"VASISTH"- An Ellipsis Resolution Algorithm for Indian Languages

Sobha, L
School of Social Sciences
Science Mahathma Gandhi University
Kottayam, Kerala, INDIA
Sobha@rocketmail.com

B.N. Patnaik
Department of Humanities and Social
Indian Institute of Technology Kanpur
Kanpur UP, INDIA
patnaik@iitk.ac.in

ABSTRACT

The paper presents an algorithm which resolves elliptical constructions for Malayalam, an Indo-Dravidian language, and then tested for Hindi, an Indo-Aryan language. This algorithm is a part of an anaphora resolution system called VASISTH. The testing was done to see if the system would apply without significant modification to another Indian language, structurally similar to Malayalam in many crucial respects. The test yielded encouraging results: only a few minor modifications are needed for the system to apply equally efficiently to Hindi. The computational grammar implemented here uses very familiar concepts such as clause, subject, object etc., which are identified with the help of morphological information, and concepts such as precede and follow. The algorithm works on a partial parser.

Introduction

Most of the available anaphora resolution systems handle particular languages and are not easily extendable to others. VASISTH, in contrast, is one which presently handles two different languages from two language families: Malayalam, from Indo-Dravidian and Hindi, from Indo-Aryan. It can easily be extended to handle other Indian languages, more generally, other morphologically rich languages. What further distinguishes VASISTH from other similar systems is that exploiting the morphological richness of the Indian languages, it makes limited use of syntax and uses only morphological markings to identify subject, object, clause etc. It uses limited parsing for parts of speech tagging, identification of clause, its subject, and person-number-gender of the NPs. VASISTH was developed and tested for Malayalam, and then modified for Hindi. It resolves referentially dependent elements such as pronominals, non-pronominals, gaps and ellipsis. This paper, howere,strictly confines itself to the handling of ellipsis alone.

Ellipsis in Malayalam and Hindi.

We deal with inter-sentential ellipsis involving *wh* constructions (*wh*-const) and the so-called "yes-no" question constructions(*q*-const) where the "o:" the question morpheme,occurs at the end of a declarative sentence. In the case of *wh* constructions the material that cannot be elided is the OBJECT ie, the "discourse focus".

Ellipsis in *wh*-cons.

Consider the following examples:

1. -ni: evite po:yi? -vi:ttil.
you where go-pst house-loc
(Where did you go?) (To the house. (=I went to the house.))

In the above sentence the *wh* word is the locative *evite* "where" which is locative. In the response, two constituents have been elided, the subject *na:n* "I" and the verb *po:yi* "went". The constituent that is not elided in the response is the locative, which is the focus of the sentence. Now consider (2):

2. - avan vanno:? - vannu.
he come-que morph come-pst
(Did he come?) (Came. (=Yes, he came.))

The response is an elliptical construction, containing only the verb. Unlike *wh*-const, in these constructions the verb cannot be elided, but the other constituents can be:

3. -ni: evite po:yi? - sku:lil.
you where go-pst school-loc
(Where did you go?) (To the school. (=I went to school.))

4. -ennane po:yi? -bassil.
how go-pst bus-loc
(How did you go?) (In the bus. (=I went by bus.))

In (3) the *wh*-word is *evite* "where" and the response to this is *sku:lil* "to the school" which is locative. In (4) the *wh*-word is *ennane* "how" and the response is again the locative *bassil* "in the bus" showing that *wh*-words *evite* and *ennane* take locative as focus.

5. -ra:man entine kantu? -pa:mbine.
raman what see-pst snake-acc
(What did Raman see?) (Snake. (=Raman saw a snake.))

6. -ra:man a:re kantu? -kRisnane.
raman who see-pst krishnan-acc
(Who did Raman see?) (Krishnan. (=Raman saw Krishnan.))

The *wh*-word *entine* "what" has the accusative as its response. In (5) *entine* "what" takes *pambine* "snake" as the response. The same holds for *a:re* "who" in (6). In the following examples, the focus is the nominal.

7. -a:ru palam va:nniccu. -ra:man.
who fruit buy-pst raman
(Who bought the fruit?) (Raman. (=Raman bought banana.))

8. -etra vila a:yi? -pattu.
how much cost copula ten
(How much is the cost?) (Ten. (=It costs ten rupees.))

In (7), the *wh*-word is *a:ru* "who" and the response is the nominative *ra:man* "Raman". The *wh*-word in (8) is *etra* "how much" and its focus is also a nominative nominal *pattu* "ten". Now consider (8).

9. -ni: entinu vi:ttil po:yi? -si:taye ka:na:n po:yi.
you why house-loc go-pst sita-acc see go-pst
(Why did you go to the house?) (Went to see Sita . (=I went to see Sita.))

Here the *wh*-word is *entinu* "why" and its focus is the accusative *si:taye* "Sita". The above examples show that the case of the *wh*-word is the case of the focus. The following table gives the information about *wh*-word in Malayalam and its focus:

Ellipsis in q-cons.

Another type of ellipsis that the language has is to be found in yes/no question constructions as the following:

10. - ni: kalicco: ? -illa.
you eat-pst-Qmorph no
(You ate? (=Did you eat?)) (No. (=I did not eat.))

Here the question is formed by adding the question morph to the verb or noun. The response will be either *illa* "no", *a:nu* "yes", *uvvu* "yes" or the verb without the question morph. Consider the following example:

11. -kalico:?   -uvvu.
eat-que morph   Yes
(Ate? (=Did you eat the food?))   (Yes. (=Yes, I ate the food))

or

-kaliccu.   -uvvu, kaliccu.
eat-pst   yes eat-pst
Ate. (=Yes, I ate the food.))   (Yes, Ate. (=Yes, I ate the food.)

In each of these responses different material is elided: in the first, only one constituent *uvvu* "yes" is present. In the second the verb *kaliccu* "ate" occurs and in the third both *uvvu* and the verb are present.

The question (11) itself has undergone ellision. If the clause is a *wh* construction, then the subject is elided and if a *q*-construction, then the subject, the object or both. The above show the following:

I. If the clause is a wh-construction, then the elided fragment in the response is of the following:
a. Subject, b. Verb, c. Both the subject and the verb.

II. If the clause is a q-const then, the elided fragment in the response can be one of the following:
d. Subject, e. Object, f Both the subject and the object.

Ellipsis in Hindi.

Turning to ellipsis in Hindi, unlike Malayalm, Hindi has only *wh*-construction and no comparable question construction (*q*-const). Consider the following examples.

12. - tum kaha gayi thi?   -sku:l
you where go-pst   school
(Where did you go?)   To school (I went to the school.)

The two constituents elided in (12) are the subject *tum* "you" and the verb *gayi thi* "went". The constituent that is not elided in the response is the object and the locative respectively ie, the focus of the sentence. Now consider:

13. -kal ko:n a:ya tha?   -ra:m
yesterday who come-pst   ram
(Who came yesterday?)   (Ram came yesterday.)
14. - voh kitna sa:l ka tha?   -das sa:l.
he how years acc pst   ten years
(How old is he?)   (He is ten years old.)

In (13) the *wh* word is *ko:n* "who" and the focus to it in the response is *ra:m* which is in nominative. In (14) the focus to the *wh* word *kitna* "how much" in the response is *das* "ten" also nominative. In the following the accusative element is the focus.

15. -ra:m ne kisko dekha tha?   -sya:m ko
ram-erg who see-pst   syam-acc
(Who did Ram see?)   (Saw Syam.)
16. -sya:m ne kya dekha tha?   -sa:p ko dekha tha.
syam-erg what see-pst   snake-acc see-pst
(What did Syam see?)   (Saw a snake.)

Consider the following sentences. In the following examples the locative is the focus.

17. -ra:m kaha gaya tha?   -sku:l.
ram where go+pst   school
(Where did Ram go?)   (To the school.)
18. -kyse gaya tha   -sku:ter se

how go-pst   scooter
(How did you go?)   (By scooter.)

All these show that the case of the *wh*-word is the case of the focus.

III. If S is a wh-construction, then the elided fragment in the response can be of the following:

g. subject h. verb i. both the subject and verb.

Algorithm for Ellipsis Resolution

The algorithm for identifying the antecedent for ellipsis in Malayalam is as follows:

```
IV 1. Create a list of words in Q and R with NPs, VPs and clauses identified.
        (from the parser)
   2. Identify the question words in the Q.
   3. Identify the focus word.
   4. For each W in Q.
        if the identical type does not occur in R
        if W is the subj in Q.
           Change W to W' and add to R (by subj change rules)
      else
           if W is the q-word in Q
               change W to W'.
           else add to R
```

The first step is to identify the structure of the question sentence (Q) and the response sentence(R) from the parser. The question words are identified in the second step. In the next step the focus for the wh words are identified. Step four identifies for each word W in Q the identical word in the response. If the word is not found it will check whether W is the subject word in Q and using subject change rule will change the subject and add to the response. If the W is a Q word then it is changed to W' and added to the response. If it is not the Q word then add W to response. Consider the examples.

19. -ni: entu kandu?   -pattiye kandu
you what see-pst   dog see-pst
(What did you see?)   (I saw a dog.)

The parsed structure

```
ni:      <N><NOM><subj><n><s><second><+human>
entu     <Wh>
kantu    <V><tran><pst>
pattiye  <N><ACC ><obj><n><s><third><-human>
```

The Wh word is entu and its response is pattiye

The W (subj) is ni: and the subject of response is na:n

The resolved response is na:n pattiye kandu

20. -ni: kalicco:?   -kaliccu
you eat-pst-Qmorph   ate-pst
(You ate(= Did you eat?))   (Ate(= I ate.))

Parsed structure:

```
ni:      <N><NOM><subj><n><s><second><+human>
kalicco: <Qword>
kaliccu  <V><tran><pst>
```

The Qword is kalicco: and the response is kaliccu.

The W subj is ni: and it is changed to na:n. The response: na:n kaliccu.

When we apply the above algorithm to Hindi, it yields cent percent success rate in case of *wh*-const. The modification required to accommodate Hindi ellipsis is minor, ie, the *q*-const identification is not required for Hindi. The modified algorithm is given below.

```
V.1. Create a list of words in Q and R with NPs, VPs and clauses identified.
        (from the parser)
     2. Identify the question words in  Q.
     3. Identify the focus word.
     4. For each W in Q.
          if the identical type does not occur in R
             if W is the subj in Q.
               change W to W' and add to R (by subj change rules)
```

The first step is to identify the structure of the question sentence (Q) and the response (R) from the parser. The question words are identified in the second step. The next step identifies focus in R. Step four identifies for each word W in Q the identical word in the response. If the word is not found, it checks whether W is the subject word in Q and using subject change rule changes the subject and add to the response. If the W is Q word, change it to W' and add to the response. If it is not the Q word, then add W to the responds.

21. -tum   kaha   gayi thi?   -sku:l
    you   where   go-pst   school
    (Where did you go?)   To school. (I went to the school.)

Parsed structure:

tum: <N><NOM><n><s><second><+human>
kaha <Wh>
gayithi <V><tran><pst>
sku:l <N><NOM><n><s><third>

Wh word: kaha
The response: sku:l
The structure of the response: me sku:l gayi thi.

Conclusion

The system works with high degree of success in the case of Malayalam and shows the same success rate for Hindi. Thus VASISTH we hope, can be extended to other Indian languages in particular and to morphologically rich languages in general.

Bibliography

Allen, J. (1987). Natural Language Understanding. *Benjamin/cumming Publishing company*, Inc. California.

Andrewskutty, A. P. (1988). "Ellipsis in Malayalam", *IJDL*, 17, No 1, 107-117.

Hart, Daniel. (1992). "An Algorithm for VP ellipsis", Proceedings of the Association for Computational Linguistics, 9-14.

Hardt, Daniel. (1994). "An Empirical Approach to VP ellipsis", *Computational Linguistics*, 16, 1.

Mitkov, Ruslan and Malgorzata Stys. (1997). "Robust Reference Resolution with Limited Knowledge: High Precision Genre-Specific Approach for English and Polish", *Proceedings of the International Conference "Recent Advances in Natural Language Proceeding"(RANLP'97)*, 74-81, Bulgaria.

Shieber, M Stuart, Fernando Pereira, and Mary Dalrymple. (1996). "Interactions of Scope and Ellipsis", *Linguistics and Philosophy*, 19, 5, 527-552.