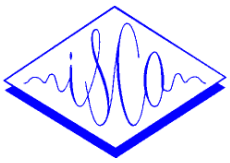


6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

GENERATION OF UTTERANCES BASED ON VISUAL CONTEXT INFORMATION*

Susanne Kronenberg, Franz Kummert
Bielefeld University, Faculty of Technology, 33594 Bielefeld, Germany
e-Mail: {susanne, franz}@techfak.uni-bielefeld.de

ABSTRACT

A major aspect of spontaneous dialogs is resembled by the collaborative process of communication where participants of a dialog cooperate by the production of utterances. Accordingly, not only independent utterances are produced but an utterance started by one agent may be continued by the other one based on the structural properties provided so far by the initial utterance. For establishing collaboration in dialogs the computer has to cooperate with the user in so far that the instructions of the user are resumed and carried on by the simulation model in that an appropriate continuation is generated by the system.

1. INTRODUCTION

Collaboration in discourse is realized in that one participant of a dialog continuous the contribution of the other one where the structural properties of the initial utterance are resumed by this continuation. Collaboration supports the production of coherent structures insofar as this evidence conceptually coordinates the various contributions of discourse [13]. Thus, collaboration establishes coherence in human-computer communication which again supports advanced human-computer interaction. The interaction of both agents is illustrated by an example of cooperative language production which is recorded from a construction dialog in German [11].

Example 1:

- I: Also jetzt nimmst du
  So now you take
- C: eine Schraube
  a bolt
- I: eine orangene mit einem Schlitz
  an orange colored one with a slit

The instructor (I) starts her utterance with *'So now you take'*. The constructor (C) continues this utterance with *'a bolt'* which leads to a completion of the utterance by the instructor given by *'an orange colored one with a slit'*. By this short clipping of the entire dialog some fundamental aspects can be observed which a computational model must take into account in order to simulate these collaborative processes:

1. Collaborative language production is often realized based on syntactic constructions where subsequent segments refer to an initial utterance. This leads to a reinterpretation of this utterance as the initial utterance is corrected or completed by the subsequent material.
2. As 'a bolt' was never linguistically introduced so far additional sources of information are necessary to complete one another's turn.

For simulating the cooperative processes in spontaneous dialogs this article presents a generation framework in that an appropriate continuation of an utterance can be generated. The generation process is based on the syntactic properties of the initial utterance, a reversed parsing strategy, and context information which can be inferred from the currently considered scene, i.e. visual semantics. Furthermore, collaboration is expressed by deriving the interpretation of the whole utterance – the parsed and generated part – where the generated part must be referred to the parsed part. Moreover, if further speech input follows this input must be processed and referred to the antecedent dialog, if necessary. As a side effect the user can correct or complete the generated or its own utterance by adding further information to it if the system has delivered an incorrect or incomplete interpretation.

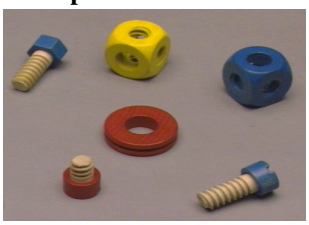
2. THE GENERATION MODEL

Collaboration has received much attention in recent computational literature (e.g. [5], [9], [12], [1], [3]). In most of this work collaboration in language has been studied mostly in problem solving scenarios similar to the considered assembly scenario where the problem is to construct a toy airplane. That collaboration is mostly studied in problem solving tasks is not surprising as in problematic situations people are forced to collaborate which implies also collaboration in communication. Collaboration considered in this approach is mostly restricted to the discrimination and determination of objects in a given scene as this is a fundamental aspect in the construction task. Accordingly, collaboration is established in that the computer resumes the last instruction of the user and delivers a more precise

*The work has been supported by the German Research Foundation (DFG) within SFB 360.

description of an object mentioned in this instruction. Following the theory of centering [4] this object description serves as a center of an utterance in that both utterances – the parsed and generated one – are linked together due this description. As the underlying syntactic constructions often used for collaboration are repairs and extrapositions which resembles a completion or correction of the source the instruction given by the user serves as the source and the generated part must resemble the target. Accordingly, the object specifications act as the reference constituent in the parsed utterance where a suitable target constituent has to be generated for. Visual context information is integrated into the generation process in order to determine the object specification which is to be generated. Verbal object descriptions are linked to object hypotheses which are generated by the vision components. On the speech side, objects can be described by TYPE (e.g. "cube"), COLOR (e.g. "red"), SHAPE (e.g. "thick"), and SIZE (e.g. "big") features. On the image side, objects are described by homogeneously colored regions – or combinations of them –, an object classification result, and a location in a two dimensional image which is defined by an enclosing polygon. In order to reflect different uncertainties in the interaction scheme a probabilistic approach is used which is modeled by a Bayesian network. Additionally, a feature structure specifying the entries of a verbally described object class is built for every constituent during the parse of an utterance [8]. In the presented approach the generation process is started if no more speech input follows. The generation process first searches in the parsed utterance for a suitable center which in extrapositions and repairs is given by a possible reference constituent. This implies, that the generation process is restricted to utterances where a reference constituent exists in the parsed utterance. Accordingly, no generation is possible for utterances missing such an object. In case of multiple reference constituents the reference constituent with the smallest distance between itself and the target constituent is in most cases the intended one. Taken this aspect into account entails that the generated part will be a correction or a completion of the last mentioned object. The last mentioned object is that part of the parsed utterance which is the first constituent on the parser stack with a non empty object specification. For this constituent an inquiry to the visual analysis is done. If a speaker is referring to an already introduced object not the full description of this object is typically used. The speaker rather use anaphoric, reduced descriptions which are distinguishing given the linguistic context [6]. Taken this aspect into account the result of the inquiry to the vision analysis is a list of all additional features, i.e. not already mentioned features, specifying the object under consideration. As described above an object is specified by the features: TYPE, COLOR, SIZE, SHAPE. If an object is already verbally described by the TYPE and COLOR feature the additional features returned from the visual analysis will specify the SIZE and SHAPE feature. The interaction between visual context and generation is demonstrated by the following examples:

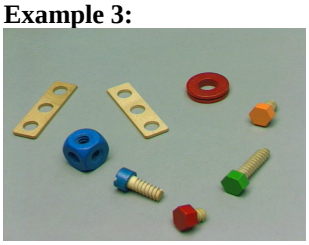
Example 2:



- I: Nimm die kurze
  Take the short one
- S: die Schraube
  the bolt
- I: ja, die rote Schraube
  yes, the red bolt

After processing the beginning of the utterance no further speech input follows which indicates that the system has to switch from parsing to generation. The last mentioned object is 'die kurze' (*the short one*). The system can infer based on the visual context information that the only object in the scene, which is short, is the red bolt. The complete specification for this bolt is: TYPE: bolt, COLOR: red, SIZE: short, SHAPE: round. As the size is already verbally specified the visual analysis result is: TYPE: bolt, COLOR: red, SHAPE: round. As mentioned above not a complete specification of the object under consideration shall be delivered by the system but rather the most discrimitive one taken the context, i.e. the already mentioned features of the reference constituent, into account. Accordingly, only the most salient additional feature is selected for the generation process based on psycholinguistic experiments [2]. The order of the feature, i.e. TYPE, COLOR, SIZE, SHAPE, expresses the frequency of the use of these features. The TYPE feature is only assumed to be most salient because it is mostly discriminative although this feature is not governed by the psycholinguistic experiments. The order of the COLOR and SIZE feature is model driven as both are equally often used for object specifications. But due to the colors of the object in the considered scenario most objects can be discriminated from other ones using the COLOR feature if the TYPE feature is already specified. Accordingly, in the example above the TYPE feature is selected for generating a description of the considered object.

Example 3:



- I: Als nächstes nimmst du
  den grünen Würfel
  Next you take the green
  cube
- S: den blauen
  the blue one

Actually, no green cube exists in the currently considered scene. The only cube which can be detected in the scene is blue. Consequently, the result of the vision analysis is: COLOR: blue, SIZE: big, SHAPE: squared. The selected feature for generation is the COLOR feature due to the order defined on these features.

2.1. OBJECT DESCRIPTIONS

As in in the underlying syntactic constructions, i.e. extrapositions and repairs, considered in this approach the reference and target constituent have an identical grammatical function the functional words of the reference constituent are used as the functional words of the generated constituent. The generation process is based on the action table of the LR(1)-parser. As described above only one feature of the vision result is selected for generation. Accordingly, it is possible to only attache this feature to the functional words already integrated into the new constituent and the generation process is done. But as in the future also more complex object descriptions shall be generated the generation process is designed in a general way to govern also this task. Accordingly, for illustration of the generation process, more precisely for illustration of the purpose of the action table of the LR(1)-parser in the generation task, it is assumed that a list of features has to be considered. Roughly spoken the action table determines the order of these features in the generated constituent and the syntactic structure of the generated constituent belonging to this order. After the integration of all functional words of the reference constituent into the new constituent the generation process is continued by searching in the action table of the LR(1)-parser for all possible actions which can be performed on the last added word. This searching process just checks the row of the action table which is given by the state of the LR(1)-parser which belongs to this word, i.e. the state of the LR(1)-parser after shifting the correspondent symbol on the parser stack. Every possible action which can be performed in this state will be considered for further processing. Simultaneously, every value which is part of the feature list returned from the visual analysis will be mapped to every inflectional form (words, for short) of this value represented in the parser lexicon. The generation process is continued by integrating these words into the new constituent. This integration is guided by the actions which can be performed from the current LR(1)-parser state taken the words as a look ahead symbol. Accordingly, the integration process is continued only on the actions where one of these words, i.e. the corresponding grammar symbol, is an appropriated look ahead symbol for. The other actions which must be performed on different grammar symbols, i.e. the actions where grammar symbols different to the words returned from the visual analysis are look ahead symbols for, are ignored. The integration process is continued by performing every possible action on these words. This implies, that the parser will switch in the next state which belongs to this action. Additionally, the action table of the LR(1)-parser is used to restrict the number of possible parses which can be derived for the list of features. The task is to produce one constituent, i.e. in this case a noun phrase, out of all features of the result list. Deriving one constituent out of all features implies that as much as possible shift actions must be performed on the considered features. This processing strategy is similar to the Right-Association principle where shift-reduce conflicts of a LR(1)-parser are resolved by preferring shift actions to reduce actions in order to attache phrases to a partial analysis as far right as possible [10]. Consequently, in case that a shift action has to be performed on one parse and a reduce action has to be performed on another one, the parse will be rejected which belongs to the reduce action. If only reduced actions can be performed on every parse the parses will be equally treated apart from the last reduce action performed on every parse. As a complete constituent must be derived out of all words all words, or more precisely stack symbols, must be reduced to one constituent. This induces, that the last reduce action must involve all symbols of the processing stack used for generation[1]. Similar to the Minimal Attachment principle [10] the parse is preferred where the reduce action is performed on most symbols, i.e. in this case all symbols. All others parses are rejected. A reduce action always implies a unification of the corresponding feature structures. A value of a feature of the vision result can be represented by several different feature structures specifying all different syntactic forms of this value. The unification of the feature structures prefers the features structures which are consistent with the already generated part of the constituent including the functional words . Accordingly, on all parses where the syntactic values of the words considered so far are inconsistent the judgment entry will be decreased [7]. Based on the best-first processing strategy the 'consistent' parses are prefered. After integrating a word in the new constituent the actions which can be performed on this word taken the remaining features of the result list as a look ahead symbol are considered in the next step of the generation process. This process is continued until all features are processed on every parse which is not rejected by this processing strategy. To derive the final interpretation of the complete dialog sequence the generated utterance will be integrated into the parsed utterance [7]. Furthermore, subsequent dialog segments will be referred to the antecedent dialog, if necessary. This enables the user to correct or complete inaccurate interpretations delivered by the system.

More than one constituent can be generated based on the generation process as outlined so far. For example if the reference constituent is 'die orange' (*the orange one*) the determiner 'die' (*the*) can be either feminine, accusative, singular or plural[2]. If for instance the vision analysis returns a female TYPE feature, e.g. bolt, two competitive

[1] Of course the parsing strategy do not know when to perform the last reduce action but at the end of the parse a constituent must be derived so that the complete processing stack used for generation is reduced, i.e. the last constituent of the stack must be adjacent to the already parsed part.

[2] The determiner 'die' can also be plural, masculine or neutrum but these cases will not considered any further due to the unification and the best-first parsing strategy.

constituents can be generated for this feature: a singular and a plural noun phrase which are both accusative and feminine. By the integration of the generated constituent into the parsed part the plural noun phrase will be judged worse as this is inconsistent for numerus with the reference constituent. Additionally, the unification of the feature structures during the generation process disambiguate cases in which several objects meets the verbal description. In the scene given in example (2) a blue cube and two blue bolts exist. If the reference constituent is given as 'die blaue' (*the blue*) the vision analysis will return a description for the cube and the bolts in absence of any linguistic knowledge. As the determiner is either either feminine, accusative, singular or plural two object descriptions are generated for the bolts: 'die Schraube' (*the blue bolt*) which is feminine, accusative, singular and 'die Schrauben' (*the blue bolts*) which is feminine, accusative, plural. Both are valid descriptions for the objects in the scene. Additionally, several descriptions are generated for the cube. But these are all judged worse as the noun *'cube'* has gender masculine and is therefore inconsistent with the feminine determiner. Thus, the descriptions of the bolts will be preferred. The integration of both constituents describing the blue bolts will lead to a preference for the singular constituent similar to the example above.

3. RESULTS

The generation model is tested on 50 utterances. This utterances are collaboration constructions recorded from German dialogs [11]. Only the source of the collaboration constructions is taken so that the generation module has to generate an appropriate target for this source. Every utterances is processed on two different scenes so that different continuations of this utterances must be generated in accordance with the visual context information. For four sentence no generation process is performed as no reference constituent exists in the utterance, i.e. no constituent exists in the source describing an object of the scenario. This is the case if the reference constituent in the source is only given by a determiner. For all other utterance a more detail descriptions is generated and this generated constituent is integrated into the source to derive the final interpretation. The length of the sentence is 6 words on average and the processing time is 114 m-sec on average using a DIGITAL AlphaStation 500/500.

4. CONCLUSION

Coordination as presented in this approach supports human-computer interaction in that the understanding process becomes transparent for the user. This implies, that in case of incorrect interpretations both the user and the computer have the opportunity to correct each other. Therefore, collaboration in human-computer communication supports problem solving together with the computer reflecting that collaboration in communication often arise when people discover that they have a problem best solved together.

5. REFERENCES

1. J. Chu-Carroll and S. Carberry. Collaborative response generation in planning dialogues. *Computational Linguistics*, 24(3), 1998.
2. H.-J. Eikmeyer, U. Schade, M. Kupietz, and U. Laubenstein. Connectionist syntax and planning in the production of object specifications. In R. M.-K. und Christiane von Stutterheim, editor, *Conceptual and Semantic Knowledge in Language Production: Proceedings of a Workshop of the Special Collaborative Research Program 245 "Language and Situation"*, volume 92, Universität Heidelberg, 1996.
3. B. D. Eugenio, P. W. Jordan, R. H. Thomason, and J. Moore. The acceptance cycle: an empirical investigation of human-human collaborative dialogues. *International Journal of Human Computer Studies*, 2000. to appear.
4. B. Grosz, A. Joshi, and S. Weinstein. Centering: A framework for modelling the local coherence of discourse. Technical Report 95-01, IRCS, University of Pennsylvania, May 1995.
5. B. Grosz and C. Sidner. Plans for discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communication*. MIT Press, 1990.
6. E. Krahmer and M. Theune. Context sensitive generation of descriptions. In *ICSLP*, volume 4, pages 1151–1154, Sydney, Australia, 1998.
7. S. Kronenberg and F. Kummert. Collaboration in human-computer communication. In *Proceedings of the 3rd Human-Computer Conversation Workshop*, pages 87–92, Bellagio, Italy, 2000.
8. S. Kronenberg, S. Wachsmuth, F. Kummert, and S. G. Disambiguation of utterances by visual context information. In *Mustererkennung 99, 21. DAGM-Symposium Bonn*, pages 338–347. Springer-Verlag, Berlin, 1999.
9. K. Lochbaum, B. Grosz, and C. Sidner. Models of plans to support communication. In *AAAI90, Proceedings of the Eighth National Conference on Artificial Intelligence*, Boston, 1990.
10. F. Pereira. A new characterization of attachment preference. In D. Dowty, L. Kartunnen, and A. Zwicky, editors, *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*, ACL Studies in Natural Language Processing, pages 307–319. Cambridge University Press, 1985.
11. G. Sagerer, H.-J. Eikmeyer, and G. Rickheit. "Wir bauen jetzt also ein Flugzeug ...": Konstruieren im Dialog - Arbeitsmaterialien. Technical Report SFB 360 "Situierte Künstliche Kommunikatoren", Universität Bielefeld, 1994.
12. M. Walker. Efficiency tradeoffs for language and action in collaborative tasks. *Language and Speech*, 39(2), 1996.
13. D. Wilkes-Gibbs. Coherence in collaboration: some examples from conversation. *Typological Studies in Language*, 31:239–267, 1995.