Identifying the mapping of semantics onto language: going beyond the text *

Donia Scott and Cécile Paris
Information Technology Research Institute
University of Brighton
Lewes Road
Brighton BN2 4AT, UK
email:{donia.scott,cecile.paris}@itri.bton.ac.uk

Abstract

From a generation point of view, our goal is to identify the appropriate mappings between the semantics to be conveyed and expressions in language, in the context of multilingual instruction generation. We study this problem focusing on the identification of the realisations of the relationships of the various components of the task the reader is being instructed about. Corpus analysis to study this issue is tricky as there is a real danger of circularity, by identifying the underlying semantic relations (or styles) based on surface features of the text, which renders any conclusions as to how these semantics are then expressed in text invalid. In this paper, we explain how it is necessary to go beyond the text to address this problem, and show how we have been able to apply this method in our work.

Introduction

An underlying goal of research in Natural Language Processing is to define the appropriate mappings between the semantics to be conveyed and expressions in language required either to understand or to generate text. At the ITRI, our concern with this issue is from the perspective of generation, in particular, the generation of pragmatically congruent instructional texts in multiple languages. To this end, we are required to identify the underlying structure of information contained in the genre of instructions and their range of possible (and preferred) expressions in the target languages.

Our work forms part of a growing body of research on the study of instructional text. Among the previous work most relevant to our goals, that of Balkanski ((1993)), Di Eugenio ((1993; 1992)) and Vander Linden ((1994; 1994)) is particularly pertinent.

Vander Linden's perspective on this problem is from the point of view of generation, attempting to characterise, within the framework of Systemic Linguistics, the decisions underlying writers' choice of particular purpose expressions. Balkanski and Di Eugenio also attempt to characterise purpose clauses, this time from the perspective of Natural Language Understanding; for Balkanski, the goal is to show that the semantics of purpose (and rationale) clauses requires specifying beliefs and intentions, whereas for Di Eugenio it is to identify the types of inferences that underlie the process of understanding purpose clauses. A common feature of the work of Balkanski and Di Eugenio is the characterisation of the grammatical forms under consideration in terms of the semantic relations between actions, first proposed by Goldman ((1970)): Generation and Enablement. In all these studies, the researchers relied totally on the surface structure on the text to identify the underlying semantic relations involved or the factors motivating a speaker's choice of a particular expression.

In our work, one issue of particular interest is the generation of the expressions appropriate to express specific semantic relationships between the component actions of a task, for each target language. We discuss how we addressed this issues, showing how we were forced to define a new methodology for studying our corpus.

Studying the mapping of semantic relations among actions and realisation

Like others, our starting point was to collect a corpus of instructional text for each language. Given our goal of multilingual generation, we ensured where possible that the chosen texts were not the product of translation. We then set about identifying from our corpus the most prevalent grammatical forms: these, not surprisingly, included purpose clause. Like Di Eugenio and Balkanski, we then tried to map these expressions against Generation and Enablement relations as defined by Goldman. Although at first glance a simple task, this process turned out to be fraught with methodological problems for the purpose of natural language generation.

First, the degree of inter- and intra-speaker variability was at a level we considered to be unacceptable. The source of this variability is three-fold: (a) Generation and Enablement tend, in the literature, to be defined in terms of canonical syntactic forms (the use of "by" as a connective between clauses expressing the component actions, and Enablement

*This work is supported by the Commission of the LRE Grant 62009 and the EPSRC Grant J19221. Tony Hartley, Judy Delin and Keith Vander Linden have made significant contributions to the work described here

with "to" or "in order to"). While using this method to identify the relations will lead to a one:one mapping from expression to relation, a range of possible other possible forms were apparent, and mapping these to the relations became problematic. (b) Even the expressions identified by Goldman turned out to be ambiguous, and it soon became evident that going only by the surface expression was not a sufficiently rigorous (or indeed, reliable) method. (c) Work on Enablement and Generation has sofar been exclusively on English, and thus no surface forms had been identified for other languages. Part of the problem is also that Enablement, as we discovered, can hold between different types of action pairs in different contexts.

Second, there is an inherent danger of circularity that arises when surface features of a text interferes with the assignment of its semantics: if one identifies the semantic relations based only on surface features of the text, then no conclusions can be drawn as to the possible expressions of these semantic expressions.

We were thus forced to study our corpus by going "beyond the text", that is beyond its surface features, and to design a more rigorous way of representing the relations between actions (and states). While going beyond the text might be considered odd and subjective, it is necessary in order to draw any conclusions about the mapping of underlying semantics and language.

To avoid the above methodological pitfalls, we analyse the instructions in our corpus by attempting to identify the different components of the task that are presented and representing them in a "plan", as a system might do if it was to produce a plan for achieving a goal. This process is only partly based on the surface features of the text. Importantly, we also rely on our understanding of the instructions, of the relationships of the stated actions in the context of the task to be achieved, and of the object under consideration. To minimize subjectivity, independent analyses are carried out, using where possible multilingual instructions.[1] The results of these analyses are then combined; disagreements (of which there are remarkably few) are then subject to further discussion.

From this new type of analysis, we have developed a way of capturing the underlying semantic information that is expressed in instructional text in terms of elements of a task. We are also able to map the semantic relations of Generation and Enablement onto these elements, thus providing a precise and robust method for capturing them. From this, we are thus able to identify without the circularity problem the mappings from Generation and Enablement to linguistic realisation.

We briefly present here the underlying representation in terms of which we represented the tasks presented in our corpus, and its relationship to Generation and Enablement. We will not, however, present our results with respect to the mappings between these relations and linguistics expressions as it has been presented elsewhere – see ((Delin *et al.* 1994)). Finally, we discuss briefly the benefits we obtained from our methodology.

The underlying representation for tasks and actions we designed to represent the information contained in instructional texts is centered around a plan. A plan comprises:

goal(s): an action (or set of actions) which motivate(s) the use of the plan;

constraints: states which must hold before a plan can be employed. Constraints cannot be achieved through planning;

side-effects: states which arise as unplanned effects of carrying out a plan;

a body: an action or action complex which executes the plan; if these are not primitive, they can themselves be achieved through another plan;

preconditions: an action or action complex which, when carried out, leads to conditions necessary for the successful execution of the plan; i.e, the body will be executable but its execution will not generate the goal (even if the constraints hold) unless the precondition is realisable. Preconditions can be planned for.[2]

Plans must minimally have a body and goal. Actions, in turn, comprise:

constraints: these have the same properties as the constraints on plans;

side-effects: these also have the same properties as the side-effects on plans;

effects: states which arise from the *bringing-about* of the action;

preconditions: like the preconditions of plans, these are either an action or an action complex and can be planned for. Unlike the preconditions of plans, however, their effects must hold before the action can take place (i.e., the action cannot result if the precondition does not hold).[3]

Actions must minimally have an effect. Finally, states can have evidences. These are phenomena which signify that the state holds. These are very important in instructional texts as they provide a way for the author to tell readers how they can verify that the actions they are performing are being done correctly.

As in all STRIPS-based planning formalisms, plans in our representation can have associated sub-plans. In our case, sub-plans arise through the body of a plan, the preconditions of a plan, or through the preconditions of an action.

Having set out this scheme, we can now proceed to identify instances of *Generation* and *Enablement* within our representation for instructions:

α *generates* β iff α is the body of a plan ϵ whose goal is β.

[1]Currently, analyses are performed by Judy Delin, Anthony Hartley, Cécile Paris, Donia Scott and Keith Vander Linden.

[2]This type of preconditions relates to Pollack's ((1986)) *generation-enabling condition*.

[3]This notion is closely related to Balkanski's ((1993)) definition of *executability condition*.

α *enables* β if α is a precondition of a plan ϵ and β is the goal of plan ϵ, or if β is the body of ϵ and α is a precondition of β.

Figure 1 shows graphically the relationship between the notion of a plan and the *Generation* and *Enablement* relations. Note that Enablement can now be clearly identified as holding between two types of action pairs.

Our refined method for identifying the relationship between the semantics and expression of multi-action sequences is thus to examine the procedural parts of our corpus in terms of their rôles in the underlying plan. With regard to applying our methodology, the following point are worth mentioning:

1. It provides a high degree of inter- and intra- coder agreement. To test the degree of reliability, we will be conducting a study of our methodology with new coders.
2. It forces us to look beyond the text and to take context into account. Going beyond the text also forces us to understand the nature of the relations expressed in the text, independently of the language it is expressed in, to derive a *language-independent* representation of the task. This is particularly important for our goal, since we are aiming at identifying the mapping of semantics to language in various languages.
3. It appears to have validity over the languages to which we have applied it: English, French, German and Portuguese. We have not encountered problems when applying it to a new language.
4. It has allowed us to identify a number of linguistic forms for expressing Generation and Enablement, beyond the purpose and rationale clause. These include conditional clauses, apposed clauses and temporal clauses.
5. It has allowed us to identify linguistic expressions which are ambiguous in terms of the semantic relations they express. This is an important point in the generation of instructions, where often, ambiguities can confuse the readers or even prevent them from successfully achieving the required task. By knowing that a possible expression is ambiguous, a generation system can either avoid the construction or plan additional text to avoid the possible confusion.
6. It has also revealed that Generation and Enablement do not always apply between actions alone, but also between actions and states. This allows us to do a more complete analysis of the corpus.
7. It provides a framework that can be applied to other relationships found in procedural texts. It is clear that Generation and Enablement are not the exhaustive list of relationships to be considered. Indeed, having identified the underlying semantic information behind Generation and Enablement with a plan framework, our need to refer to them at all, as opposed to relations between parts of the underlying plan, becomes questionable.

Given that are still working from textual data, it is clearly difficult to avoid completely the possibility of circularity; however, we do reduce the possibility significantly. Of the variability that does occur, most instances arise from the well known problem in planning of determining whether a given action is the precondition or substep of a plan. We are currently designing tests that can be done to determine the relation between two actions.[4]

Summary and Further Work

When trying to determine the mappings between semantic features and syntactic features, one need to be wary of the danger of circularity. In some cases, it is necessary to go beyond the text and attempt to understand the underlying semantic it conveys, without relying on the specific linguistic realisations. While this methodology may appear subjective, we have been able to obtain good results even in independent analyses. We are complementing this type of analysis with a coding and statistics approach to strengthen the results. Furthermore, we plan to refine the methodology in order to be able to ask subjects to perform the analysis to further diminish the risk of subjectivity.

References

Balkanski, C. T. 1993. *Actions, Beliefs and Intentions in Multi-Action Utterances*. Ph.D. Dissertation, Harvard University.

Delin, J.; Hartley, A.; Paris, C.; Scott, D.; and Vander Linden, K. 1994. Expressing procedural relationships in multilingual instructions. In *Proceedings of the Seventh International Workshop on Natural Language Generation,* Kennebuckport, MN, 21–24 June 1994.

Di Eugenio, B. 1992. Understanding natural language instructions: The case of purpose clauses. In *Proceedings of the Annual Meeting of Association for Computational Linguistics,* Newark, DE, 120–127.

Di Eugenio, B. 1993. *Understanding Natural Language Instructions: A Computational Approach to Purpose Clauses*. Ph.D. Dissertation, University of Pennsylvania. also available as IRCS Report 93-52.

Goldman, A. I. 1970. *A Theory of Human Action*. Englewood Cliffs, NJ: Prentice Hall.

Pollack, M. E. 1986. *Inferring Domain Plans in Question-Answering*. Ph.D. Dissertation, University of Pennsylvania. SRI Technical Report SRIN-403.

Vander Linden, K., and Martin, J. 1994. Expressing local rhetorical relations in instructional text: A case-study of the purpose relation. *Computational Linguistics*. to appear.

Vander Linden, K. 1994. Generating precondition expressions in instructional text. In *Proceedings of the 32nd Annual Meeting of Association for Computational Linguistics,* June 27–30, Las Cruces, NM, 42–49.

[4] Note that again, we have to be careful to avoid the circularity problem when designing our tests.

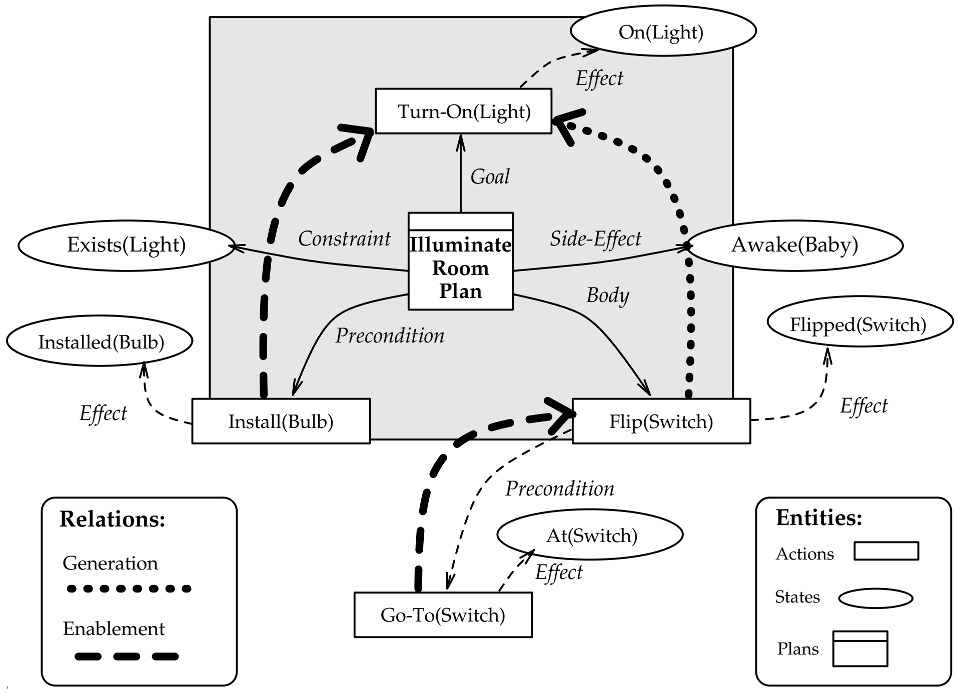

Figure 1: Generation and Enablement in the Plan for Turning on the Light