DEALING WITH CONJUNCTIONS IN A MACHINE TRANSLATION ENVIRONMENT

Xiuming Huang
Institute of Linguistics
Chinese Academy of Social Sciences
Beijing, China*

ABSTRACT

A set of rules, named CSDC (Conjunct Scope Determination Constraints), is suggested for attacking the conjunct scope problem, the major issue in the automatic processing of conjunctions which has been raising great difficulty for natural language processing systems. Grammars embodying the CSDC are incorporated into an existing ATN parser, and are tested successfully against a wide group of "and" conjunctive sentences, which are of three types, namely clausal coordination, phrasal coordination, and gapping. With phrasal coordination the structure with two NPs coordinated by "and" has been given most attention.

It is hoped that an ATN parser capable of dealing with a large variety of conjunctions in an efficient way will finally emerge from the present work.

0 INTRODUCTION

One of the most complicated phenomena in English is conjunction constructions. Even quite simple noun phrases like

(1) Cats with whiskers and tails

are structurally ambiguous and would cause problem when translated from English to, say, Chinese. Since in Chinese all the modifiers of the noun should go before it, two different translations in Chinese might be got from the above phrase:

(1a) (With whiskers and tails) de (cats) ("de" is a particle which connects the modifiers and the modifieds);

(1b) ((With whiskers) de (cats)) and (tails).

Needless to say, a machine translation system should be able to analyse correctly among other things the conjunction constructions before high quality translation can be achieved.

As is well known, ATN (Augmented Transition Network) grammars are powerful in natural language parsing and have been widely applied in various NL processing systems. However, the standard ATN grammars are rather weak in dealing with conjunctions.

In (Woods 73), a special facility SYSCONJ for processing conjunctions was designed and implemented in the LUNAR speech question-answering system. It is capable of analysing reduced conjunctions impressively (eg, "John drove his car through and completely demolished a plate glass window"), but it has two drawbacks: first, for the processing of general types of conjunction constructions, it is too costly and too inefficient; secondly, the method itself is highly non-deterministic and easily results in combinatorial explosions.

In (Blackwell 81), a WRD AND arc was proposed. The arc would take the interpreter from the final to the initial state of a computation, then analyse the second argument of a coordinated construction on a second pass through the ATN network. With this method she can deal with some rather complicated conjunction constructions, but in fact a WRD AND arc could have been added to nearly every state of the network, thus making the grammar extremely bulky. Furthermore, her system lacks the power for resolving the ambiguities contained in structures like (1).

In the machine translation system designed by (Nagao et al 82), when dealing with conjunctions, only the nearest two items of the same parts of speech were processed, while the following types of coordinated conjunctions were not analysed correctly:

(noun + prep + noun) + and + (noun + prep + noun);
(adj + noun) + and + noun.

(Boguraev in press) suggested that a demon should be created which would be woken up when "and" is encountered. The demon will suspend the normal processing, inspect the current context (the local registers which hold constituents recognised at this level) and recent history, and use the information thus gained to construct a new ATN arc dynamically which seeks to recognise a constituent categorially similar to the one just completed or being currently processed. Obviously the demon is based on expectations, but what follows the "and" is extremely uncertain so that it would be very difficult for the demon to reach a high efficiency. A kind of "data-driven" alter-

* Mailing address:
Cognitive Studies Centre
University of Essex
Colchester CO4 3SQ, England.

native which may reduce the non-determinism is to try to decide the scope of the left conjunct retrospectively by recognising first the type of the right conjunct, rather than to predict the latter by knowing the category of the constituent to the left of the coordinator which is "just completed or being currently processed" -- an obscure or even misleading specification.

I CASSEX PACKAGE

CASSEX (Chinese Academy of Social Sciences;University of Essex) is an ATN parser based on part of the programs developed by Boguraev (1979) which was designed for the automatic resolution of linguistic ambiguities. Conjunctions, one major source of linguistic ambiguities, however, were not taken into consideration there because, as the author put it himself, "they were felt to be too large a problem to be tackled along with all the others" (Boguraev 79, 1.6).

A new set of grammars has been written, and a lot of modifications has been made to the grammar interpreter, so that conjunctions could be dealt with within the ATN framework.

II PARSING MATERIALS

The following are the example sentences correctly parsed by the package:

Ex1. The man with the telescope and the umbrella kicked the ball.

Ex2. The man with the telescope and the umbrella with a handle kicked the ball.

Ex3. The man with the telescope and the woman kicked the ball.

Ex4. The man with the telescope and the woman with the umbrella kicked the ball.

Ex5. The man with the child and the woman kicked the ball.

Ex6. The man with the child and the woman with the umbrella kicked the ball.

Ex7. The man with the child and the woman is kicking the ball.

Ex8. The man with the child and the woman are kicking the ball.

Ex9. The man with the child and the umbrella fell.

Ex10. The man kicked the ball and the child threw the ball.

Ex11. The man kicked the ball and the child.

Ex12. The man kicked the child and the woman the ball.

Ex13. The man kicked the child and threw the ball.

Ex14. The man kicked and threw the ball.

Ex15. The man kicked and the woman threw the ball.

III ELEMENTARY NP AND EXPANDED NP

The term 'elementary NP' is used to indicate a noun phrase which can be embedded in but has no other noun phrases embedded in it. A noun phrase which contains other, embedded, NPs is called 'expanded NP'. Thus, when analysing the sentence fragment "the man with the telescope and the woman with the umbrella", we will have four elementary NPs ("the man", "the telescope", "the woman" and "the umbrella") and two expanded NPs ("the man with the telescope" and "the woman with the umbrella"). We may well have a third kind of NP, the coordinated NP with conjunction in it, but it is the result of, rather than the material for, conjunction processing, and therefore will not receive particular attention. In the text followed we will use 'EL-NP' and 'EXP-NP' to represent the two types of noun phrases, respectively.

LEFT-PART will stand for the whole fragment to the left of the coordinator, and RIGHT-PART for the fragment to the right of it. LEFT-WORD and RIGHT-WORD will indicate the word immediately precedes and follows, respectively, the coordinator. The conjunct to the right of the coordinator will be called RIGHT-PHRASE.

VI CSDC RULES

Constraints for determining the grammaticalness of constructions involving coordinating conjunctions have been suggested by linguists, among which are (Ross 67)'s CSC (Coordinate Structure Constraint), (Schachter 77)'s CCC (Coordinate Constituent Constraint), (Williams 78)'s Across-the-Board (ATB) Convention, and (Gazdar 81)'s nontransformational treatment of coordinate structures using the conception of 'derived categories'. These constraints are useful in the investigation of coordination phenomena,but in order to process coordinating structures automatically, some constraint defined from the procedural point of view is still required.

The following ordered rules, named CSDC (Conjuncts Scope Determination Constraints), are suggested and embodied in the CASSEX package so as to meet the need for automatically deciding the scope of the conjuncts:

1. Syntactical constraint.

The syntactical constraint has two parts:

1.1 The conjuncts should be of the same syntactical category;

1.2 The coordinated constituent should be in conformity syntactically with the other constituents of the sentence, eg if the coordinated constituent is the subject, it should agree with the finite verb in terms of person and number.

Acoording to this constraint, Ex8 should be analysed as follows (the representation is a tree diagram with 'CLAUSE' as the root and centred around the verb, with various case nodes indicating the dependency relationships between the verb and the other constituents):

```
(CLAUSE
  (TYPE DCL)
  (QUERY NIL)
  (TNS PRESENT)
  (ASPECT PROGRESSIVE)
  (MODALITY NIL)
  (NEG NIL)
  (V (KICK ((*ANI SUBJ)
            ((*PHYSOB OBJE)
             ((THIS (MAN PART)) INST) STRIK))*
     (OBJECT ((BALL1 ...))
             (NUMBER SINGLE)
             (QUANTIFIER SG)
             (DETERMINER ((DET1 ONE)))
     (AGENT
      AND
      ((MAN ...)
       (NUMBER SINGLE)
       (QUANTIFIER SG)
       (DETERMINER ((DET1ONE)))
       (ATTRIBUTE ((PREP (PREP WITH))
                   ((CHILD ...)
                    (NUMBER ...)
                    ...
      ((WOMAN ...)
       ...
```

while Ex7 (and the more general case of Ex5) should be analysed roughly as:

```
(AGENT
 ((MAN ...)
  (NUMBER SINGLE)
  (QUANTIFIER SG)
  (DETERMINER ((DET1 ONE)))
  (ATTRIBUTE ((PREP (PREP WITH))
              AND
              ((CHILD ...)
               (NUMBER ...) ...)
              ((WOMAN ...)
               ...
```

2. Semantic constraint.

NPs whose head noun semantic primitives are the same should be preferred when deciding the scope of the two conjuncts coordinated by "and". However, if no such NPs can be found, NPs with different head noun semantic primitives are coordinated anyhow.

According to rule 2, Ex1 should be roughly represented as 'The man with (AND (telescope) (umbrella))'; Ex2, 'The man with (AND (telescope) (umbrella with a handle))'; Ex3, '(AND (man with telescope) (woman))' and Ex4, '(AND (man with telescope) (woman with umbrella))'.

3. Symmetry constraint.

When rules 1 and 2 are not enough for deciding the scope of the conjuncts, as for Ex5 and Ex6, this rule of preferring conjuncts with symmetrical pre-modifiers and/or post-modifiers will be in effect:

Ex5. ...with (AND (child) (woman)) ...

Ex6. (AND (the man with ...) (the woman with ...))...

4. Closeness constraint.

If all the three rules above cannot help, the NP to the left of "and" which is closest to the coordinator should be coordinated with the NP immediately following the coordinator:

Ex9. The man with (AND (child) (umbrella)) fell.

V THE IMPLEMENTATION

The seemingly straightforward way for dealing with conjunctions using the ATN grammars would be to add extra WRD AND arcs to the existing states, as (Blackwell 81) proposed. The problem with this method is that, as (Boguraev in press) pointed out, "generally speaking, one will need WRD AND arcs to take the ATN interpreter from just about every state in the network back to almost each preceding state on the same level, thus introducing large overheads in terms of additional arcs and complicated tests."

Instead of adding extra WRD AND arcs to the existing states in a standard ATN grammar, I set up a whole set of states to describe coordination phenomena. The first few states in the set are as follows:

```
(CONJ/                          At the moment only
  ((JUMP AND/)                  "and" is taken into
   (EQ (GETR CONJUNCTION)consideration.
       'AND))
  ......)

(AND/
  ((JUMP S/)                    Try to analyse RIGHT
   LEFT-PART-IS-CLAUSE)         -PART as a clause,if
                                LEFT-PART is one.
  ((JUMP S/)                    This arc is for such
   (AND (EQ LEFT-WORD-          cases as Ex15.
         CAT 'VERB)
        NPSTART)
   ((SETQ BUILD-RIGHT-CLAUSE-FIRST 'T)))
  ((PUSH NP/) (NPSTART)         Try phrasal coordi-
   ((SENDR SUBJNP T)            nation.
    (SETR RIGHT-PHRASE *)
```

---

* Cf (Wilks 75).

```
      (SETR RIGHT-PHRS-SMNTC-CAT
            (HEAD (CAAR *)))
      (IF NMODS-CONJ THEN
        (SETQ **NP-STACK
          (REVERSE **NP-STACK)))) The role of
      (TO AND/NP/PREPARE))        **NP-STACK
                                  will be ex-
                                  plained la-
                                  ter.
   ((JUMP S/NP)                   For cases
    (EQ (GET CURRENT-WORD 'CAT)   like Ex14.
        'VERB)
    ((SETQ BUILD-RIGHT-CLAUSE-FIRST 'T))))

(AND/NP/PREPARE
  ((JUMP AND/NP) T
   (SETQ **TOP-OF-NP-STACK (POP **NP-STACK))))

(AND/NP
  ((JUMP AND/NP/MATCH) T
   ((SETR LEFT-PHRASE (CAR (GETR **TOP-OF-
                                 NP-STACK)))
    (SETR LEFT-PHRASE-SYN (CAR (REVERSE
              (GETR **TOP-OF-NP-STACK))))
    (SETR LEFT-PHRS-SMNTC-CAT (HEAD (CAAR
              (GETR **TOP-OF-NP-STACK)))))))

(AND/NP/MATCH
  ((JUMP AND/NP/COORD)
   (EQ (GETR LEFT-PHRS-SMNTC-CAT)  To imple-
       (GETR RIGHT-PHRS-SMNTC-CAT))ment se-
    ...)                           mantic
  ((JUMP AND/NP)                   constaint.
   (NOT (NULL **NP-STACK))
   (SETR **TOP-OF-NP-STACK (POP **NP-STACK)))
  ((JUMP AND/NP/COORD) T)
   ...)
```

The CONJ/ states can be seen as a subgrammr which is separated from the main (conventional) ATN grammar, and is connected with the main grammar via the interpreter.

The parser works in the following way.

Before a conjunction is encountered, the parser works normally except that two extra stacks are set: **NP-STACK and **PREP-STACK. Each NP, either EL-NP or EXP-NP, is pushed into **NP-STACK,together with a label indicating whether the NP in question is a subject (SUBJ) or an object (OBJ) or a preposition object (NP-IN-NMODS).

The interpreter takes responsibility of looking ahead one word to see whether the word to come is a conjunction. This happens when the interpreter is processing "word-consuming" arcs, ie CAT, WRD, MEM and TST arcs. Hence no need for explicitly writing into the grammar WRD AND arcs at all.

By the time a conjunction is met, while the interpreter is ready to enter the CONJ/ state, either a clause (Ex10-13) or a noun phrase in subject position (Ex1-9) would have been POPed, or a verb (Ex14-15) would have been found. For the first case, a flag LEFT-PART-IS-CLAUSE will be set to true, and the interpreter will try to parse RIGHT-PART as a clause. If it succeeds, the representation of a sentence consisted of two coordinated clauses will be outputted. If it fails, a flag RIGHT-PART-IS-NOT-CLAUSE is set up, and the sentence will be reparsed. This time the left-part will not be treated as a clause, and a coordinated NP object will be looked for instead. Ex10 and Ex11 are examples of coordinated clauses and coordinated NP object, respectively. One case is treated specially: when LEFT-PART-IS-CLAUSE is true and RIGHT-WORD is a verb (Ex13), the subject will be copied from the left clause so that a right clause could be built.

For the second case, a coordinated NP subject will be looked for. Eg, for Ex4, by the time "and" is met, an NP "the man with the telescope" would have been POPed, and the state of affairs or the **NP-STACK would be like this:

(((MAN ...) (NUMBER ...) (QUANTIFIER ...)(DETERMINER ...) (ATTRIBUTE ((PREP (PREP WITH)) ((TELESCOPE ...))) SUBJ) ((TELESCOPE ...)NP-IN-NMODS))

After the excution of the arc ((PUSH NP) (NP-START)), RIGHT-PHRASE has been found. If it has an PP modifier, a register NMODS-CONJ will be set to the value of the modifier. Now the NPs in the **NP-STACK will be POPed one by one to be compared with the right phrase semantically. The NP whose formula head (the head of the NOUN in it) is the same as that of the right conjunct will be taken as the proper left conjunct. If the NP matched is a subject or object, then a coordinative NP subject or object will be outputted; if it is an EL-NP in a PP modifier, then a function REBUILD-SUBJ or REBUILD-OBJ, depending on whether the modified EXP-NP is the subject or the object, will be called to re-build the EXP-NP whose PP modifier should consist of a preposition and two coordinated NPs.

Here one problem arises: for Ex5, the first NP to be compared with the right phrase ("the woman") would be "the man with the child" whose head noun "man" would be matched to "woman" but, according to our Symmetry Constraint, it is "child" that should be matched. In order to implement this rule, whenever NMODS-CONJ is empty (meaning that the right NP has no post-modifier), the **NP-STACK should be reversed so that the first NP to be tried would be the one nearest to the coordinator (in this case "the child").

For the third case (LEFT-WORD is a transitive verb and the object slot is empty, Exs 14 and 15), right clause will be built first, with or without copying the subject from LEFT-PART depending on whether a subject can be found in RIGHT-PART.Then, the left clause will be completed by copying the object from the right clause, and finally a clausal coordination representation will be returned.

In the course of parsing, whenever a finite verb is met, the NPs at the same level as the verb and having been PUSHed into the **NP-STACK should be deleted from it so that when constructing possible coordinative NP object, the NPs in the subject position would not confuse the matching. Ex11 is thus correctly analysed.

VI DISCUSSION

The package is written in RUTGERS-UCI LISP and is implemented on the PDP-10 computer at the University of Essex. It performs satisfactorily. However, there is still much work to be done. For instance, the most efficient way for treating reduced conjunctions is to be found. Another problem is the scope of the pre-modifiers and post-modifiers in coordinate constructions, for the resolution of which the Symmetry constraint may prove inadiquate (eg, it cannot discriminate "American history and literature" and "American history and physics").

It is hoped that an ATN parser capable of dealing with a large variety of coordinated constructions in an efficient way will finally emerge from the present work.

ACKNOWLEDGEMENTS

I would like to thank Prof. Wilks of the Department of Language and Linguistics of the University of Essex for his advice and his patience in reading this paper and discussing it with me. Any errors in the paper are mine, of course. I would also like to thank Dr. Boguraev and my colleague Fass for part of their parsing programs.

REFERENCES

Blackwell, S.A. "Processing Conjunctions in an ATN Parser". Unpublished M.Phil. Dissertatation, University of Cambridge, 1981.

Boguraev, B.K. "Automatic Resolution of Linguistic Ambiguities". Technical Report No. 11, University of Cambridge Computer Laboratory, Cambridge, 1979.

Boguraev, B.K. "Recognising Conjunctions within the ATN Framework". Sparck-Jones, K. and Wilks, Y. (eds), Automatic Natural Language Parsing, Ellis Horwood (in press).

Gazdar, G. "Unbounded Dependencies and Coordinate Structure". Linguistic Inquiry 12, 155-84,1981.

Nagao, M., Tsijii, J., Yada, K., and Kakimoto, T. "An English Japanese Machine Translation System of the Titles of Scientific and Engineering Papers". In Horecky, J. (ed), COLING 82, North-Holland Publishing Company, 1982.

Radford, A. Transformational Syntax. Cambridge University Press, London, 1981.

Ross, J.R. Constraints on Variables in Syntax. Doctoral Dissertation, MIT, Cambridge, Massachusetts, 1967. Also distributed by the Indiana University Linguistics Club, Bloomington,Indiana 1968.

Schachter, P. "Constraints on Coordination," Language 53, 86-103, 1977.

Wilks, Y.A. "Preference Semantics". In Keenan(ed), Formal Semantics of Natural Language, Cambridge University Press, London, 1975.

Wilks, Y.A. "Making Preferences More Active". AI 1978.

Williams, E.S. "Across-the-Board Rule Application", Linguistic Inquiry 9, 31-34, 1978.

Winograd, T. Understanding Natural Language, Academic Press, N.Y., 1972.

Woods, W. "An Experimental Parsing System for Transition Network Grammar". In Rustin, R.(ed), Natural Language Processing, Algorithmic Press, N.Y., 1973.