Karlsruhe Institute of Technology

Predicting Clarification Questions in a Social Dialog System

Bachelor's Thesis of

Xizhe Lian

at the Department of Informatics
Institute for Anthropomatics and Robotics

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

8. July 2015 – 7. November 2015

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, 7. Nov. 2015

. . . . . .

(Xizhe Lian)

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text. I have followed the respectively valid KIT statutes for safeguarding good scientific practice.

Karlsruhe, 7. Nov. 2015

. . . . . .

(Xizhe Lian)

Abstract

Spoken dialog systems have a broad application in human-machine-interaction, which enable direct conversations between human and machines. Due to immense technical limitations, spoken dialog systems encounter with enormous errors from the automatic speech recognition component. Hence, the process to repair those errors is triggered frequently. On the other hand, it has been researched for long how to generate "natural" dialogues under the restriction of incompetent speech recognition. Since social dialogue, on the contrary to goal-oriented dialogue, is the tendency in field spoken dialog system, many processes, models and engaging techniques are at starting and relative immature. Therefore we focus on clarification issues in social dialog systems in this thesis.

The topic of this thesis is to identify different classes of clarification questions in social dialogue and thereby to predict a clarification question in social spoken dialog systems. The "naturalness" of a social dialog system should be enhanced accordingly. The approach in this thesis involves language modelling and classification. Several language models were built according to different classes defined in this thesis. Two classifiers were trained, one was for question/statement prediction, the other was for subdivision of clarification questions. Both procedures require corpora as training and test data, thus the OpenSubtitles corpus is exploited in this thesis. To examine our approach, a user study is performed. The user study is constructed with query groups, each query group has a pattern and each class of the clarification question has a question serving as reaction to the pattern. The participant of the study needs to score how appropriate is each questions in the situation.

The results of each sub-class in the study were close to each other, they were overall slightly better than neutral. However, the study reveals that vague and brief questions were appropriate in the most of the situations in the study.

Zusammenfassung

Sprachdialogsysteme ermöglichen direkte Konversation zwischen Menschen und Maschine und finden dadurch breite Anwendung in der Mensch-Maschine-Interaktion. Auf Grund von Technischen Einschränkungen müssen Sprachdialogsysteme mit Fehlern der Automatischen Spracherkennungskomponente zurecht kommen. Dadurch greift regelmäßig ein Prozess ein der diese Fehler korrigiert. Auf der anderen Seite wird nach Möglichkeiten gesucht natürliche Dialoge trotz der vielen Einschränkungen der Automatischen Spracherkennung zu generieren. Da der sozial Dialog, im Gegenteil zum zielorientierten Dialog, ist die Tendenz auf dem Gebiet Sprachdialogsystem. Da in Sprachdialogsystemen die Tendenz zu „sozialen“ Dialogen geht, im Gegensatz zu zielorientierten Dialogen, sind viele Prozesse, Techniken und Modelle noch relativ unausgereift. Aufgrund dessen konzentrieren wir uns in dieser Arbeit auf das Problem der Klärungsfragen in sozialen Sprachdialogsystemen.

In dieser Arbeit geht es darum verschiedene Klassen von Klärungsfragen in sozialen Dialogen zu identifizieren und dadurch Klärungsfragen in sozialen Sprachdialogsystemen vorherzusagen. Die ”Natürlichkeitëines Sozialdialogsystem soll dadurch verbessert werden. Der Ansatz in dieser Arbeit beinhaltet Sprachmodellierung und Klassifizierung. Verschieden Sprachmodelle wurden nach unterschiedlichen Klassen, die in dieser Arbeit definiert sind, gebaut. Zwei Klassifikatoren wurden trainiert, einer für die Unterscheidung von Frage/Aussage und der zweite zur Unterscheidung von Typen von Klärungsfragen. Beide Prozesse brauchen eine Trainings- und Testdatengrundlage. Dafür wurde in dieser Arbeit der OpenSubtitles corpus benutzt. Um unseren Ansatz zu bewerten wurde eine Benutzerstudie durchgeführt. Die Studie ist so aufgebaut, dass eine Musteraussage mit jeweils einer Frage aus den Klaerungsfragenklassen gegeben ist. Der Teilnehmer konnte die Frage nach ihrer Angemessenheit in der jeweiligen Situation bewerten.

Die Ergebnisse der einzelnen Unterklasse waren einander sehr ähnlich, insgesamt waren sie etwas besser als neutral. Die Studie zeigte jedoch das vage und kurze Fragen in den meisten Situationen als passend erachtet wurden.

Acknowledgments

First of all, I would like to thank my supervisor Maria Schmidt, who has guided me with patience and flexibility and given valuable comments and suggestions during the thesis.

Second, it is really appreciated that I received so many helps from my colleges and friends, Jingping Lin, Meng-meng Yan, Rebecca Seelos, Alexandar Andonov, Chunho Wu, Diego Langarica Fuentes and Luiz Henrique S.Silva. Without your suggestions and encouragements, this thesis will not be achieved.

Also to the people who has finished the user study, your response provides a valuable sociologic and linguistic support.

Lastly, I would like to thank my parents, Xiaoping Lin and Jianhua Lian, who have been always standing behind me.

Contents

List of Figures

List of Tables

1. Introduction

One of the popular human-machine interaction (HMI) means is human-machine communications with *Spoken Dialog System* (SDS). To enable HMI by speaking, a SDS needs to recognise and understand user utterances, as following is to interpret within context, then to decide what to answer and finally generate a speech segment as response. Due to the limitation of speech recognition and ambiguity of natural language, *Dialog Manager* (DM), the dialog flow controller, frequently encounters with *errors*, for instance *non-understandings* and *misunderstandings*. *Non-understanding* indicates that the speaker cannot interpret his interlocutor's utterance, meanwhile *misunderstanding* implies the other interpretations than the interlocutor meant (Skantze, 2007). Generally, when a non-understanding or a misunderstanding occurs in a dialog, the effort to repair made by the speakers is called *error handling*. Hence, *error handling* becomes an essential part of a smooth dialog.

1.1. Clarification Issues in Social Dialog Systems

In this thesis, *social dialogues* are human-machine dialogues, which are rather communicational than task-based. Social dialogue is different from most applied spoken dialog systems nowadays. Therefore, it requires clarification techniques which are more "natural" than those in goal-oriented dialog systems(DSs). By "natural" we mean that error handling techniques and answer generations resemble the human-human communication. The most common applied clarification technique in goal-oriented DSs is *generic* clarification, which requires *repeating* or *rephrasing* from the user. Nevertheless, according to the study in human error-handling strategies (Skantze, 2004), *implication* is preferred by speakers to repair occurred or upcoming errors. Moreover, it is assumed in this thesis that in social dialogue, *non-understanding* or *misunderstanding* must not *always* be corrected due to the flexibility and generality of social dialogues. Flexibility and generality refer to frequent topic switches and an unlimited number of topics, which probably happens in social dialogues.
To have an intuitive impression of social dialogue, here is a possible example.

(1) U: I have been to the Orsay **museum** in Paris this weekend.
S: Sounds nice, how was it?
U: That was awesome ...

The above example shows a scenario, which we called *keep talking without complete recognition*. The system (S) does not recognises the word "Orsay", but it recognises "museum", resulting in an assumption that the user (U) was in some kind of museum. The system can

encourage the user to tell his/her impression about this experience instead of signalling a non-understanding to the word "Orsay". However, the user would not realise that the system did not entirely understand his/her utterance. The dialogue seems to keep flowing. In this case, signalling non-understanding is not necessary, since the communication can be continued without a clarification. As Skantze (2007) attemped to avoid signalling non-understanding in a route-planning dialogue, the result confirmed the assumption that the users from the experiment "reported that they were almost understood, despite the numerous non-understandings."

As attempted by Tim Paek and Eric Horvitz (2000), a spoken dialog system called *Quartet* is built with the concept *conversation under uncertainty*. In *Quartet*, grounding process is handled as decision making under uncertainty, where key uncertainties are characterised by Bayesian networks. Local expected utility and value-of-information analyses are used to determine actions that can maximise mutual understanding before boosting grounding.

An interesting question is: when does the **necessity** arise to prompt a clarification question. The non-understandings accumulate to a level that the system is unable to keep the dialog flowing without sufficient recognition support. Besides, excessive clarification questions, arisen mainly by the imperfect speech recognition, discriminate the satisfaction of user experience. In the opinion of Paek and Horvitz (1999), a dialog system is supposed to maintain the conversation with immature speech recognition and imperfect language understanding. Thus, this thesis is aimed to find out the necessary moment to contribute a clarification question in social dialogues and thereby improving the naturalness of spoken social dialogues. The technique *talking under uncertainty* is applied to reduce unnecessary clarification questions approaching to the HM social conversation in a legitimate manner.

The thesis is aimed to detect a proper timing for generating a clarification question and thereby moderating a clarification question in a proper form.

1.2. Thesis Overview

The rest of thesis consists of a **Background** chapter, an **Implementation** chapter called *Predicting Clarification Question in Social Dialogs*, an **Evaluation** chapter and a **Conclusion**. In *Background*, detailed description of the clarification issues are presented and researches providing theoretical support or inspirations are reviewed. The *implementation* chapter delineates comprehensively how to predict clarification questions in social dialogues. To test and verify, two evaluations are accomplished in objective or subjective manner. A *conclusion* is presented at the end, to summarise the evaluations with implementation.

2. Background

As a subcategory of *error handling*, **clarification** is a technique used to repair misunderstanding or non-understanding in dialogues by uttering questions or requests. In some literature Hirohiko Sagawa and Nyberg, 2004; Kazunori Komatani and Okuno, 2004, the process to get uncertain information confirmed is referred to **confirmation**, which belongs to *error handling* in SDS as well. In this thesis, these two concepts are equivalent. The term *clarification* implies a *clarification question*, concurrently *confirmation* is used to refer to a *confirmation strategy*. Generally there are two confirmation strategies, *implicit* and *explicit*. For instance,

(1) U: I need a train ticket to Paris.

S1: Which city do you want to go? explicit confirmation
S2: At what time do you need to arrive in Bali? implicit confirmation

In this example, the word *"Paris"* has a low confidence score from speech recognition. As a result, the system assigns a non-understanding to the word but a partial understanding to the whole utterance here. S1, S2 are two system answers with different strategies. From the example we can see that *explicit confirmation* specially states which point the system does not understand. Whereas, *implicit confirmation* uses related information calculated by system with related keywords and associate topics from database to cover the lack of information in the dialog, moving on at the same time. However, if the related information provided by implicit confirmation is wrong, the user needs to correct it in the next turn, which means one more turn with clarifying information. A *turn* here refers to a conversation turn, which consists of one utterance from each speaker (Skantze, 2004). When non-understanding occurs, the system should pose relevant questions to the user instead of directly specifying non-understanding, which corresponds to implicit confirmation strategy and targeted clarification.

When referring to *Clarifications*, *targeted* questions are preferred in human-to-human dialogues using contextual relevant confirmations, while *generic* clarifications used commonly by SDS such as *please repeat* or *please rephrase*, to indicate non-understanding directly. This implies that in HM social dialogue, implicit confirmation should be exploited as much as possible.

2.1. Spoken Dialog Systems

To enable HMI by speaking, a SDS needs to recognise and interpret user utterance, and then decide how to respond and return an audio segment. Therefore, a SDS contains several components: an *Automatic Speech Recogniser (ASR)*,a *Natural Language Understand-*

ing (NLU) engine, a *Dialog Manager (DM)*, a *Natural Language Generation (NLG)* engine, and a *Text to Speech (TTS)* engine. These components work sequentially as shown in figure 2.1., which can also be seen as a pipeline or a working flow. Figure 2.1. is inspired by the SDS model from F.McTear (2004). Each component takes its predecessor's output as input. SDS depends on ASR transferring voice input into text format. However, the result from ASR is usually unsatisfying with massive mistakes or unrecognised words. Thus a robust *error handling* is the core in a user-satisfying SDS under restriction of ASR.

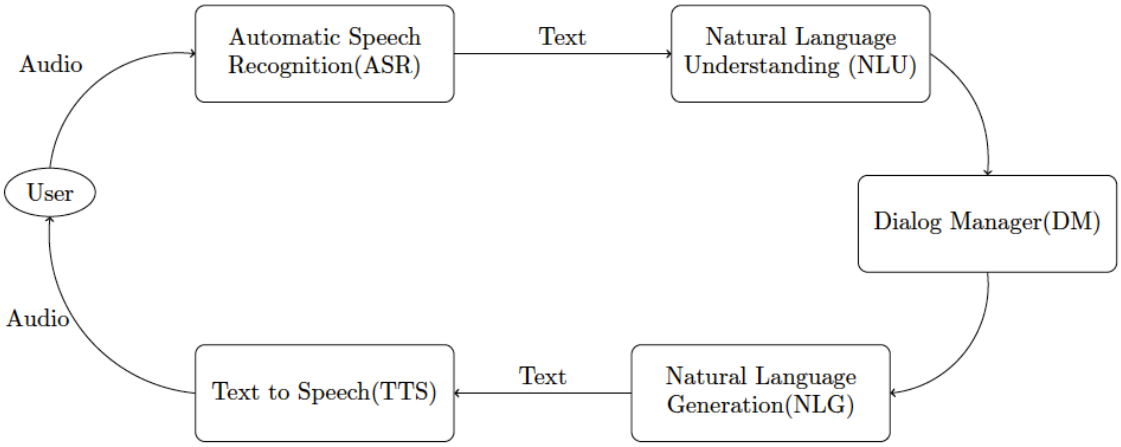

Figure 2.1.: Flow of SDS with five components

Dialog systems (DSs) may be classified into two categories according to their functionality: one is **goal-oriented DSs**, the other is **non-goal-oriented DSs** such as **social DSs**. The difference between them is from the purpose of HMI system and the technique of answer generation, which leads to a variation of confirmation strategies.

2.1.1. Goal-Oriented Dialog Systems

Goal-Oriented DSs usually serve as voice portals for certain interaction applications, which are task-oriented, command-based and constrained to certain specialised topics. The conversation flow in goal-oriented DS usually begins with a command or enquiry from user. For example, a user asks the system to buy a train ticket. The flow will have some certain sections and a clear criteria about when to end the conversation. In this example, the system will ask the information of departure train station, destination and departure time. The flow ends when the system succeeds to buy a ticket for the user. Otherwise the system will repeat to gather the key information. Though, it will halt when the maximum repeat count is reached.

2.1.2. Social Dialog Systems as an Instance of Non-Goal-Oriented Dialog Systems

Unlike a goal-oriented DS, a social DS does not have a preset goal, but intents to simulate the social dialogue found in human-human communication. Commonly, interpersonal goals are foregrounded in social dialogue and task goals - if they exist- in the background (Bickmore and Cassell, 2005). Compared with goal-oriented dialogue, a social dialogue can be more general, that it may have more conversation turns, and contain multiple various topics, but have no clear criteria when to end the conversation unless the user aborts it.

2.2. Clarification Questions in Dialog Systems

According to social studies (Clark and Schaefer, 1989), in the course of conversation, common grounds are accumulated by participants uttering right sentences at the right time. When the mutual belief between participants differs, a process called *grounding* gets activated to update the common ground properly in which also clarification questions are involved. (Clark and Brennan, 1991)

Analogous to human-to-human conversation, HMI also requires grounding, including an essential part: clarification/confirmation questions. Moreover, due to the bottleneck resulting from speech recognition, SDS needs to deal with additional system errors which would less frequently appear in natural human-to-human conversation. These system errors from speech recognition can be caused by speaker variability such as accents, background noise, or unexpected language usage. Thus, the system never really *understands* user utterance, and can only make *hypotheses*. (Skantze, 2007)

When the system assumes that the mutual understanding between user and system differs or the system could not interpret the user utterance, a clarification or repair step occurs. The output from the speech recogniser has a confidence score for each word and phrase, which reveals how strong the system believes that this word or phrase is literally what the user said. However, the system prompts a clarification question every time when the confidence score is low due to the weakness of speech recognition would lead to too many clarification or confirmation questions. This situation would have a negative impact on the smoothness of the dialog flow, which may reduce the user satisfaction. Therefore, it is important to detect the necessity of prompting a clarification question when the system assumes there exists a non-understanding or misunderstanding. Analogously, people will look for *negative evidence* in a conversation (Clark and Brennan, 1991). Once there is no negative evidence showing that speakers have misheard or misunderstood, people can assume that they have understood everything properly by default.

As stated in psychological and linguistic researches, *least collaborative effort* is spent on contributing to a dialogue, which is treated as principle by Clark et al. (1986) To avoid a potential misunderstanding, speakers would rather repair their own utterances than let

interlocutors prompt them to do it. Besides, when the speakers realise that the effort to generate a proper utterance is greater than presenting a provisional utterance and enlist their interlocutors' help, they will apply the non-perfect utterance and rise intonation on specific noun phrases to ask for confirmation. To approximate this character of natural dialogues computationally, a social DS needs to avoid a high number of clarification questions triggered by the speech recognition but still keep the dialogue flow in the right way. Thus, how to *recover from non-understanding* has a great meaning in error handling. Moreover, Colman and Healey (2011) have attempted to discover differences between general or called *ordinary* dialogues and task-oriented dialogues held by humans and thereby gain a distribution of *repairs* in natural dialogue. Their research reveals that people prefer more direct corrections in task-oriented dialogues than in ordinary dialogues. Whereas in ordinary dialogues, the occurrence of self-repairs such like *repeats, reformulations* and *repeats with articulation* is higher.

In *HCRC Map Task Corpus* from Anderson (1991) are dialogues between two speakers, who were sitting on the opposite site of the table and each had a copy of a map. One speaker, the giver, designates the instruction of the route marked on his/her map. Whereas the follower is the speaker following the instruction with no routes in his/her map. Expect the *HCRC Map Task Corpus*, which serves as the corpus for *task-oriented dialogue*, the *British National Corpus (BNC)* for *social dialogue* is also studied by Colman and Healey. Dialogues in the *BNC* are tape recordings of spontaneous conversational dialogues. The tapes recording were contributed by volunteers sampled "demographically" in ages, regions and social classes.

Examples for different types of repair techniques preferred in social dialogues from Colman and Healey (2011) are :

- *Repeat* with *Articulation*
  Follower: Which is due we– **due west**?

- *Clarification* and *Follow-up*
  Follower: So you want me to go ... **east ... then south**?
  Giver: **No, south then east,** we may have a different map

- *Reformulate*
  Giver: Right now, have you got the hot wells?
  Follower: They're over a bit
  Giver: **or hot springs?**

In this paper from Colman and Healey (2011) it is specified that, direct correction is preferred over clarification requests in task-oriented dialogues. See examples below:

- (direct) *Correction*
  Giver: Right the very end of ... paper

Follower: **The very end of the map**?

- *Clarification Request*
  Giver: Past a forge on your right?
  Follower: **Past a what**?

Comparing the *repair* frequency in both corpora, Colman and Healey stated that there are substantial more repairs in task-oriented corpus than in SD corpus. Inspired by this conclusion, the first concern in the following thesis becomes *when is it natural to prompt a question in SD*.

3. Predicting Clarification Questions in Social Dialogue

To gain insight in human social dialogues, the *OpenSubtitles* corpus is exploited for analysing and extracting properties of social dialogues. According to those properties, sentences from the corpus are used to train *Language Models* (LMs).

3.1. Data

3.1.1. OpenSubtitles Corpus

The *OpenSubtitles* corpus[1] is an open source collection of over 20,000 film subtitles in more than 30 languages (e.g. English, Spanish, Chinese, etc.) (Tiedemann, 2009). In this thesis only English subtitles are used. Since it is a large, parallel corpus, it is suitable for machine translation as well, as mentioned by Müller and Volk (2013). The corpus is divided into 30 genres, for instance, action, comedy, documentary. The number of films differs across genres, for example, *comedy* has the most films (378 films), whereas in *adventure* there are only 97 films. To maintain the amount of sentences and thereby ensuring the LM quality, genres *Action, Adventure, Animation, Comedy, Crime, Drama, Horror* with a total number of 1417 files are selected, which are the six genres containing the most films.

3.1.2. Preprocessing and Data Division

The first step in *preprocessing* is **data cleaning**: subtitle sentences together with time-stamps are extracted from the original **XML** data from the corpus and are written to **text** files, meanwhile film duplicates are removed. In original XML data from the corpus, sentences are performed as list of *tokens* with *begin times-tamps* and *end time-stamps*. A *token* can be a single word, a word combination linked with a *hyphen* sign or a punctuation. The list of tokens between a pair of begin and end timestamps forms a sentence of a speaker. If during the same scene there are several speakers, then there will exist multiple pairs of timestamps with *enumeration*. The *question mark* at the end of the sentence is used to distinguish questions and statements. Following this, dialogue pairs are extracted and divided into two categories, one is a pair that ends with a statement, while the other is a pair that ends with a question (general and clarification questions are both included). Either a statement or a question can appear as the first sentence of the dialogue pairs. *General questions* here are referred to questions that do not intent to state a confirmation,

[1] http://www.opensubtitles.org/en/search

whereas *clarification questions* are those made for repairing the *misunderstanding* or *non-understanding* among the speakers. Here we assume that the sentences appearing in the same scene together form a dialogue. To determine if a question is a general or a clarification question, the appearing position of a sentence in the dialogue is considered. If a question is the first sentence in the dialogue then it is a general question. On the other hand, questions appearing afterwards during the dialogue belong to the clarification question catalogue.

Through manual inspection of the dialogue pairs of **clarification questions** it is found that, *WH-questions* dominate the question types. In linguistics, questions can be roughly divided into *yes-no questions*, formally known as *polar questions*, and *WH-questions*, which begin with **interrogative words**, *what, why, where, who, which (five Ws)* or *how* in English (Dryer, 2013; Kearsley, 1976). In particular, short WH-questions, which contain maximally two words , have different usage comparing to general WH-questions. Short or *one-word WH-questions* can be used to express surprise, to indicate non-understanding or just to respond when the follower is called. Among the *five Ws*, the word "*what*" is a questioning form, while the others are *category-specific* question words (Schegloff, 1997). Hence the functionalities between them are slightly different. However, the cardinality of *what?* is not sufficient enough for an individual sub-class. Therefore *one-word WH-questions* is seperated as an individual sub-class from WH-questions.

Moreover, questions beginning with *what* account for nearly half of all WH-questions, as illustrated in Figure 3.1. As a result, *what-questions* with at least three words are treated as a separate sub-class here. The rest of *WH-questions*, those with at least three words and not beginning with "*what*", are divided into *(logical) what-exclusive WH-questions*. The word *logical* denotes that, such questions are usually contextually coherent, as shown in the example[2] below:

(1) A: They are dumb. They say nothing.
B: So how can you understand? You are just making up excuses.

In the example, speaker B asks "*how can you understand*" as clarification to say nothing instead of asking for a strategy such as *what are you going to do?* The word understand implies that speaker A understands "them", but speaker B is not persuaded by A's utterance nor A's behaviour. Following this suspicion, B states his/her conclusion of the situation. In this example, using **how** question is more implicit than **what** question. This sub-class is called *logical what-exclusive WH-question* in order to summarise this property of implicitness.

Another notable phenomenon in the corpus is the predominance of short questions consisting of *really, yeah* or backchannel words like *huh*, which are frequently exploited to signify surprise, to indicate attendance in conversation or to signal mishearing/non-understanding. More to the questioning term "*huh*" with upward intonation, it is not

[2] in this capital examples are from OpenSubtitles Corpus

an item in dictionary, but an articulated and intoned item. Schegloff regards it as a *repair initiation* for pursuit of response. (1997) Thus the "*huh*" is specified in the class name. Furthermore, those *really / yeah / huh* questions do not contribute to the dialog flow, since they are merely providing acknowledgement of attendance or giving approval. As a consequence, *really / yeah / huh* are separated as a sub-class.

(2) A: She's gotta do it with her teeth momma.
B: Huh?
A: You got to put it in her mouth.

In example (2), *huh* signals non-understanding from B. Speaker A adds an explanation to repair the non-understanding, meanwhile the dialogue stays at the phase *explanation of a process*. Unlike *really/yeah/huh*, *confirmation to an assumption (in question manner)* generally provides related information to the dialogue and thereby contributes to unfold the dialogue. *Questions in confirmation to an assumption* are without *interrogative words questions* and without words/phrase *repetition*. Such a question can be a **A-not-A** question offering two possibilities for the answer, ordinarily with distinguishing word "**or**". (Dryer, 2013) It can also be a *polar question* ending with a *rhetorical word*.

(3) A: He's neighbour of mine.
B: Do you want to call him, or should I give him your number?

In this example, speaker A utters a fact, "they are neighbours", but speaker B leads the conversation to a new phase, namely *phone numbers exchanging*. From B's utterance it can be implied that, in B's opinion, A and "he" *should* have each others numbers and make a phone call, which is entirely not mentioned by A. With two alternatives of phone number exchanging, the dialogue comes to a new phase than before. Thus, the sub-class of clarification questions is called *confirmation to an assumption* due to its functionality.

There is one more situation left, namely *Phrase Repetition*. A repeat with question intonation expresses surprise or disbelief according to Norrik (2009), as displayed in following example:

(4) A: We have more to do here.
B: No , we are finished.
A: **Finished**?

Apparently speaker A is quite surprised and disbelieves as well through the upward intonation of the repeated word *finished*. Despite this situation is an illocutionary clarification, invoking *phrase repetition* is still a popular manner in clarification questions. *Repetition* here is referred to rephrasing of all or part of some preceding turn, "most commonly the immediately preceding turn of another". (Schegloff, 1997) A repetition can be elicited by the same speaker or by his/her interlocutors. In this thesis, repetition by interlocutors, a.k.a. *Second-speaker repetition* (Norrik, 2009), is focused. Another methodology used by

phrase repetition in clarification questions is that the second-speaker adds extension to the restatement in order to complete his/her assumption to the interlocutors' intention, as shown below:

(5) A: What about this banana?
B: One **banana** for three of us?
A: Yes, I am starving.

In example (5), speaker B is uncertain about speaker A's intention mentioning "one banana". Speaker B prompts a confirmation as consequence, with his/her own interpretation while rephrasing. Speaker A affirms it directly so we have a completed confirmation process here. As found by manual inspection into the dialogue pairs of the questions, *repetition by interlocutors* represent a clarification question. Hence, *Phrase Repetition* is handled as a sub-class of clarification questions.

We now have six sub-classes of *clarification questions* in this thesis. These are **(one-word) WH-question, (short) phrase repeat, (general) What-question, logical what-exclusive WH-question, confirmation (questions) to an assumption and *Really/Yeah/Huh***.

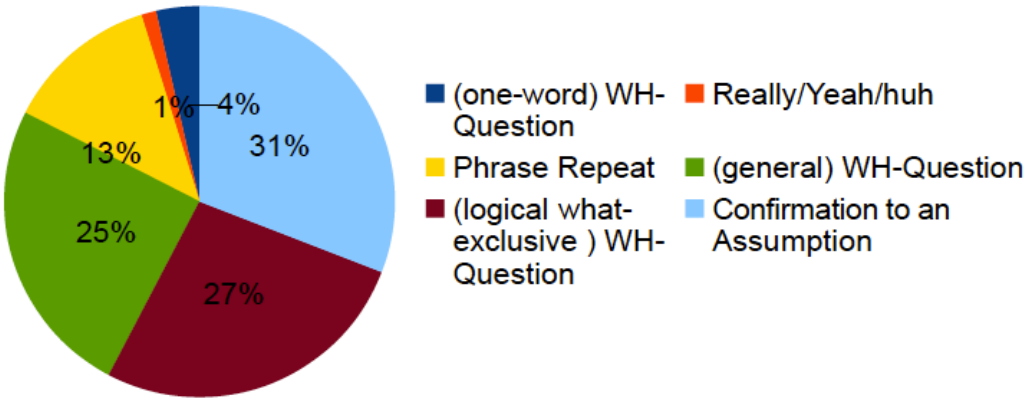

Figure 3.1.: Distribution of sub-classes of clarification questions in the corpus

Examples of the question categories described above:

(one-word) WH-question:

- non-understanding:
  A: I'm sorry.
  B: What?

  A: We have to go.
  B: Why?

- surprising:
  A: He died.
  B: What?

- seeking clarification:
  A: There' s side effects.
  B: Like what?

(short) Phrase repeat:

- A: Sorry, sir, nothing.
  B: **Nothing**?

- with extension:
  A: I owe you a life.
  B: A dog's **life** or human **life**?

Really/Yeah/Huh:

- surprising:
  A: People are trying to kill me.
  B: Really?

- attendance in conversation:
  A: It was really good.
  B: Yeah?

- signal misheard/ non-understanding:
  A: You know , you're a remarkable man.
  B: Huh?

In comparison to *one-word WH-questions*, *general what questions* or *logical WH-questions* are more grammatically complete sentences, which contain a subject, a verb and an object at least. Regarding context, the more complex sentences contribute to more dialogue unfolding, since the second speaker uses related information in the dialogue. A distinct feature in the corpus is the **phrase repetition** which mostly consists of just a word or a phrase, primarily a noun combination. Such partial phrase repeats also correspond to the *Least Collaborative Effort principle* by Clark and Brennan (1991).

General what-question:

- A: Now , so whenever these guys call.
  B: What if it' s during a game ?

- A: I couldn't decide.
  B: Between what and what?

Logical WH-question:

- A: You had a bad day.
  B: Why do you say that?

- A: Down in Florida, it' s 500 an hour.
  B: You like it down there so much, why don't you go buy some oranges?

Confirmation to an assumption:

- A: I am not lying.
  B: So you're going to tell me that he's lying about James too, right ?

- A: I know exactly where she is.
  B: You looking for a date?

From the above examples we can see that *logical what-exlcusive-WH-questions* are more implicit than the what-questions. Questions in *confirmation to an assumption* are mostly logically related and also contribute the most to the dialogue unfolding, as more associate information is provided through the question. Besides, *rhetoric* is also a natural manner to prompt an assumption by stating out a sentence. Adding a rhetorical word in a sentence indicates confirmation, like the word *right* in the first example of **Confirmation to an assumption** above.

Those dialogue pairs provide references about how to construct proper clarification questions in DSs. The dialogue pairs are extracted and divided into these six classes. Meanwhile the first sentence of each pair is also extracted for language model training as the next step. The **interrogative words** (the *five Ws and how*) are used to distinguish *polar questions* and *interrogative questions*. Among the *interrogative questions*, sentences with less than three words are divided as *one-word WH-question*. *What-questions* with sentences longer than three are determined as *general what-questions*, meanwhile the rest of interrogative questions with sentence lengths at least three are mapped to *logical what-exclusive WH-questions*. Questions with less than four words and contains *really, yeah* or *huh, uh huh* are assigned to *Really/Yeah/Huh* group. If a non-pronoun word or a word combination appears in both sentences in a dialogue pair, this dialogue pair belongs to *Phrase Repetition*. The rest of questions, which do not contain *interrogative words* nor *phrase repetitions*, are *Confirmation (questions) to an assumption*. Moreover, each class is partitioned into training and testing sets, with proportion 9 to 1.

3.2. Question Prediction

After the *data preprocessing*, first sentences of dialog pairs are grouped. Each group may have its discernible linguistic properties, which can serve as features for *classification*. Thus, *language model* is employed to capture the eventually existing linguistic properties, i.e. *word distributions*. Subsequently, *classifiers* are trained for question prediction.

3.2.1. Language Model

A statistical **Language Model** (LM) is a probability distribution over word sequences. It assigns a probability to the given word sequence, which reveals the relative likelihood between phrases. More precisely, a LM is a **n-gram** model, i.e., it predicts a word in a given word sequence based on previous (n-1) words. When given a word sequence with four words, say, $\omega_1\,\omega_2\,\omega_3\,\omega_4$. P($\omega_1\,\omega_2\,\omega_3\,\omega_4$) would be

$$\mathrm{P}(\omega_1\,)\mathrm{P}(\omega_2\,)\mathrm{P}(\omega_3\,)\mathrm{P}(\omega_4\,)$$

in **uni-gram** models, whereas the probability would be

$$\mathrm{P}(\,\omega_1\,)\mathrm{P}(\,\omega_2\,|\,\omega_1\,)\mathrm{P}(\,\omega_3\,|\,\omega_2\,\omega_1\,)\mathrm{P}(\,\omega_4\,|\,\omega_3\,\omega_2\,\omega_1\,)$$

in **four-gram** models. For any *n-gram* models except *unigram*, **Bayes' Theorem** is involved, which delineates the probability of an event with dependency on other events. To maintain the consistency, *random variables* are used in our probability cases instead of using *events*. Since the term *random variable* delineates more accurately the linguistic issue than the term *event* does. The theorem is stated mathematically as follows :

$$P(\mathcal{A}|\mathcal{B}) = \frac{P(\mathcal{B}|\mathcal{A})P(\mathcal{A})}{P(\mathcal{B})} \tag{3.1}$$

where $\mathcal{A}$ and $\mathcal{B}$ are random variables. This can be derived from the definition of **conditional probability**:

$$P(a|b) = \frac{P(a,b)}{P(b)} \tag{3.2}$$

$$P(b|a) = \frac{P(a,b)}{P(a)} \tag{3.3}$$

$$\Rightarrow P(a,b) = P(a|b)P(b) = P(b|a)P(a) \tag{3.4}$$

where a $\in \mathcal{A}$ and b $\in \mathcal{B}$. P(a,b) is their *joint density*, i.e., if $\mathcal{A}$ and $\mathcal{B}$ are independent, their joint density can be calculated via :

$$P(a,b) := P(a)P(b) \tag{3.5}$$

With equation (3.2)(3.3)(3.4), conditions between a pair of random variables can be *reversed*, which is made use of by LM calculation.

The LMs in this thesis are trained by **four-gram** models, i.e., for any given sentence $\omega_1 ... \omega_n$, where ω_i are words for i = 1 ... (n-1) and ω_n = *STOP*, the probability under our LMs are

$$P(\omega_1 ... \omega_n) = \prod_{i=1}^{n} q(\omega_i | \omega_{i-3}, \omega_{i-2}, \omega_{i-1}) \tag{3.6}$$

where $\omega_0 = \omega_{-1} = \omega_{-2} = *$ is defined, and the parameter q (ω_i $|\omega_{i-3}, \omega_{i-2}, \omega_{i-1}$) is calculated by the LMs.

LM is broadly used in many *natural language processing* applications, for instance in *machine translation* and *speech recognition* (Christopher D. Manning and Schütze, 2008). Inspired by the idea that a particular type of questions may occur after some certain phrases, which can be achieved by LMs. Here *KenLM*[3] is exploited to train the LMs. *KenLM* is an efficient open source LM toolkit developed by Kenneth Heafield et al. In order to gain a word distribution of different classes, LMs are trained separately on the first sentence of each dialogue pair from its own class. Before training, **punctuations** are removed to ensure the distributions only depend on word sequences.

KenLM offers a python interface, which is used to score sentences. Scores are given as negative float numbers, the closer the score is to 0, the higher the similarity between the tested sentence and the LM. With the python interface we can test the sentences in *test sets*. The test sets of statement/question classes are tested respectively by statement and question LMs. Test sets from clarification sub-classes are tested respectively by sub-classes LMs. During the test, the LM scores each sentence, which indicates the likelihood between the sentence and the LM. Analyses on the scores show that there exists a difference between *LM scores its own test* and *LM scores other classes' tests*, see Table 3.2 and Table 3.3. Further discussion regarding these tables is found in the section 5, *Conclusion*.

		question LM	statement LM
on statement test	mean	-20.4183	-17.5812
	standard deviation	13.0817	9.8685
	maximum	-2.2324	-2.3569
on question test	mean	-15.8241	-19.5741
	standard deviation	7.6718	12.5116
	maximum	-2.2324	-2.3569

Table 3.1.: Question/statement LMs on question/statement tests

[3]https://kheafield.com/code/kenlm/

		assumption LM	logical w-question LM	one word w-question LM	phrase repeat LM	really/ yeah/huh LM	general what-question LM
on assumption test	mean	-13.404	-16.890	-16.776	-17.084	-15.187	-16.860
	std.dev.	7.1567	11.017	11.109	10.911	10.544	11.016
	maximum	-1.775	-1.770	-1.682	-1.871	-1.223	-1.743
on logical w-question test	mean	-18.940	-15.123	-18.788	-19.095	-17.233	-18.811
	std.dev.	13.247	8.473	13.128	12.957	12.648	13.034
	maximum	-1.775	-1.770	-1.682	-1.871	-1.223	-1.743
on one word w-question test	mean	-16.713	-16.669	-14.380	-16.856	-15.024	-16.659
	std.dev.	10.347	10.093	8.325	10.140	9.751	10.105
	maximum	-4.434	-4.187	-4.141	-4.134	-3.829	-4.299
on phrase repeat test	mean	-24.078	-23.942	-23.863	-19.156	-22.102	-23.876
	std.dev.	15.882	15.777	15.898	11.821	14.984	15.716
	maximum	-4.423	-4.527	-4.767	-3.820	-4.559	-4.620
on really/ yeah/huh test	mean	-17.084	-16.938	-16.999	-17.219	-12.880	-17.142
	std.dev.	11.246	10.931	11.154	10.969	7.937	11.092
	maximum	-5.022	-5.219	-5.251	-5.476	-3.912	-5.323
on general what-question test	mean	-18.431	-18.294	-18.253	-18.514	-16.774	-14.643
	std.dev.	13.043	12.754	12.953	12.683	12.362	8.362
	maximum	-1.775	-1.770	-1.682	-1.871	-1.223	- 1.743

std.dev. stands for standard deviation

Table 3.2.: Clarification LMs on Clarification Tests

3.2.2. Handling Different Classes of Questions

When a particular sentence is given, we want to know after this sentence whether a statement or a question is more likely to appear. Furthermore, when the DM determines to prompt a question, the DM needs to decide what kind of question should be prompted at first. Such decision problems can be handled as **classification problems** as well. There is a large number of algorithms solving the classification problem, for instance, *Naives Bayes, Perceptron, K-nearest Neighbour* (Smola and Vishwanathan, 2008), later in this thesis will provide further details to the *maximum entropy estimation*.

3.2.2.1. Classifier

In computer science or *machine learning* a **classifier** refers to the algorithm implementing **classification**. A classifier solves the classification problem of mapping a new instance to one of the given categories, and is trained by a set of category-mapped instances, a.k.a. *training data*. *Classification* can be either *binary* or *multi-class*. A *binary classifier* assigns the given instance either to class A or class B, whereas a *multi-class* classifier identifies among at least three classes. The following figure is taken from the textbook *Introduction to Machine Learning* by Smola and Vishwanathan, 2008.

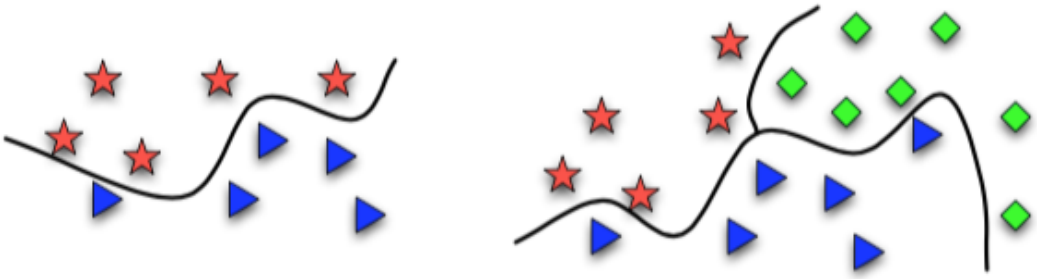

Figure 3.2.: Left: binary classification. Right: 3-class classification

For this thesis, the *MegaM Classifier*[4] is used. The *MegaM Classifier* is a *maximum entropy model* and its algorithm is an implementation of *maximum likelihood*. **Maximum entropy** models, which are inspired by the *Principle of maximum entropy*, provide an elegant possibility to estimate a certain linguistic class occurring with a certain linguistic context, which is popular in *Natural Language Processing* (NLP) (Ratnaparkhi, 1997). Concretely, we want to know the probability of class statement(s)/question(q) occurring in dialog context b, denoted as p(s,b)/p(q,b). More accurately, for clarification questions, the probabilities of clarification question sub-classes occurring in context b are desired.

Maximum Entropy

Maximum-entropy estimation is a variety of statistical inference offering probability distributions on the dependence of partial knowledge. It is "maximally noncommittal with

[4]http://www.umiacs.umd.edu/ hal/megam/

regard to missing information" (Jaynes, 1957). More explicitly, if p(a,b) is a correct distribution which maximises entropy, let $\mathcal{A}$ denote the set of possible classes and $\mathcal{B}$ denote the set of possible context (in our case it is the set of dialog pairs), then p should maximise the entropy H(p):

$$\text{H (p)} = \text{-} \sum_{a \in \mathcal{A}, b \in \mathcal{B}} p(a,b)\, ln\, p(a,b)$$

The consistency of *partial information* should be thereby maintained (Ratnaparkhi, 1997).

Maximum Likelihood

The *Maximum Likelihood principle* is to maximise the *joint probability distribution* P(x| ϕ) via selecting the value of ϕ, with x = $(x_1, ..., x_n)$ being a vector in the sample from given random variables and ϕ is a parameter from some parameters space. If the observations $x_1, ..., x_m$, a.k.a. *training set*, are sampled independently and identically distributed, then the maximum likelihood principle is equivalent to finding the maximum of:

$$\phi^* = argmax_\phi \prod_{i=1}^{m} P(x_i|\,\phi) = argmax\, log \prod_{i=1}^{m} P(x_i|\,\phi) = argmax \sum_{i=1}^{m} log\, P(x_i|\phi) \quad (3.7)$$

The logarithm is used for effective calculation, i.e, it turns multiplicative into additive. It is proven that *maximum entropy* and *maximum likelihood* are *duals* of each other (Shashua,2008). I.e., if an exponential form is assumed, the result of searching for the *most likely distribution* agrees with the result for the *maximum entropy distribution.*

The MegaM classifier maximises a *posterior* optimisation model of parameters. Let x be an instance out of a random variable set $\mathcal{X}$, let p be the *sampling distribution* of x, let ϕ denote a parameter, then the function :

$$\phi \;\mapsto\; p\,(\,x\,|\,\phi\,) \quad (3.8)$$

is known as the *likelihood function* of ϕ. Now assume that a *prior distribution* over ϕ exists, denotes as p(ϕ). This allows that ϕ can be treated as a *random variable* in our case. Through **Bayes' Theorem**, the prior distribution (p(ϕ)), can be converted to *posterior* probability (p(ϕ | x)) :

$$p(\,\phi\,|\,x\,) = \frac{p(\,x\,|\,\phi\,)\,p(\,\phi\,)}{p(\,x\,)} \quad (3.9)$$

where p(x | ϕ) is the maximum likelihood function. The *Maximum a-posterior (MAP)* estimate is defined as:

$$\hat{\phi}_{MAP} \;=\; \underset{\phi}{\text{argmax}}\; p(\,\phi\,|\,x\,) \quad (3.10)$$

Note that ϕ is a seen parameter, hence p(x) does not depend on ϕ, with equation (3.10) (3.11) we have :

$$\begin{aligned}\hat{\phi}_{MAP} &= \operatorname*{argmax}_{\phi} p(\phi \mid x) \\ &= \operatorname*{argmax}_{\phi} \frac{p(x \mid \phi)\, p(\phi)}{p(x)} \\ &= \operatorname*{argmax}_{\phi} p(x \mid \phi)\, p(\phi)\end{aligned}$$

which infers via choosing the value of ϕ, the model can be maximised by known prior distributions (Robinson,2012).

Feature

In terms of describing individual instances, an instance can be seen as a set of *features*. Features are measurable properties of instances, e.g., *binary* features "male" or "female", *categorical* features "A", "B", "AB" or "C" for blood type. With the given feature set of the instance, a classifier can predict the class of the instance comparing the feature parameters of the instance with those of each class. Some features are complementary, for instance, "male" and "female" (assuming that there are only two genders), a human can be either male or female. More explicitly, when a human is not a male then she must be a female. Some features are more vague, for instance, a persons name ending in "a" as a feature for female names but there are exceptions, for example "Joshua". In this name example,*vagueness* infers that, even though knowing a human's name is ending in "a", we can still not absolutely confirmedly say that it is a girl's name (Steven Bird and Loper, 2009).

Given a sentence, the scores from each LM are taken as features to train the classifiers in this thesis. A binary classifier for question/statement and a multi-class classifier for the six clarification questions sub-classes are trained. E.g., for the binary classifier between questions and statements, a possible segment in the training set is displayed as below:

0 F1 -21 F2 -24

where 0 represents the question class; F1 stands for *Feature one*, following the score of the question LM; and F2 stands for *Feature two*, following with the score (the -24 at the rightest position) from statement LM. The first number "0" indicates that this sentence is an instance of *question class*, analogously "1" would symbolise a statement. Meanwhile the sentence is scored by question LM with -21 and by statement LM with -24. The feature pair forms a *tuple*, consisting of a feature name and its value. A feature vector for a sentence for clarification questions has thirteen elements, which are one symbol for the class and six feature pairs. A list of such vectors with five elements is used to train the binary classifier, analogously a list of vectors with thirteen elements is applied to train the multi-class clarification question classifier. Moreover, for the guarantee of classification quality, sentences in the training set of a classifier contain at least 5 words. For the binary classifier *question/statement*, respective LMs were trained with 4-gram *KenLM* models, which trains substantially the data with from unigram model to 4-gram model.

3-gram *KenLM* models were employed, since a 4-gram model could not applied on the class *Confirmation question to an assumption*. It is because the amount of sentences with at least 4 words is not sufficient for *KenLM* to train a 4-gram model. Sentences with at lease 5 words contain entire information of our LMs, i.e., eventually 4-gram models can applied on those sentences, hence no information of LMs is faded away. Sentences account for training the *question/statement classifier* 98,048; concurrently the training set for *clarification questions sub-classes classifier* possesses 50,620 sentences.

The **MegaM** algorithm calculates a distribution model based on the given data. The model for the question/statement binary classifier has the following parameters:

BIAS	-0.02846392802894115448
F1	-0.01892893761396408081
F2	0.01714214496314525604

Note: F1 standes for *question*, F2 for *statement*

Table 3.3.: Distribution for question/statement classifier

The distribution model for sub-classes is as displayed below:

BIAS	0.0000	-0.0099	0.0014	-0.0095	-0.0117	-0.0094
F1	0.0000	0.0389	-0.1835	-0.0114	0.0578	0.0044
F2	0.0000	0.0246	-0.1597	-0.0077	0.0448	-0.0001
F3	0.0000	0.0036	0.6535	0.0708	0.1363	0.0903
F4	0.0000	-0.0331	-0.0799	-0.0037	0.0158	-0.0012
F5	0.0000	-0.0623	-0.0933	-0.0485	-0.0245	-0.0549
F6	0.0000	-0.0500	-0.0768	-0.0368	-0.0187	-0.0460

Note: the numbers are cut down in 4th digit after float comma, but in the original file these contains with 20 digits after float comma.
F1 , … , F6 correspond to the six sub-classes, respectively.

Table 3.4.: Distribution for clarification question sub-classes

3.2.2.2. From Different Classes to Prediction

With the distributions estimated by *MegaM*, we can predict after the given sentence whether a question or a statement should follow. Explicitly, the question/statement classifier first identifies the given sentence's class. If the binary classifier predicts that it should be followed by a question, the DM analyses further what kind of question should follow up. When a clarification question is supposed to occur, the multi-class classifier can prognosticate which question sub-class should occur based on the given sentence.

4. Evaluation

For demonstrating the prediction of clarification questions in social dialogues, both as **objective evaluation** (i.e. the result from **Classification**) and as **subjective evaluation**, i.e. a *user study*, are exploited. The *objective evaluation* offers a demonstration of how to implement the *clarification question prediction* in a dialog system; whereas the *subjective evaluation* provides a sociological and linguistic support to the thesis.

4.1. Objective Evaluation of Classification (MegaM)

Two classifiers are built in this thesis, a binary classifier for the prediction between questions and statements and a multi-class classifier for the further subdivision of clarification questions. For evaluation, a random test and a last test for each classifier are generated. More precisely, sentences in the *random test* are sampled randomly within the last 2% of sentences of each category, samples in the *last test* are the sentences having at least five words in last 2% sentences in the training set of each class. These sentences are cut out of the training data and only applied for test. Generally, a longer sentence reflects more actual accuracy of the result of classifier than a short sentence. Since our LMs are trained by from unigram to 3-gram or 4-gram models. A longer sentence contains valuations of 3-gram or 4-gram models, which reflexes the word combinations better than the valuations of unigram or bigram models. Therefore the training data for classifiers are filtered to ensure the minimum length of the training sentence. Moreover, the numbers of sentences in the random test and the last test should not have great difference, so that the results remain comparable to each other. The cardinality of test exemplars are approximately evenly distributed in last tests, and are evenly distributed in random tests.

Each test input (i.e., a sentence) is scored by different LMs at first, while the original class symbol is written to an extra file called “class”. Note that the order of the classes symbols corresponds to the order of the sentences in the input file. The LM scores serve as a feature value in a file called “test.data”. Together with the sentence’s class symbol, features and their values form an entry for test data. A test data entry of a binary classifier has the following structure :

0 F1 -21 F2 -24

note that it is the same in training data.

The classifier evaluates such data, and generates a result entry with its expected class symbol and its probability. A result entry of a binary classifier may look like this :

0 0.48875877714877202784

Comparing the predicted class symbol and its original symbol in the file "*class*", an error rate is thereby calculated. For the binary question/statement classifier some error rates are presented in the Table 4.1. :

random test:	1593	/	3860	=	0.412694
last test:	289	/	3845	=	0.0751625

Table 4.1.: Error rate of binary classifier question/statement

For the multi-class clarification question subdivision classifier the following error rates are obtained :

random test:	847	/	1050	=	0.806667
last test:	904	/	1109	=	0.815149

Table 4.2.: Error rate of multi-class classifier clarification question subdivision

Comparing the evaluation results from the binary classifier in the Table 4.1., the error rate of the last test is "astonishingly good". One possible explanation is that, the sentences in last test are at least 5 words whereas there was no restriction on the sentence length in the random test. Moreover, the sentences for training the classifier are with at least 5 words, hence, the result in the last test is accordingly better than in random test. Additionally, the result in random test is better than *random*, i.e., in our case, a sentence has 50% probability to be classified as a question.

However, when we look at the Table 4.2., where the results are rather "unsatisfying". One argumentation of the high *error rate* is that the samples for the class *ryh* and *oww* are not sufficient to present their classes properties. The sentences applied for training LMs and classifiers among different classes may not contain distinguishable differences. I.e., from the samples there is no convinced evidence of their classes features. For instance, the sentence: *I do not know*, appears in all the sub-classes training and test data. Such overlapping data, which are not rare in the corpus, burdens the LMs model estimating the real distributions. The result from LMs for clarification subdivision is presented as below:

LM	score
asp	-9.681313514709473
log	-9.632256507873535
oww	-9.883711814880371
rep	-9.948943138122559
ryh	-10.11093616485596
wht	-9.890035629272461
mean	-9.857866128285727
std. dev.	0.176699104481545

Abbreviations of clarification sub-classes:

asp :	confirmation question to an assumption
log :	logical what-exclusive WH-question
oww:	one-word WH-question
rep :	phrase repetition
ryh :	really / yeah / huh
wht :	general what-question

Table 4.3.: A sentence with similar scores from different LMs

Table 4.3. explains Table 3.2. (the scores by six LMs respectively on six classes). The immensely small standard deviation in Table 4.3. reveals there is no intention on classification given by the scores from LMs. Besides, such phenomenon is not extraordinary in our classifiers and LMs. In this thesis there are six sub-classes, it may be not common that the same sentence appears in all six classes, however the probability that it appears in three or four training sets is not low. Besides, overlapping can be indicated statistically by Table 3.2. as well, take the class *one-word WH-question (oww)* and *really/yeah/huh (ryh)* for instances. The *oww* LM, i.e., sentecens in the *oww* test are followed by an instance of the class *oww* and for test analogously, has the maximum -5.251 on the *ryh* test, in the corpus. While the *ryh* LM has an arithmetic mean of -12.880, and a standard deviation of 7.937 on *ryh* test, the maximum of the *oww* LM locates in the interval of (-12.880-7.937, -12.880+7.937), a.k.a. (-20.817, -4.943), thus the distribution of *oww* LM and the of *ryh* LM is not discrete to each other. Partial overlapping in training data discriminates the differences between features, which leads to unsuccessful classifiers.

Moreover, the distribution of classes is rather unbalanced in the corpus, the data perplexes the classifier. In the ideal case, each class should possess the same proportion in training data for classifiers and LMs. Nonetheless, if the minimum of sentence cardinalities is applied, say, class A owns 1,000 sentences, whereas the other classes possess 20,000 for each, then 1,000 sentences are taken from each class. So the generated distribution may not approximate the real distribution, since the rest of classes, i.e. 1/20 of the data, may not be enough to extract the features, which can objectively describe the properties

of the classes. However, for the rest of classes 20,000 instances are all taken to the training set for a classifier and from class A are only 1,000 available for training. Statistically, for the classifier class A does not "exist", since the proportion of class A is significantly small thus can be statistically omitted.

Another hindrance from the corpus to a successful classifier is that, there is no allusion provided whether the ensuing sentence of the current sentence is from the same speaker or from his/her interlocutor. This hinders a correct extraction of dialogues. Concretely in our case, there is no evidence on a clarification question, revealing whether it is a *self-repair* or a *clarification on his/her conversation partner*. *Self-repair* and *other-initiated repair* are two distinct categories of *repair* in dialogues (Schegloff, 1997). *Clarification questions* in spoken dialog system belong to *other-initiated repair*. Some *self-repair* linguistic distinctions in training data confuse the LMs and the classifiers for*clarification questions* for dialog system use.

4.2. Subjective Evaluation

To see whether the results from classifiers correspond to the user expectation, a user study is performed. The user study has 15 query entries. The first sentence of a query is given either a question or a statement, followed by 6 alternatives corresponding to the six sub-classes of clarification questions as the responses to the first sentence. Thus the first sentence and each of the six questions form respectively six dialogue pairs. The study participants should score each response's appropriateness in 5 ranges. In the study, the scores range from 1 (i.e. *inappropriate*), through 3 (i.e. *neutral*), to 5 (i.e. *appropriate*).

Query sentences in the study are chosen from the *OpenSubtitles* corpus in the following described manner: firstly dialog pairs which are likely to have six different responses are selected into a set. For instance, *Greetings* occur frequently in the corpus, commonly seen as in the following manner:

(1) A: Hello.
B: Hi, how are you?

Example (1) does not fulfil the criterion of a clarification question, since *how are you* is not asking for any kind of clarifications. However, some clarification questions in greetings are still found:

(2) A: Well, hello there, little fella.
B: Do I know you?

Speaker B's utterance in example (2) denotes a clarification question. Considering *greetings* with the small talk triggered by greetings are common in social dialogues, example

(2) is used in the study. For five more reactions, dialog pairs containing the word *hello* is selected. Along those filtered dialog pairs, if a *hello* appears in the first sentence and the second sentence denotes a clarification questions, which states different class to the already used class, then the second sentence is used for further reaction to example (2). A query in the study has the following form:

		1	2	3	4	5
A:	Well, hello there, little fella.					
B1:	Where have you been ?					
B2:	Hello ?					
B3:	Yeah, what's up ?					
B4:	Yeah ?					
B5:	Do I know you ?					
B6:	Who ?					

Table 4.4.: An example of a query in user study

In Table 4.4., B1's utterance is an instance of the class*logical what-exclusive WH-question*; B2 prompts a clarification with *phrase repetition*; B3'question is a *general what-question*; B4 states an instance of *really/yeah/huh* class; B5's reaction, also the original one in corpus, belongs to the class *confirmation to an assumption*; meanwhile the sixth reaction corresponds the to class *one-word WH-question*.

Queries in this study are constructed analogously to the *greeting* situation: a dialog pair in a certain context (for instance greeting, apology) from the corpus was taken as the initial pattern; next, pairs in the same or similar contexts were searched through. If the first sentence of the pair was similar to the first sentence in the fixed pair, then the second sentence of the newly found pair would be added to the query set; the process is repeated to find dialog pairs in similar context until six different responses are gathered, one per clarification class. If there is no suitable existing question to the pattern of a class, a clarification question will be generated manually in case of need. Hence, a query set is obtained. During the construction of queries, the variety of topics is maintained as much as possible. Besides, involved dialogue pairs are maintained "social" by attempts. "*Social*" hereby means that such contexts are frequently seen in all-day dialogues, for instance, *greetings, apology*, or are related to social relation or social life. With these two mentioned properties, the user study provides a realistic aspect revealing how people evaluate these six clarification sub-classes in various social scenes.

We have received 18 completed responses, a rough analysis is displayed in Table 4.5., details on raw survey data can be found in the *First Appendix Section*.

	mean	std. dev.	min	max*
asp :	3.1953	0.7591	2	4.22
log :	3.0293	0.8212	1.94	4.11
oww:	3.6113	0.8006	2.28	4.72
rep :	3.5367	0.4468	2.89	4.28
ryh :	3.6293	0.5474	2.61	4.39
wht :	3.1107	0.7767	1.61	4.39
		arithmetic mean**		3.3521
		standard deviation***		0.2702

Abbreviations of clarification sub-classes:

asp :	confirmation question to an assumption
log :	logical what-exclusive WH-question
oww:	one-word WH-question
rep :	phrase repetition
ryh :	really / yeah / huh
wht :	general what-question

* : here is the minimum/maximum of each class's means from each query
** : here is the mean of the six means of the six sub-classes
it is also the mean of all scores in the study
***: here the std. dev. is for means

Table 4.5.: Analyse on the user study

The Figure 4.1. is illustrated based on the Table 4.5., the columns are the mean values, the lines on the columns are the standard deviation.

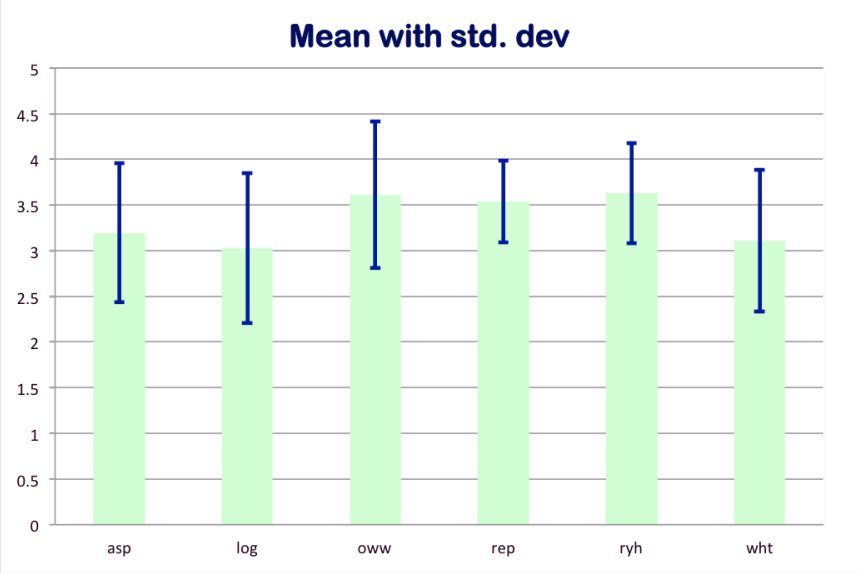

Figure 4.1.: Descriptive chart for the means with std. dev. of the user study based on Table 4.5

From the the Table 4.5. or the more descriptive Figure 4.1., we can see that the means of the six classes are close to each other, with their *standard deviation* 0.2702. The *arithmetic mean* of each class is the average of all scores of the same class from the 15 queries in the study. The six means vary from 3.0 to 3.7, which means that the participants regarded them generally as slightly better than "so-so". The class *Really / Yeah / Huh (RYH)* has the highest mean and second-highest maximum, while the class *logical WH-question (log)* has the lowest mean and global minimum. The query entry with highest average score (i.e. mean) is this pair:

(3) A: I'm sorry.
B: Why ?

Whereas the entry with lowest average score is the example (4):

(4) A: Take as many as you like.
B: Where are you going ?

The second sentence in example (4) is picked up by the query constructing process described above. The original pair, that contains the utterance of B, the word "*take*" is

included as well, and the situations are similar. Thus, B's reaction is considered as a possible machine generated clarification question and therefore is selected into the query. Though B's reaction is not that intuitive, the dialog pair in example (4) can still occur under certain circumstances, hence example (4) bases strongly situational. When we are taking a look back to example (3), utterances in example (3) are relatively brief. Though there is no sufficient evidence in linguistics or sociology that reveals that the briefness of an utterance is positively related to its generality (i.e., a brief utterance is proper for more different situations than a longer utterance), the results of the user study indicate this tendency. The class *one-word what (oww) question* and the class *ryh* possesses the two highest average scores and two maxima. Sentences in these two classes contain no more than three words. Especially in the class *ryh*, the sentences there do not refer to a concrete object, hence the utterances from class *ryh* are rather general.

Despite the fact that *ryh*-questions suit to the most of situations, WH-questions with more accuracy are preferred if they are uttered in a correct manner.

(5)	A:	I have to talk to you.	
	B1:	Huh ?	*ryh*
	B2:	What are you going to talk about ?	*general what-question*

In example (5), B2's reaction is scored with an average of 4.39, while B1's brief backchannel utterance "*Huh?*" is scored with 3.33 in average. In this scene the general what-question can overtake the backchannel utterance owing to B2 shows more attendance than B1. In fact, the *huh* may indicate the absence in conversation in some occasions. B2's utterance presents the attendance in conversation via a re-statement, which corresponds to the first level of conversation , i.e. *committing attendance*, according to the *four-level model*, a.k.a *collaborative / conversational model*, from Clark and Wilkes-Gibbs (1986). Calling back to the brief introduction about *confirmation strategies* in Section **Background**, B2's utterance is exactly an *explicit* confirmation. Example (5) inspires that if there is adequate evidence supporting to state a clarification question explicitly on the non-understanding position, a *general what-question* would be a better choice than a *ryh*-question for the DM .

4.3. Results of Evaluations

Despite the high error rate of the clarification questions sub-class classifier, the result from the classifier agrees roughly with the result from the user study. It is noticeable in the user study that there exists no distinct difference among classes regarding *appropriateness*. Taking a look at the feature set of the classifiers, we found that the differences among classes are small as well, see Table 3.2., *Clarification LMs on clarification sub-classes*. Our sub-class classifier identifies willingly to a certain class repeatedly. The user study infers a tendency that concise question can be applied in most situations, whereas if possible, explicit clarification question is preferred. When we take a look back to the pie

chart in Section 3.1, in Chapter **Prediction Clarification Questions in SD**, Figure 3.1., *WH-questions* dominate in the corpus. *WH-questions* except the *one-word WH-questions* are mainly explicit clarification questions, since in WH-questions, the speaker needs to know *exactly* what to ask. Despite the enormous amount of WH-questions, during the construction of the user study, it is rare to find a *relatively suitable* WH-question to each query owing to its *explicitness*. Accordingly, the average scores from WH-questions, i.e, *logical what-exclusive WH-questions* and *general what-questions* are ranked at the bottom. In contrast, *ryh* and *one-word WH-questions (oww)* can be added after most of the sentences in the user study, which are holding the first two places in average score ranking. Though they are less probable to be found in the corpus, since *ryh* corresponds merely to 1% of the whole clarification questions while *oww* holds 4%.

More explanation to *explicitness* here, regarding B2's utterance in example (5), "*What are you going to talk about?*", some information can be inferred. Information 1: A, denotes B2's conversation interlocutor, wants to talk with B2, which is exactly what A requires in the example; information 2: B2 is probably not eager to *talk to A*, otherwise there is no need to ask *what will be talked*. To put it in another way, if B2 knows what A is going to talk about and why A is so eager (can be inferred from *have to*), B2 can respond with an affirmation, such as *ok*. Nevertheless, from B1's reaction, the backchannel *huh*, there is no further information on the context. There is no clue indicating which part of A's utterance confused B1. Does B1 not understand why A uses *have to*, or does B1 have no idea about what A is going to talk about, or does B1 not understand the whole situation, that *A is eager to talk to her/him*? It might also be that B1 was not listening to A, so the *huh ?* here might also serve as an informal alternative for *pardon ?* Many possibilities can explain why B1 utters a *huh* as clarification question. But through B2's utterance, exact non-understanding reference is revealed. Hence we say B2's question is *explicit*, while B1's question is *general* or *vague*.

One participant gave the feedback that he found the user study is *kind of vague* thus the questions are *strange* and *somehow impolite*, which seems a paradox to the results of the user study. However, the best average score does not indicate that users prefer vague questions the most. It merely presents the fact that, the main group of the participants considers brief and general questions suitable in most of the situations, though they do not have to be (and actually are not) the most preferred alternative in clarification issues. Along this, the phenomenon that the class *general what-question* (wht) has the second best maximum meanwhile the second worst average. Once a *general what-question* suits the situation, then it is with higher user satisfaction as well. But if the *wht*-question is presented inappropriately, then the user satisfaction will be less than average, which happens unfortunately more often than the satisfying case.

If we think in a reversed way, *huh* can be applied to all these situations, indicating non-understanding to partial information without explicit information, indicating absence in the conversation, indicating confusion on the whole situation (here, *why A have to talk with B1*), which corresponds to class *ryh* having the best average score in the user study.

Moreover, as stated by Schegloff, *huh* presents the least information of the trouble-source to its recipient, e.g., where is the source located and what is the trouble with it. Besides, it is so "powerful" that nothing more is needed to deploy the question; "even a *putative* trouble-source is adequate to deliver the problematic utterance" (Schegloff , 1997). Whereas B2's utterance can only be applied in situation, *A and B2 need to talk*. Furthermore, during the construction of query set, we have noticed that generating a *ryh* entry is easier than finding an applicable *general what-question* or *logical wh-question*.

Nevertheless, a frequent occurrence of *really?/ yeah?/ huh?* diminishes the interest from the interlocutor continuing the dialogue, since this provides a negative evidence of engaging in the interaction. Once a speaker considers his/her interlocutor has few interest in the conversation, the speaker may end the dialogue earlier than usual. If the system is regarded as lack of interest in the dialogue, it leads to a lower user satisfaction which has been attempted to avoid.

Though the query set is constructed to remain the *naturalness* with effort and concern, feedbacks about *unnaturalness* are still given by 5 participants out of 18 responses. Considering the query set is not moderated completely by machine, the negative feedbacks draw a brief overview to the difficulties with generating natural human-machine-dialogue, detailed discussion can be found in Chapter **Conclusion**.

A proper utterance of *explicit clarification question* is commonly desired, however, the ability of DS limits the presentations of accurate clarification questions. I.e, the effort invested by DM to generate an accurate *WH-question* is greater than the effort to generate a vague clarification question from class *ryh* or class *one-word WH-question* (oww). Precisely, take class *ryh* and class *logical WH-question* for instances, the DM simply needs a non-understanding indication, then eventually decides which of the three possibilities (really ?, yeah ?, huh ?) suits the best, which is in most context no significant difference. However, if the DM is for generating a *logical WH-question*, the DM needs accurate information about which part of the context is not interpretable, then chooses an *interrogative* word accordingly, and finally organises an interrogative question. Reviewing the restricted ability of *ASR*, multiple non-understandings to the user input might be notified, then the DM encounters with a decision problem which non-understanding should be informed to the user. In fact, encountering with multiple non-understandings in one user utterance, informing a non-understanding to the whole utterance repairs the dialogue most effectively. For instance, the system may send a *repeat request* on the user, such as *I don't understand, can you please repeat it?* Such utterance seems *artificial*, since it rarely occurs in the corpus. Instead of a explicit *repeat request*, people prefer a brief *huh?* or *what?* in informal occasions, namely *social dialogue*.

Back to our case, *ryh* and *oww* are so pragmatic in general usage that they should remain as individual classes. Thus *how to construct a classifier in such case* is the key to reduce the high error rate of the classifier. More detailed discussion is followed in Section **Conclusion**.

5. Conclusion

As stated in Section **Introduction** and Section **Background**, conversations under *uncertainty* are not extraordinary in human-human-dialogues. Besides, the non-understanding or misunderstanding reported by the immature *ASR* are more frequently than common dialogues between humans. Towards a more recognition-error-tolerant social dialog system, experiments on *clarification issues* are performed in this thesis. Explicitly, an experiment on prediction of the occurrence of a question or a statement depending on a given sentence. Meanwhile, the other is on classification of the sub-class of a clarification question based on a given sentence.

The first step of the experiments is the *data preprocessing*. The original data from the corpus was extracted as dialog pairs. The dialog pairs ending with a question were grouped to the class *question*, while the pairs ending with a statement were mapped to the class *statement*. In the question class, dialog pairs engaged to a clarification question would be selected as the clarification question class. Especially, in the six sub-classes of clarification questions, thereby, the clarification question dialog pairs were partitioned into the six sub-classes.

In the aftermath of *preprocessing*, 8 LMs (i.e., the question, the statement and the six clarification sub-classes LM) and 2 classifiers (i.e., one for classifying questions/statements, the other for identifying the sub-class among the six clarification sub-classes) were trained by the first sentences of the corresponding class. Particularly, the *features* used to train classifiers were the scores of the LMs valuing on classifier-training sentences.

For a subjective evaluation as well as a linguistic provision of our experiments, a user study on clarification questions in different situations was involved. As discussed in the chapter **Evaluation**, though vague and brief question sub-classes have better average in the user study, there are feedbacks form the study participants meaning the queries were *strange* or *vague*. Hence, it is inspired that in a "natural" human-machine social dialogue, a balanced proportion of vague and brief clarification as well as accurate and proper explicit clarification questions should be maintained. However, the experiment result for the subdivision clarification questions (i.e., the result from MegaM) could be improved in a better way.

5.1. Future Work

As mentioned above, a balanced combination of vague and explicit clarification questions can improve the "naturalness" of the system, which could be implemented in the near future.

One potential procedure for classifying clarification questions is that, before the subdivision, first determine whether a *vague* or an *explicit* clarification question should be initiated. The class *ryh* and *oww* have the same magnitude of training set, whereas the class *general what-question*, *confirmation question to an assumption* and class *logical what-exclusive WH-question* have similar data volumes. A binary classifier for *vague/explicit* clarification questions can be constructed analogously to the classifier for *question/statement*. Finer subdivision under each clarification classes (*vague*/*explicit*) can be processed analogously to the multi-class classifier on the six classes in the thesis, though with relative even distributed data volumes.

To eliminate the differences of data volumes among classes, more corpora can be exploited. Precisely, for class *ryh* and class *oww*, data from other social dialogue corpora (for instance, the *British National Corpus*[1]) can be involved, so that the two classes *vague* and *explicit* can be approximately of the same magnitude.

Since the results from our binary classifier are relatively delighting, which inspires us to train six more binary classifiers instead of one multi-class classifier. Explicitly for finer subdivision of clarification questions, complementary binary classifiers can be trained, classifier for *ryh* or *not-ryh*, for*general what-question* or *not-general-what-question*, and so forth. With six results from each complementary binary classifier, the sub-class with best result can be selected as the decision for this classification process.

[1] http://www.natcorp.ox.ac.uk/

Bibliography

[1] A Anderson. *HCRC Map Task Corpus*. http://groups.inf.ed.ac.uk/maptask/. 1991.

[2] Timothy Bickmore and Justine Cassell. *Social Dialongue with Embodied Conversational Agents*. 2005.

[3] Prabhakar Raghavan Christopher D. Manning and Hinrich Schütze. *Language models for information retrieval*. Cambridge University Press, 2008.

[4] Herbert H. Clark and Susan E. Brennan. "Grounding in Communication". In: *American Psychological Association* (1991).

[5] Herbert H. Clark and Edward F. Schaefer. "Contributing to Dicourse". In: *Cognitive Science* 3 (1989), pp. 259–294.

[6] Herbert H. Clark and Deanna Wilkes-Gibbs. "Referring as a collaborative process". In: *Cognition* 22 (1986).

[7] Marcus Colman and Patrick G.T. Healey. "The Distribution of Repair in Dialogue". In: *Cogsci 2011* (2011).

[8] Matthew S. Dryer. "Polar Questions". In: *The World Atlas of Language Structures Online* (2013).

[9] Michael F.McTear. *Spoken Dialogue Technology*. Springer, 2004.

[10] Teruko Mitamura Hirohiko Sagawa and Eric Nyberg. "A Comparison of Confirmation Styles for Error Handling in a Speech Dialog System". In: *ICSLP, 8th International Conference on Spoken* (2004).

[11] E.T. Jaynes. "Information theory and statistical mechanics". In: *The physical review* (1957).

[12] Tatsuya Kawahara Kazunori Komatani Teruhisa Misu and Hiroshi G. Okuno. "Efficient Confirmation Strategy For Large-Scale Text Retrieval Systems With Spoken Dialogue Interface". In: *International Conference on Computational Linguistics* (2004).

[13] GregP Kearsley. "Questions and question asking in verbal discourse: A cross-disciplinary review". In: *Journal of Psycholinguistic Research* (1976).

[14] Mathias Müller and Martin Volk. "Statistical Machine Translation of Subtitles:From OpenSubtitles to TED". In: *Language Processing and Knowledge in the Web* (2013).

[15] Neal R. Norrik. "Functions of repetition in conversation". In: *Interdisciplinary Journal for the Study of Discourse* 7 (2009).

[16] Tim Paek and Eric Horvitz. "Uncertainty, Utility, and Misunderstanding: A Decision-Theoretic Perspective on Grounding in Conversational Systems". In: *AAI Fall Symposium on Psychological Models of Communication in Collaborative Systems* (1999).

[17] Tim Paek and Eric Horviz. "Conversation as Action Under Uncertainty". In: *UAI '00 Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence* (2000), pp. 455–464.

[18] Adwait Ratnaparkhi. *A Simple Introduction to Maximum Entropy Models for Natural Language Processing*. Tech. rep. Computer and Information Science, 1997.

[19] Peter N Robinson. *Parameter Estimation: ML vs MAP*. 2012. URL: http://www.mi.fu-berlin.de/wiki/pub/ABI/Genomics12/MLvsMAP.pdf.

[20] Emanuel A. Schegloff. "Practices and actions: Boundary cases of other-initiated repair". In: *Discourse Processes* 23.3 (1997), pp. 499–545. DOI: 10.1080/01638539709545001.

[21] Amnon Shashua. *Lecture 3: Maximum Likelihood/ Maximum Entropy Duality*. 2008.

[22] Gabriel Skantze. "Error Handling in Spoken Dialog System". In: *Doctoral Thesis* (2007).

[23] Gabriel Skantze. "Exploring Human Error Handling Strategies: Implications for Spoken Dialogue Systems". In: *Speech Communication* 45 (2004). DOI: 10.1016/j.specom.2004.11.005.

[24] Alex Smola and S.V.N. Vishwanathan. *Introduction to Machine Learning*. Cambridge university press, 2008.

[25] Ewan Klein Steven Bird and Edward Loper. *Natural Language Processing with Python*. O'Reilly, 2009.

[26] *The British National Corpus*. 2007. URL: http://www.natcorp.ox.ac.uk/.

[27] Jörg Tiedemann. "News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces". In: *Recent Advances in Natural Language Processing*. Ed. by N. Nicolov et al. Vol. V. Borovets, Bulgaria: John Benjamins, Amsterdam/Philadelphia, 2009, pp. 237–248. ISBN: 978 90 272 4825 1.

A. Appendix

A.1. First Appendix Section

Raw data of survey "Clarification Questions in Social Dialogues"

Each line corresponds to a query entry in the user study.
Each entry is scored in range from 1 to 5, at the end of the class a mean, a standard deviation, the maximum of the means and the minimum of the means are presented.

Confirmation to an assumption:

		scores				
1	2	3	4	5	mean	std. dev.
1	6	6	4	1	2.89	1.02
7	5	3	3	0	2.11	1.13
5	5	3	2	1	2.28	1.18
1	2	1	3	11	4.17	1.29
4	2	4	6	2	3	1.37
1	2	0	9	6	3.94	1.16
1	1	3	7	6	3.89	1.13
2	4	3	3	6	3.39	1.46
5	4	5	3	1	2.5	1.25
0	3	1	9	5	3.89	1.02
7	4	7	0	0	2	0.91
2	5	4	3	4	3.11	1.37
0	4	5	6	3	3.44	1.04
6	2	4	6	0	2.56	1.29
0	2	3	2	11	4.22	1.11

mean : 3.1953
std. dev. : 0.7591
min : 2
max : 4.22

One-word WH-question :

scores						
1	2	3	4	5	mean	std. dev.
0	0	0	5	13	4.72	0.46
2	0	6	7	3	3.5	1.15
0	0	2	1	5	4.72	0.67
3	5	5	4	1	2.72	1.18
2	4	7	2	3	3	1.24
0	1	4	6	7	4.06	0.94
0	0	1	6	11	4.56	0.62
1	1	1	5	10	4.22	1.17
3	7	3	2	7	2.72	1.36
0	1	2	8	7	4.17	0.86
1	4	4	5	4	3.39	1.24
5	5	6	2	0	2.28	1.02
1	4	2	5	6	3.61	1.33
5	1	8	3	1	2.67	1.24
2	1	2	6	7	3.83	1.34

mean : 3.6113
std. dev. : 0.8006
min : 2.28
max : 4.72

Phrase Repetition :

scores						
1	2	3	4	5	mean	std. dev.
0	3	2	6	7	3.94	1.11
1	0	2	5	10	4.28	1.07
1	1	4	4	8	3.94	1.21
2	3	3	8	2	3.28	1.23
3	5	4	3	3	2.89	1.73
2	2	3	5	6	3.61	1.38
2	5	5	4	2	2.94	1.21
2	2	5	4	5	3.44	1.34

2	1	4	8	3	3.5	1.2
1	2	4	6	5	3.67	1.19
0	1	2	7	8	4.22	0.88
4	2	3	8	1	3	1.33
4	1	1	6	6	3.5	1.58
5	1	4	4	4	3.06	1.55
2	2	0	8	6	3.78	1.35

mean : 3.5367
std. dev. : 0.4468
min : 2.89
max : 4.28

Really / Yeah / Huh ? :

	scores					
1	2	3	4	5	mean	std. dev.
1	1	5	8	3	3.61	1.04
0	4	7	4	3	3.33	1.03
1	3	7	3	4	3.33	1.19
1	5	4	6	2	3.17	1.15
0	4	3	7	4	3.61	1.09
1	5	4	5	3	3.22	1.22
1	1	0	4	12	4.39	1.14
1	4	7	4	2	3.11	1.08
2	2	6	3	5	3.39	1.33
1	2	5	6	4	3.56	1.15
0	1	2	6	9	4.28	0.89
3	5	7	2	1	2.61	1.09
0	2	2	4	10	4.22	1.06
0	1	1	7	9	4.33	0.84
1	0	1	7	9	4.28	1.02

mean : 3.6293
std. dev. : 0.5474
min : 2.61
max : 4.39

General What-question :

		scores				
1	2	3	4	5	mean	std. dev.
3	5	3	4	3	2.94	1.39
5	3	5	2	3	2.72	1.45
0	0	1	9	8	4.39	0.61
1	1	1	7	8	4.11	1.13
4	5	4	1	4	2.78	1.48
1	0	7	6	4	3.67	1.03
3	3	7	2	3	2.94	1.3
1	1	1	3	12	4.33	1.19
2	0	4	8	4	3.67	1.19
6	5	3	3	1	2.33	1.28
9	7	2	0	0	1.61	0.7
3	3	7	2	3	2.72	1.27
4	3	3	5	3	3	1.46
4	6	4	2	2	2.56	1.29
3	0	11	4	0	2.89	0.96

mean : 3.1107
std. dev. : 0.7767
min : 1.61
max : 4.39

Logical what-exclusive WH-question :

		scores				
1	2	3	4	5	mean	std. dev.
6	2	2	4	4	2.89	1.64
7	5	4	2	0	2.06	1.06
4	3	4	2	5	3.06	1.55
2	2	4	5	5	3.5	1.34
1	0	5	2	10	4.11	1.18
4	8	4	1	1	2.28	1.07
0	3	1	8	6	3.94	1.06
6	5	4	2	1	2.28	1.23

3	1	10	2	2	2.94	1.16
1	2	2	4	9	4	1.28
9	6	3	0	0	1.67	0.77
1	3	2	6	6	3.72	1.27
2	4	3	6	3	3.22	1.32
9	4	2	3	0	1.94	1.16
1	1	3	8	5	3.83	: 1.3

mean : 3.0293
std. dev. : 0.8212
min : 1.94
max : 4.11

A.2. Second Appendix Section

The user study:

Clarification Questions in Social Dialogues

In this questionair, dialogue pairs are extracted from **Social Dialogues,** which means that the conversation participants are just talking generally (or with no specific goal). The dialogue participants **A** and **B** can be freinds, partners or even strangers . Note: the following dialogues are independent situations to each other. Speaker A and B can change their identities based on each situation.

Each utterance from Speaker A comes along with six different possible questions uttered by Speaker B. Please score how appropriate the questions fit in the dialogues and situations (from 1 to 5; with 1 = unfitting, 5 = very fitting).

There are 15 questions in this survey

Fragments of Social Dialogues

[]

A: I'm sorry.

*

Please choose the appropriate response for each item:

	1	2	3	4	5
B : Why ?	○	○	○	○	○
B : Sorry ?	○	○	○	○	○
B : Huh ?	○	○	○	○	○
B : Why did you agree to marry me then ?	○	○	○	○	○
B : What' s the matter with you ?	○	○	○	○	○
B : Something' s happening , isn' t it ?	○	○	○	○	○

[]A: I got you. *

Please choose the appropriate response for each item:

	1	2	3	4	5
B: What' s all this about ?	○	○	○	○	○
B: You' ll catch me ?	○	○	○	○	○
B: What ?	○	○	○	○	○
B: All right, where is it ?	○	○	○	○	○
B: Uh ?	○	○	○	○	○
B: You got me, and then?	○	○	○	○	○

[]A : I have to talk with you. *

Please choose the appropriate response for each item:

	1	2	3	4	5
B : About what ?	○	○	○	○	○
B : With me ?	○	○	○	○	○
B : What are you going to talk about ?	○	○	○	○	○
B : Huh ?	○	○	○	○	○
B : Do you think we can communicate with each other ?	○	○	○	○	○
B : Why do I have to talk with you ?	○	○	○	○	○

[]A : Well , hello there , little fella . *

Please choose the appropriate response for each item:

	1	2	3	4	5
B : Do I know you ?	○	○	○	○	○
B : Where have you been ?	○	○	○	○	○
B : Hello ?	○	○	○	○	○
B : Yeah, what's up ?	○	○	○	○	○
B : Yeah?	○	○	○	○	○
B : Who ?	○	○	○	○	○

[]A : I' m wasted . *

Please choose the appropriate response for each item:

	1	2	3	4	5
B : Have you eaten enough ?	○	○	○	○	○
B : Why are you wasted ?	○	○	○	○	○
B : What' s wrong with celebrating sobriety by getting drunk ?	○	○	○	○	○
B : Really ?	○	○	○	○	○
B : What ?	○	○	○	○	○
B : Wasted ?	○	○	○	○	○

Fragments of Social Dialogues II

Yeah, almost done ! ;D

[]A : But you can only have it if you' ll marry me . *

Please choose the appropriate response for each item:

	1	2	3	4	5
B : Are you' re sure you' re not just drunk ?	○	○	○	○	○
B : What ?	○	○	○	○	○
B : Uh ?	○	○	○	○	○
B : You really want to marry ?	○	○	○	○	○
B : Why is that funny ?	○	○	○	○	○
B : What did you say ?	○	○	○	○	○

[]A : She's lost her boyfriend. *

Please choose the appropriate response for each item:

	1	2	3	4	5
B : You know how that feels , right ?	○	○	○	○	○
B : Really ?	○	○	○	○	○
B : Why ?	○	○	○	○	○
B : Where do you know this ?	○	○	○	○	○
B : Boyfriend ?	○	○	○	○	○
B : What are you talking about ?	○	○	○	○	○

[]A : Maybe something to drink first. *

Please choose the appropriate response for each item:

	1	2	3	4	5
B : More tequila ?	○	○	○	○	○
B : Why would you drink ?	○	○	○	○	○
B : What can I get you to drink ?	○	○	○	○	○
B : You drink ?	○	○	○	○	○
B : What drink ?	○	○	○	○	○
B : Yeah?	○	○	○	○	○

[]A : Some things are true whether you believe them or not. *

Please choose the appropriate response for each item:

	1	2	3	4	5
B : How ?	○	○	○	○	○
B : Why should we care ?	○	○	○	○	○
B : Are you OK ?	○	○	○	○	○
B : Really ?	○	○	○	○	○
B : What are you talking about ?	○	○	○	○	○
B : Do you believe them ?	○	○	○	○	○

[]A : You should play some music, you would feel better. *

Please choose the appropriate response for each item:

	1	2	3	4	5
B : I would ?	○	○	○	○	○
B : Does it work ?	○	○	○	○	○
B : What music ?	○	○	○	○	○
B : Yeah ?	○	○	○	○	○
B : Why not sing a song for me ?	○	○	○	○	○
B : What do you mean by play some music ?	○	○	○	○	○

Fragments of Social Dialogues III

[]A : Maybe I love her a little bit. *

Please choose the appropriate response for each item:

	1	2	3	4	5
B : A little bit ?	○	○	○	○	○
B: Really ?	○	○	○	○	○
B: What are you doing here ?	○	○	○	○	○
B: Is that beautiful or what ?	○	○	○	○	○
B: What happened ?	○	○	○	○	○
B: How come it works ?	○	○	○	○	○

[]A : I just came by to wish you good luck. *

Please choose the appropriate response for each item:

	1	2	3	4	5
B: I'll see you later ?	○	○	○	○	○
B: Yeah ?	○	○	○	○	○
B: What ?	○	○	○	○	○
B: Good luck ?	○	○	○	○	○
B: What are you talking about ?	○	○	○	○	○
B: Why not stay for a drink ?	○	○	○	○	○

[]A : More Americans suffer heart attacks from lack of exercise. *

Please choose the appropriate response for each item:

	1	2	3	4	5
B: You know why ?	○	○	○	○	○
B: Really ?	○	○	○	○	○
B: Why ?	○	○	○	○	○
B: Do you exercise a lot ?	○	○	○	○	○
B: You are talking about exercise stuff ? Like gym stuff ?	○	○	○	○	○
B: What do you do for exercise ?	○	○	○	○	○

[]A : Take as many as you like. *

Please choose the appropriate response for each item:

	1	2	3	4	5
B: Is it completely honest ?	○	○	○	○	○
B: What ?	○	○	○	○	○
B: Really ?	○	○	○	○	○
B: What's that supposed to mean ?	○	○	○	○	○
B: So many ?	○	○	○	○	○
B: Where are you going ?	○	○	○	○	○

[]A : Well, I am not so sure I even want to go to college. *

Please choose the appropriate response for each item:

	1	2	3	4	5
B: Since when ?	○	○	○	○	○
B: Really ?	○	○	○	○	○
B: Is there something happend ?	○	○	○	○	○
B: What are you talking about ?	○	○	○	○	○
B: Give me one good reason why ?	○	○	○	○	○
B: Not so sure ?	○	○	○	○	○