with each module. The modules considered in the current system are explained in this section. Figure 4.1 shows the diagram of the architecture of our system. In the following paragraphs we describe in detail each of the modules utilized in the most important processes.

In our system, the input is the textual representation of a previous sub-dialogue (i.e. one or a few of the utterances contributing to the same dialogue). This input serves as a representation of the dialogue context. Processing this dialogue context, the implemented system is supposed to provide suggestions for sentence(s) which are relevant to the context and appropriate for a dialogue contribution (e.g. including some added-information). The output will be in the form of a ranked list of sentence(s) believed to be suitable off-activity talks (see 3.3). Each item in the ranked list is accompanied by a score.

Concerning the output, we disregard the task of answer paraphrasing or answer generating. We extract the final answer by means of its sentence-based representation. In other words, we only use the top-ranked sentences retrieved from online resources as the output for our conversational agent without any further manipulation.

FIGURE 4.1: System architecture diagram



4.1.1 Query Processing Module

Syntactically, we have to distinguish those parts of the input sentences that represent the most content-bearing concepts (i.e. *focus terms*) to use them as the keywords in later modules. This module detects *focus terms* based on the POS tags and named entities extracted in its NLP sub-modules. Two procedures are considered to construct the queries. In the first one, the query is created based on part of speech tags and in a complementary method, the query is created based on a specified set of named entities. The terms extracted in this modules are to be employed in later modules to construct the queries for the information retrieval engine.

This layer also includes a pipeline of preprocesses each of which applies a sub-process over the query: tokenizing (extracting individual words, ignoring punctuation and case), removing stop words (removal of common words such as articles and prepositions), normalizing, word stemming (reduction of inflected or derivational words to their root form), case folding (capital letters) and lemmatizing.

In order to go beyond surface-based keyword extraction, some topic modeling techniques are also employed in a number of layers of the architecture. Topic modeling modules interact with each other by feeding the input or using the output of one another. The role of the topic modeling sub-module in this layer is to load/train a topic model and use it to detect the topic of the dialogue context. The output of this topic detection process is further used to formulate or expand the keyword set. The functionality of topic modeling modules are explained in more detail in sections 4.1.7 and 4.2.

Table 6.1 shows some examples of how this module works in action. Columns A and B of this table represent the output of the keyword selection process over a number of test queries (see table 5.3 for query set). Column A represents the outcome of applying POS-based procedure and column B represents the results of applying topic modeling techniques.

4.1.2 Query Expansion Module

A possible query expansion can takes place using the WordNet lexical database. In this method, synonyms of the terms contained in the input sentence(s) are extracted from WordNet. (see 3.7.4) However, the engine we used in the final version of the system for relevant online content retrieval (i.e. Bing) includes such query expansion functionalities. This makes the necessity of a word-net sub-module questionable. Therefore we disabled it in the latest version.

Another strategy for query expansion is developed in this module by taking advantage of topic modeling techniques. The role of the topic modeling sub-module in this layer is to use the information stored in the trained topic model and also the information from topic inference process which is done in the last layer to choose some relevant terms to be used for expansion of the current keyword set. The advantage of this expansion technique over a WordNet-based expansion method is that it goes beyond the kind of relations which can be found in a WordNet (e.g. synonyms). This method helps finding terms which are topically related to the dialogue context. For example, if the initial keyword set contains the term “breakfast”, this expansion method returns terms such as: “food”, “cooking”, and “make”. The functionality of topic modeling modules are explained in more detail in section 4.1.7 and 4.2.

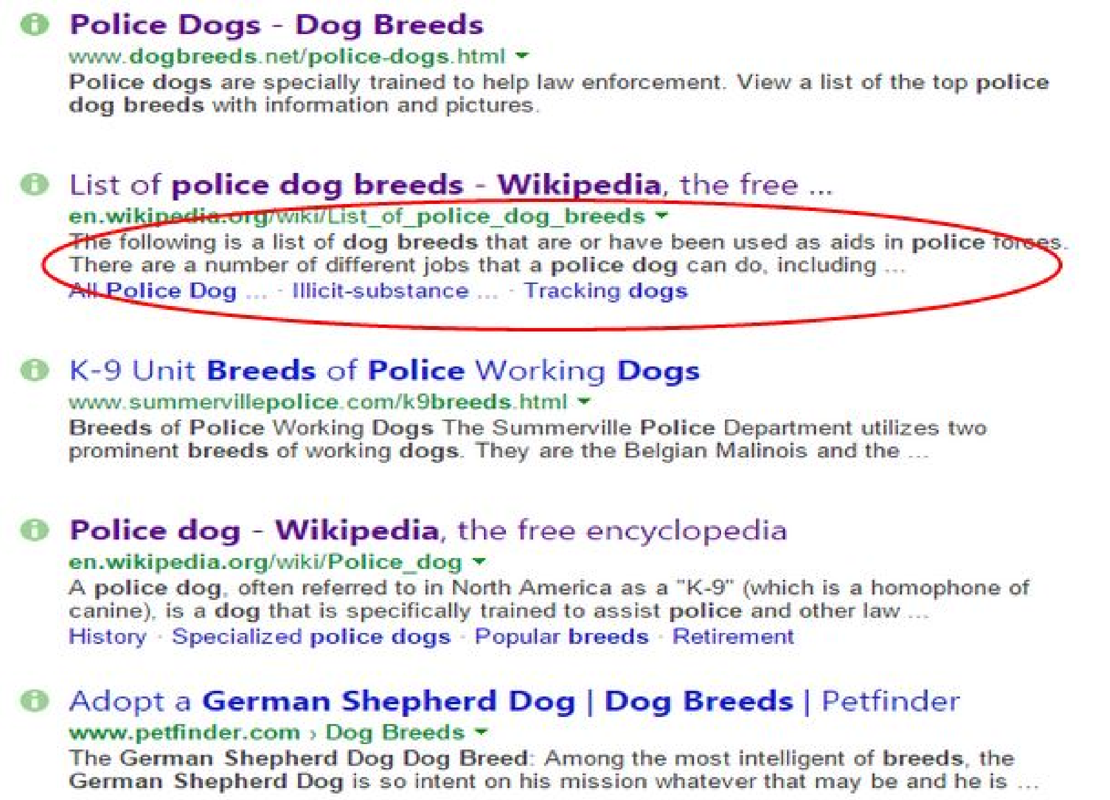
Table 6.1 lists some examples of how query expansion module works in action. Columns C and D of this table represent the output of the keyword expansion process over a number of test queries (see table 5.3 for query set). Both columns C and D represent the outcome of applying topic modeling techniques for query expansion; the first in an offline mode and the latter in an online mode. (see 4.2.1 for online and offline mode of topic modeling techniques)

4.1.3 Content Retrieval

At this stage, we query Microsoft Bing search engine with the keywords extracted in the previous step. Figure 4.2 is a graphical representation of a retrieved list of items from a search engine. The result is a ranked list of items in the form of snippets each including a title and the url address to the web document. The fragment of the result page that summarizes the context of the searched keywords in that page is also presented as part of the query results. A number of top-ranked webpages are fetched from the corresponding web servers.

4.1.4 OAT selector/extractor

FIGURE 4.2: OAT selection example [1]



Police Dogs - Dog Breeds
www.dogbreeds.net/police-dogs.html
Police dogs are specially trained to help law enforcement. View a list of the top police dog breeds with information and pictures.

List of police dog breeds - Wikipedia, the free ...
en.wikipedia.org/wiki/List_of_police_dog_breeds
The following is a list of dog breeds that are or have been used as aids in police forces. There are a number of different jobs that a police dog can do, including ...
All Police Dog ... · Illicit-substance ... · Tracking dogs

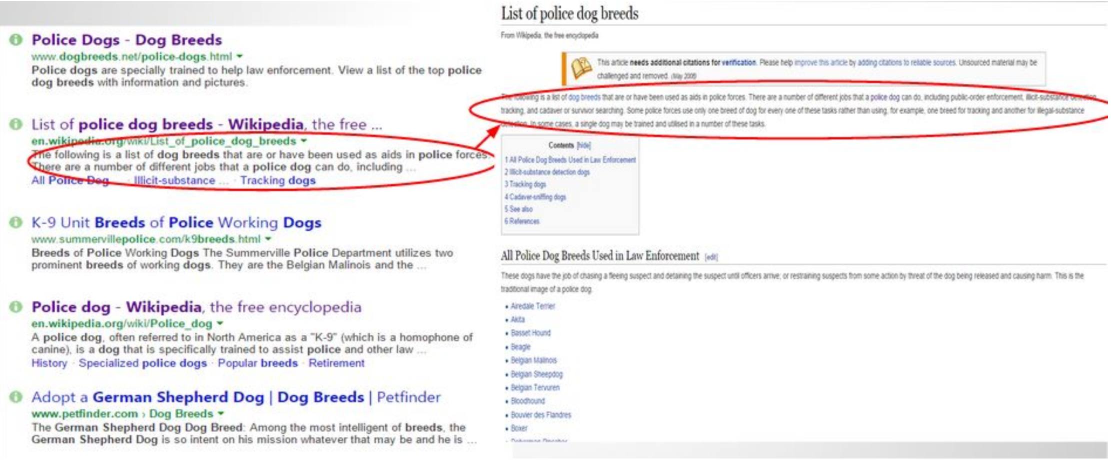
K-9 Unit Breeds of Police Working Dogs
www.summervillepolice.com/k9breeds.html
Breeds of Police Working Dogs The Summerville Police Department utilizes two prominent breeds of working dogs. They are the Belgian Malinois and the ...

Police dog - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Police_dog
A police dog, often referred to in North America as a "K-9" (which is a homophone of canine), is a dog that is specifically trained to assist police and other law ...
History · Specialized police dogs · Popular breeds · Retirement

Adopt a German Shepherd Dog | Dog Breeds | Petfinder
www.petfinder.com › Dog Breeds
The German Shepherd Dog Dog Breed: Among the most intelligent of breeds, the German Shepherd Dog is so intent on his mission whatever that may be and he is ...

Having the documents retrieved from the search engine, we need to extract appropriate parts of these documents to be used as a context-relevant off-activity talk.

FIGURE 4.3: OAT selection example [2]



Police Dogs - Dog Breeds
www.dogbreeds.net/police-dogs.html
Police dogs are specially trained to help law enforcement. View a list of the top police dog breeds with information and pictures.

List of police dog breeds - Wikipedia, the free ...
en.wikipedia.org/wiki/List_of_police_dog_breeds
The following is a list of dog breeds that are or have been used as aids in police forces. There are a number of different jobs that a police dog can do, including ...
All Police Dog ... · Illicit-substance ... · Tracking dogs

K-9 Unit Breeds of Police Working Dogs
www.summervillepolice.com/k9breeds.html
Breeds of Police Working Dogs The Summerville Police Department utilizes two prominent breeds of working dogs. They are the Belgian Malinois and the ...

Police dog - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Police_dog
A police dog, often referred to in North America as a "K-9" (which is a homophone of canine), is a dog that is specifically trained to assist police and other law ...
History · Specialized police dogs · Popular breeds · Retirement

Adopt a German Shepherd Dog | Dog Breeds | Petfinder
www.petfinder.com › Dog Breeds
The German Shepherd Dog Dog Breed: Among the most intelligent of breeds, the German Shepherd Dog is so intent on his mission whatever that may be and he is ...

List of police dog breeds

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. (May 2008)

The following is a list of dog breeds that are or have been used as aids in police forces. There are a number of different jobs that a police dog can do, including public-order enforcement, illicit-substance detection, tracking, and cadaver or survivor searching. Some police forces use only one breed of dog for every one of these tasks rather than using, for example, one breed for tracking and another for illegal-substance detection. In some cases, a single dog may be trained and utilised in a number of these tasks.

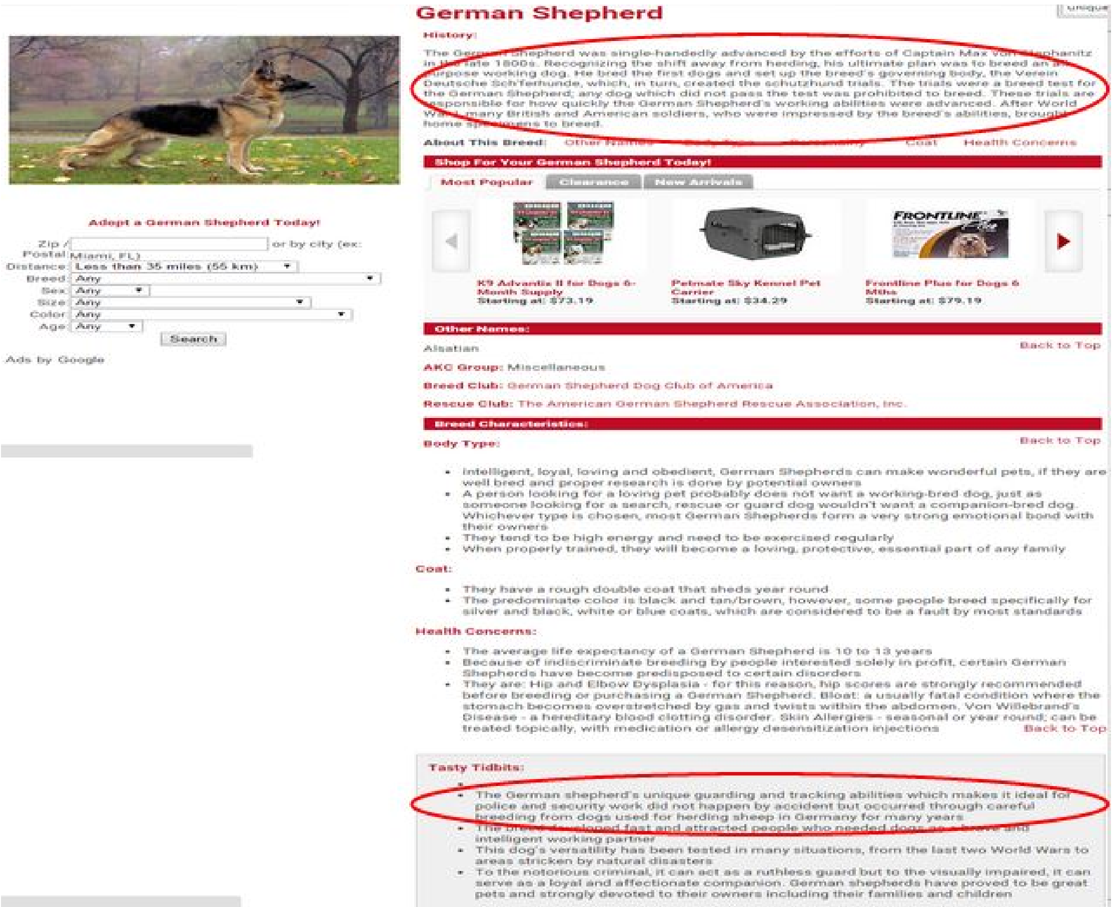
Contents [hide]
1 All Police Dog Breeds Used in Law Enforcement
2 Illicit-substance detection dogs
3 Tracking dogs
4 Cadaver-sniffing dogs
5 See also
6 References

All Police Dog Breeds Used in Law Enforcement [edit]

These dogs have the job of chasing a fleeing suspect and detaining the suspect until officers arrive; or restraining suspects from some action by threat of the dog being released and causing harm. This is the traditional image of a police dog.

- Airedale Terrier
- Akita
- Basset Hound
- Beagle
- Belgian Malinois
- Belgian Sheepdog
- Belgian Tervuren
- Bloodhound
- Bouvier des Flandres
- Boxer

Figure 4.4: OAT selection example [3]



German Shepherd

History:

The German Shepherd was single-handedly advanced by the efforts of Captain Max von Stephanitz in the late 1800s. Recognizing the shift away from herding, his ultimate plan was to breed an all-purpose working dog. He bred the first dogs and set up the breed's governing body, the Verein Deutsche Sch'ferhunde, which, in turn, created the schutzhund trials. The trials were a breed test for the German Shepherd; any dog which did not pass the test was prohibited to breed. These trials are responsible for how quickly the German Shepherd's working abilities were advanced. After World War I, many British and American soldiers, who were impressed by the breed's abilities, brought home specimens to breed.

About This Breed: Other Names Body Type Personality Coat Health Concerns

Shop For Your German Shepherd Today!

Most Popular | Clearance | New Arrivals

K9 Advantix II for Dogs 6-Month Supply
Starting at: $73.19

Petmate Sky Kennel Pet Carrier
Starting at: $34.29

Frontline Plus for Dogs 6 Mths
Starting at: $79.19

Adopt a German Shepherd Today!

Zip / Postal: or by city (ex: Miami, FL)
Distance: Less than 35 miles (55 km)
Breed: Any
Sex: Any
Size: Any
Color: Any
Age: Any
Search

Ads by Google

Other Names:

Alsatian

Back to Top

AKC Group: Miscellaneous

Breed Club: German Shepherd Dog Club of America

Rescue Club: The American German Shepherd Rescue Association, Inc.

Breed Characteristics:

Body Type:

Back to Top

- Intelligent, loyal, loving and obedient, German Shepherds can make wonderful pets, if they are well bred and proper research is done by potential owners
- A person looking for a loving pet probably does not want a working-bred dog, just as someone looking for a search, rescue or guard dog wouldn't want a companion-bred dog. Whichever type is chosen, most German Shepherds form a very strong emotional bond with their owners
- They tend to be high energy and need to be exercised regularly
- When properly trained, they will become a loving, protective, essential part of any family

Coat:

- They have a rough double coat that sheds year round
- The predominate color is black and tan/brown, however, some people breed specifically for silver and black, white or blue coats, which are considered to be a fault by most standards

Health Concerns:

- The average life expectancy of a German Shepherd is 10 to 13 years
- Because of indiscriminate breeding by people interested solely in profit, certain German Shepherds have become predisposed to certain disorders
- They are: Hip and Elbow Dysplasia - for this reason, hip scores are strongly recommended before breeding or purchasing a German Shepherd. Bloat: a usually fatal condition where the stomach becomes overstretched by gas and twists within the abdomen. Von Willebrand's Disease - a hereditary blood clotting disorder. Skin Allergies - seasonal or year round; can be treated topically, with medication or allergy desensitization injections

Back to Top

Tasty Tidbits:

-
- The German shepherd's unique guarding and tracking abilities which makes it ideal for police and security work did not happen by accident but occurred through careful breeding from dogs used for herding sheep in Germany for many years
- The breed developed fast and attracted people who needed dogs as a brave and intelligent working partner
- This dog's versatility has been tested in many situations, from the last two World Wars to areas stricken by natural disasters
- To the notorious criminal, it can act as a ruthless guard but to the visually impaired, it can serve as a loyal and affectionate companion. German shepherds have proved to be great pets and strongly devoted to their owners including their families and children

As mentioned before, each result of the search engine includes description part which is a fragment of the result page that summarizes the context of the searched keywords in that page. In order to provide a sneak preview of a webpage, the search engine usually truncates few of the sentences from the document and appends them together. This is more likely as the number of the keywords in the query increases. At an early stage, we take this textual preview into account, but any truncated sentence in this preview is omitted. It will not be surprising if this omission step results in an empty set of sentences. In order to resolve this issue we go through the document to find the sources of the sentences which are mentioned in the preview.

For this purpose, we first clean the HTML tags from the retrieved web page and then divide the document into several chunks each of which includes one or a few sentences as the candidate OATs. This is done with regard to the structure of the document (e.g. paragraphs, list items, etc).

In the next step, we calculate an overlap similarity measure using Szymkiewicz-Simpson coefficient and compare each chunk with the corresponding snippet retrieved from the search engine (see: 3.8.1 for definition of snippet). Those chunks satisfying a similarity threshold are extracted. This threshold can be tuned as a coefficient for the relaxedness of the string similarity requirement of the chunks. We tried to tune it according to a number of observations we made in a way that on average 3-5 chunks are returned from each web page even though this number can vary greatly from one web page to another. In the current implementation of the system, this threshold is set as 0.5. The similarity measurement is implemented by considering the snippet and the candidate chunk as two sets of keywords. First, the stop words are removed from each set, then the intersection of the two is computed. If size of the intersection set is bigger than half of the size of the snippet set, the chunk will be selected as a candidate chunk of that web page. This relaxed similarity measurement helps us to attain certain level of recall from the web pages that we retrieve from the web search engine. Since this high recall comes at the cost of lower precision in retrieving relevant information, we employ a ranking mechanism in the end of the pipeline to assure of the precision (see 4.1.6).

An example of how this component works is illustrated in figures 4.2, 4.3 and 4.4. The first shows an example of the output of the underlying search engine. Say, the second result is of our focus. The web page addressed in green is fetched from the corresponding web server. It is chunked and then the snippet is compared to each chunk of the web page. Figure 4.3 shows the exemplary chunk which is selected. The figure also illustrates the span and location of such a chunk in the web page. The number of chunks which are extracted from the page can be higher than one. As an example for a web page with more than one relevant snippet extracted, figure 4.4 shows a web page marking two selected chunks with different span sizes and from different parts of the web page. Chunks can come from different HTML tags; as in this example, the first chunk is a paragraph and the second is a list element.

In order to limit the length of a chunk, two threshold coefficients are defined: character-wise and sentence-wise. These coefficients are tuneable concerning the application. In the final implementation of the system which is used for the evaluation, the length limit coefficients are defined as maximum 300 characters for string length and a maximum of 2 sentences. A logical conjunction of the two limits is used for defining of the condition. Any chunk exceeding the length limits goes through a coreference resolution process and then will be split to its constituent sentences. The coreference resolution process helps with avoiding unresolved mentions after splitting the sentences.

4.1.5 Sentence Filtering

From the set of results returned from the OAT selector, sentence filtering component is to remove unsuitable sentences. The grain-size of the inputs to this component is of sentence-level. Incomplete sentences, ungrammatical sentences, imperative sentences, questions and also sentences including unresolved coreferent mentions are considered as inadmissible. Table 4.1 lists some sample sentences filtered by each of these filters. Section 4.2.5 explains how these filters are implemented in our system. Concerning the application of such a system, one might allow or disallow utilization of questions as OATs. For the evaluation of this thesis, questions are allowed.

It is worth mentioning that these components work as in a pipeline. So, the output of OAT selecting component is the input of sentence filtering. While the first process does not manipulate the text in any way and just selects some chunks based on their similarity to the snippet, the later processes chunks sentence by sentence and removes improper sentences.

Table 4.1: Some examples of filtered sentences

Filter	**Sample Filtered Sentence(s)**
Incomplete sentences	”There are a number of jobs, including ...”
Ungrammatical sentences	”German Shepherd Dog Puppies for Sale”
Imperative sentences	”Find great deals on eBay for german shepherd police dog.”
Unresolved references	1. ”**This** article needs additional citations for verification.” 2. ”In some cases, a single dog may be trained and utilised in a number of **these** tasks.” 3. ”**It** was badly damaged during fighting between separatists and Russian troops.”
Questions	”How many meals should you eat a day?”

4.1.6 OAT Ranking Layer

The aim of this component is to rank the OATs according to their suitability to the dialogue context. We define and implement two strategies for ranking. In the first strategy, ranking of candidate responses is performed by computing the cosine similarity metric into a vector space model representation in which each candidate is represented by a vector. The features in this case are term frequencies. *Apache Lucene* libraries were utilized for this purpose. As for the second strategy, we employ topic modeling techniques. In this case, each candidate is represented by a vector of topic inference probability distribution. First, each candidate OAT is processed in a topic modeling inference method over the topic model from the first layer (see 4.1.1). This inference process returns distribution of probabilities for each topic. Similarly, the dialogue context vector is also generated in the second layer when it is processed the same way(see: 4.1.2). Analogous to the first ranking method, for each candidate, OAT cosine similarity between topic modeling vectors of the candidate and the dialogue context is calculated. This value is used to rank the candidates. The more topically similar the candidate is to the dialogue context, the higher it gets in the ranking. The functionality of topic modeling modules are explained in more depth in sections 4.1.7 and 4.2.1.

4.1.7 Topic Modeling Modules

We try to include topic modeling techniques in our tool in order to gain better performance with regard to both *relevance* and *informativity* of the resulting OATs. Figure 4.1 depicts how the implementation of these techniques is included in system architecture. The analysis is done in three sub-components (portrayed in blue). First one is responsible for loading an offline or training an online topic model. This occurs in Query Processor module. Second sub-component is in charge of determining the topic of the textual representation of the previous sub-dialogue (i.e. the input to the system). This part is done in Query Expansion module in order to detect the key concepts of the input and consequently to finalize the query. Third, topic inference of the OAT candidates which occurs in OAT Ranker component in order to compare latent relations of the key concepts in the input (previous sub-dialogue) to the ones in OATs.

For this purpose, we employ the implementation of Latent Dirichlet Allocation in MALLET java-based package(see 3.7.3). Latent Dirichlet Allocation [Blei et al., 2003] is a generative probabilistic model which is prevalently used for topic modeling. LDA treats a document in a bag-of-words scheme. Each document is modeled as a finite mixture over an underlying set of various topics and each topic is modeled as a probability distribution over an underlying set of topic probabilities. A topic in LDA has probabilities of generating various words with a distribution assumed to have a Dirichlet prior. A lexical word may occur in several topics with a different probability. The features of this model make it a good candidate to be used for topic modeling. (for more on LDA see 3.6.1).

In MALLET package, LDA is mainly implemented in a class called *ParallelTopicModel*. The number of topics is a fixed parameter to set before training the topic model. In the current work, this parameter is set to 100 topics in all cases.

The two hidden variables to be estimated in this generative process are:

- $P_{TM}(w|t_i)$: The probability of term w being generated by topic t_i.
- $P_{TM}(t_i|q)$: The probability of topic t_i generating document q.

Using MALLET, we can train a topic model over a corpus by calculating $P_{TM}(w|t_i)$ for each given term in the vocabulary and each topic. Note that a topic is defined as a probability distribution over all the terms of the vocabulary. Usually the most likely terms of this distribution can be used as a representation of a topic. This is the task of the first blue component of the process in figure 4.1.

In order to take advantage of the trained topic model, we follow two mechanisms:

1. Query expansion: to calculate a query-related topic by using topic models and to employ the representation of that topic for query expansion.

Given an input dialogue to the system, topic inference is done in the second blue component as a sub-component of query expansion. This inference process is the calculation of $P_{TM}(t_i|q)$ for each predefined topic and the given dialogue. Therefore, the inference process results in a probability distribution over all the predefined topics.

Having the highest relevant topics to the context, we can choose the top topic and use its most weighted terms for query expansion. Approaches B and C follow this scheme.

Another way to investigate whether word w can be used for query expansion is to calculate a topic model based relevance [Yi and Allan, 2009]:

$$P_{TM}(w|q) = \sum_{t_i} P_{TM}(w|t_i) * P_{TM}(t_i|q) \tag{4.1}$$

Intuitively, this way we rank each topic t_i by its probability of generating the query q, then we use the words in high-ranked topics (not only the first) to calculate a query-specific topic for query expansion. Approach D takes advantage of this scheme.

2. OAT ranking: to represent each utterance by the topics to which they belong and to use these representations for ranking OATs.

Table 4.2 provides an example of two OATs and the dialogue context (i.e. system input) represented by their topic distribution vectors. As it can be seen, we use the topic probability distribution for each candidate OAT ($P_{TM}(t_i|q)$) as a feature vector. We use these vector representations to compare each candidate OAT to the dialogue context. In order to compare the vectors, for each of the OAT vectors, the dot product to the context vector is calculated and the resulting value is used for ranking the candidate OATs. In the example of table 4.2, the dot product of the first OAT to the context is 0.00846 and the dot product of the second OAT to the context is 0.5113 which is interpreted as higher topical similarity of the second OAT to the context.

Table 4.2: An example of representing OATs as topic distribution vectors

P($t_i|\mathbf{D}$)	t_1	t_2	t_3	t_4	t_5
OAT 1	0.001	0.020	0.030	0.249	0.70
OAT 2	0.55	0.25	0.10	0.05	0.05
Context	0.90	0.049	0.030	0.020	0.001

By using different topic models trained over different corpora, we will have a family of topic-based query expansion and OAT ranking methods. Since performance of the above-mentioned techniques is profoundly dependent on the quality of the topic model (e.g. the corpus over which the model is trained), it will not be surprising to observe different levels of performance with different topic models. However, one specific corpus is used to carry out evaluations of offline approaches in this thesis. Because of the variance of the performance of a topic inference process with different inputs and topic models, the author has tried to estimate the quality of the topic-based techniques before they contribute to the result. This quality estimation process is explained in more detail in section 3.6.2. The estimated value is used to assign a weight in order to control the influence of the techniques on the output.

We follow the formula below for smoothed integration of topic modeling techniques with the initial architecture:

$$P'_{TM}(w|q) = \lambda P_{TM}(w|q) + (1-\lambda)P(w|q) \tag{4.2}$$

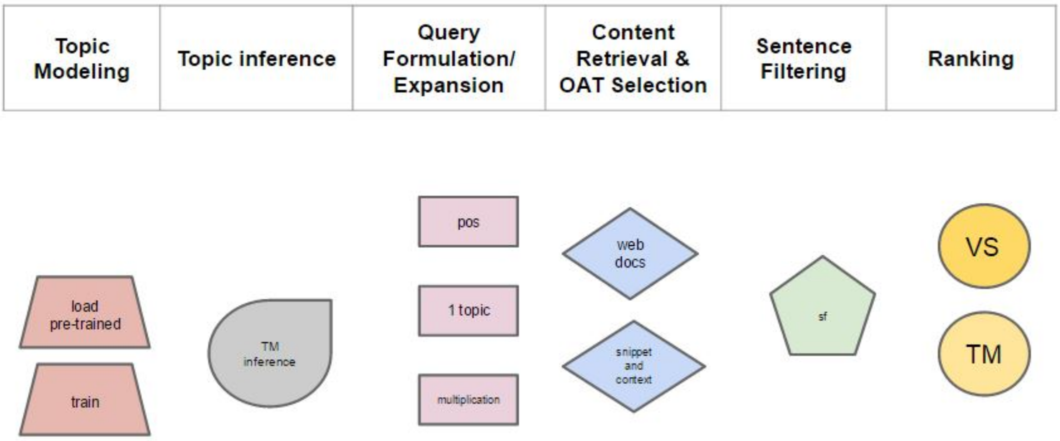
while $P(w|q)$ stands for the original query expansion method and $P'_{TM}(w|q)$ for the integrated method. λ is the smoothing coefficient which can be a fixed value or a function of the quality of topic modeling inference(see 3.6.2). This smoothing mechanism can be analogously applied to topic-based OAT ranking processes.

4.2 Pipeline Building Blocks

Through the architecture explained in the last section, four different approaches are implemented in different processing pipelines. In this section, the sub-processes are explained as the building blocks of the pipelines. Each pipeline is illustrated separately in the next section.

The overall procedure for context-relevant dialogue contribution using online resources can be divided into several sub-processes which will be employed within a processing pipeline. Each of the sub-processes is possibly implemented in different methods. Figure 4.5 is a sketch of these sub-processes and different methods by which each sub-process is implemented. The same chart will be used in the coming sub-sections to illustrate pipelines of the four approaches. Following comes the explanation of each sub-process and the methods to implement each.

FIGURE 4.5: Sub-processes schema



4.2.1 Topic modeling sub-process

Topic modeling sub-process is implemented in two ways:

1. Load a topic model trained in an offline manner on a specific corpus.
2. Train a topic model in an online process over the given documents.

As explained in 4.1.7, training a topic model is the process of calculating $P_{TM}(w|t_i)$ for each given term in the vocabulary and each topic. This sub-process needs to be done over a specified corpus. In one of the approaches(D), this corpus is collected from online sources while in others(B,C), the topic model is trained on an offline corpora as a pre-process and is loaded into memory when the server starts. For the experiments in this thesis, we trained our offline topic model over *New York Times* corpus [Sandhaus, 2008]. In the latter case (B,C), processing time is excessively less than the others, but the topic model is a general one which might not necessarily lead to a confident topic inference. In the other case employed in approach D (see 4.3.4), the process is executed in query-time that makes the process slower but instead the resulting topic model is more specialized for the current domain.

4.2.2 Topic inference sub-process

This sub-process is to infer topics for given input (i.e. dialogue context). Topic inference is implemented in a single method using MALLET toolbox. As explained in 4.1.7, topic inference is the process of calculating $P_{TM}(t_i|q)$ for each topic and the given dialogue context. In other words, for each topic, the probability of that topic belonging to the context is calculated. Having the topic detected this way, the result of this process will be further used to formulate or expand the query.

4.2.3 Query formulation/expansion sub-process

There are three methods for keyword extraction implemented in the system which will be used for query formulation or query expansion in different approaches:

1. **POS based:** Terms tagged with the following set of specified part of speech tags are accepted: "NN", "NNS", "NNP", "NNPS", "VBG", "JJS". *Stanford POS tagger* is used for this purpose.

2. **Only first topic:** Having the highest relevant topics to the dialogue context, the most highly ranked terms of the most likely topic is chosen. The number of terms to pick is in correlation with both topic modeling inference confidence value (see 3.6.2) and $p_{TM}(w|T1)$ while $T1$ is the highest ranked topic. If the task is expansion (in contrast to formulation) which is the case in *approach C* , the size of the initial keyword set is also taken into account.

3. **Multiplication technique:** To investigate whether word w can be used for query expansion, a topic model based relevance value [Yi and Allan, 2009] is calculated as in formula 4.2. Intuitively, this way we rank each topic t_i by its probability of generating the query q, then we use the most weighted words in high ranked topics (not only the first) to calculate relevant terms for query expansion. The number of terms to pick is in correlation with topic modeling inference confidence value (see 3.6.2) and size of the initial keyword set. This method is used in *approach D* for query expansion. The reason for this decision is that, in *approach D* the corpus over which the topic model is trained is of specific domain, therefore the statistically defined topics are closely related, as opposed to the offline general topic model which is used for *approaches B* and *C*. As a consequence, it is rational to take the whole topic model into account to select keywords rather than just the first topic.

4.2.4 Content retrieval and OAT selection sub-process

In order to retrieve relevant content, Microsoft Bing search engine is employed. In figure 4.5 the top blue diamond shape stands for retrieving the content of the top ranked web pages which Bing suggests. This content will be cleaned and normalized to be used as an input to a topic model training in approach D (see 4.3.4).

The bottom diamond shape in the figure stands for using the search snippets returned from Bing in order to select relevant parts of the web page (OAT selection) . For this purpose, each web page is chunked based on its structure. In other words, even though HTML tags will be cleaned from the web page, some specific tags such as $< br >$, $< h >$, $< div >$, $< p >$, etc will be utilized as delimiters for chunking the web page. Then, the string similarity of each chunk to the snippet is computed and those chunks satisfying a similarity threshold are extracted. This extraction procedure is explained with an example in section 4.1.4. This sub-process feeds sentence filtering sub-process.

Content retrieval sub-process is one of the bottle-necks of the system. The reason is that in order to fetch many web pages from different web servers, the system requires to connect to those servers and read streams of data over network from each. *Parallel programming* paradigms are applied to boost this process as much as possible.

4.2.5 Sentence filtering sub-process

Chunks of the web page which are extracted in the last sub-process go through a linguistic process in order to decide which sentences are inappropriate for dialogue contribution. There is a single sentence filtering module implemented which applies several predefined syntactic rules. Incomplete sentences, ungrammatical sentences, imperative sentences, questions and also sentences including unresolved coreferent mentions are filtered out. Sentences including specific erroneous terms (e.g. ”below” , ”following”, ’hyperlink’, etc) are also removed. Mentions (i.e. pronouns) referring to entities in a pruned sentence are replaced by their reference. A limitation on length of an OAT both on character and sentence level is applied as well.

A natural language processing pipeline is used for this purpose which includes processes such as tokenization, sentence splitter, POS tagger, lemmatization, named entity recognition, sentence parser and coreference resolution. *Stanford NLP Toolkits* is employed for this purpose. Section 4.1.5 includes some examples of how this filtering process prunes inappropriate sentences.

This process is relatively computationally expensive. In order to prevent it from drawing back the system performance as a whole, *parallel programming* techniques are used to maximize the utilization of available processing capacity of the hardware which our software is running on. Processor permitting, each sentence is processed in a separate *thread*.

4.2.6 Ranking sub-process

Ranking is implemented as calculation of cosine similarity of feature vectors of the dialogue context and each OAT candidate. Feature vectors are defined in two ways:

1. A vector space model with term frequency values as the features.
2. Topic modeling inference result as the feature space. Feature i is the probability of topic t_i having generated the candidate ($P_{TM}(t_i|q)$).

More about these two ranking methods is explained in section 4.1.6.

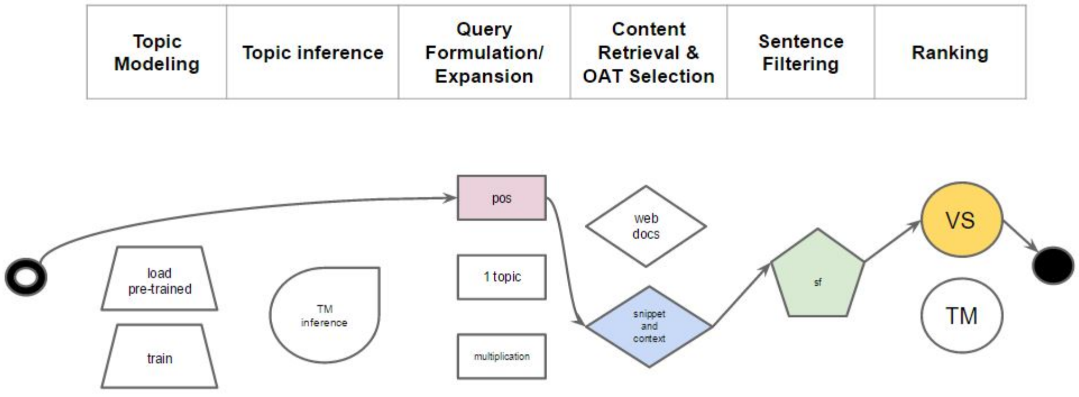
4.3 Four Approaches: Four Processing Pipelines

The four approaches in this thesis are implemented in terms of different processing pipelines. The first approach is solely based on POS tags for query formulation. The second, entirely uses topic modeling to form the query. The third and the fourth are combinations of the two, one in a parallel mode and the other in a sequential mode. Using the building blocks introduced in the last section, we will show how each pipeline works.

4.3.1 Approach A

Approach A solely relies on POS (and partly on NER) for query formulation. So, no topic modeling technique is employed. Figure 4.6 illustrates which building blocks and in what order form this approach.

FIGURE 4.6: *Approach A* pipeline



First, keywords are extracted from the dialogue context based on part of speech tags and NER (4.2.3). Content retrieving, OAT extracting (4.2.4), and sentence filtering (4.2.5) are the subprocesses which respectively execute afterward. Ranking (4.2.6) is the last step in this processing pipeline.

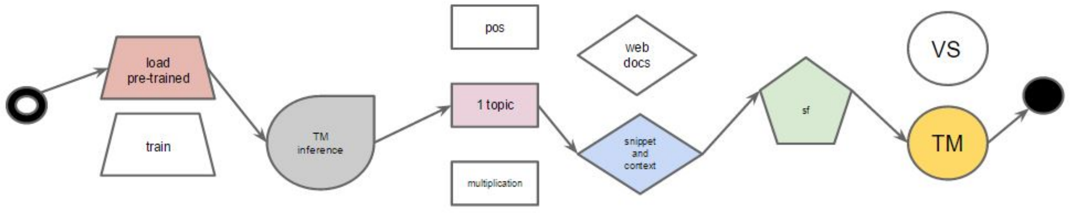
4.3.2 Approach B

Approach B merely relies on topic modeling techniques for query formulation. Figure 4.7 depicts which building blocks (and in what sequence) form this approach.

The first step in this pipeline is to load a pre-trained topic model (4.2.1). Then, topic inference is executed to define the topic of the input dialogue(4.2.2). Afterwards, the most likely topic (only first topic 4.2.3) is used to select the keywords and form the query. Next, content retrieving,

FIGURE 4.7: *Approach B* pipeline

Topic Modeling	Topic inference	Query Formulation/ Expansion	Content Retrieval & OAT Selection	Sentence Filtering	Ranking



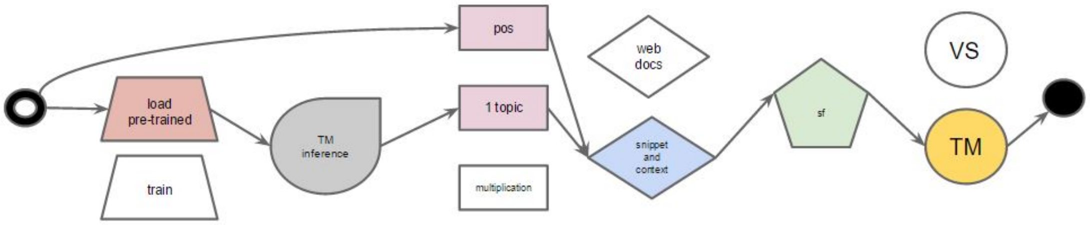
OAT extracting (4.2.4) and sentence filtering (4.2.5) are the sub-processes respectively executed. Ranking (4.2.6) is the last step in this processing pipeline.

4.3.3 Approach C

In *approach C*, both POS and TM strategies contribute alongside (i.e. in parallel) to formulate the query. Our topic modeling confidence measure (see 3.6.2) defines the weight of each strategy. In other words, the number of keywords that TM strategy adds to the keyword set is in direct correlation to the confidence of the TM inference (see 4.2.3). Similarly to *approach B*, the most likely topic (only first topic) is used to select the keywords and form the query. Figure 4.8 depicts which building blocks (and in what sequence) form this approach.

FIGURE 4.8: *Approach C* pipeline

Topic Modeling	Topic inference	Query Formulation/ Expansion	Content Retrieval & OAT Selection	Sentence Filtering	Ranking

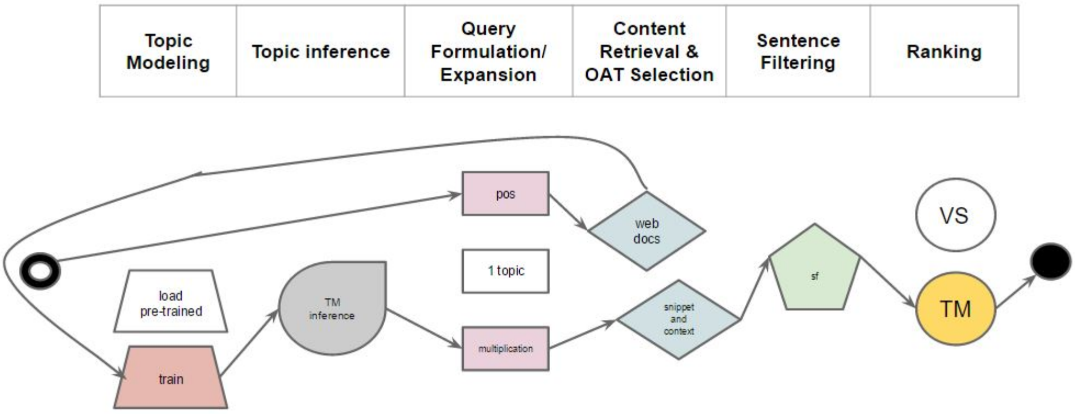


4.3.4 Approach D

In *approach D*, both POS and TM strategies contribute in the process, but not alongside. Here, a similar mechanism as *pseudo-relevance feedback* (also known as *blind relevance feedback*) is designed. The POS strategy forms a query which results in an initial set of retrieved results.

These results are a number of relevant texts (feedback documents) out of which the system computes a topic model. This feedback procedure makes the process slower but instead the resulting topic model is specialized for the current domain. In this approach, multiplication technique (see 4.2.3) is employed. Table 6.1 shows the result of the keyword extraction of each approach over the test set. Figure 4.9 depicts which building blocks and in what sequence form this approach.

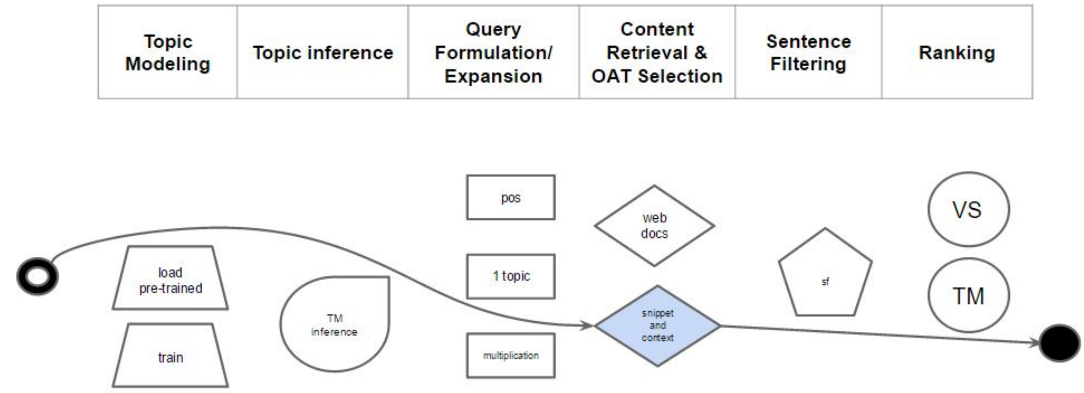
FIGURE 4.9: *Approach D* pipeline



4.4 The Baseline Method

In section 3.8, some possible candidates for a baseline solution for our problem are discussed. One main factor we took into account for baseline selection was to keep the baseline as simple and reproducible as possible. Our baseline method relies mainly on the output of a web search engine. Figure 4.10 depicts the baseline which consists of one single building block: *snippets and their context* (see 4.2.4).

FIGURE 4.10: Baseline method



Same as in our four approaches, Microsoft Bing search engine is employed for content retrieval in the baseline method. The search snippets returned from Bing are used in order to select relevant parts of the web page. For this purpose, each web page is chunked based on its structure (see 4.2.4). Then, the string similarity of each chunk to the snippet is computed and those chunks satisfying the similarity threshold are extracted. The extraction procedure is the same as other approaches (explained in section 4.1.4) with this difference that the similarity threshold is set to 1. This means that while processing the web page no relaxation is applied, instead the same (or the most similar) sentences to the ones in the snippet are selected as the output. Even though no relaxation is applied, this process is necessary to avoid returning occasional truncated sentences of snippets as the output of the baseline method. As it can be seen in Figure 4.10, no syntactic filtering is applied in the baseline method. The search engine is fed with the dialogue context and OAT selection is simply done by string comparison. No extra ordering is applied in our baseline which means the ranking provided by the search engine is preserved.

Chapter 5

Experimental Setup

A dialogue system can be evaluated in various styles. The evaluation approach can be either subjective or objective [Walker et al., 1997]. Evaluation metrics can be derived from questionnaires or log files [Paek, 2001]. The scale of the metrics can vary from the utterance level to the whole dialogue [Danieli and Gerbino, 1995][Kamm et al., 1999]. The dialogue system can be treated as a "black box" or as a "transparent box" [Hone and Graham, 2000]. This variety of styles beside lack of any agreed-upon standards in the research community and incompatibility of evaluation methods make evaluation of dialogue systems a challenge.

The main goal of the experimental evaluation in this thesis is to assess the potential of our proposed system in contributing to a dialogue with two distinct partners: a questioner and an answerer. In this quiz-like setting, the output of the system is supposed to be an utterance by a conversational agent which follows the dialogue by elaborating on that and by providing relevant information in an appropriate way for a real-world dialogue contribution.

In this chapter the experimental setting is introduced to answer the following questions:

- Is it really beneficial to employ techniques introduced in this thesis (e.g. keyword selection, topic modeling, syntactic filtering, ranking) for dialogue contribution?
- If the answer to the above-mentioned question is yes, which approach is more effective among the four we proposed.

Current chapter starts with some discussion about our evaluation strategy. This is followed by definition of the evaluation aspect we employed. A later section details the format of the query set we employed for our experiment. Next, the design of the experiment is discussed. In the final part, test questions as a quality control mechanism is introduced. The results, error analysis, and result discussion are given in the next chapter.

5.1 Embedded or Stand-alone Strategy?

An ideal strategy to test a tool which automatically suggests *off-activity talks* for dialogue contribution is embedding the developed component in a test-bed application (e.g. Aliz-E) and assessing the change in the usability of that application by carrying out a subjective evaluation of user satisfaction. In Aliz-E, a set of experiments is yearly carried out with the integrated system with real users at San Raffaele hospital in Milan in order to evaluate certain aspects of the system and collect data for further development[Kruijff-Korbayová et al., 2012]. Unfortunately, because of time constraints it was not possible to carry out the experiments of the current thesis in this embedding scheme. Having the above-mentioned strategy left out, we aimed to assess our application in a stand-alone scheme. However, there are some advantages to assessing with a stand-alone scheme instead of an embedded one. As an example, by employing a stand-alone scheme, we avoid problems which might occur during the interaction with the embedding system (e.g. Aliz-E in our study). Such problems can influence the usability results while we have no control over them. Therefore, the stand-alone strategy, in our case, brings the benefit to focus specifically on the quality of the OATs.

In order to compare the different approaches defined in this thesis, we assess the usefulness of the suggested OATs by each approach independently of the employer dialogue system. In other words, we observe the inputs and outputs of the system and while treating the system as a "black box", we measure the plausibility of each approach in terms of the relevance of their outputs to the corresponding input dialogue. This also means that the measurement scale in our evaluation scheme is at utterance level, in contrast to dialogue level.

5.2 Subjective or Objective Evaluation?

As stated before, evaluation of a dialogue system can be done with an objective or subjective approach. In case of objective evaluation, metrics like resources used (e.g. time, turns, user attention, etc) or the number of errors the system makes or inappropriate utterances made by the system can be mentioned. In some cases, a number of specified definitions of task success is used as an objective metric. However, it is not always easy to define task success in an objective way. The subjective measurement of the acceptance of an application or technology belongs to the group of usability evaluation. Usability evaluation focuses on users and their needs. Through usability evaluation, we want to figure out if a system can be used for the specific purpose from the user's point of view, and if it allows the users to achieve their goals, meeting their expectations. The most important criterion for measuring usability is user's satisfaction. In order to compare the different approaches defined in this thesis, we assess the usefulness of the suggested OATs from each approach by measuring user satisfaction with regard to a specific factor. That means, our evaluation strategy falls into the category of subjective evaluation.

Information about user satisfaction is usually gathered through interviews and questionnaires in the end of a session of interaction with a dialogue system. In principle, our component is not meant to establish a dialogue with the human, but to be integrated in a dialogue system and provide suggestions to the dialogue manager of such a system. That is why, we cannot evaluate user's satisfaction in the end of a dialogue session. Alternatively, we decided to ask participants to qualify the suggested OATs separately and regardless of any potential preceding or following dialogue. In other words, the experiment is not designed as a dialogue; instead, pairs of inputs and outputs of the system are presented to participants and they are asked to define which output is more satisfactory regarding some well-described aspects.

In contrast to objective evaluation techniques which are fairly well-established, subjective measurements are not as structured and straightforward. A difficulty which arises here is that dialogue systems and their users sometimes have inconsistent attitudes toward a dialogue. As a consequence, we need to take an extra care to make sure that the participants of our experiments have a correct sense of what we need to measure in a qualification process. We define different factors with regard to which we want to measure the quality of an OAT. So, we need to provide accurate instructions to make sure the participant has comprehended the differences and is able to distinguish between these factors. We also use some test questions as a means to make sure that the participants have read and clearly understood the instructions (see 5.6).

5.3 Target Evaluation Aspect: Relevance

In order to judge the usefulness of an OAT, we need to define the aspects of evaluation. The type of application determines the aspects that are important for a usability evaluation. Test parameters usually cover aspects such as the efficiency in reaching a goal and the effectiveness of single system characteristics. This is the case in the ISO standard 9241/10, for example. This standard is intended to calculate the usability summing-up values for effectiveness (probability of achieving a goal), efficiency, and user's satisfaction. Efficiency parameters include time required to reach a goal, the error rate, and the amount of effort needed to achieve a goal. Another commonly applied usability test is SUMI (Software Usability Measurement Inventory), the industry usability evaluation standard for analyzing users' opinions towards software products. SUMI covers most of the principles described in the ISO standard but focuses on the dimensions of efficiency, effect, helpfulness, control and learnability [Klüwer, 2015]. In this thesis, we define *relevance to the context* as the target aspect for evaluation of our system.

We take a seemingly objective notion as the target aspect to evaluate the quality of the output. We employ *relevance* with the assumption that an off-activity talk is useful to the degree of its relevance to the context. Few attempts have been made to quantify *relevance*. It is discussed in section 2.3 that *relevance* is a subjective concept even though, at first, it might not seem to be the case. Defining *relevance* in a dialogue is a very difficult task for many reasons: *relevance* is a subjective and time-varying concept; the dialogue is heterogeneous and highly dynamic; dialogue partners have different expectations and goals. Grice's maxim of relation [Grice, 1975] implicitly defines *relevance* as a relation between a set of propositions and a discourse-topic. Berg views *relevance* as "usefulness with regard to the conversational goals" [Berg, 1991]. In our case, conversational goals can be defined as sustaining long-term interaction and user's engagement while providing *relevant* information. Van Dijk defines *relevant* information as that information which is worth hearer's attention [Van Dijk, 1979].

Table 5.1: Rules and hints for relevance judgment in our experiment

A high ranked follow-up	A low ranked follow-up
Follow-up is related to the sub-dialogue.	The relevance of the follow-up and sub-dialogue is limited.
Follow-up and sub-dialogue are on the same topic.	There are major problems in the coherence between follow-up and sub-dialogue.
Follow-up contains discussion of people or objects mentioned in the sub-dialogue.	The transition is awkward.

Table 5.2: Relevance evaluation scale

Grade	Description
1	Irrelevant; not related in any way
2	Partially relevant, but little or no coherence between the dialogue context and the response.
3	Mostly relevant, but with major problems in the coherence with the context.
4	Relevant, but the transition is somewhat awkward.
5	Relevant and perfectly fluent

In our experiment, the participants are asked a key question to consider while assessing the results regarding *relevance* aspect: "*Is the follow-up relevant to the sub-dialogue given above?*" In the instructions, we provide a section of *rules and tips*, including table 5.1. This table shows some statements which must be true for a high-ranked or a low-ranked follow-up.

Given the hints in table 5.1, the participants are asked to assess the follow-up regarding its level of relevance to the previous sub-dialogue. This is done through a drop-down box which includes 5 levels of relevance. Table 5.2 shows the grading scheme that we used in our user evaluation.

5.4 Query Set

Table 5.3 displays the set of queries we used as our test set. All of the queries are in the format of a fact-based question plus the answer to that question. This is meant to mimic the dialogue context in quiz-game scenario of Aliz-E project. For each of these queries several suggestions by the system as candidate follow-ups are generated and later judged in the experiment in order to evaluate the performance of each of the solutions we developed in the course this thesis work.

In table 5.3, queries Q1 to Q10 are randomly selected from the test set of the question answering task at CLEF-2003 (The Cross-Language Evaluation Forum)[Magnini et al., 2004]. The test set of CLEF-2003 is in the form of 200 factual question-answer pairs derived from a document collection of over 1.5 million documents in nine European languages which has been expanded gradually over the years. However, since the current implementation of our system is only applicable for English language, we focused on the English version of their collection. The queries in CLEF-2003 are meant to simulate user information needs in various topics. In fact, the question answering task of CLEF is intended to "represent the real-world problem of an ordinary user posing a question to a system"[Magnini et al., 2004]. The available test set of CLEF includes manually retrieved exact answer of each of the open-domain questions together with the text snippet containing the answer string. Since this document is available on the Web[1], we actively exclude this URL when our system is fetching the relevant content from web servers in the content retrieval sub-process (see 4.2.4).

In order to maintain the connection between this work and Aliz-E project, which was the original motivation to begin this thesis work, and also to show how the system works on specific topics, we added 7 question-answer pairs in health-related topics as well (Q11-Q17). For this purpose, we considered question-answer pairs from some existing online sources of quizes in the health domain.

Table 5.3: Query set

Name	**Text**
Q1	What is the capital of Chechnya? Grozny
Q2	When was Elvis Presley's first record recorded? on July 5, 1954
Q3	What is the capital of Somalia? Mogadishu
Q4	Who is the Italian Prime Minister? Silvio Berlusconi
Q5	Who won a wrestling state title in 1990? Ken Stegall
Q6	In what year did Crimea become part of the Ukraine? 1954
Q7	Where is Edwin Tang from? New York City
Q8	When was the Berlin Wall built? in 1961
Q9	Where did Ayrton Senna have the accident that caused his death? in Imola, Italy
Q10	Where was the 1991 Copa America played? in Chile
Q11	How many portions of fruit and vegetables should we try to eat? At least five a day.
Q12	Which of the following does not count as one portion of our 5 fruit and vegetables a day? A potato
Q13	How many glasses of water should we try to drink each day? 6-8 tall glasses
Q14	How many teaspoons of sugar are there in a 250ml (tall) glass of cola? Nearly 5 and a half teaspoons
Q15	To keep our teeth healthy which of the following should we try not to have between meals? Fizzy drinks
Q16	Which of the following is not another word for sugar? Malt extract.
Q17	How often should we eat breakfast? Every day

It is worth mentioning that the input to the system does not necessarily need to be in the form of a question-answer pair, however because of the initial definition of the task of this thesis work (i.e. suggestion of follow-up talks to contribute in dialogues in a quiz-game setting like Aliz-E), we decided to preserve this structure in our test set. As a future work, different forms of inputs other than QA-pairs can also be tried with the system in order to mimic a non-quiz sub-dialogue.

The queries can be thought of as a simulation of a sub-dialogue between a questioner and an answerer, one or none of whom can be our *relational agent*. Correspondingly, the output of the system is a simulation of an elaboration on this sub-dialogue or a provision of relevant information in an appropriate way for a real-world follow-up to this sub-dialogue. We told the evaluators that our virtual agent is posed to a sub-dialogue and comes up with an utterance which it believes to be a suitable follow-up for that dialogue. For each follow-up, they should decide if it is relevant or not.

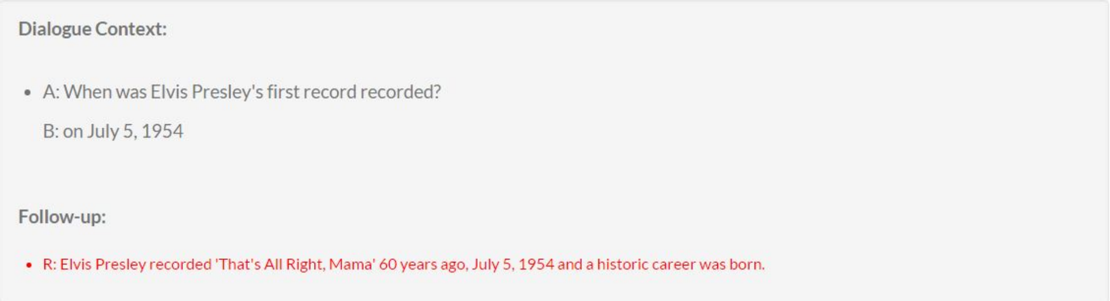
[1] `http://clef-qa.fbk.eu/2005/down/data/clef03/CLEF03_QA_cl_ENG_ENG_with-answers.xml`

5.5 Experiment Structure

The presentation of the experiment to the participants includes a section for instructions and several assessment units. In the instruction section, an explanation about the experiment is provided with some rules and hints (see 5.3). The assessment part is a list of judgment units consisting of the two most highly ranked results from each of the five strategies (4 approaches + 1 baseline). So for each query, 10 candidate suggestions (5 approaches * 2 suggestions from each approach) are to be judged and compared in separate judgment units (see 5.4 for query set). Every judgment unit consists of three steps:

1. Presenting a *dialogue context* and a *follow-up* suggested by the system. Figure 5.1 shows an example. The participant is asked to read the *dialogue context* and the *follow-up* well before they begin their judgment. In this setting, A and B could be either a human or a conversation system, but the follow-up (R) is always given by our conversation system.
2. First assessment step: evaluating the relevance through a drop-down box with 5 levels of relevance (see table 5.2). Figure 5.2 depicts this step.
3. In the last step, we collect possible comments by the participant. We provide a text-box in which they can optionally put their comments. An exemplary comment could be an explanation why they give a low ranking to some item. This could be informative for us to understand the results. Figure 5.1 depicts this step.

Figure 5.1: First step of a judgment unit



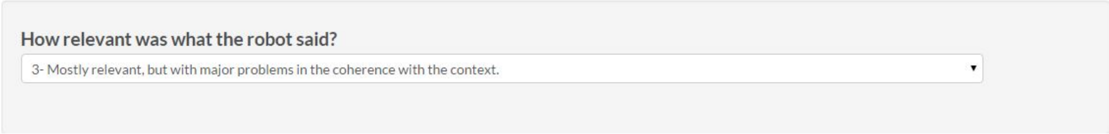
Dialogue Context:

- A: When was Elvis Presley's first record recorded?
  B: on July 5, 1954

Follow-up:

- R: Elvis Presley recorded 'That's All Right, Mama' 60 years ago, July 5, 1954 and a historic career was born.

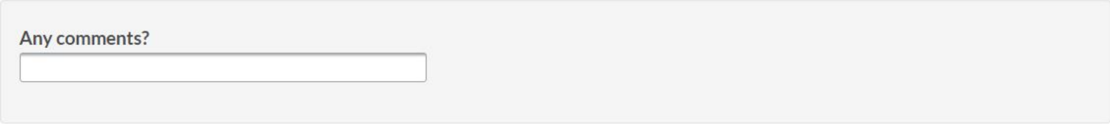
Figure 5.2: Second step of a judgment unit



How relevant was what the robot said?

3- Mostly relevant, but with major problems in the coherence with the context.

Figure 5.3: Third step of a judgment unit



Any comments?

We employed a crowd-sourcing platform named CrowdFlower[2] which is a software-as-a-service platform to access on-demand, scalable workforce to coordinate the use of human intelligence and intuition to perform tasks that computers are currently unable to do. For quality control, test questions are included to ensure the quality of the assessment work (see 5.6). For the evaluation, we had 10 judgments over each unit by several anonymous users. This is other than the untrusted judgments which are the ones done by those participants who fail to gain an accuracy of higher than 75% over our test questions (see 5.6).

5.6 Quality Control Mechanism

As a mechanism for quality control in our experiment, we used a set of test questions. We designed 20 "hidden" test questions with predetermined answers that are used to test contributors and calculate their individual trust scores. The trust score is a measure of the contributor's accuracy over all of the test questions they have submitted. We utilize test questions to ensure high quality judgments by the participants. They are designed to be based only on the instructions and as objective as possible. In order to attain higher levels of objectivity, all possible answers to a test question are allowed (e.g. a range of acceptable grades for relevance). In fact, they are carefully designed to be a little more lenient rather than overly strict. In the current work, test questions are designed to be hard enough to test a contributor's performance but easy enough for the participants who are carefully following the instructions to pass. In a number of test questions, the distribution of contributor responses were also reflected in the range of acceptable answers. This helps with detecting the outliers in the answers. Special care is taken to make sure that no prior knowledge about the data is required to answer test questions. Every test question is equipped with an instructive feedback such that if the contributor gives a wrong answer, clear reasons for that will be explained. As an example of test questions in our experiment, for a dialogue context like the following one:

- A: What is the capital of Somalia?
- B: Mogadishu

with a follow-up such as the below one:

- R: The Capital City of Somalia is the city of Mogadishu.

the acceptable range for a relevance assessment in our experiment is 3 and above (see 5.2 for the grading scale). As another example, for the below dialogue context:

- A: Which of the following does not count as one portion of our 5 fruit and vegetables a day?
- B: A potato

[2] `http://www.crowdflower.com`

and a follow-up like:

- R: you can keep your food above the safe temperature of 140 °F by using a heat source like a chafing dish, warming tray, or slow cooker.

the acceptable range is 2 and below.

Our experiment setup includes a quiz mode before the actual work mode which is entirely composed of test questions. We employ this quiz mode as a means to make sure that the participant has read and understood the instructions. The quiz mode is made up of 5 test questions and the participant needs to meet the requirements of at least 4 of them to pass the quiz. In case of a pass, the contributor is allowed to proceed to the work mode which is the actual judgment phase. This ensures that only contributors who prove they can complete the job accurately will be able to begin the work mode. Contributors who fail the quiz task are permanently disqualified from working on the experiment. In the experiment, 47 % of the participants who took the quiz failed it.

In the work mode also, one out of every 5 judgments is randomly a test question. Ideally, contributors will behave exactly the same on test questions as they do on regular questions. To accomplish this, test questions are indistinguishable from the overall data set. Indistinguishable test questions make sure that contributors cannot tell the difference. This means that contributors will not change their behavior on test questions. Test questions have an appropriate answer distribution in order to train contributors on every possible answer instead of biasing them towards one answer. Regarding the number of available test questions, the idea is to make sure that there are enough test questions that contributors see each test question only once while working on the experiment. For contributors whose trust score falls below 75% accuracy within the job (i.e. fail test questions for more than 75% of the times), they will be prevented from working on the experiment and all their work will be discarded. From the total judgements in our experiment, 0.204% were identified as untrusted according to this quality control mechanism.

Chapter 6

Results and Discussion

In this chapter we present the results obtained from the system using a set of queries in order to assess the strength and utility of the different approaches implemented within our system. We carried out an experiment in order to compare the *relevance* of the results from each approach. In the beginning of this chapter, the results of the keyword selection process for each of the approaches are comparatively discussed. In a later section, we provide a qualitative comparison of the OATs generated by different approaches implemented in our system and also by the baseline. This will be followed by a discussion of the results of the subjective evaluation in the final part of this chapter.

6.1 Results of The Keyword Selection Process

Table 6.1 shows the output of the keyword extraction process in each approach. It is worth noting that the baseline approach does not include a keyword extraction process because the dialogue context is directly used to query the search engine. Approaches C and D extend the keywords used in approach A. So in order to save the space, we avoid repeating the same keywords below corresponding columns and just mention the extension set.

In approach A, all the keywords are extracted from the sub-dialogue itself, with the assumption that some specific parts of speech represent the most content-bearing concepts (i.e. focus terms).

In approach B, the highest ranked terms of the most relevant topic are selected as the query. These keywords represent a general topic which the sub-dialogue most probably belongs to. As an example, for 6 out of 7 health-related queries we have in our query set (i.e. Q11-Q17), exactly the same set of keywords are suggested (e.g. food, add, cooking). In other words, this approach is not able to differentiate between dialogues of closely related topics. This means that approach B has the potential to cause the flow of the dialogue diverge from the current topic to a general topic, for example from the first Elvis Presley's song to music in general. The number of keywords in this approach is fixed at 3. In some cases such as Q13, the topic inference process fails to suggest convincingly relevant keywords.

Approach C, combines the keyword selection process of approaches A and B. The result will be the keywords from the surface of the query extended by keywords from the most relevant general topic of the offline topic model. The number of keywords in this approach is in correlation to the number of keywords selected from sub-dialogue and the confidence of the topic model inference. This is to keep the balance between surface-based keywords and topic-based keywords. For example, in case of Q8, since the surface-based keyword set is small (e.g. two), only one more keyword is added as the extension while in case of Q9 with six surface-keywords four extending keywords are used.

Approach D also has a combinatory method for keyword selection. In this approach also the same surface-based keywords are used, but the topic inference is applied over a topic model which is trained in an online mode. A number of relevant texts (feedback documents) extracted from web are used to train the online topic model. This online procedure slows down the process, instead the resulting topic model is specific to the current domain. Resulting specific topic model causes a more domain-specific set of keywords. Comparing the keywords in approach D to the one in approach C, one can clearly notice that approach D generates keywords from more specific topics. As for Q11-Q17, the keywords selected are different while they were the same in approaches B and C.

Approach D sounds more promising at selecting relevant keywords. For example, despite the offline topic inference approaches, approach D does not fail to offer a convincingly relevant keyword set for Q13. This is mainly because of the specific-domain corpora that is collected in an online manner for each query. However, there is a possibility that approach D deviates from the topic because of the diverging feedback documents of the web search engine. An example of this case occurs for Q7. In this case, the sub-dialogue is about “Edwin Tang” and “New York”, but the extension keyword set includes keywords about “hotels” and “reviews” which have occurred frequently in the feedback documents.

To sum up, one can conclude that keyword selection process of approach D restricts the flow of the dialogue in more specific topics while approaches B, expands the topic to a more general span. In case of B there is a risk of failure because of unseen topics and in case of D, there is a risk of deviation because of diverging feedback documents.

TABLE 6.1: Keyword extraction/expansion results

Query	**A**	**B**	**C**[1]	**D**
Q1	capital, Chechnya, Grozny	film, films, jan	..., film, films	..., russian, chechen
Q2	Elvis, Presley, record, July	music, album, songs	..., music, album	..., music, time
Q3	capital, Somalia, Mogadishu	afghanistan, taliban, pakistan	..., afghanistan, taliban	..., somali, city
Q4	Prime, Minister, Silvio, Berlusconi	government, argentina, country	...,government, argentina, country	..., italian, italy, milan
Q5	wrestling, state, title, Ken, Stegall	game, team, players	..., game, team, players	..., college, ncaa, school
Q6	year, Crimea, part, Ukraine	government, people, africa	..., government, people	..., russia, russian, crimean
Q7	Edwin, Tang, New, York, City	art, museum, york	..., art, museum, york, show	..., hotels, reviews, reviewed
Q8	Berlin, Wall	city, park, street	..., city	..., east
Q9	Ayrton, Senna, accident, death, Imola, Italy	injuries, injury, head	..., injuries, injury, head, trauma	..., williams, formula, steering
Q10	Copa, America, Chile	government, argentina, country	..., government, argentina	..., argentina, peru
Q11	portions, fruit, vegetables, least, day	food, add, cooking	..., food, add, cooking	..., eating, health, eat
Q12	portion, fruit, vegetables, day, potato	food, add, cooking	..., food, add, cooking	..., healthy, eat, eating, health
Q13	glasses, water, day	city, park, street	..., city, park, street	..., drink, drinking, glass
Q14	teaspoons, sugar, glass, cola, half	food, add, cooking	..., food, add, cooking, minutes	..., calories, health, food, fruit, body
Q15	teeth, meals, drinks	food, add, cooking	..., food, add	..., health, care
Q16	word, sugar, Malt, extract	food, add, cooking	..., food, add	..., beer, brewing
Q17	breakfast, day	food, add, cooking	..., food, add	..., make

6.2 Results of The OAT Suggestion Process

The output of the keyword selection process which was discussed in the last section is employed in processing pipelines to suggest OATs. In this section, we provide a qualitative discussion about the OATs suggested by each of the approaches implemented in our system and also from the baseline.

6.2.1 Rephrasal follow-ups

In some cases, the follow-up is barely a rephrase of the dialogue context without provision of any further development of the topic or any addition of relevant information. For example, a follow-up for Q3 suggested by the baseline approach is: "The Capital City of Somalia is the city of Mogadishu.". Another example for the same query by approach A is: "The capital of Somalia is Mogadishu.". Even though such follow-ups are, by definition, the most relevant content to the dialogue context and might sound as a natural response in the sense of a confirmation to the correct answer, they are usually considered as boring follow-ups by the participants. In fact, these follow-ups are not actual elaborations to the dialogue and do not provide added-information. These follow-ups count as acceptable ones to keep on the dialogue and avoid pauses, but they are not desirable in applications with education and entertainment purposes. Table 6.2 shows the frequency of such cases by each approach. As it is clearly shown in the below table, rephrasal follow-ups are frequent in baseline and approach A. The reason is that in approaches B, C, and D, the ranking algorithm applied does not simply rely on term-frequency, but on topic distributions. The two ranking methods we employed in this work are explained in section 4.1.6.

Table 6.2: Frequency of rephrasal follow-ups in the test-set

Approach	**Frequency (out of 34)**
Baseline	5
App. A	3
App. B	0
App. C	0
App. D	1

6.2.2 Sentence filtering performance

One of the advantages of the four approaches developed in this thesis over the baseline is the sentence filtering functionality. It is meant to remove sentences which are not suitable for dialogue contribution; such as incomplete sentences, ungrammatical sentences, imperative sentences, and also sentences including unresolved coreferent mentions (see 4.1.5). Table 6.3 shows the frequency of these unsuitable sentences in the output of the baseline method. Evidently, these undesirable cases only occur in the baseline method and our system was competent enough to avoid them.

Table 6.3: Frequency of unsuitable sentences by the baseline method

Filter	**Frequency (out of 34)**
Incomplete sentences	4
Ungrammatical sentences	3
Imperative sentences	2
Unresolved references	3

6.2.3 Occasional shortcoming of approach B

In 2 cases (out of 34), approach B is not able to provide an utterance for the given test query. The reason to this is the fact that approach B tries to expand the topic by choosing general terms from the topics which are trained over an offline corpus. While using general terms for querying the web search engine, in some cases, the topic-based ranking mechanism of approach B rejects all the fetched OATs as not convincingly relevant. In such situations, the system returns a fixed phrase “nothing to say”. In both cases, all participants, with no exception, rated the relevance as the lowest (i.e. one). Ninety percent of the participants judged these cases as “boring” or “frustrating”.

6.2.4 Repetitive follow-ups by approach B

In 6 out of 7 queries from health-related topics (i.e. Q11-Q17), approach B returns the same candidate follow-up. This clearly shows that this approach categorizes any given dialogue context into one of the previously trained topics and returns some follow-ups from a general topic. In other words, this approach is not able to differentiate between dialogues of closely related topics. This issue did not occur in case of other approaches within the test set.

6.3 Results of The Relevance Assessment

As part of the evaluation of our system, we assessed each of the OATs suggested by the proposed approaches regarding their relevance to the corresponding dialogue context. In this section, the results of this experiment with human participants will be presented and comparatively discussed. As mentioned in 5.5, we employed a crowd-sourcing platform to coordinate the use of human intelligence to perform a subjective evaluation of the relevance of each suggested OAT to the corresponding input dialogue. This platform enabled us to collect human judgments from anonymous online participants all over the world in exchange for a specific amount of payment for each set of judgment units. The quality of these anonymous judgments is controlled via a set of hidden test questions (see 5.6). Through this service, we collected 10 judgments for each unit by several anonymous users.

The results for the relevance judgment is shown in tables 6.4 and 6.5. The first table shows the results over the open-domain queries (Q1-Q10) and the second over the health-related queries. For table 6.4, there are 10 open-domain queries assessed over each of our approaches, and for table 6.5 there are 7 domain-specific queries. For each query, 10 candidate suggestions (5 approaches

Table 6.4: Relevance assessment results: open domain (Q1-Q10)

Approach	**Relevance Results**	**Variance**	p-value	p<0.05
Baseline	4.14	0.838	—	—
Approach A	4.37	0.941	0.075899	
Approach B	1.03	0.117	2.31e-12	*
Approach C	2.20	0.956	0.000135	*
Approach D	2.82	0.848	0.001549	*

Table 6.5: Relevance assessment results: health-related domain (Q11-Q17)

Approach	**Relevance Results**	**Variance**	p-value	p<0.05
Baseline	2.31	0.944	—	—
Approach A	2.87	1.063	0.138	
Approach B	1.14	0.275	7.04e-05	*
Approach C	1.89	0.803	0.157	
Approach D	2.62	0.853	0.441	

* 2 suggestions from each approach) are judged. This means that in table 6.4, each number below relevance column stands for the average of 200 judgments (= 10 participants * 10 queries * 2 most highly ranked OATs by the corresponding approach for each query). Corresponding numbers in table 6.5 stand for the average of 140 judgments (10 participants * 7 queries * 2 OATs by the corresponding approach). The measurement unit is the relevance scale defined in table 5.2. Variance of each set of judgments in addition to the significance test results obtained from a linear mixed model test is also mentioned in both tables of relevance assessment results (i.e. tables 6.4 and 6.5).

The results show satisfying performance for approach A in both health-related and open domain test set. This means that inclusion of the components we developed for POS-based keyword extraction, sentence filtering and term-frequency ranking altogether produced follow-ups with higher level of relevance over the baseline method.

In case of the open-domain test set, the three approaches involving topic modeling techniques (i.e. approaches B, C, and D) significantly resulted in lower levels of relevance comparing to the baseline and approach A. One possible reason to that could be the relatively low number of topics trained in each model. Concerning the computational capacities available to the author, in each of the training processes the number of trained topics is fixed at 100. Observing the results, this number turns out to be far lower than a reasonable one for an open domain dialogue contribution.

In case of the restricted-domain (i.e. health-related topics) test set, the outcome of the experiment shows that the performance of the two approaches based on offline topic model (i.e. approaches B and C) is not satisfactory. The reason is that the offline corpus we used for training these topic models (i.e. *New York Times* corpus) is a general-purposed corpus and not one for the specific domain we are targeting. As it can be seen in table 6.1, this resulted in a poor selection of keywords for both approaches B and C and, as a consequence, low levels of relevance were obtained.

The only exception in approaches involving topic modeling techniques is approach D over the specific-domain test set. As presented in table 6.5, even though approach D shows lower performance than approach A, it outperforms the baseline. This shows that the process pipeline we proposed in approach D, which includes topic modeling techniques for keyword selection and OAT ranking, has the potential to outperform the baseline in specific domains. To achieve such results, it is required to train a proper number of topics over a specialized corpora for the target domain (e.g. collected by a relevance feedback process). However, it is worth mentioning that the overhead of collecting a specialized corpus and an online topic training makes this approach to a high extent slower than all the others.

Chapter 7

Conclusions

This thesis set out to propose a system able to provide dialogue contribution suggestions which are relevant to the context, yet out of the main activity of the dialogue (i.e. *off-activity talk*). Following the work done in Aliz-E project, we defined *off-activity talk* as a comparable concept to *small talk*s with the same aim of breaking out of the fixed structure of task-bound dialogues. In the context of the current thesis, an *off-activity talk* is a verbal reaction which is required to be contextually *relevant* to the content of the previous interaction and preferably includes a provision of some added-information. This verbal reaction can consist of one or more sentences and is extracted from the freely available online resources. The output of our tool is supposed to be employed as an utterance by an artificial agent (i.e. a virtual character or a robot) in a single dialogue turn. The sub-tasks addressed by our system are to analyze the dialogue context, to detect the topic, and to provide a ranked list of appropriate candidate utterances to be employed by the dialogue manager in a conversation system. The outcome of this thesis is a satisfactory point of entry to investigate the hypothesis that adding automatically generated *off-activity talk* feature to a conversational agent can lead to building up engagement of the dialogue partners.

We proposed a modular architecture that allows exploring different solutions. This architecture takes a textual representation of a previous sub-dialogue as the input and returns a ranked list of sentence(s) believed to be suitable off-activity talks. It consists of modules for dialogue context processing, query expanding, relevant content retrieving, OAT selecting from online content, syntactic filtering, and OAT ranking. Although the system we proposed was fed with English input, the same system architecture can easily be used in other languages by replacing few of the components, concretely: the POS tagger, the NER, and the trained topic model.

We implemented four novel approaches through our modular architecture in the form of four different processing pipelines. The first approach is solely based on shallow surface processing for query formulation and term frequency vectors for ranking. The second one entirely relies on topic modeling to form the query. The third and the fourth approaches are combinations of the first two. The third is in a parallel mode with an offline topic model, and the fourth is in a sequential mode with a topic model trained over relevant documents fetched from online resources. The last three approaches employ a topical ranking scheme.

We tested our proposed approaches and evaluated their performance. The outcome of the evaluation was improvement in the results for one of the approaches over the baseline method, both in the open-domain and the restricted-domain test set. This can be explained on the ground that inclusion of the components we developed for POS-based keyword extraction, sentence filtering and term-frequency ranking altogether produced follow-ups with higher levels of relevance. The three other approaches we proposed involved topic modeling techniques for keyword extraction and OAT ranking. Excluding one exception, these approaches did not achieve satisfactory results either in the open domain, nor in the health-related domain. The only exception to this was approach D which outperformed the baseline over the specific-domain test set. Out of this observation, we conclude that topic modeling techniques proposed in this thesis for keyword selection and OAT ranking have the potential to outperform the baseline in specific domains. However, to achieve this goal specific-domain corpora and a carefully tuned number of trained topics are required.

We showed that employing topic inference algorithms can be profitable for extending the set of focus terms of the conversation. Having detected the focus terms of the dialogue, topic inference can be helpful to come up with extra keywords which are different from the initial focus terms, but still on the same or similar topics. However, beside the benefits of employing topic models for this application, there are some clear issues: the topic models are so sensitive to the input data that small changes to the stemming/tokenization algorithms can result in completely different topics; topics are “unstable” in the sense that adding new documents can cause significant changes to the topic distribution (less of an issue with large corpora). As a wrap up, we can conclude that employment of topic models for applications with goals similar to the ones of our system requires special care and adequate tuning.

On the whole, the results suggest that our system provides an applicable and effective architecture to process dialogue context and suggest relevant off-activity talks, which was the main goal of this thesis. Additionally, it provides an organized framework to perform further research in each one of the sub-tasks of the system: modeling the topic of the dialogue, retrieving appropriate sentences from online resources, ranking candidate OATs, filtering certain types of inappropriate sentences for dialogue contribution, etc.

Chapter 8

Future Work

Although satisfactory results were obtained with the current architecture of the system and the methods proposed in each one of these modules, there is still space for improvement in different parts of the system. This chapter presents these possible improvements as reasonable extensions to pursue the current work.

Since the POS-based query analysis module identifies focus terms as the ones with specific POS tags, it returns lengthy lists when the input dialogue is a long sequence of tokens (e.g. when the input consists of more than two utterances). This can lead to a confusion for the search engine or make the domain too restricted. The first case is more troublesome, but the second also results in an output by the system which is too similar to the input and not much informative. Therefore, it is desirable to employ additional techniques to avoid this issue. To do so, named-entities can be used instead of POS-tags when the dialogue is large. An alternative could be defining a weighing mechanism to emphasize on the most recent utterances instead of the whole given dialogue. Semantic techniques for focus term detection can be also an option. As a cross-linguistic study, the system can be further enhanced by adding support to other languages.

Since Bing web search engine, which was employed as our online content retrieval engine, includes query expansion functionalities using synonyms of the terms, we disabled our offline searching and query expanding modules in the latest version. However, it is encouraged to test offline search functionalities of our system. This will study the flexibility of our system to use offline sources as well. For this purpose, the Lucene library is used for searching and WordNet for query expansion. Using offline search mechanisms instead of Bing can also help to avoid the influence of personalized, click-based features of Bing.

With the presented research as a starting point, advantages or disadvantages of including automatically retrieved OAT turns in conversation systems can be studied. It is encouraging to investigate the effect of embedding our system on engagement of the users with a conversation system. It is also desirable to investigate whether automatically retrieved OAT turns can promote users' trust in knowledgeability of the agent and their perception of the agent's intelligence.

Integration of the two ranking methods is also an option to investigate further possibilities for improvement. This can be done by taking the inference confidence measure into account as a smoothing factor, similarly to what we did in approach C to integrate topic modeling keyword extraction and POS-based keyword extraction.

As explained in section 3.5.2, contrary to the similarity in definitions, there is a fundamental difference between the purpose and implementation of an OAT in this thesis and the one in Aliz-E project. As in Aliz-E, OAT is used in form of prerecorded questions around predefined topics to encourage the user to talk about those topics and elicit information from them while in this thesis, we focused on automatic generation of OATs as follow-up added information relevant to the context of the previous interaction with the aim of encouraging users' engagement without any deliberate direct information elicitation. This difference can also be investigated in order to study the correlation of these characteristics with the overall effects of OAT such as users' perception of intelligence and their interest in having further sessions of interaction with the agent.

As another possible future work, different forms of input other than QA-pairs (e.g. longer dialogue snippets) can also be tried with the system in order to mimic non-quiz sub-dialogues. Factual text-snippets can be used as an alternative source of queries. One might also think of twitter corpora as a source of informal utterances even though it will be a big challenge for a conversation system to respond to.

Bibliography

Doris M Dehn and Susanne Van Mulken. The impact of animated interface agents: a review of empirical research. *International journal of human-computer studies*, 52(1):1–22, 2000.

Timothy W Bickmore and Rosalind W Picard. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12 (2):293–327, 2005.

Timothy Bickmore and Justine Cassell. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 396–403. ACM, 2001.

Tina Klüwer. Social talk capabilities for dialogue systems. 2015.

Justine Cassell and Timothy Bickmore. External manifestations of trustworthiness in the interface. *Communications of the ACM*, 43(12):50–56, 2000.

Ivana Kruijff-Korbayova, Elettra Oleari, Ilaria Baroni, Bernd Kiefer, Mattia Coti Zelati, Clara Pozzi, and Alberto Sanna. Effects of off-activity talk in human-robot interaction with diabetic children. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 649–654. IEEE, 2014.

Ivana Kruijff-Korbayová, Heriberto Cuayáhuitl, Bernd Kiefer, Marc Schröder, Piero Cosi, Giulio Paci, Giacomo Sommavilla, Fabio Tesser, Hichem Sahli, Georgios Athanasopoulos, et al. Spoken language processing in a conversational system for child-robot interaction. In *WOCCI*, pages 32–39, 2012.

Ulli Waltinger, Alexa Breuing, and Ipke Wachsmuth. 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–11)*, 2011.

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

Alejandro Figueroa, Günter Neumann, and John Atkinson. Searching for definitional answers on the web using surface patterns. *IEEE Computer*, 42(4):68–76, 2009.

Alejandro Figueroa and Günter Neumann. Language independent answer prediction from the web. In *Advances in Natural Language Processing*, pages 423–434. Springer, 2006.

Günter Neumann and Bogdan Sacaleanu. Experiments on robust nl question interpretation and multi-layered document annotation for a cross–language question/answering system. In *Multilingual Information Access for Text, Speech and Images*, pages 411–422. Springer, 2005.

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*, pages 428–431. IEEE Computer Society, 2010.

Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611, 2007.

Peter Schönhofen. Identifying document topics using the wikipedia category network. *Web Intelligence and Agent Systems*, 7(2):195–207, 2009.

M Höppner, W Horstmann, S Rahmsdorf, Alexander Mehler, and Ulli Waltinger. Enhancing document modeling by means of open topic models: Crossing the frontier of classification schemes in digital libraries by example of the ddc. *Library Hi Tech*, 27(4):520–539, 2009.

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*, pages 95–102. Association for Computational Linguistics, 2002.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Jay M Ponte and W Bruce Croft. Text segmentation by topic. In *Research and Advanced Technology for Digital Libraries*, pages 113–125. Springer, 1997.

Xing Yi and James Allan. A comparative study of utilizing topic models for information retrieval. In *Advances in Information Retrieval*, pages 29–41. Springer, 2009.

Victor Lavrenko and W Bruce Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127. ACM, 2001.

Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410. ACM, 2001.

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*, pages 32–35. Association for Computational Linguistics, 2008.

Sisay Fissaha Adafre and Josef Van Genabith. A hybrid filtering approach for question answering. 2009.

Boris Katz, Sue Felshin, Deniz Yuret, Ali Ibrahim, Jimmy 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*, pages 230–234. Springer, 2002.

Alejandro G Figueroa and Günter Neumann. Genetic algorithms for data-driven web question answering. *Evolutionary computation*, 16(1):89–125, 2008.

Davide Buscaldi and Paolo Rosso. Mining knowledge from wikipedia for the question answering task. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 727–730, 2006.

Abhinav Sethy, Panayiotis G Georgiou, and Shrikanth Narayanan. Building topic specific language models from webdata using competitive models.

Daniel Magarreiro, Luísa Coheur, and Francisco S Melo. Using subtitles to deal with out-of-domain interactions.

Rafael E Banchs and Haizhou Li. Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42. Association for Computational Linguistics, 2012.

Mehran Sahami and Timothy D Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web*, pages 377–386. AcM, 2006.

Tim Paek. Empirical methods for evaluating dialog systems. In *Proceedings of the workshop on Evaluation for Language and Dialogue Systems-Volume 9*, page 2. Association for Computational Linguistics, 2001.

Morena Danieli and Elisabetta Gerbino. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI spring symposium on Empirical Methods in Discourse Interpretation and Generation*, volume 16, pages 34–39, 1995.

Kate S Hone and Robert Graham. Towards a tool for the subjective assessment of speech system interfaces (sassi). *Natural Language Engineering*, 6(3&4):287–303, 2000.

Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics, 1997.

Candace Kamm, Marilyn A Walker, and Diane Litman. Evaluating spoken language systems. In *Proc. of AVIOS*. Citeseer, 1999.

Rukmini M Iyer. *Improving and predicting performance of statistical language models in sparse domains*. PhD thesis, Citeseer, 1998.

Hoa Trang Dang, Diane Kelly, and Jimmy J Lin. Overview of the trec 2007 question answering track. In *TREC*, volume 7, page 63, 2007.

H Paul Grice. Logik und konversation. *1979*, pages 243–256, 1975.

Jonathan Berg. The relevant relevance. *Journal of Pragmatics*, 16(5):411–425, 1991.

Teun A Van Dijk. Relevance assignment in discourse comprehension. *Discourse processes*, 2(2): 113–126, 1979.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27. Association for Computational Linguistics, 2009.

Azzah Al-Maskari, Mark Sanderson, and Paul Clough. Relevance judgments between trec and non-trec assessors. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 683–684. ACM, 2008.

Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction*, 19(1): 61–84, 2004.

Juan Fasola and Maja J Mataric. Using socially assistive human–robot interaction to motivate physical exercise for older adults. *Proceedings of the IEEE*, 100(8):2512–2526, 2012.

Cory D Kidd and Cynthia Breazeal. A robotic weight loss coach. In *Proceedings of the national conference on artificial intelligence*, volume 22, page 1985. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.

Timothy Bickmore, Daniel Schulman, and Langxuan Yin. Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence*, 24(6):648–666, 2010.

Ivana Kruijff-Korbayová, Georgios Athanasopoulos, Aryel Beck, Piero Cosi, Heriberto Cuayáhuitl, Tomas Dekens, Valentin Enescu, Antoine Hiolle, Bernd Kiefer, Hichem Sahli, et al. An event-based conversational system for the nao robot. In *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*, pages 125–132. Springer, 2011.

Tony Belpaeme, Paul E Baxter, Robin Read, Rachel Wood, Heriberto Cuayáhuitl, Bernd Kiefer, Stefania Racioppa, Ivana Kruijff-Korbayová, Georgios Athanasopoulos, Valentin Enescu, et al. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, 1(2):33–53, 2012.

Ivana Kruijff-Korbayov, Elettra Oleari, Clara Pozzi, Stefania Racioppa, and Bernd Kiefer. Analysis of the responses to system-initiated off-activity talk in human-robot interaction with diabetic children. In *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue*, pages 90–97. Heriot Watt University, 2014.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

Brian Goetz. The lucene search engine: Powerful, flexible, and free. *JavaWorld. Available http://www. javaworld. com/javaworld/jw-09-2000/jw-0915-lucene. html*, 2000.

Michael McCandless, Erik Hatcher, and Otis Gospodnetic. *Lucene in Action: Covers Apache Lucene 3.0*. Manning Publications Co., 2010.

Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.

Paolo Ferragina and Antonio Gulli. A personalized search engine based on web-snippet hierarchical clustering. *Software: Practice and Experience*, 38(2):189–225, 2008.

Silviu-Petru Cucerzan and Matthew R Richardson. Systems and methods that enable search engines to present relevant snippets, March 31 2009. US Patent 7,512,601.

Günter Neumann and Sven Schmeier. A mobile touchable application for online topic graph extraction and exploration of web content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, pages 20–25. Association for Computational Linguistics, 2011.

Evan Sandhaus. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752, 2008.

Bernardo Magnini, Simone Romagnoli, Alessandro Vallin, Jesús Herrera, Anselmo Penas, Víctor Peinado, Felisa Verdejo, and Maarten de Rijke. The multiple language question answering track at clef 2003. In *Comparative Evaluation of Multilingual Information Access Systems*, pages 471–486. Springer, 2004.

Martha Palmer, Daniel Gildea, and Nianwen Xue. Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103, 2010.

Ronakkumar Patel, Anton Leuski, and David Traum. Dealing with out of domain questions in virtual characters. In *Intelligent Virtual Agents*, pages 121–131. Springer, 2006.

Andrew Turpin, Yohannes Tsegay, David Hawking, and Hugh E Williams. Fast generation of result snippets in web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 127–134. ACM, 2007.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2):257–287, 2008.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106, 2005.

Sudeep Gandhe, Andrew S Gordon, and David Traum. Improving question-answering with linking dialogues. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 369–371. ACM, 2006.

Mark A Stairmand. Textual context analysis for information retrieval. In *ACM SIGIR Forum*, volume 31, pages 140–147. ACM, 1997.

Zhiping Zheng. Answerbus question answering system. In *Proceedings of the second international conference on Human Language Technology Research*, pages 399–404. Morgan Kaufmann Publishers Inc., 2002.