Dialogue-based disambiguation
Koeling, Robert Wietse

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2002

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 25-04-2020
Chapter 3

Short answer interpretation in context

3.1 Introduction

During the first two years of the OVIS project, a grammar for spoken language for the public transportation information domain was developed. The grammar was evaluated with real-world input. After optimising the grammar, a number of user utterances was still not analysed correctly. A certain number of errors were due to choosing the wrong hypothesis from the wordgraph. That is, although the right hypothesis was available, the parser favoured some other. It is expected that the addition of contextual information in the interpretation process can avoid (at least some of) these types of errors.

At an earlier stage some experiments were performed with a simple mechanism for testing the adequacy of an answer (e.g. preference for yes/no answers on yes/no questions). These experiments taught us that we could benefit from taking contextual information into consideration. Other researchers also reported increasing performance (Hanrieder and Görz, 1995; Drenth, 1996). These experiments however all used quite simple dialogue models. They used only the preceding system question as context and did not take sources of linguistic information into account. In this chapter I report on the first steps towards using a more sophisticated view of context. In this section, I first give an idea of the system we are working on and explain which problem I would like to solve. The approach I suggest in this chapter is based on a dialogue framework developed by Jonathan Ginzburg. In the next section I will explain how Ginzburg’s account for short answer interpretation can be used for my means. And finally I discuss the current implementation.

3.2 Moving on the dialogue gameboard

Before the possibility of employing contextual information can be discussed, we will need to define context. In this application it is quite clear what the context consists of. When we look at the example presented in the previous
section, it seems to be the case that most constraints on what makes sense in this particular dialogue state, are given by the previous system question. If other examples of OVIS dialogues are considered, we find out that there are several situations in which larger contexts are helpful. If the user for example tries to correct something the system misunderstood, earlier system questions are often referred to. For example:

(3.1) \[ S1 > \text{From which station to which station do you want to travel?} \]
\[ U1 > \text{I want to go from Groningen to Haren} \]
\[ S2 > \text{When do you want to travel from Groningen to Haarlem?} \]
\[ U2 > \text{To Haren!} \]

It will be clear that in most cases full dialogue context will be preferred. However, in applications like this (where the system asks the questions and the user will typically just answer these questions) I think that not much information is lost if attention is restricted to just the previous system question when the full dialogue context is not available. Later I will come back to this issue.

Before I can say something about the appropriateness of an utterance in a dialogue I need to define a dialogue model. For example, I need to be able to identify the state the dialogue is in at any moment. In a cooperative dialogue only a certain number of moves make sense in a certain state.

### 3.2.1 The dialogue game board

Later in this chapter I describe the approach which I propose to solve the problem sketched in the previous section. I will use a proposal by Jonathan Ginzburg for the analysis of short answers. He builds his analysis on a dialogue model he describes in terms of a *Dialogue Game Board*. Here I have chosen to adopt his view on dialogue structure. In this section I will briefly review the basic concepts that Ginzburg defines in (Ginzburg, 1996a; Ginzburg, 1994) and presupposes in his analysis of short answers (Ginzburg, 1996c).

Ginzburg identifies that on the one hand each dialogue participant (DP) must have some version of the *common ground*. This is information that is believed to be shared with the other DP's. He calls this component the DP's *gameboard*. It specifies in fact the space in which each DP can make his/her moves (i.e. you will not introduce a topic that makes no sense to the other DP's). On the other hand, a DP will of course also have his or her individual (mental) information that the other DP's do not have access to (e.g. general inferential capabilities). This is what Ginzburg calls the DP's *unpublished mental situation* (UNPUBMS(DP)). In what follows we only have to deal with the former. The DP's view about the state a dialogue is in at a certain moment (represented by the gameboard) is given by the values of a number of attributes. Ginzburg argues that at least the following three attributes are needed:

- **FACTS**: set of commonly agreed upon facts
• QUD The 'questions under discussion': set of questions that currently can be discussed. If a question \( q \) is topmost in QUD it is permissible to provide any information specific to \( q \) (i.e. that might be information about \( q \) or questions on which \( q \) depends).

• LATEST-MOVE: constrains the set of possible next moves a DP can make.

It is possible to straightforwardly describe an OVIS dialogue along these lines. The goal of the dialogue is to find out about a certain number of facts. To achieve this, the system poses questions that will be added to QUD. The system asks implicitly for confirmation of the information given by the user by echoing the relevant parts of the user utterance. The user is given the possibility to correct the information understood by the system. If the user just answers the new question, it is assumed that the user agrees upon the echoed facts. Facts can then be modified by adding these facts and QUD can be modified by removing the question(s) corresponding to these facts and by moving the new question to the topmost position.

3.2.2 Short answer interpretation

One of my observations with respect to the dialogues of our system was the fact that people tend to give short answers (i.e. utterances in which only the focus (Vallduvi, 1990) is realized). These utterances then refer to some preceding system question. In most cases, the referenced question immediately precedes the utterance, but that's not obligatory.

Ginzburg gives an analysis for short elliptical answers in a dialogue (Ginzburg, 1996b). Although interpretation of short answers is not really a problem in this application, Ginzburg makes some useful observations for answering the questions I discuss in this chapter. I am looking for information in the context that will constrain the possible dialogue moves in a certain dialogue state. He claims that, although psycholinguistic data shows that syntactic information as a whole decays rapidly, certain syntactic dependencies are maintained across the dialogue (I refer to Ginzburg (1996c) for evidence and an extensive discussion of this claim). He observes that context can influence the form of an utterance, at least that part of the utterance often dubbed the focus. Ginzburg concludes first that the focus of an utterance must match the specifications provided for it by the question it belongs to (which is called a syntactic presupposition). And second, when a question is raised in a dialogue (and not yet conclusively answered), it is incorporated in the dialogue as a semantic entity. He concludes that we need to have available at arbitrarily long distance the syntactic information of the argument role associated with the interrogated wh-phrase. Before I explain how we can use these observations for our means, I first sketch how short answers are interpreted in Ginzburg's approach.

Abstracts with restrictions

The idea The basic idea taken advantage of here is simply that an interrogative utterance creates a context in which it is allowable to answer with a
short answer, but only when the short answer satisfies the conditions set by the question.

The lambda abstract created by scoping wh- phrases in a question possesses argument-roles that carry appropriateness restrictions on the category of utterances associated with them. This can best be illustrated with an example:

(3.2) \[ \lambda X,Y,.<\text{RELY},\text{reli}_\text{er}X,\text{reli}_\text{ee}Y> \]

Restrictions: (utterances related with) X: NP [+nom]; Y: PP [+on]

In words, this lambda abstract expects arguments that must obey certain (syntactic) restrictions in order to be acceptable (in this example, the utterance whose content is X must be a nominative noun phrase and the utterance whose content is Y must be a prepositional phrase headed by the preposition 'on'). When this question is answered with for example the utterance: 'John on Mary', we can follow Ginzburg’s general scheme for short-answer interpretation:

(3.3)

a. $S \rightarrow (\text{ADV}), (\text{XP}_1), \ldots, (\text{XP}_2)$
b. Context: $\mu = \lambda \text{Abstr}(\text{MAX}_\text{QUD})$
c. Cont(s)(context) =

\[
\text{Cont(ADV)}[\text{Branch}_\text{closure}(<\text{Re}y_1, r_1; \text{Cont}(\text{XP}_1), \ldots, r_n; \text{Cont}(\text{XP}_n)>)]
\]

If an utterance is analysed as a series of constituents (a.) and the context is defined by the lambda abstract associated with the element that is maximal in QUD (b.), then the interpretation of the utterance in a particular context is given by applying the content of the utterance to the context function (as defined in b.) (c.). Applied to my example:

(3.4)

a. $S \rightarrow \text{NP}_\text{John}, \text{PP}_\text{onMary}$
b. Context: $\mu = \lambda X,Y,.<\text{RELY},\text{reli}_\text{er}X,\text{reli}_\text{ee}Y>$

Restrictions: (utterances related with) X: NP [+nom]; Y: PP [+on]

c. Cont(s)(context) =

\[
\text{closure}(\lambda X,Y,.<\text{RELY},\text{reli}_\text{er}X,\text{reli}_\text{ee}Y>, \text{NP}_\text{John}, \text{PP}_\text{onMary})
\]

(Restrictions: (utterances related with) X: NP [+nom] Y: PP [+on])

Ginzburg concludes that this approach has some advantages over purely semantic as well as reconstruction-based approaches to the analysis of short answers. I mention them here without going into details: 1) there is only one interpretation possible, 2) the syntactic dependency is stated only once (i.e. in the question), 3) there is no need for appealing to any unmotivated hybrid entities and 4) there is an explanation for emergence of syntactic presuppositions. 5) the syntax of the source does not play a role in constructing the interpretation of the short answer.

---

1 In case $n > 1$, and quantification is involved in more than one XP, scoping of the quantifiers needs to be taken care of. Ginzburg (1996c) explains why a branching-quantificational closure instead of a cumulative quantificational closure is needed.

2 There are no quantifiers involved, therefore scoping is not a problem here.
3.3. **Effecting Syntactic Presuppositions**

**Implementation** Before I can say something about the way this idea is used for our means, I first sketch Ginzburg’s proposal for implementation. First a new concept to describe the abstracts with associated restrictions is introduced: *restricted objects*.

(3.5) **Restricted objects:** given any object X and proposition p

\[ X \overline{p} = X, \text{if} \ p \text{ is true} \]
\[ = \text{undefined, if} \ p \text{ is false} \]

The next thing is to modify the application operation for these objects: *application of restricted objects to an assignment:*

(3.6) **Application of restricted objects:** given an abstract with restrictions, \( \mu \overline{p} \), and an assignment \( f \):

\[ [\mu \overline{p}]_f = [\mu]_f, \text{if} \ [p]_f \text{ is true} \]
\[ = \text{undefined, if} \ [p]_f = \text{false} \]

Now the example can be rewritten as:

(3.7) \[ \lambda u_1, X, u_2, Y < RELY, reli_{\lambda X, \lambda Y} : X, reli_{\lambda X, \lambda Y} : Y > \]

\[ \overline{\sqrt{\langle \text{CONTENT}, u_1, X \rangle } \land \langle \text{CAT}, u_1, \text{NP}[+nom] \rangle } \]
\[ \land \langle \langle \text{CONTENT}, u_2, Y \rangle \land \langle \text{CAT}, u_2, \text{PP}[+on] \rangle \rangle \]

Or the redundant parameters X and Y can be removed to get a notation that uses only utterance parameters:

(3.8) \[ \lambda u_1, u_2 < RELY, reli_{\lambda X, \lambda Y} : X, reli_{\lambda X, \lambda Y} : Y > \]

\[ \overline{\sqrt{\langle \text{CONTENT}, u_1, X \rangle } \land \langle \text{CAT}, u_1, \text{NP}[+nom] \rangle } \]
\[ \land \langle \langle \text{CONTENT}, u_2, Y \rangle \land \langle \text{CAT}, u_2, \text{PP}[+on] \rangle \rangle \]

And subsequently (\( \beta c \)) can be rewritten as:

(3.9) \[ \text{Cont(s)(context)=} \]
\[ \text{closure}( < \lambda u_1, u_2, < RELY, reli_{\lambda X, \lambda Y} : X, reli_{\lambda X, \lambda Y} : Y > \]
\[ \overline{\sqrt{\langle \text{CONTENT}, u_1, X \rangle } \land \langle \text{CAT}, u_1, \text{NP}[+nom] \rangle } \]
\[ \land \langle \langle \text{CONTENT}, u_2, Y \rangle \land \langle \text{CAT}, u_2, \text{PP}[+on] \rangle \rangle , \]
\[ \langle \langle \text{CAT}, u, \text{NP} \rangle , \langle \text{CONTENT}, u, \text{John} \rangle , \]
\[ \langle \langle \text{CAT}, u, \text{PP}[+on] \rangle , \langle \text{CONTENT}, u, \text{on Mary} \rangle \rangle ) \]

3.3 **Effecting syntactic presuppositions**

The idea I investigate in this chapter is how these *restrictions* can be used to structure the hypotheses in the wordgraph. Therefore I have to identify certain syntactic information that can help me to distinguish those hypotheses that I think are more acceptable than others.
3.3.1 Adding Syntactic Presuppositions to the OVIS grammar

The analysis of short answers is originally described in a situation semantics framework. Ginzburg further shows how to implement these ideas in a feature based (i.e. HPSG-like) framework. Although the current OVIS grammar is suitable for such a feature based implementation to a large extent, practical considerations made me decide not to follow this track. In our grammar we already used a QLF-like semantic representation (Alshawi, 1992). In the QLF-formalism there is space reserved in the formulas for syntactic features. I extended QLF’s CAT feature (reserved for syntactic information) with the feature SYN_REST to store information about syntactic restrictions. In (3.10) I give an example question from our system and I present what Ginzburg’s abstracts with restrictions will look like. In Figure 3.1 I give the corresponding QLF.³

(3.10)

\[ \lambda u_1, X. \text{want}(ev_A, you, travel(ev_A, you)) \land \text{destination}(ev_A, Assen) \land at(ev_A, X) \land (\text{content}, u_1, X) \land (\text{cat}, u_1, \text{mod} [+\text{temp}]) \]

\[ \begin{array}{c}
\text{pred whq} \\
\text{arg want}(ev_A, you, travel(ev_A, you)) \land \text{destination}(ev_A, Assen) \\
\land \text{at} \left( \begin{array}{c}
\text{cat} \\
\text{syn_rest} \\
\text{rest} \quad X
\end{array} \right)
\end{array} \]

Figure 3.1: QLF with syntactic restrictions

The value of SYN_REST must be of type syn_rest which is defined in figure 3.2. The major idea is that phrases are compatible if they agree in type and features. Let me give an example here. Consider the the user utterance following the first system question in a dialogue from our test corpus:

System: 'Van welk station naar welk station wilt u reizen?'
(From which station to which station do you want to travel?)

The wh-phrases 'Van welk station' and 'Naar welk station' of the system question are associated with the restrictions: PP(+loc, van) (locative PP headed by preposition 'van') and PP(+loc, naa, naa) (locative PP headed by preposition 'naar') respectively. The user actually uttered:

User: 'Van Hoogeveen naar Assen' (From Hoogeveen to Assen)
PP(+loc, van) PP(+loc, naar)

³In this simplified feature structure I have omitted many features irrelevant to the example
The user's reply is thus perfectly compatible with the restrictions given by the question. Apparently there was a lot of noise, because the wordgraph representing the user utterance is large (too large to display it here in a readable way). The NLP component offers a large number of hypotheses. The method nlp_speech for example, yields the following hypotheses (best one first):

1. 'Op vrijdag ochtend' (at friday morning)
   PP(+temp,op)

2. 'Een uur achteren' (Eighteen past one)
   NP(+temp)
   ...

7. 'Van Hoogeveen naar Assen' (From Hoogeveen to Assen)
   PP(+loc, van); PP(+loc, naar)
   ...

The first hypothesis is compatible in type (pp), but not in the features FORM and PFORM. The next five (I only listed 2.) are neither compatible in type nor in features. The context constraint demands that the focus of the answer meets the restrictions given by the wh-phrase in the corresponding question. The seventh hypothesis is the first one that is compatible with the context constraint. If I incorporate the context constraint with the method nlp_speech the hypotheses are ordered as follows (method nlp_speech_context):

1. 'Van Hoogeveen naar Assen' (From Hoogeveen to Assen)
   PP(+loc, van); PP(+loc, naar)

2. 'Van Hoogeveen' (From Hoogeveen)
   PP(+loc, van)

3. 'Hoogeveen naar Assen' (Hoogeveen to Assen)
   NP(+loc); PP(+loc, naar)
   ...

In this example the additional constraint works out perfectly; however, in practice things might be more complicated. For example the case of the adverbial wh-phrase (e.g. Wanneer (when) or Waarheen (where to)). Answers consisting of PP's are equally well formed as answers consisting of adverbs (but of course only when they do agree in their features). Another case can be demonstrated by looking at the previous example. Here, the question could also be answered with for example the utterance Hoogeveen naar Assen' (it is in fact one of the hypotheses). The first part of the utterance, Hoogeveen, is an NP. This would result in incompatible types (the type of the corresponding wh-phrase is PP). To account for these cases, a number of additional compatibility checks are formulated. If we have for example a wh-phrase of type PP, and an utterance whose focus consists of an NP, the context constraint is satisfied if the NP (i.e. the
focus of the user utterance) agrees with the NP inside the PP (i.e. the wh-phrase of the question). This check is straightforwardly formulated as follows:

\[(3.11) \quad \% \text{pp / np}
\]

\[
\begin{align*}
\text{compatible}_{sr}(\text{WH}_{SR}, \text{A}_{SR}) :&=& \\
\text{pp}_{sr}(\text{WH}_{SR}, \text{NP}_{Arg}, \_ , \_ , \_ ), \\
\text{np}_{sr}(\text{NP}_{Arg}, \text{Case}, \_ , \_ , \text{Form}), \\
\text{np}_{sr}(\text{A}_{SR}, \text{Case}, \_ , \_ , \text{Form}).
\end{align*}
\]

The \text{WH}_{SR} and the \text{WH}_{A} semantic representations are compatible in case the \text{WH}_{SR} would be the semantic representation of a prepositional phrase with an NP object, whose semantic representation is \text{WH}_{A}.

Another problem is the fact that a particular PP wh-phrase might be answered with a PP with a somewhat different preposition; e.g. S:‘Van waar wilt u naar Assen reizen?’, U:‘Van Venlo’ or U:‘Vanuit Venlo’ or U:‘Uit Venlo’ are equally well-formed answers (the fact that they all would be translated with ‘From Venlo’ in English, already implies that these preposition belong to some class of equivalent prepositions). Therefore I have defined ‘equivalence relations’ for some of the prepositions in our lexicon.

### 3.3.2 The current implementation

**Context includes only corresponding system question**

Of course I would prefer to take into account the complete dialogue. I have already noticed that OVIS dialogues can be described easily in terms of Ginzburg’s Dialogue Game Board. It seems also to be desirable to have more context than just the preceding question in cases such as corrections and user utterances that follow *yn-questions*. In the later implementations I foresee using a wider context; practical considerations however, forced me (at this stage) to use the second notion. The test data I have at this moment is recorded with an older version of the current speech recognizer for Dutch. The recorded data is subsequently
used for improving speech recognition, which resulted in much better performance. Many of the ‘not-understood-please-repeat’ questions can be omitted now. When we are looking at the original dialogues, we see many unnatural dialogue moves, since questions are re-asked while they should be removed (or at least not stay on top) from QUd.

Although I expect that even in this application the results can be improved by extending the notion of context, I think that in these kinds of very restrictive dialogues, most benefit will be a result of the direct context (i.e., the system utterance preceding the user utterance).

Matching

It is very easy to test if a hypothesis meets the context constraint. In general, the SYN_REST feature of the utterance must match the SYN_REST feature of (the wh-phrase of) the question (unification). In practice some additional combinations of types are also acceptable (as noticed and taken care of in the previous section).

Context score

**fit/nofit** I think that it is not a good idea to make the context constraint a fixed one, but just to let it be an indication of the appropriateness of a hypothesis. That is, in a situation where a hypothesis that does not obey the context constraint scores so much better on other grounds (e.g., acoustic or bigram), it might be favoured above a hypothesis that does obey this constraint. It is probably the best strategy to combine all different scores. It is however not a trivial job to find out the weight of a certain factor. It has been difficult to determine the best way to combine the scores so far. Adding one extra must of course be possible, but will need extra work. In the following two chapters I will investigate how statistical models can help to favour one hypothesis over another. The chosen statistical framework also gives an answers to the above mentioned score combination problem. However, in those chapters I will not elaborate on integrating contextual information in a dialogue model, but integrate it in the language model instead.

**Context gets highest priority** In the example below I first list all the hypotheses that obey the context constraint. This decision has two unwanted side effects. First, we noticed that the system gets too restrictive sometimes. Especially in the case of corrections (in which the focus of the user utterance does not match the specification of the wh-phrase of the question), we found out that often the right hypothesis was not chosen due to the context constraint. This is demonstrated by the following example from our corpus:

```
(3.12) SystemQuestion  'U wilt dus om tien uur 's ochtends vertrekken?
So you want to leave at ten o'clock in the morning?
Uttered          Om een uur (At one o'clock)
Possible          Om een uur
nlp_speech        Van een uur
nlp_speech_context Neen (No)
```
Second, I saw a tendency to 'hallucination'. For example in case of a yes/no question, any hypothesis containing a yes/no element is favoured above others. In very many wordgraphs there exists some path containing such an element. Hallucination became a serious problem. An example:

\[(3.13)\]

\begin{align*}
\text{SystemQuestion} & \quad \text{kunt u nog eens zeggen of u op dinsdag drieentwintig januari wilt reizen?} \\
\text{Uttered} & \quad \text{ik wil dinsdag drieentwintig januari reizen} \\
\text{Possible} & \quad \text{ik wil dinsdag drieentwintig januari reizen} \\
\text{nlp\_speech} & \quad \text{ik wil dinsdag drieentwintig januari reizen} \\
\text{nlp\_speech\_context} & \quad \text{nee ik wil dinsdag drieentwintig januari reizen}
\end{align*}

Relaxing the constraint

One way to overcome this problem, would be to solve the problem of the fit/no fit score. I decided however to follow a different track first. I relaxed the context constraint slightly. Instead of considering appropriate only those hypotheses whose focus matches the wh-phrase of the question, I now also look at other phrases in the question. If the focus of the utterances matches the specification of some of the other phrases, the utterance is also acceptable. This strategy prevents the following example from a possibly faulty analysis:

\[(3.14)\]

\begin{align*}
S: & \quad \text{U wilt dus morgen om twee uur vertrekken?} \\
& \quad \text{(So you want to leave tomorrow at two o'clock?)} \\
U: & \quad \text{Om tien uur!} \\
& \quad \text{(At ten o'clock!)}
\end{align*}

Relaxing the constraint this way succeeds in minimizing the errors mentioned in the previous section. But there is a price to pay. The constraint also loses a great deal of its strength. Just consider a question like: \text{Wanneer wilt u van Assen naar Groningen reizen?} ('When do you want to travel from Assen to Groningen'). According to the new constraint utterances matching the phrases \text{Wanneer} ('when'), \text{Van Assen} ('from Assen') and \text{Naar Groningen} ('to Groningen') are equally acceptable. It might again be a good idea to favour a direct answer (thus matching the wh-phrase) above corrections. Statistics about combinations of speech acts can probably give useful information about this.

Yes/no questions

In the previous example I used a yes/no question. In Ginzburg's analysis of short answers it is predicted that yes/no questions do not syntactically constrain the possible answers. This seems to be the case for the pure yes/no questions such as (56) in dialogue 4.5. Other questions that ask for confirmation of information (e.g. the previous example) would follow that prediction perfectly if context more extended than just the previous system question was available (i.e. by means of the correction, the user makes clear that he does not accept
the last system question. Nothing is yet added to facts and before the new
system question could be added to qud the current topmost question of qud
must be properly answered.)

3.4 Some experiments

3.4.1 Data

For a first experimental version of the ovis system, a German system made
by Philips in Aachen (Aust et al., 1995) was adapted for Dutch. This experi-
mental system (which was not open to the public) is used to train the speech
recognizer and to collect data. As a result we have at our disposal a collection
of about 30,000 wordgraphs together with a registration of the dialogues these
user utterances were taken from. About 10,000 of the user utterances are given
a syntactical and semantical annotation and collected in a corpus (Scha, Bod,
and Bonnema, 1996).
The collection of 30,000 wordgraphs gave us the opportunity to improve the
ovis grammar. The performance of the grammar was evaluated in terms of
word accuracy and sentence accuracy (Boros et al., 1996). When the corpus of
annotated wordgraphs became available a better evaluation method (based on
semantic rather than syntactic similarity) could be used: concept accuracy.

3.4.2 Maximal results

Several methods for finding the best path in the wordgraphs were developed.
Methods solely based on acoustic scores (speech), grammatical analysis (nlp),
ngrams (bigram, trigram) were compared. The next step was to test combina-
tion of methods (e.g. speech_nlp, speech_nlp_bigram, etc).

An informal evaluation of the results I found for a subset of 1000 randomly
chosen wordgraphs from the corpus was conducted to estimate the maximal
results I could establish by incorporating contextual information. I collected
and counted the analyses (using method nlp-speech)\footnote{Note that this was not the best scoring method. Other methods, such as those that combine \textit{nlp}, \textit{speech} and \textit{ngram} methods, gave better results. The fact that it is not obvious to define a score function that yields the best result given a certain score for each of the methods, I decided to do my first experiments with just the \textit{nlp-speech} method and augment this method with a \textit{context} component. Later I will report on first experiments with adding the \textit{context} component to the best scoring method.} for which the concepts
correctly found differed from the annotation. For these 1000 utterances I found
that 139 received a erroneous analysis. Inspecting these 139 wordgraphs taught
me that in 52 cases the analysis included an error of the form:

\begin{equation}
\begin{array}{ll}
\text{(3.15)} & \\
\text{Uttered} & \text{Ik wil naar Haren} \\
& \text{I want to go to Haren} \\
\text{Understood} & \text{Ik wil naar Haarlem} \\
& \text{I want to go to Haarlem} \\
\end{array}
\end{equation}

Even if the right path exists in the wordgraph, these kind of errors typically can
not be corrected by a context dependent analysis. In 51 cases I found out that
the wordgraph is either such a mess that it is not possible to find a correct path in the wordgraph at all (31), or the results were partially right and could not be improved (20). In three cases I found that although the analysis differed from the annotation, I think the analysis is right. In five cases I could not decide what caused the error.

So, finally there were 28 wordgraphs left that might be improved. In 10 cases it was clear that contextual information really was needed to do the disambiguation (e.g. cases such as 3.16). In the other 18 cases some progress could possibly be expected (e.g. 3.17).\(^5\)

(3.16)

<table>
<thead>
<tr>
<th>SystemQuestion</th>
<th>Wilt u nog eens zeggen wanneer u naar Eindhoven wilt reizen?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uuttered</td>
<td>Zaterdag (Saturday)</td>
</tr>
<tr>
<td>Possible</td>
<td>Zaterdag (Saturday)</td>
</tr>
<tr>
<td>nlpspeech</td>
<td>Zaandam dank u (Zaandam, thank you)</td>
</tr>
</tbody>
</table>

(3.17)

<table>
<thead>
<tr>
<th>SystemQuestion</th>
<th>Wilt u dat ik de verbinding nog eens herhalen?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uuttered</td>
<td>Nee hoor, dat hoeft niet (No, that's not necessary)</td>
</tr>
<tr>
<td>Possible</td>
<td>Nee hoor, goed niet (No, good not)</td>
</tr>
<tr>
<td>nlpspeech</td>
<td>Nee hoor, half tien (No, nine thirty)</td>
</tr>
</tbody>
</table>

It is difficult to decide beforehand which analyses would suffer from incorporating contextual information. There is reason to fear some 'hallucination' if you decide to give too much weight to contextual information (as I noticed in section 3.2.3).

It seems to be the case that in this application I can gain little by adding contextual information. Just 20% of the errors (which corresponds to only 3% of the utterances) might benefit more or less from adding this constraint and that at most 10% of the errors can really be corrected (maximal gain; not taking possible loss into account). I think that this result is not surprising. The lexicon of the speech recognizer is rather small. Moreover it contains mostly city and station names (i.e. words with equal properties). I saw that a substantial part of the errors are caused by wordgraphs containing parallel paths with station/city names.

### 3.4.3 Some results

In a first test I measured the effects of adding the context constraint using method \(nlpspeech\). I knew already that only a small quantity of improved analyses could be expected. I did not know in how many cases the results would become worse. The table in figure 3.4.3 shows the results of adding the

\(^5\)Possible denotes a path through the wordgraph that is as close to the actually uttered sentence as possible.
3.5. DISCUSSION

After 999 sents 3357 words 10118 trans (3.01 transitions/words):

<table>
<thead>
<tr>
<th>Method</th>
<th>Word (%)</th>
<th>Sentence (%)</th>
<th>Concept (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Accuracy</td>
<td>match</td>
</tr>
<tr>
<td>best_possible</td>
<td>92.672</td>
<td>84.2</td>
<td>-</td>
</tr>
<tr>
<td>speech</td>
<td>78.642</td>
<td>63.4</td>
<td>-</td>
</tr>
<tr>
<td>spoken</td>
<td>-</td>
<td>-</td>
<td>97.60</td>
</tr>
<tr>
<td>nlp_speech</td>
<td>82.395</td>
<td>70.8</td>
<td>80.78</td>
</tr>
<tr>
<td>nlp_speech_context</td>
<td>82.812</td>
<td>71.1</td>
<td>81.58</td>
</tr>
</tbody>
</table>

Figure 3.3: Effects of adding context constraint

constraint I suggested. A small improvement can be noted. I also checked those cases that I marked as potentially improvable. I found out that in 10 cases the analysis did improve and in 1 case a worse result was obtained. I did the same test before I decided to relax the context constraint (see section 3.2.4). Then I got a better result in 16 cases, but I also found a worse result in 5 cases. As I hoped, I managed to decrease the undesirable side effects, but I had to pay a price for it. By relaxing the constraint, it became less informative.

3.5 Discussion

I think that it might be a good idea to follow this track, because it models the relationship between a question and an answer, without any appeal to information not yet available. Intuitively the set of possible answers to a question is constrained. Hopefully the constraints are within reach, but maybe it is mostly constrained by common sense or information that falls within the class of (difficult to model) 'knowledge of the world'.

I noticed that other accounts for dealing with this problem rely on highly application-specific information. I tried to isolate available information that discriminates non-interpretable hypotheses. But does this make a general account? As an account for modeling the interpretation of short answers, maybe it does, but as a means for constraining the number of paths in a wordgraph, it does not. It makes a rather detailed assumption on the nature of the data (i.e. on the existence of informative distinguishing features). These informative features are dominant in this domain (e.g. the temporal and locative features), but not generally in every domain. Mostly the relevant features are too subtle too be distinguishing (e.g. case marking is hardly (if at all) noticable for a speech recognizer in for example spoken German).

If it is not a general account, is it helpful in this environment? Although there is an improvement in performance, I am not yet convinced that it is. I think that it could be helpful if there is more space for improvement in the data. But the
informal study of the errors taught me that, although there still is a considerable error rate, many of the errors are possibly not correctable by the approach discussed in this chapter. Partly this is not surprising. The fact that the speech recognizer is trained for this application, implies that there will be a tendency to yield hypotheses that are semantically close to each other (i.e. it is expected that there will be confusion between the city names Haarlem and Haren, but there will not often be confusion between two semantically completely different, but phonologically rather similar words such as Haren and harem).

Conclusions and further work

I have reported on the implementation of a constraint that takes contextual information into account for analysing a wordgraph. The implementation is incorporated into the NLP component of an experimental spoken language system. The implementation is tested on real world data. Although an informal evaluation of the data taught us that not much could be gained (in this data set), the first tests imply that it performs quite well in the cases where improvement can be expected.

I did my first tests on the nlp_speech method. This is however not the best scoring method at this moment. Combination of this method with trigram scores gives better results. To be conclusive, one would have to investigate whether context adds different information than for example trigram.

I already mentioned the fact that I am not yet satisfied with the way the score of the context method is combined with the other scores. With a more subtle interaction of scores, I could probably keep the strength of the original context constraint, while avoiding the unwanted side effects.

Finally, I really want to work out the effects of the full dialogue context instead of just the previous system question. Although it is possible to describe a typical ovis dialogue straightforwardly in terms of the Dialogue Game Board, it is difficult to foresee what the consequences will be for the analysis of wordgraphs. New data (full dialogues) will be needed to say more about this. It would also be interesting to see what the effect of 'bigger' wordgraphs (i.e. more transitions per word) would be on the effects of the context constraint.