Abstract. A major problem of understanding language in spoken dialog systems is to detect recognition errors in the output of a speech recognizer. Such a capability is the basis of implementing repair strategies that allow a dialog system to handle communication about misunderstandings similarly to other clarifications. In this paper we present a two-phase approach that combines chunk and dependency parsing and takes the global syntactic structure of recognizer output into account. This enables us to identify dependencies between chunks and detect syntactical errors caused by word confusions in case dependency constraints are violated. Finally, we apply these diagnostics to dialog modeling and discuss how the resulting error information can be used by clarification strategies.

1 Introduction

A major problem of understanding language in spoken dialog systems is to detect recognition errors in the output of a speech recognizer. While the signal processing community works on improvements of feature extraction algorithms in order to decrease the average word error rate of a speech recognizer, less attention is paid to the linguistic side of errors and how to detect them.

The capability to locate errors in a speech recognizer output (word chain, word lattice, or word confusion matrix) enables a dialog system to engage in clarification dialogs on acoustic misunderstandings. Our paper presents a study how to identify these misunderstandings.

As example application, we choose a natural language dialog system for controlling digital equipment for home entertainment, such as TV sets or DVD players. In this domain, a typical utterance is

\textit{Die Sparte Serie auf RTL ist ausgewählt?}

Normally, speech recognizers will misrecognize some words producing hypotheses which may be inconceivable to a human hearer and—in particular—to an automatic dialog system. Figure 1 lists hypotheses for the above utterance. As discussed in detail later, they are ill-formed and every native speaker will ask a number of questions in order to clarify the misunderstandings. The contribution of this paper is a parsing algorithm that computes error diagnostics similar to those a native speaker would find. Then we sketch how such diagnostics can be used in a dialog system for clarification sub-dialogs.

The paper is structured as follows: First, we discuss some corpus studies that exemplify the way in which humans communicate about misunderstandings on different levels. In the section to follow, we report on other computational approaches to localize and identify errors in ill-formed utterances. Then we explain our two-phase approach that combines chart-based chunk parsing with constrained-based dependency parsing to a global syntactic analysis providing possible readings of the input and detailed error diagnostics. We conclude with a presentation of the system’s performance, a discussion of the results and an outlook to open issues and future work.

2 Clarification of Misunderstandings

Current (commercial) state-of-the-art systems do not implement strategies revealed in corpus analyses, but rather heuristic approaches to clarifications. Often, feature extraction that is needed for implementing recovery or clarification strategies is difficult to compute. Secondly, the features are unclear themselves. Mostly, they are not specified semantically, so it is very hard to compare them.

In contrast to such ad-hoc technical solutions, our approach is to find operational semantics for clarification strategies and for algorithmic detection of features that allow to diagnose misunderstanding errors at run-time. In our view, clarification has to be implemented differently and separately on all levels of perception:

- Acoustic and syntactic errors are hard to distinguish in speech recognizer output. Often, they are generated during the speech recognition process—either because the speech recognizer did not classify correctly or because the user transgressed the lexical and grammatical limitations of the recognizer’s language model:

\textit{ich möchte eine komödie laufen werden}\(^4\)

These types of error are the focus of our paper.

- Another type is misunderstanding on the semantic level:

\textit{ich möchte eine komödie auswählen} (I want to select a comedy)
\textit{ich möchte eine komödie aufnehmen} (I want to record a comedy)

Such misunderstandings may be caused by acoustic problems as well, but cannot be detected on syntactic level as both utterances are grammatically perfect. They are only detectable with information about context and therefore beyond the scope of this paper.

- The same holds for pragmatic misunderstandings. For clarifications of pragmatics the context or state of the dialog has to be considered in parallel to the state of the application which is addressed in the interaction. Pragmatic misunderstandings are discussed e.g. in [11] or [12] and are beyond our scope as well.

3 Error Classification

In a corpus of spoken user input collected with the EMBASSI dialog system (see [8]), the example utterance from section 1 is recognized
die sparte gelaufen ab elf Uhr gewählt 5
die sparte gelaufen ab elf Uhr ausgewählt 6
die sparte gelaufen das will vox gewählt 7
die sparte gelaufen was wählte vox gewählt 8

die sparte gelaufen und wählte vox gewählt

Figure 1. 5 best hypotheses for Die Sparte Serie auf RTL ist ausgewählt

as shown in Figure 1. None of the five hypotheses computed by the
speech recognizer contains the transcription (see above).

Though some words are recognized correctly, no native speaker of
German would consider any hypothesis out of the five to be conceiv-
able as a complete sentence. He would argue that in each hypothesis
there are sequences of words conceivable for themselves, but along
with the other ones they cannot be integrated in a way that a hearer
can make sense out of the whole. The reason for that observation is
that in a sentence there are several grammatical functions that have
to be fulfilled by certain phrases. Having identified these, the hearer
is able to construct a hypothesis for the meaning of the utterance.

However, in the hypotheses in Figure 1 some of the functions are
missing (e.g. predicate in hypothesis 1 and 2), some are fulfilled more
than once (e.g. the predicate in hypothesis 3, 4, and 5), and some are
located in unusual positions (e.g. what in hypothesis 4). These facts
are obstructive for the hearer to understand the hypothesis.

In most dialog systems, there is a work-around: sequences that
have meaning in isolation are extracted and syntactic functions are
ignored. Whenever an recognition error affects a word that is crucial
for the meaning of the utterance, this work-around fails. When errors
lead to conflicting information (e.g. VOX vs. ARD in hypothesis 5),
it is even harder for the dialog system to take a decision.

4 Related Work

In literature, the analysis of clarification dialogs has attracted the
interest of many researchers. While there are many corpus studies
on which patterns are used for clarification in human-human dia-
log (Ginzburg [7] provides an overview of types of clarifications
which may follow an utterance), it is difficult to define an efficient
and tractable decision procedure for error classification and selection
of an repair strategy. Indeed, there is even no consensus about an
appropriate feature set. Different ones are used e.g. in [5, 2, 16].

Our approach differs from the cited publications in that we want
to classify errors in natural language output of a speech recognizer
by employing as much knowledge about language use as possible
instead of abstracting immediately to data that involves the (error
prone) interpretation of the recognizer output. Another main differ-
ence is that the cited approaches lead to an acceptance or rejection
of the user utterance as a whole. However, we are interested in find-
ning the type of misunderstanding in order to provide a computational
basis for (interactive) clarification also on partial utterances.

In the area of speech recognition, confidence scores are defined
by an entropy-based measure of confusion in a word graph. While
the idea of word confusion graphs goes back at least to [13] for the
purpose of comparing different hypotheses from a speech recognizer,
Hakkani-Tür and Riccardi [9] report about experiments to use
posterior probabilities or posterior entropy as a confidence measure
and to localize errors in positions with low confidence. In [3] anti
models for phones are proposed to compute confidence scores.

5 The Two-Phase Parsing Approach

The parsing is composed of two phases: In a preprocessing step a
chunk parser generates a chart graph with all possible chunks (cf.
Fig. 2) from a word confusion graph (see Fig. 3). In a second phase
a dependency parser searches the chart graph for a path that can be
parsed according to a rule-based linguistic model. If no path can be
found that fits the model, it chooses a path as close to the model
as possible and describes the deviation on the basis of the conflicts
arisen during the parsing process. This description serves as input to
a subsequent diagnostics procedure specified in section 6.

The underlying linguistic model is based on a dependency gram-
mar combined with a topology model. The grammar contains sub-
categorization rules and lexical subcategorization information to de-
scribe admissible dependencies. Our approach lifts dependency rela-
tions from word level to a higher level, the chunk level. Thus, partial
parses computed by the chunk parser can be utilized as short cuts for
the search process and abstractions within parse trees.
As dependency grammars do not make statements on word order, we need a topology model describing admissible linearizations of parse trees. It is based on topology field theory in German linguistics that subdivides sentences into five fields (see [1]). Similarly to [6], who suppose that every word (here: chunk) induces a new sub-field, we defined a dynamic hierarchical field model capable of handling complex phenomena like subordinate clauses, scrambling or partial verb fronting. While target field rules such as

\[(\text{predicative}_\text{noun}, \text{VL}, x) \rightarrow (R, \text{left})\]

\[(\text{predicative}_\text{noun}, \neg \text{VL}, \text{finite}_\text{Verb}) \rightarrow (\text{RB}, \text{right})\]

\[(\text{predicative}_\text{noun}, \neg \text{VL}, \neg \text{finite}_\text{Verb}) \rightarrow (R, \text{left})\]

\[(\text{predicative}_\text{noun}, \neg \text{VL}, x) \rightarrow (\text{VF}, \text{right})\]

specify in which field a dependent can be located (depending on its function, verb order and regent’s properties), precedence rules like

\[
\text{NP}_{\text{es}} < \text{NP}_{\text{akk}} : 0
\]

\[
\text{NP}_{\text{Nom}} < \text{NP}_{\text{Dat}} : 2
\]

constrain the relative order of chunks within a field. Unlike common ones, our topology model works interactively since field boundaries can be extended outwards when a new chunk is assigned to the field. Thus, whenever a partial parse is to be extended by a new dependency, we can immediately validate that the new dependent’s position fits the topological structure built up by all previous dependencies.

The dependency parser works top down using an A* search algorithm to find a path fitting the linguistic model best. For this, it must work fault-tolerant as to a certain degree also erroneous dependencies must be taken into account. The deviation from the linguistic model can be described by a set of conflicts of the following types:

1. Incongruency: A dependency is presumed while features of the dependent do not exactly meet the specifications. E. g. in

   he sleep

   he can be interpreted as subject to sleep though regent and dependent do not correspond in number.

2. Vacant grammatical function: A mandatory grammatical function is not fulfilled because there is no adequate chunk that is not occupied by a previously assigned dependency. E. g. the hypothesis

   he gives to her

causes a conflict because there is no chunk like it that can serve as direct object to gives.

3. Spare chunk: At the end of the parsing process a chunk remains whose function could not be identified. E. g. the hypothesis

   he sleeps it

contains the chunk it that cannot be incorporated into a semantic interpretation of the whole sentence.

Efficient parsing as postulated in section 4 can be achieved by relaxing the parser’s precision: As we only need to find out if the input is erroneous, exact resolution of ambiguities is only required for error-related sequences. For all other ones it is sufficient to validate that there are admissible interpretations, no matter which one to choose.

6 Error Diagnostics

The system analyzes the conflicts detected by the parser and decides whether the speech recognizer output is accepted or rejected or whether an error model can be used to identify misunderstandings. Assuming that the linguistic model contains the original utterance, all those conflicts are caused by classification errors of the following types (the examples refer to the utterance I don’t understand you):

- Simple confusion: A word is replaced by another one:
  I don’t understand who
- Omission: A word is omitted or replaced by a break:
  don’t understand you
- Insertion: A word not contained in the utterance is inserted:
  I don’t understand it you
- Contraction: A word not contained in the utterance is inserted:
  I don’t understanding
- Separation: A single word is replaced by a sequence of words:
  I don’t thunder stand you

In order to find out location and type of the confusion we need to infer confusions from conflicts. However, in most cases there is no clear relation between cause and effect: A single confusion can cause more than one conflict. Often conflicts are bidirectional, that is either the dependent or the regent can be erroneous. In the end, several confusion types can cause the same conflict. For an adequate handling of these ambiguous relations, we need a statistical error model.

7 Worked Example

In this section, we discuss step by step how the 5 best hypotheses in Fig. 1 are processed to localize and analyze the detected errors.

The first step is to transform the single hypotheses into a joint word confusion graph as shown in Fig. 3. In such a graph, sub-sequences of words can be identified that are typical for ill-formed input. The result is a disambiguated word chain with markers for critical regions (see Fig. 4). In these regions, syntactic errors are likely to be found.

Then a chunk parsing step tries to find all chunk readings for the given hypotheses or the whole word confusion graph. Fig. 2 shows an extract of the chunk graph for the input in Fig. 4. This chart graph is the basis for the following dependency analysis. It aims at identifying how the chunks are related taking topological fields for German and the subcategorization frames of the involved words into account.

For going step by step through the second parsing phase we discuss another example from the home entertainment domain illustrating clearly the linguistic background and how error diagnostics can be made. The original utterance and the 5 best hypotheses in Fig. 1 are processed to localize and analyze the detected errors.

In this section, we discuss step by step how the 5 best hypotheses in section 4 are processed to localize and analyze the detected errors.

The first step is to transform the single hypotheses into a joint word confusion graph as shown in Fig. 3. In such a graph, sub-sequences of words can be identified that are typical for ill-formed input. The result is a disambiguated word chain with markers for critical regions (see Fig. 4). In these regions, syntactic errors are likely to be found.

Then a chunk parsing step tries to find all chunk readings for the given hypotheses or the whole word confusion graph. Fig. 2 shows an extract of the chunk graph for the input in Fig. 4. This chart graph is the basis for the following dependency analysis. It aims at identifying how the chunks are related taking topological fields for German and the subcategorization frames of the involved words into account.

For going step by step through the second parsing phase we discuss another example from the home entertainment domain illustrating clearly the linguistic background and how error diagnostics can be made. The original utterance and the 5 best hypotheses for it were
In this example, parsing the word confusion graph would result in automatic correction because it includes the transcription as a path. As no other path can be parsed without any conflicts, the search algorithm will definitely choose the transcription as best interpretation.

Parsing each word chain independently illustrates how error detection works if the error cannot be eliminated. First of all, the dependency parser performs the following steps:

1. Possible sentence readings and verb orders are taken into account. One of them is statement with Verb-Zweit-Stellung (verb-second order). This reading calls for the finite verb to constitute the left bracket. Finite verb candidates are möchten, fernsehen and anschauen. The former obtains the highest priority as the left bracket is expected to be early in the sentence.

2. The verb form möchten is an first-person singular modal auxiliary with subcategorization frame

   \textit{verbal\_part(infinite\_tive), subject(first\_person singular NP)}

   The topology model allows the infinitive to be in the right bracket or the Vorfeld. As the left bracket has already been located in step 1, the Vorfeld is constituted by ich, which cannot be an infinitive. Thus fernsehen and anschauen are the candidates for the infinitive constituting the right bracket. This time the latter is prioritized.

3. It has the subcategorization frame object(NP).

   For the transcription, ich would be assigned to the subject slot and einen Krimi to the object slot while morgen abend would be interpreted as time supplement to the verb and im fernsehen as location adjacent to either the verb or the object.

   However, for each hypothesis one of these assignments can only be done at the cost of a conflict. Hence the A* search tests the other alternatives of step 1 and 2, but as none of them provides a conflict-free solution, the primary interpretation is rated best.

For the individual hypotheses, the following conflicts are detected:

H1: Incongruency: subject mich has ACC case instead of NOM
H2: Incongruency: subject ich differs from the verb in person/number
H3: Spare chunk: ersehnt (desired)
Vacant gramm. no noun to preposition im (on) function:
H4: Spare chunk: sorgen (sorrow)
H5: Spare chunk: haben (have)

For H1 the error can be localized in the word mich (me). An adequate repair strategy was either to guess that it was originally nominative or to ask the speaker for the subject: Wer oder was? (who or what?)

For H2 the system is unsure if the error is located in the subject or in the verb. Either both were originally first-person singular or both were second-person plural. If an error marker (compare Figure 4) was found at exactly one of both positions, it is possible to disambiguates this error diagnosis, with the same conclusions as for H1. If not, the risk is too high to choose the wrong alternative. Therefore it is better to ask for the subject.

For H3 it is likely that the spare chunk ersehnt originally fulfilled the vacant grammatical function (and thus described the location of the detective story). The adequate question to the speaker was either wo? (where?) or in wem oder war? (on whom or what?)

Diagnostics are more complicated for H4. As both chunks are adjacent, it is likely that both words form a single word or chunk in the original utterance. But as they serve as supplement (which is facultative and therefore not included in the subcategorization frame) their grammatical function is unknown. Hence a specific question cannot be formulated. Only if the system expected the user to tell something about time, it is reasonable to guess that the missing phrase is about time. In this case it could try to ask: wann? (when?)

However, the latter does not work for H5 because here a part of the time specification (tomorrow) is complete apart from the other one (night). Hence an adequate strategy was either to ignore the missing chunk (that is to suppose it to be irrelevant for the meaning of the utterance), to reject the whole sentence (Sorry, could you repeat?) or to formulate a position-oriented question (Sorry, could you repeat the word you said after morgen?).

8 Evaluation on a Small Test Corpus

For evaluating the parser on word chains, we used a subset of the EMBASSI corpus [8] containing 5 best hypotheses (cf. Figure 1) for each of 53 utterances. 87% of the transcriptions were recognized as error free while 80% of the wrong hypotheses were classified as erroneous. The latter is due to the fact that not all confusions by the speech recognizer result in syntactic errors. Leaving out hypotheses that are grammatically correct but contain semantic misunderstandings, 96.6% are recognized as syntactically erroneous. Taking only one best hypothesis per utterance, both recognition rates rise to 94%.

Localizing the error on the basis of syntactic conflicts turns out to be more difficult. Even if we only take those hypotheses into account that cause a single conflict, for only 49% the error position can correctly and unambiguously be identified. In contrast, for 36% the conflict is not tangent to the misunderstood word.

Taking a look at the individual utterances we can clearly see that it is hardly possible to find the error by means of parsing when the stem of the finite verb is affected by the confusion. This is due to the fact that the finite verb determines the most important semantic relations in a sentence and that its subcategorization frame differs for different verbs. Hence the syntactic interpretation of an utterance tends to be completely wrong if the finite verb could not be identified correctly.

Besides, error diagnostics that are useful for repair strategies seem hardly practicable if an hypothesis contains more than one misunderstanding. Particularly in short utterances multiple errors often effectuate that a substantial part of the semantic structure is not accessible.

9 Open Issues

More detailed diagnostics—particularly if more than one conflict occurs—would require an elaborate statistical error model trained by large amounts of data. In order to provide more significant indications for error positions, syntactical information given by conflicts should be combined with statistical information given by error localization labels as shown in Figure 4. An important question to answer is if this could raise the rate of correctly localized errors to a level that makes clarification worthwhile compared to complete rejection.

Additionally, an evaluation is to be done on word confusion graphs as their use could lead to automatic correction provided that they contain every word of the utterance.

Up to now, our work was confined to independent utterances. However, in dialog systems contextual information is essential for interpretation of utterances. Considering context means to integrate our approach with a semantic analysis. Thereby also considerable improvements could be achieved both in efficiency and error detection if dependencies between chunks are only considered if there is an adequate semantic relation according to a semantic model.
10 Further Work on Clarification Strategies

In this section we sketch briefly how the error diagnostics help the dialog system in initiating a clarification dialog. The following table shows what information the dialog system can extract from the (parsed) word confusion graph (see Fig. 3, 4, and 2):

chunk interpretation

(0,2) DP chunk is the only hypothesis. It is grammatically correct.
(2,3) PPART chunk is the only hypothesis, but in this position it is grammatically ill-formed.
(3,6) PP chunk ab elf Uhr is correct, but there are several alternatives; the risk for choosing it is very high even if the acoustic score of this reading is the highest available (see Fig. 1).
(6,7) The readings for the PPART chunk are semantically synonymous. So, there is little risk in choosing any of them.

The topological information computed in phase two of the parsing process delivers the following diagnostics:

(0,2) DP chunk may be the subject in the utterance. However, it is semantically incomplete as genre lacks an apposition, which is another indication that chunk (2,3) is wrong.
(2,3) PPART chunk may be the (elliptical) predicate. However, none of its case frames can be filled by any other chunk. So, the risk is high for this interpretation.
(6,7) PPART chunk may be the (elliptical) predicate. The risk for this option is lower as (0,2) fills the subject case frame.

While deciding how to react on the utterance, the dialog system must be aware that (3,6) is very confusing. This indicates that a large part of the whole utterance may be misrecognized. The critical part even extends to (2,3) as discussed above. Therefore, a cautious strategy would be to reject the utterance as a whole and to answer:

Sorry, I didn’t understand you. Could you repeat, please?

A risky strategy would assume the subject (0,2) and the elliptical predicate (6,7) to be stable and try to clarify the genre in chunk (2,3):

Which genre do you want to select?

This strategy includes a clarification of (3,6) as it assumes the chunk to be a complement to the DP or the predicate. It further assumes the user to repeat this part of the utterance as well. However, the risk for something different to happen is high. Therefore, this strategy potentially increases the risk for the whole clarification to fail.

11 Summary

The discussion of clarification strategies indicates further research directions: Do native speakers really interpret ill-formed sentences in a similar way to that presented here? Do they similarly react in case they feel there is some misunderstanding?

While these are open issues for dialog research, the paper shows that progress can be achieved in the parsing phase: Our approach constructs a possible reading of user input and computes error diagnostics that are not related to artifacts of parsing strategy or grammar (formalism), but to a model of how native speakers analyze and understand utterance. The resulting analysis eases the tasks for modules in later steps of the natural understanding process. We showed that the analysis delivers valuable criteria for a dialog module. In case of misunderstandings it can generate more “natural” continuations than common slot-filling approaches do. The reason is again that the diagnostics are not related to artifacts of the technical aspects of the dialog system, but to human usage of spoken language.

The evaluation of our approach proved to be reliable. Almost all grammatical errors can be localized as indicated by Figure 5.

10 References