The Value of Weights in Automatically Generated Text Structures

Dana Dannélls

NLP Research Unit, Department of Swedish Language,
University of Gothenburg, Sweden
dana.dannells@svenska.gu.se

Abstract. One question that arises if we want to evolve generation techniques to accommodate Web ontologies is how to capture and expose the relevant ontology content to the user. This paper presents an attempt to answer the question about how to select the ontology statements that are significant for the user and present those statements in a way that helps the user to learn. Our generation approach combines bottom-up and top-down techniques with enhanced comparison methods to tailor descriptions about a concept described in an ontology. A preliminary evaluation indicates that the process of computing preferable property weights in addition to enhanced generation methods has a positive effect on the text structure and its content. Future work aims to assign grammar rules and lexical entries in order to produce coherent texts that follow on from the generated text structures in several languages.

Keywords: NLG, Ontology, Semantic Web.

1 Introduction

The ability to generate natural language text from web ontology languages and more generally knowledge bases that are encoded in RDF (Resource Description Framework) imposes new demands on natural language generators that aim to produce written text either for textual presentation or for eventual use by text-to-speech system. One of these demands concerns the process of text planning. Text planning, also referred to *Document Planning* [20], is the process responsible for producing a specification of the text's content and structure. The fact that aspects such as the user characteristics, e.g., cognitive state, desires, the background domain knowledge, and linguistic properties must be taken into account and computed simultaneously during planning makes this process computationally hard and so far there has been little success in computing a general model with a suitable structure for generating from ontologies in general and from web ontologies in particular. This brings a need to find alternative strategies to generate knowledge from ontology languages, or alternatively to adapt previously presented ideas to the new emerging technology standards.

Recent attempts to develop natural language generators that support the Web Ontology Language (OWL) and similar Semantic Web languages,[1] treat the

[1] http://www.w3.org/TR/

A. Gelbukh (Ed.): CICLing 2009, LNCS 5449, pp. 233–244, 2009.

class hierarchy as kind of directed graph that is utilised to produce a coherent text [3, 4] with the most common algorithms including top-down approaches. To enhance personalisation and improve the clarity of the text content describing an object in a hierarchy, these approaches have been combined with comparison methods whose goal is to facilitate learning by relating new concepts to a user's existing knowledge [12, 17]. Yet, one of the main questions that arises in this context is how to capture and expose the relevant ontology content to the reader.

In this paper we present a text planning technique that has been developed to explore the value of assigning weights to ontology properties in addition to comparison methods. The generation technique is optimised to tailor descriptions about a concept from the rich logical structure of Web ontologies and was implemented as a part of a question-answering system. It combines top-down and bottom-up algorithms with enhanced comparison methods to produce a personalised text structure. To test the method performance we run the system on a range of user queries with different user preferences. The generation results indicate that the process of computing preferable property weights in addition to known generation techniques has a positive effect on the text structure and its content. An experiment was conducted to evaluate the generation results using human subjects. The evaluation results show the benefits of manipulating the ontology knowledge on the basis of pre-assigned property weights.

The remainder of this paper is structured as follows. In section 2 we describe the prior approaches in more detail. In section 3 we present the methodology of the generation machinery and the motivation behind the implementation. In section 4 we describe the implementation and the text planning approach. In section 5 we report on the experimental setup and present the evaluation results. In section 6 we discuss their implications and we conclude with section 7.

2 Background

2.1 Semantic Web Ontologies

An *Ontology* is defined as a representation of a shared conceptualisation of a specific domain and plays a vital role in the Semantic Web [2] as it provides a shared and common understanding of a domain that can be communicated between people and heterogeneous, distributed application systems.

Web ontology languages are built upon the RDF and RDF Schema.[2,3] The basic RDF data model contains the concepts of resource in terms of named properties and their values. It is an object-property-value mechanism, which can be seen as forming a graph where each edge represents a *statement* that the resource at the starting end of the edge, called the *Subject* of the statement has a property called the *Predicate* of the statement with a value called the *Object* of the statement. This is shown in Figure 1. Every elliptical node with a label corresponds to a resource and every edge in the graph represents the property of the resource. Formally:

[2] http://www.w3.org/RDF/
[3] http://www.w3.org/TR/rdf-schema/

Definition 1. An ontology O=(G,R) where G is a labeled graph and R is a set of rules. The graph G=(V,E) comprises a finite set of nodes V, and a finite set of edges E. An edge e belonging to the set of edges E is written as (n_1,α, n_2) where n_1 (the subject) and n_2 (the object) are labels of two nodes belonging to a set of nodes V and α is the label (the predicate) of the edge between them.

2.2 Planning the Text Structure from Web Ontologies

The fact that the RDF's abstract syntax can be represented as a directed graph which corresponds to the structure of a coherent text was exploited by various authors who utilise top-down approaches to generate natural languages [4, 18, 22]. As pointed out by these authors, selection methods which follow the ontology graph structure pose several difficulties on the task of planning the text content. One of those is the fact that web ontologies are described as resources and are identified with URIs. This means that they can act as fields not just in the local store but anywhere they appear; when generating natural languages from ontologies it is not always clear where to begin to acquire knowledge about the concept that will be described. Recently, a new approach to content planning has been suggested by [16] who impose a bottom-up method to identify appropriate text contents. They follow an approach that is associated with conversational maxims to select and plan consistent and informative contents [16, 23]. Our approach is most closely in line with [16, 23], however our goals are different. While Mellish and Pan [16] aim to find optimal axioms that are language motivated by inducing new inferences, we aim to improve the text content and structure by combining different generation approaches with preferred property weights. Here we describe an attempt to enhance the input of an NLG system with some domain specific preferences in a way that is adaptive to the task at hand and test whether the generation results actually improve.

2.3 Tailoring the Content and Form of the Text

Bontcheva [3] extends the approach presented in [4] towards portability and personalisation. She presents an approach for producing tailored summaries by accounting for the user preferences that are imposed during the last generation phase, mostly to adapt the length of the generated text. No weights are computed to distinguish what should be included in the text content, and thus there is no adaptation in terms of the contextual information. In M-PIRO [1], it is the user himself who chooses the information that should be included in the generated text and specifies his/her preferred language. This is accomplished through an authoring tool that makes the properties of the object visible to the user. The specified preferences are stored in a user model that is consulted during generation. Similarly to ILEX [18] their user model contains scores indicating the educational value of the chosen information as well as how likely it is for him/her to find a particular type of information interesting. Our approach adds an addition feature to those as it allows to define a set of properties with higher

weights which can be interleaved with the user model and computed during the comparison process.

3 Methodology

3.1 Conveying Semantic Information

To make certain predictions that will help us to convey an ontology content and will allow the system to generate certain continuities in the text structure, there are several questions that are asked, these are: what statements must occur; where can they occur; how often must they occur. Answers to these questions which guide our generation approach depend on the statement's property weight, the ontology content, the user preferences, the context, etc. Let us introduce the following text.

Text T1
U: What is Ghm156?
S: Ghm156 is titled "Vid Rya Strand". Ghm156 was painted by Ekholm Gideon.
U: Who is Ekholm Gideon?
S: Ekholm Gideon is a painter. Ekholm Gideon was born in Sweden.

Text T1 is an example of a successful interaction sequence with the user, the user model in this context was: UM={a18,eN,g2,lS}, following the UM attributes described in section 3.2. The four statements that were generated by the system have received the highest property weights, given the ontology content. A fragment of the ontology from which the ontology statements were generated is shown is Figure 1, in this ontology four domain ontologies are emerged.

We consider this interaction to be a successful one since in the text sequence produced by the system the generated ontology statements that are relevant to the topic of the conversation are presented. Thereby following the Grice's conversational maxims of quantity, i.e., the contribution to the conversation is informative. There is no abundance of information and the generated statements allow the user to ask back on one of the new concepts given in the generated description, e.g., the title, the painter place of birth, etc., from which the system can generate new descriptions relevant to natural language presentation.

To find an adequate sequence of statements about a concept described in the ontology and be able to present the related statements that are relevant in the context, are relevant to the user and eases the user understanding about it, we implemented a stepwise text planning (described in section 4). The planning procedure combines top-down and bottom-up algorithms with comparison techniques to generate relevant content about a concept described in an ontology. In addition it is possible to specify a set of properties with higher weights that can be computed during the comparison process.

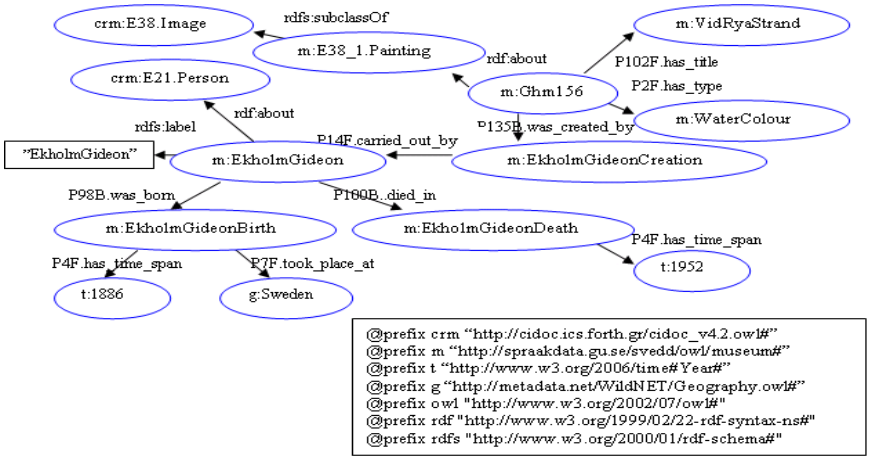

Fig. 1. Classes and properties represented in RDF syntax

3.2 Tailoring the Ontology Content

We aim to establish the rhetorical text content that supports reader and listener preferences. This is accomplished with the help of two modules: (1) the *User Module (UM)*, holds metadata information about the user's: age $a \in \{$ 7-16, $\geq 17\}$; expertise $e \in \{$ expert, non-expert $\}$; generated facts per sentence $g \in \{1, 3, \geq 4\}$ preferred textual complexity $l \in \{simple, complex\}$. (2) the *Memory Module (MM)*, represents the user knowledge, filters out repetitive RDF statements and ranks the selected statements. As the discourse evolves the memory increases; depending on the user module, statements in the memory might receive higher selection priority (section 4.2). This information characterise the user specific part of the input to a single invocation of the generation system.

Similarly to [3, 21], we utilise the names of the ontology concepts and properties to generate the lexicon and produce the text content. Our point of departure is the English language in which the ontology information is given. However, we intend to map each concept and property to its appropriate lexical entry in languages other than English and implement a grammar that makes use of those entries to generate natural language contents.

4 Implementation

4.1 The Generation Machinery

Our approach was implemented within a question-answering system where users can access a knowledge base of information in natural language [13]. The system architecture is introduced in [9]. The initial input data to the process are an ontology file and a user profile file. The user profile holds the user preferences that are stored in the UM. The ontology knowledge is held in a Jena store from

which information is retrieved.[4] The system generates a description about the concept described in the ontology that was chosen by the user. The output is a set of content elements describing the input concept for the given case. It is a subset of verbalised statements describing the input concept.

4.2 Stepwise Text Planning

The text planning module is decomposed into two distinct phases [5], it is a flexible approach that allows to exploit text possibilities [15]:[5] (1) *rhetorical representation*, deciding on how to select and organise the data (see below); (2) *document representation* (also called surface realisation) distributing the available data among sentences, paragraphs and perhaps vertical lists in the hope that it will permit a coherent realization as text. Here we take a simple approach to complete the generation process, i.e., concepts are assumed to be lexicalised as nouns and properties as verbs.[6]

The rhetorical representation module acquisition problem is decomposed in two main steps: Content selection and Content organisation.

Content Selection. Content selection operates over a relevant data that has been identified within the generator, see (1a), Table 1. Given the user query, the user model, the memory model, the ontology knowledge-base and a set of scored properties (edges) the task is to select the informative statements that meet the user request and that eases the user understanding about it.

First, all the edges in which the concept n appears in are selected. Second, every concept, i.e.,n_new other than the input one that has a path from n in G is selected. The selected edges are added to a subgraph *G'*, the prim sign ' indicating a subset.

Scoring Equation. Scores are computed for every selected edge according to the equation presented in (1b) Table 1 that was partially inspired by [12].
W_α: the edge property weight;

$Hier_n$: hierarchical distance between the selected concept and the compared resource (i.e., the subject node of the edge in focus);
$Hist_n$: historical distance, i.e., the amount of generated edges after the edge in focus was presented to the user, 0 if it was never presented.

Content Organisation. In this phase we assume there is no useful organisation to the taxonomic information of the selected subgraphs, or alternatively that

[4] `http://jena.sourceforge.net/`

[5] This process of text planning is equivalent to the two processing modules: *Content Determination* and *Content Planning* that were proposed by [19].

[6] Although there appears to be similarities between lexical entries and concepts, in linguistics and philosophy the term *concept* is defined as a nonlinguistic psychological representation of a class of entities in the ontology, where verbs distinguish what properties it has.

Table 1. Content selection algorithm, following the formal ontology Definition 1, section 2.1.

(1a) Statement selection:
function SELECT(n, G)
Input a node n, and an ontology graph G
For $n \in$ V
Add (V_n,E_n) to G'
For $n\_new \in$ V
Add (V_{n_new},E_{n_new}) to G'
return G'

(1b) Score selected statement:
procedure SCORE(E,p)
Input a set of edges E, and a set of properties p
For $e \in$ E
Score(e)= W_α + $Hier_n$ + $Hist_n$

such organisation as there is, follows the ontology structure. Given a set of scored edges that cover the input query, the task is to look for the relevant ones and organise them accordingly to generate the final output. This step is carried out by a stochastic search method [14, 15]. The stochastic search method is a form of heuristic search that executes the following generic algorithm:

1. Randomly pick one or more edges from the set, in such a way as to prefer items with the highest scores.
2. Use these to generate one or more new random variations.
3. Add these to the set, possibly removing less preferred edges in order to adapt the size to the user requirements.

5 Evaluation

5.1 The Domain Ontology

Our domain ontology follows the CIDOC Conceptual Reference Model (CRM) thesaurus standard.[7] It is a conceptual model that subscribes an object-centred view of the CH domain. The model comprises 81 relations and 244 concepts and covers the semantic field of hundreds of schemata [10].

The domain ontology was created from the Carlotta database,[8] which is designed to be equally applicable to the CIDOC-CRM and covers objects from cultural history, photos, literature, archaeology, theatre, etc. It was enhanced with about 150 new concepts and properties, each of which was assigned with a *rdfs:label* that links its lexical string-name. Figure 1 illustrates a fragment of the data represented as RDF graph (in this graph only one label is made visible).

[7] `http://cidoc.ics.forth.gr/`
[8] `http://carlotta.gotlib.goteborg.se/pls/carlotta/welcome`

Table 2. Property list for scoring edges (w1-less valuable, w2-valuable, w3-most valuable).

View	Property name	Weight
Artist name	carried_out_by	w3
Artwork titles	has_title	w3
Artist's birth date	was_born	w2
Artist's death date	died_in	w1
Artist birth place	took_place_at	w2

5.2 Adjusting the Domain Properties

The experiment presented here was carried out within the cultural heritage domain and the heuristics for selecting and ordering properties are based on the results from pilot studies that were conducted in order to gauge user needs and preferences with regards to browsing and exploring material related to artists and artwork [6, 7].

As a result of these studies four main information views that correspond to user needs were identified. We learned that the four views that were found most helpful and interesting for the variety of users correspond to five properties in our domain ontology. The identified views, their corresponding properties and their weights are specified in Table 2.

5.3 Experiment and Result

We run the system on a range of user queries with different user preferences and generated two outputs for each query, one that applies the method described here (Our method), and another that does compute user preferences but does not utilise the scores and the ordering component, i.e., the results follow the ontology structure (Simple method). Below follows one example of the generation results.

Our method

Text T2
(1) U: What is Ghm156?
(2) S: Ghm156 is a painting. Ghm156 has title "Vid Rya Strand". Ghm156 was carried out by Ekholm Gideon.
(3) U: Who is Ekholm Gideon?
(4) S: Ekholm Gideon is a painter. Ekholm Gideon's birth took place in Sweden. Ekholm Gideon birth date has time span 1886.

Simple method

Text T3
(1) U: What is Ghm156?
(2) S: Ghm156 is a painting. Ghm156 has type water colour. Ghm156 has title "Vid Rya Strand".

Table 3. Evaluation results. Quality (Q), Coherency (C), Interest (I)

	Our method	Simple method
Q	3.6	1.8
C	2.3	2.8
I	3.8	1.9

(3) U: What is Ghm156?
(4) S: Ghm156 has time span 1924. Ghm156 was created by Ekholm Gideon's creation. Ghm156 was carried out by Ekholm Gideon.

In text T2 that was produced using our approach, three of the most important properties (according to our property set) are presented already after the first enumeration question, which enables the user to precede with the next question about the new concept, e.g., "Ekholm Gideon". When we employed the simple method approach, text T3, the user needs to repeat on the query about the same concept, e.g., "Ghm156" since the information provided after the first enumeration question does not contribute with informative knowledge. In this case the generated statements are not consist and violet Grice's maxim.

Fourteen interaction sequences, similar to the above examples, were generated and presented to non-experts human subjects, in total eleven subjects participated in the evaluation. Each participant was asked to evaluate the usefulness of each interaction sequence in terms of: (a) Quality (Q), whether the content of the generated statements were relevant and helpful in describing the required object; (b) Coherency (C), whether the generated text structures were coherent and made sense; (c) Interest (I), whether the presented statements (facts) invoked the user interest. For this evaluation a five-point scale (0-poor, 5-excellent) was used. We calculated the mean value of results, these are summarised in Table 3.

A closer look at the generated text structures that were presented in different points of the interaction sequences showed there were cases where the generated content contained a mixture of statements describing different concepts, yet that are all related to the required concept. This may explain why the simple method is superior in "coherency". On the other hand in "quality" and "interest", our method outperforms over the simple approach, which is encouraging.

6 Discussion

Though the idea of exploiting properties of the ontology concepts for generation tasks is not new, the approach here is new in regards to testing in practice how the choice of the property weights effects the text structure and its content with the aim to promote insights into generating knowledge from web ontologies. The fact that content determination is not bounded to the ontology structure makes it possible to gradually present information that accommodates to different contextual degrees and user needs.

The choice of employing a generation approach such as the one presented here that is compatible with employing a domain-specific ontology is based on the relative ease by which such knowledge might provide solutions for building domain-independent generators. Currently it is assumed that a task-specific approach such as the one presented here is tied to the domain ontology and operates at the object level, however, when merged with other ontologies it may operate on meta-level [22].

The approach presented here was only tested on a small ontology with, where only a few subjects participated in the evaluation, a question that comes to mind is how well does it scale [11]. From our observation we anticipate that operating on larger ontologies may give raise to several modifications, for example the selection strategy may result in a large content when retrieving all knowledge about an object, this might be limited by putting an exceeds threshold on the depth and length of the required graph.

The growing body of research that generates from structured databases has directed different methods towards comparisons to enhance comprehension and improve the clarity of texts for the end-user [8, 17, 18]. Comparison methods can reveal the patterns of contrast and similarity [12] and have proven to be useful to remove redundant information, a problem that is exhibited in RDF's [4]. The specific selection strategy adopted here accommodates to these approaches and has proven these methods feasibility for generating from a web ontology.

7 Conclusion and Future Work

In this paper we presented a generation approach to text planning that has been developed to explore the value of assigning weights to domain specific properties. The generation method combines bottom-up and top-down approaches with enhanced comparison techniques to accommodate for the complex structure of Web ontologies. It was implemented within a question-answering framework where the primary goal was to tailor descriptions about a concept described in an ontology to a variety of users. The generated results show the benefits of assigning preferred property weights to enhance the quality and relevance of the generated content elements. A preliminary evaluation indicates that when several factors are enforced during planning, users' interest about the content describing an ontological concept seems to increase.

Although this study focused on a domain specific ontology, and conclusions were drawn based on a small amount of generation results, the findings and technical principles behind the presented methodology could likely to be generalised to other domains. Furthermore, an evaluation can potentially be repeated to confirm the generation results and to test how well does the method scales. Future work aims to assign grammar rules and lexical entries in order to produce coherent texts from the generated text structure elements. In this paper we emphasised mainly the text structure and rhetorical content, but it is necessary to cover linguistic aspects to motivate the chosen text structures for producing grammatically correct texts.

References

1. Androutsopoulos, I., Kokkinaki, V., Dimitromanolaki, A., Calder, J., Oberl, J., Not, E.: Generating multilingual personalized descriptions of museum exhibits: the m-piro project. In: Proceedings of the International Conference on Computer Applications and Quantitative Methods in Archaeology (2001)
2. Berners-lee, T.: Semantic Web Road Map. W3C Design Issues (October 1998), `http://www.w3.org/DesignIssues/Semantic.html`
3. Bontcheva, K.: Generating tailored textual summaries from ontologies. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 531–545. Springer, Heidelberg (2005)
4. Bontcheva, K., Wilks, Y.: Automatic report generation from ontologies: The MI-AKT approach. In: Meziane, F., Métais, E. (eds.) NLDB 2004. LNCS, vol. 3136, pp. 324–335. Springer, Heidelberg (2004)
5. Bouayad-Agha, N., Power, R., Scott, D.: Can text structure be incompatible with rhetorical structure? In: Proceedings of First International Natural Language Generation Conference (INLG), pp. 194–200 (2000)
6. Capra, R., Marchionini, G., Oh, J.S., Stutzman, F., Zhang, Y.: Effects of structure and interaction style on distinct search tasks. In: JCDL 2007: Proceedings of the 2007 conference on Digital libraries, pp. 442–451. ACM Press, New York (2007)
7. Clough, P., Marlow, J., Ireson, N.: Enabling semantic access to cultural heritage: A case study of tate online. In: Proceedings of the ECDL. Workshop on Information Access to Cultural Heritage, Aarhus, Denmark (2008) ISBN 978-90-813489-1-1
8. Dale, R., Reiter, E.: Computational interpretations of the gricean maxims in the generation of referring expressions. Cognitive Science 19, 233–263 (1994)
9. Dannélls, D.: A system architecture for conveying historical knowledge to museum visitors. In: Proceedings of the 12th European Conference On Research And Advanced Technology For Digital Libraries (ECDL) Workshop on Information Access to Cultural Heritage, Aarhus, Denmark (2008) ISBN 978-90-813489-1-1
10. Doerr, M., Ore, C.E., Stead, S.: The cidoc conceptual reference model: a new standard for knowledge sharing. In: ER 2007: Tutorials, posters, panels and industrial contributions at the 26th International Conference on Conceptual Modeling (2007)
11. Hardcastle, D., Scott, D.: Can we evaluate the quality of generated text? In: The 6th edition of the Language Resources and Evaluation Conference (LREC 2008), Marrakech, Morocco (2008)
12. Isard, A.: Choosing the best comparison under the circumstances. In: Proceedings of the International Workshop on Personalization Enhanced Access to Cultural Heritage (PATCH 2007) (2007)
13. Johansson, P., Degerstedt, L., Jönsson, A.: Iterative development of an information-providing dialogue system. In: Proceedings of 7th European Research Consortium for Informatics and Mathematics (ERCIM) Workshop (2002)
14. Mellish, C., Knott, A., Oberlander, J.: Experiments using stochastic search for text planning. In: International Conference on Natural Language Generation (1998)
15. Mellish, C., Oberlander, J., Knott, A.: An architecture for opportunistic text generation. In: The Ninth International Workshop on Natural Language Generation (1998)
16. Mellish, C., Pan, J.Z.: Natural language directed inference from ontologies. Artificial Intelligence 172, 1285–1315 (2008)
17. Milosavljevic, M.: Content selection in comparison generation. In: The 6th European Workshop on Natural Language Generation (6th EWNLG) (1997)

18. O'Donnell, M.J., Mellish, C., Oberlander, J., Knott, A.: ILex: An architecture for a dynamic hypertext generation system. Natural Language Engineering 7, 225–250 (2001)
19. Reiter, E.: Has a consensus NLG architecture appeared and is it psycholinguistically plausible? In: Proceedings of the Seventh International Workshop on Natural Language Generation, Kennebunkport, Maine, pp. 163–170 (1994)
20. Reiter, E., Dale, R.: Building Natural Language Generation Systems. MIT Press and The McGraw-Hill Companies,Inc., Singapore (2000)
21. Wilcock, G.: Talking owls: Towards an ontology verbalizer. In: Human Language Technology for the Semantic Web and Web Services, Sanibel Island, Florida, pp. 109–112 (2003)
22. Wilcock, G., Jokinen, K.: Generating responses and explanations from rdf/xml and daml+oil. In: Knowledge and Reasoning in Practical Dialogue Systems IJCAI, Acapulco, pp. 58–63 (2003)
23. Young, M.R.: Using grice's maxim of quantity to select the content of plan descriptions. Artificial Intelligence 115, 215–256 (1999)