Domain Dependent Natural Language Understanding

Klaus Heje Munch

Dept. of Computer Science

Technical University of Denmark

DK-2800 Lyngby, Denmark

A natural language understanding system for a restricted domain of discourse - thermodynamic exercises at an introductory level - is presented. The system transforms texts into a formal meaning representation language based on cases. The semantical interpretation of sentences and phrases is controlled by case frames formulated around verbs and surface grammatical roles in noun phrases. During the semantical interpretation of a text, semantic constraints may be imposed on elements of the text. Each sentence is analysed with respect to context making the system capable of solving anaphoric references such as definite descriptions, pronouns and elliptic constructions.

The system has been implemented and succesfully tested on a selection of exercises.

1 Introduction

This paper describes a natural language understanding system for the domain of naive thermodynamics. The system transforms exercises formulated in (a subset of) Danish to a somewhat "adhoc" chosen meaning representation language. Given the representation of an exercise, a problem solver shall deduce its solution in a subsequent computation.

The weakest demand on the system is that it transforms texts into representations which are "equivalent" to the texts. The ultimate demand on the system and the problem solver is of course that exercises are solved correctly.

The system consists of three parts dealing with respectively *morphology*, *syntax* and *semantics*. The morphological and syntactical analyses are domain independent and only related to the natural language. The semantical analysis is dependent on both the natural language and the specific domain. During the semantical analysis of an exercise, syntactic structures are transformed into a set of logical propositions arranged as (implicitly) in the exercise. After having completed the semantical analysis, a language independent representation exists. The semantic component does not include an inferential mechanism for deducing the progress in thermodynamic experiments. Therefore, it may regard a text as being ambiguous. For instance, it may not be possible to determine the referent of an anaphora unambiguously without considering common sense reasoning. However, such ambiguities will be solved by the problem solver, which uses domain-dependent knowledge as well as commonsense knowledge (see e.g. (Hobbs, Moore 1985)), and operates with an interval-based representation of time (Allen 1984).

This paper considers only the semantical interpretation of sentences. The semantical analysis is based on a compositional principle similar to the one used by Katz and Fodor (1963). It claims that the semantical interpretation of a sentence is obtained by replacing its words or phrases with their semantic representations and combining these according to the syntactic structure of the sentence as well as the context. The interpretation is controlled by a case grammar, which consists of case frames. The case frames relate syntactic structures to a case system and place semantic constraints on their constituents. In examining if constraints are fulfilled during the analysis, a static worldknowledge is used. The most important component of the worldknowledge is an is-a hierarchy which organizes all concepts in the domain of discourse. The worldknowledge is called "static", since it does not contain "dynamic" information such as implications or preconditions of actions.

During the semantical interpretation, the context in a text is considered. Connections between elements of the text is established by resolving anaphoras due to definite nouns and pronouns. The system resolves ellipses too.

2 Meaning Formation in Sentences

Semantically, a text is regarded as a specification (denotation) of a series of propositions. In natural languages, propositions can be expressed not only by sentences, but also by other syntactic structures such as noun groups, infinitive phrases and embedded sentences. Thus a single sentence may express several propositions. The goal in understanding a text is to extract its propositions and specify them in a formal language.

A sentence can be characterized as the basic independent structure in the language. Relating language to formal logic, the meaning of a sentence can be described by a predicate which is identified by the head verb of the sentence. The arguments of the predicate are denoted by the constituents of the sentence. Such a representation is the basis for both systemic (Winograd 1983), case (Fillmore 1968) and lexical-functional (Bresnan 1981) grammars.

3 Meaning Representation

The meaning representation language is based on a case system (Bruce 1975) inspired by Fillmore's notion of deep cases. Basically, a text is represented by a list of propositions, each consisting of a proposition type corresponding to a predicate name, and a list of cases corresponding to the arguments of the predicate. Contradictory to Fillmore's notion, proposition types are not verbs, but abstract concepts defined in the case frames of a case grammar. Furthermore, cases show semantic relationships between proposition types and abstract concepts. The case system (set of cases) is chosen in a somewhat adhoc way. The cases, which are supposed to be necessary in order to describe the domain of thermodynamic exercises, are included. The cases and their use are explained below :

object	object being affected by an action or event, or being described.
instr	instrument for action.
stuff	materia
force	the thing or event forcing an action.
action	action being referred to.
attr	physical attribute.
referent	the object being referred to in a description.
reason	reason for event.
direction	direction of change.
descr	description of object.
spatial_loc	"spatial location", includes a object, which describes a physical location, a relation to the location and a direction (from/to/at).
value_loc	"value location", as spatial_loc, but including a value.
temporal_loc	"temporal location", includes an indication of time and a relation to this.

Besides the case system, the meaning representation language includes elements which make it possible to introduce or define physical objects, to connect or refer to objects, actions or events, and to relate propositions temporally. The syntax of the meaning representation language is :

```
Meaning       = Meaning_elem*;
Meaning_elem  = obj_def(Obj_iden,Context_rel,Obj_type) !
                proposition(Prop_iden,Prop_type,Cases);
Obj_iden,
Prop_iden     = identifier
Context_rel   = refer ! introd;
Prop_type     = string;
Obj_type      = stuff_obj(string)    ! single_obj(String);
Cases         = Case*;
Case          = object(Obj_iden)     ! instr(Obj_iden) !
                action(Prop_iden)    ! attr(string) ! ...
```

Notice that it is possible to reference propositions and object definitions through their identifiers. 'Context_rel' specifies whether an object is introduced in the text (introd) or being referred (refer).

As an example of the representation language consider the sentence : "The calorimeter contains 100 g water with the temperature 50 C". The corresponding representation is :

```
obj_def(C,refer,single_obj(calorimeter))
obj_def(W,introd,stuff_obj(water))
proposition(1,obj_attr_val,<object(W), attr(temperature),
                            value_loc(at,equal,50 C)>)
proposition(2,contain,<object(C), referent(W)>)
```

Here the calorimeter and the water are defined as physical objects and denoted by the identifiers C and W respectively. The calorimeter is in definite form, it is referring, and the referent cannot be found. The proposition type "obj_attr_val" relates an attribute of an object to a value or quantity. Finally, the proposition type "contain" relates an object, which contains, to an object which is contained.

4 Relating Syntax to Cases

The transformation from syntactic structures to the meaning representation language is controlled by a case grammar. The case grammar specifies the correspondence between syntactic representations, based on surface grammatical roles of phrases and sentences, and case representations.

The semantical analysis of a sentence is based on its head verb, while the analysis of a noun group is be based on the head noun and also on adjective descriptors, genitive determiners and prepositional phrases. For each head verb, head noun, etc., the case grammar contains a case frame. A case frame consists of the following parts : cases, selection, constraints, extract and presence. The "cases-part" states what a phrase shall be transformed into by means of proposition types and cases. The "selections" relate elements in the case frame to syntactic constituents. "Constraints" contain semantic constraints on elements of the case frame. "Extract" makes it possible to extract elements from compound, or complex, semantic elements, and finally, "presence" specifies whether constituents are mandatory, elliptic or optional.

As an example of a case frame consider the verb to "rise" in combination with "temperature" or any other physical attribute. Some examples of sentences containing "rise" are :

"The temperature rises 5 degrees"
"The temperature of the liquid rises from 50 to 55 degrees"
"The temperature rises"

Observe that in the first sentence, the object with the mentioned temperature is denoted by an elliptic reference. In the analysis of the sentence, it has to be found using the context, i.e. the previous sentences. A case frame for "rise" is shown below. Here the selections "subject", "sdir" and "prep" refer respectively to the subject, the direct object and prepositional phrases in a sentence. The constraint 'is_a(x,y)' means that x is of type y according to the taxonomy. 'has_attr(o,at)' means that the object o has the attribute at.

```
rise : proposition change
  cases       : object(O), attr(At), value_loc(to,equal,Rv),
                value_loc(from,equal,Sv),
                value_loc(relative,equal,Gv).
  selection   : subject(Subj), sdir(Rv),
                prep(from,Sv), prep(to,Gv), prep(with,Rv).
  constraints : is_a(At,physical attribute),
                is_a(O,physical object),has_attr(O,At)
  extract     : ex_attr(Subj,At), ex_obj(Subj,O).
  presence    : obligatory(At), elliptic(O),
                optional(Sv), optional(Gv), optional(Rv).
```

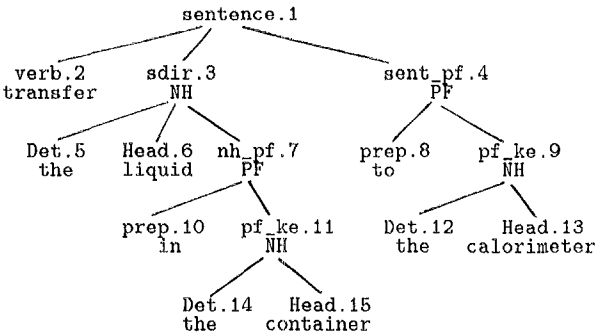
Notice that the subject, which besides being a compound structure consisting of an attribute and an object, may alternatively take form of an attribute only (because O is ellipsed). The constituents of the subject are extracted by the *ex_attr* and *ex_obj* predicates.

The semantical analysis of a syntactic structure is carried out in a mixed bottom up - top down way. The formation of the meaning of a phrase progresses bottom up, while the control of its constituents (selection of them and constraints on them) progresses top down. Generally, when a case frame is applied in the analysis of a phrase, the elements specified in the selection-part are matched with the constituents of the phrase. If an element has the same syntactical role as a constituent, the constituent is analysed, while possible constraints are imposed on it. The result of the analysis is a list of propositions derived from the phrase as well as the semantic element which the phrase denotes.

To illustrate the semantical analysis consider the sentence : "the liquid in the container is transferred to the calorimeter". Suppose the sentence is analysed in isolation, so that the definite descriptions cannot be solved. The case frames needed to analyse the sentence are :

```
transfer : proposition transfer
  cases       : object(O), spatial_loc(goal,in,G)
  selection   : sdir(O), prep(to,G).
  constraints : is_a(O,physical object), is_a(G,container).
  presence    : obligatory(O), elliptic(G).
calorimeter   : object(calorimeter).
container     : object(container).
liquid        : stuff(liquid).
in : proposition contain
  cases       : object(O), referent(R)
  selection   : head(R), prep(in,O)
  constraints : is_a(O,container), is_a(R,physical object)
  presence    : obligatory(O), obligatory(R).
```

The syntactic structure of the sentence can be depicted as a tree where each node is numbered :



The sentence analysis selects the case frame for the verb "transfer". The case frame claims the mandatory occurence of a direct object O which must be a physical object. Thus O

matches by its syntactical role the constituent identified by node 3. When analysing the corresponding noun group, the case frame for the head noun "liquid" is used at first. The constraint of the noun group (being a physical object) is fullfilled, thus the analysis proceeds. The determiner combined with the head noun determines the context dependency of the liquid as referring. Since the referent cannot be found, an object definition is generated, though marking the object as referring.

The prepositional phrase of the noun group is analysed by first selecting the case frame for the preposition "in". In this frame, it is claimed that the head noun must be a physical object. This is already known to be true. Furthermore, the head of the prepositional phrase must specify an object in the class "container". Thereafter the container is defined as an object and a proposition of type "contain" is generated. The result of the analysis of the noun group is the object identifier denoting the liquid and the "contain" - proposition.

The rest of the analysis will not be commented. The representation of the sentence is :

obj_def(L, refer,stuff_obj(liquid)),
obj_def(Co,refer,single_obj(container)),
proposition(1,contain,<object(Co), referent(L)>),
obj_def(Ca,refer,single_obj(calorimeter)),
proposition(2,transfer,<object(L),spatial_loc(goal,in,Ca)>).

5 Context-Dependent Analysis

The context-dependent analysis covers resolution of the most important types of anaphoric references. The system resolves the following types of references in a text : identic, synonymous, pronominal, adverbial, modifying and some of the elliptic references. Examples of these references are :

Identic	**a calorimeter** contains helium, and **the calorimeter** ...
Synonymous	a calorimeter contains **helium**, and **the gas** ...
Pronomial	**a calorimeter** contains helium, and **it** ...
Adverbial	**in the calorimeter** is gas, and **there** is also ...
Modifying	the calorimeter is **heated** to 50 C, and **the heating** ...
Elliptic	the calorimeter **contains** gas and the bucket (**contains**) water.
Elliptic	**a calorimeter** contains **water**. The temperature (**in the calorimeter** or **of the water**) is 50 C.

During the semantical analysis, the references are resolved as soon as they are met. In order to be able to do this, the leftmost context of a text must be reachable when analysing a phrase. The leftmost context is all propositions derived from the text so far.

The system uses no special features for delimiting the scope of referred objects. When a reference is to be solved, the objects and events specified in the leftmost context are examined. An object or event, which fullfills the constraints specified in the case frame and which matches possible syntactic features (gender and number), is claimed to be the token referred to. The resolution of synonymous references (for instance of gas in : "A container contains helium, and the gas ... ") uses the is-a hierarchy.

6 Example

The following exercise is considered :

"A cupper calorimeter with the heatcapacity 75 J/K contains 300 g paraffin. The temperature is 18 C. A cupper block with the mass 100 g is heated to 100 C, whereupon it is transferred to the calorimeter, where the temperature becomes 22 C. The specific heat of cupper is 387 J/kg*K. Find the specific heat of paraffin."

The system generates the representation shown below. The propositions are separated into time - dependent and time - independent propositions, the former are related temporally.

object specifications :
1. obj_def(C,calorimeter)
2. obj_def(P,stuff_obj(paraffin))
3. obj_def(L,block)

constant attributes :
4. consist_of(object(C),stuff(cupper))
5. obj_attr_val(object(C),attr(heatcapacity),quant(75,J/K))
6. obj_attr_val(object(P),attr(mass),quant(300,g))
7. consist_of(object(L),stuff(cupper))
8. obj_attr_val(object(L),attr(mass),quant(100,g))
9. obj_attr_val(stuff(cupper),attr(spec_heat), quant(387,J/kg*K))
10. obj_attr_val(object(P),attr(spec_heat),quant(question))

time-dependent propositions :
11. contain(object(C),referent(P))
12. obj_attr_val(sp_loc(at,in,C), attr(temperature),quant(18,C))
13. heat(object(L),value_loc(to,equal,quant(100,C))
14. block_transfer(object(L),sp_loc(goal,in,C))
15. obj_attr_val(sp_loc(at,in,C), attr(temperature),quant(22,C))

Some points worth of noticing are the resolution of the ellipsed object (location) in the second sentence of the exercise (proposition 12) and the resolution of the identic reference (the calorimeter) as well as of the pronoun (it) in the third sentence (proposition 14).

7 Conclusion

The system described in this paper transforms thermodynamical exercises expressed in Danish into a formal meaning representation language. In order to accomplish this, morphology, syntax and semantics are considered. Most important is the application of the case grammar formalism, in which semantic constraints can be imposed on phrases, causing ambiguities in a text to be removed. The case grammar have a clear, well-defined structure and is easy to extend, also to other domains.

For varied selections of thermodynamical exercises, the system has derived correct meaning representations. Thus the goal has been accomplished. Currently, the problem solver is under development.

References

Allen, J. F. 1984 Towards a General Theory of Action and Time. Artificial Intelligence 23: 123-154.

Bresnan, J. 1981 An Approach to Universal Grammar. Cognition 10: 39-52.

Bruce, B. 1975 Case Systems for Natural Languages. Artificial Intelligence 6: 327-360.

Fillmore, C. 1968 The Case for Case. In: Bach, E. and Harms, R., Universals in Linguistic Theory. Holt, Rinehart and Winston, New York.

Hobbs, J. R. and Moore, R. C. (eds.) 1985 Formal Theories of the Commonsense World. Ablex, Norfolk, New Jersey.

Katz, J. J. and Fodor, J. A. 1963 The Structure of a Semantic Theory. Language 39: 170-210.

Winograd, T. 1983 Language as a Cognitive Process. Addison-Wesley, Reading, Mass.