Ellipsis Generation in Communicative Dialogues

Kristiina Jokinen
Graduate School of Information Science, Nara Institute of Science and Technology*

1 Introduction

A common approach to the generation of elliptical utterances is to construct a semantic representation for a full sentence and then remove those concepts that are already known to the user or otherwise clear in the context (cf. [8, 3]).

We raise two objections to this approach. First, from the point of view of communication in general (Communicative Activity Theory, [1]), communicators share the assumption that everything the partner says develops the joint purpose of the dialogue, and it is not necessary to explicitly refer to all information that is to be communicated.[1] Consequently, if an elliptical contribution appropriately conveys the new information, is a *conversationally* full contribution, regardless of its *syntactic* incompleteness.

Second, it is a fallacy that conceptual specifications mostly correspond to propositions to be realised as clause-like chunks. As pointed out by [4], the underlying content of a text cannot be expressed as a set of composable facts, since the facts stand in relations and dependencies: whether a fact is explicitly expressed or not, depends not only on whether the hearer knows the fact, but also on a complex reasoning process with respect to the context.

In this paper we present a new way to plan conversationally appropriate contributions and generate elliptical utterances in natural language dialogue systems. We regard contributions as *referring* expressions: their generation must ensure that the correct conceptual situation is identified. Only new information is necessary in the contribution, and related information is added to make the reference accurate, valid, consistent and free from false implicatures. The planning process is governed by pragmatic rules which determine successfulness of an utterance in a given dialogue context. Success is measured by a preference function which partially orders the possible contributions with respect to how successfully they convey the communicative goal. This approach also tackles the problematic area between content planning and surface generation: it builds a bridge over the "generation gap" by using communicative knowledge to interleave reasoning about information content with reasoning about linguistic expression in planning conversationally adequate contributions.

The paper is organised as follows. We first discuss the distinction between explicit and implicit information and their relation to ellipsis. We then introduce the response planner algorithm, and finally we work through an example that shows how the framework has been applied in an implemented dialogue manager.

2 Explicitness and Implicitness

According to *Communicative Activity Theory* ([1]), speakers behave as rational motivated agents and trust the partner to behave in a similar way. In [5], this is formalised into two principles on which the speakers' communicative competence is based on:

(1) **The Responsiveness Principle**: Report the new information that results from the evaluation of the partner's contribution.

(2) **The Minimalism Principle**: Add contextual information only as needed to convey the whole goal, to avoid false implicatures, and to obey syntactic constraints.

The Responsiveness Principle accounts for the fact that communication takes place at all. The Minimalism Principle (a variation of Grice's Maxim of Quantity) accounts for elliptical contributions.

We make separate distinctions between explicit and implicit information on the one hand and between elliptical and complete sentences on the other hand. Explicitness and implicitness deal with concepts to be communicated to the partner, while ellipsis deals with grammatical realisation.

A relevant concept is a concept which is a part of the conceptual representation of the contribution.

An explicit concept is a relevant concept which is lexically realised on the surface level. NewInfo must always be explicit.[2]

An implicit concept is a relevant concept which is not lexically realised on the surface level, but can be inferred from the context.

*This work was started at the Centre for Computational Linguistics, UMIST, Manchester, UK, and continued at Computational Linguistics Laboratory, Nara Institute of Science and Technology, where the author is a JSPS visiting researcher. I would like to thank Yuji Matsumoto, Graham Wilcock and John Phillips for helpful and stimulating discussions on the topic.

[1] If the speaker repeats facts already known in the immediate dialogue situation, an implication can be drawn that there is an important reason why the facts are repeated.

[2] If NewInfo is unrealisable, replanning must take place since the system is unable to express the result of the evaluation.

Ellipsis denotes syntactic incompleteness. A contribution is elliptical if some of the syntactically obligatory arguments of the main verb are not lexically realised (*Rent.*), or if it does not contain a main verb (*In Bolton. Where? 12*).

Concepts are represented as world model concepts, and a contribution can realise a concept if there is a mapping from the concept to a lexical predicate. Elliptical realisation is subject to the linguistic constraints of the particular language. A concept which could be implicit in the conceptual representation may explicitly appear in the surface contribution, if it is required by language specific syntactic constraints. Conversely, implicit responses need not be elliptical. For instance, *S2* in (3) carries implicit information that the car hire companies are located in Bolton, and that the user's wish to rent a car is linked to the system's ability to give information about car hire companies. The user can infer the link from the initial setting of the dialogue, but if this were in doubt, the link would be made explicit to prevent the user from making false implicatures.

(3) Welcome to the Electronic Yellow Pages Information Service System. Please enter your request.
U1: I want to rent a car.
S1: Where?
U2: In Entwistle
S2: Where is Entwistle?
U3: In Bolton.
S2: Ok. Here is a list of car hire companies: < list >

3 The Response Planner

The Dialogue Manager (DM) and Natural Language Engine (NLE) share a Conceptual Lexicon (CL) which maps between NLE semantic predicates and World Model concepts. Thus DM reasons on language-independent conceptual representations, while NLE operates only on linguistic information and in generation. DM gives a fully specified semantic representation to NLE for surface generation.[3] The key resource of the Dialogue Manager is the Context Model, a dynamic knowledge base containing information about contributions, discourse referents, Central Concept, NewInfo, goals, and expressive, evocative and evoked attitudes. The DM also accesses the application backend (in our case: Yellow pages database), world model knowledge base, and the communicative principles. A more detailed description of the content of the different components is given in [5]. In the World Model every instantiated conceptual object has a unique index and the conceptual objects are organised into a subsumption hierarchy. Explicit concepts are also used to refer to a disjunction D of the concepts A and B (`disj(D,A,B)`), a set S of referents of the concept type C (`setOf(C,S)`), and a cardinality C of the set X (`cardinality(X,C)`). Designated concepts are mapped to application model headings.

We assume that the user goal has been recognised, expressive and evocative attitudes inferred, and that the system has formulated its own goal with a specified NewInfo. More detailed description of goal formulation can be found in [5].

The response planner produces the minimal representation for a system intention which successfully refers to the concepts to be communicated, includes NewInfo and conveys no false implicatures. The algorithm is based on four Relevance Criteria (cf. [7]): the contribution must be Accurate (represent the speaker's goal truthfully), Valid (indicate that the partner's evocative attitudes have been addressed), Consistent (the concepts must form a connected graph in the World Model and be linguistically realisable), and Free From False Implicatures (FFI, the contribution must not trigger unwanted implicatures). The Relevance Criteria are preference functions which define a partial order among the possible contributions, and the preferred contribution is the one which is among the maximal elements of each preference function.

The algorithm resembles Reiter's algorithm [7] to generate successful referring expressions for object-type entities. Reiter's algorithm is based on a user's domain knowledge, and determines a minimal set of attributes which are to be included in the object description, so that the description distinguishes the intended object from other objects in the context, is minamal and and free from false implicatures.[4] [6] formalises conversational implicature as a preference function which orders object descriptions according to their ability to successfully refer to the intended object. The function is decomposed into separate preference rules that cover each type of implicature, and the description which is the maximal element under the preference function is considered free from false implicatures. It is assumed that the preference rules do not conflict.

Our algorithm differs from Reiter's in two respects: we allow preference functions to conflict, and we exploit the fact that a partial order may have several maximal elements. We do not use Minimality as a preference

[3] The chosen task division between the two components is radical and more research is needed to draw appropriate and practical border-line between the reasoning and genertion processes. This paper assumes that DM is the main system component, and it has indirect access to linguistic information via CL. It has control over such border-line tasks as the check if a concept is linguistically realisable, if some linguistic constraint requires the planned set of concepts to be augmented with more concepts, if a particular lexical element carries some extra connotations, and the choice between ambiguous lexical entries.

[4] Minimality refers to the minimal number of conceptual componets of a contribution and not to the length of actual surface expression, thus the task is polynomial, as pointed out by [6].

criterion as such, but have incorporated it into the basic setting of the task: any of the contributions rendered maximal by a preference function can be used as a successful referring expression, and the smallest one is chosen only if it is maximal according to other preference functions as well. The maximal element satisfies all the Relevance Criteria simultaneously and can be found in the intersection of the maximal elements of the four relevance criteria. If such an element cannot be found, the goal cannot be successfully expressed in the dialogue context.

The algorithm is given below, with the following abbreviations: *Agenda* is the set of chosen concepts, *GC* is the set of goal concepts, *NI* is NewInfo, *DR* is the set of (known) discourse referents, *EEC* is the set of concepts that form the content the partner's explicit evocative attitudes, *IEC* is the set of concepts that form the content the partner's implicit evocative attitudes, *LRC* is the set of linguistically required concepts, and *LUC* is the set of linguistically unrealisable concepts (that need to be replaced by sub- or super-concepts).

(1) Initialise Agenda with NewInfo.

(2) Check Accuracy: AccAgenda contains those GCs which are not known in the context:
$AccAgenda = Agenda \cup GC \setminus DR$.

(3) Check Validity:
(a) If $sysGoal = want(s, know(u, P))$, check the difference between explicitly evoked user expectations and the planned response:
- Collect the concepts of the previous explicit evocative user attitudes (EEC).
- $Val = EEC \setminus GC$
- If $Val = 0$, then check the user's evocative intentions:
If $userGoal = want(u, know(u, P))$, then AccAgenda matches the evocative intentions and $ValAgenda = AccAgenda$.
If $userGoal = know(s, P)$, then evocative intentions and the evoked response have no common concepts since NI is based on the re-evaluation of the previous user goals, and the GCs which are not evoked by the evocative intentions must be added: $ValAgenda = AccAgenda \cup (GC \setminus EEC)$
- If $Val \neq 0$, then $ValAgenda = AccAgenda \cup Val$.
(b) If $sysGoal = want(s, know(s, P))$, check the difference between implicitly evoked user expectations and the planned response:
- Collect the concepts of the previous implicit evocative user attitudes (IEC).
- $Val = IEC \setminus GC$
- If $Val = 0$, then no unreachable implicit concepts: $ValAgenda = AccAgenda$
- If $Val \neq 0$, then unknown concepts are added (if the user responded with an elliptical utterance, the implicit concepts are known)
$ValAgenda = AccAgenda \cup (Val \setminus DR)$ (4) Check Consistency:
(a) Connectedness: there exists a path between each concept in the ConcAgenda
If a path exists, then $Connected = 0$
If not, then $Connected$ = intermediate concepts that make the graph connected
(b) Linguistic constraints:
- Map concepts in $ValAgenda$ to semantic predicates via Conceptual Lexicon. If such a mapping exists, $Conc = 0$.
If no mapping, then
if $type(Concept) = object$, map from super- or subconcept, and if this mapping exists, $Conc = ValAgenda \setminus \{Concept\} \cup \{super/subConcept\}$, else fail.
if $type(Concept) = event$ map from partOfPlanConcept.
If this mapping exists, $Conc = ValAgenda \setminus \{Concept\} \cup \{partOfPlanConcept\}$, else fail.
- Grammaticality constraints:
LRC = obligatory arguments of the main verb
LRC = full complements of explicit arguments
$ConcAgenda = ValAgenda \cup Connected \cup Conc \cup LRC$
(5) Check Freedom From False Implicatures (FFI):
if `shift(PrevCC,CurrCC) & notClosed(PrevCC)`,
$FFIAgenda = ConcAgenda \cup \{CurrCC\}$
if `CurrNI = PreviousNI & notSurprise(NI)`,
$FFIAgenda = ConcAgenda \cup \{CurrCC\}$
(6) Send FFIAgenda to the surface generator to realise.

4 An Example

Consider example 3. The context after the first user utterance are given in Fig. 1. The relevant part of the World Model contain the dashed concepts in Fig. 2.

EXPRESSIVE ATTITUDES OF INPUT:
explicit want(u,know(s,[user(u),wantE(w,u,r),
rentE(r,u,c),car(c)]))
EVOCATIVE ATTITUDES OF INPUT:
explicit want(u,want(s,know(s,[user(u),wantE(w,u,r),
rentE(r,u,c),car(c)])))
EVOKED ATTITUDES FOR RESPONSE:
know(s,[user(u),wantE(w,u,r),
rentE(r,u,c),car(c)]),
SYSTEM GOAL FOR RESPONSE:
want(s,know(s,[user(u),wantE(w,u,r),
rentE(r,u,c),car(c),location(r,_)]))
CENTRAL CONCEPTS:
topic(1,rentE(r,u,c))
topic(2,rentE(r,u,c))
DISCOURSE REFERENTS:
dr(wantE(w,u,r)), dr(rentE(r,u,c)),
dr(car(c)), dr(user(u))

Figure 1: The content of the Context Model after the first user contribution *I want to rent a car.*

The system goal is to know what is the location of the renting event.[5] NewInfo is `location(r,_)` and Agenda is initialised with this. The Accuracy rule prefers this Agenda over other possibilities, and so does

[5]The database contains several car hire companies in several locations, and the system wants to know which of them are given to the user.

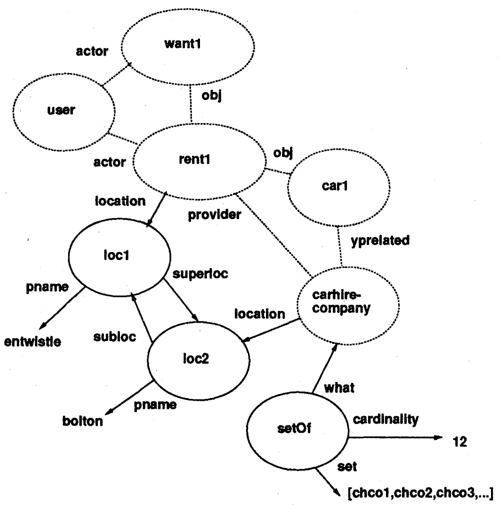

Figure 2: Relevant part of World Model for dialogue 3.

Validity Consistency and FFI. It is chosen as the preferred representation and given to the surface generator to realise.

The situation is different when the system plans a similar type of question $S2$. NewInfo is `superloc(e,_)`. The Agenda, initialised with this concept, is again preferred by the Accuracy and Validity rules. Consistency fails since there is no mapping from `superloc(e,_)` to a lexical predicate, and the superconcept, `location(e,_)`, is tried instead. This mapping is possible, and the concept `superloc(e,_)` is replaced by the concept `location(e,_)` in ConsAgenda. However, FFI does not prefer this Agenda because of false implicatures: the topic is shifted from `rent(r,u,c)` to `location(r,e)` and since the previous topic is not closed (the dialogue can continue with the topic), the current topic must be explicit to prevent the user from interpreting NewInfo as related to the previous topic (the assumption is that elliptical contributions continue the previous topic unless this is closed). Also, since the system does not want to convey surprise of the given location and still repeats NewInfo about location, the topic is added to the Agenda. The preferred FFIAgenda is thus `{location(e,_),pname(e,entwistle)}`, which is realised as *Where is Entwistle?*

5 Limitations and future work

The presented framework provides an intuitively appealing approach to response generation. It regards ellipses as natural utterances which are generated as a side-effect of the speaker fulfilling the obligations of communicative responsiveness. The approach can be compared to [8] who want to generate cooperative responses by over-answering yes-no questions. However, our system is not designed from the point of view of over-answering to extend a response with information that would prevent follow-up questions, but rather, to provide an appropriate and relevant responses on the basis of communicative activity principles. Thus we interleave the planning and realisation processes so that the decisions on the relevant concepts to be included in the response and the realisation of these concepts is flexibly controlled, and

Future work will include extension of the model to more complex 'elliptical' constructions like gapping, and refinement of the pragmatic rules. Also the criterion of consistency, the border-line between DM and NLE, needs more attention.

References

[1] J. Allwood. *Linguistic Communication as Action and Cooperation.* Department of Linguistics, University of Gothenburg, 1976. Gothenburg Monographs in Linguistics 2.

[2] J. G. Carbonell and P. J. Hayes. Recovery strategies for parsing extragrammatical language. *Computational Linguistics*, 9(3-4):123–146, 1983.

[3] N. Green and S. Carberry. A hybrid reasoning model for indirect answers. In *Annual Meeting of the Association for Computational Linguistics*, pages 58 – 65, 1994.

[4] H. Horacek. An integrated view of text planning. In R. Dale, E Hovy, D. Rösner, and O. Stock, editors, *Aspects of Automated Natural Language Generation*, pages 29–44. Springer-Verlag, Berlin, 1992.

[5] K. Jokinen. *Response Planning in Information-Seeking Dialogues.* PhD thesis, University of Manchester Institute of Science and Technology, 1994.

[6] E. Reiter. The computational complexity of avoiding conversational implicatures. In *Proceedings of the ACL*, pages 97–104. 28th Annual Meeting, University of Pittsburgh, Pittsburgh, Pennsylvania, 1990.

[7] E. Reiter. Generating descriptions that exploit a user's domain knowledge. In R. Dale, C. Mellish, and M. Zock, editors, *Current Issues in Natural Language Generation*, pages 257–285. Academic Press, London, 1990.

[8] W. Wahlster, H. Marburger, A. Jameson, and S. Buseman. Over-answering yes-no questions: Extended responses in a NL interface to a vision system. In *Proceedings of the 8th IJCAI Conference*, pages 643–646, 1983.