Chapter 7

Sanskrit compound paraphrase generator

7.1 Introduction

Paraphrase or **विग्रहवाक्यम्** [1] is an expression providing the meaning of a compound. In Sanskrit tradition, two types of paraphrases are discussed : 1) **स्वपदविग्रहः** and 2) **अस्वपदविग्रहः** ।

The paraphrase involving only the components of a compound, is known as **स्वपदविग्रहः**[2]. For instance for the compound **नगरगतः** (The man who has gone to the city) , the paraphrase would be **नगरं गतः** and this meaning is expressed by the components which are already available in the compound **नगरगतः**. When the paraphrase is not expressed by the components of a

[1]वृत्त्यर्थावबोधकं वाक्यं विग्रहः ।

[2]समासस्यार्थः समासघटकैः पदैर्यदि वर्ण्यते तदा स्वपदविग्रहः । तथा च - समासघटकपदसहितं वाक्यं स्वपदविग्रहः इत्यर्थः । (स्व(=समास)पदैः विग्रहः ।) -- समासः

compound, it is known as अस्वपदविग्रहः[3]. For instance the paraphrase of a compound उपकृष्णम् (Near to Krishna) is कृष्णस्य समीपम्. Here, the word समीपम् is not a component of a compound उपकृष्णम्.

Three different ways have been observed of expressing the paraphrase of a compound in Sanskrit. For instance for तपस्स्वाध्यायनिरतम् [4], the paraphrase can be expressed as :-

a. तपः च स्वाध्यायः च = तपस्स्वाध्यायौ,
तपसि च स्वाध्याये च निरतः = तपस्स्वाध्यायनिरतः,
तं तपस्स्वाध्यायनिरतम् or as

b. तपः च स्वाध्यायः च = तपस्स्वाध्यायौ,
तपस्स्वाध्याययोः निरतः = तपस्स्वाध्यायनिरतः,
तं तपस्स्वाध्यायनिरतम् or as

c. तपः च स्वाध्यायः च = तपस्स्वाध्यायौ,
तयोः निरतः = तपस्स्वाध्यायनिरतः,
तं तपस्स्वाध्यायनिरतम्

Since the second type of paraphrase is more popular and found in most of the Literature, we decided to generate the paraphrase following the second type.

7.2 Paraphrase generator

A paraphrase generator takes a well formulated tagged compound as an input and produces its paraphrase as an output. A semantically analysed compound has the following syntax.

[3]समासस्यार्थः समासघटकपदानि विहाय पदान्तरैर्यदि वर्ण्यते तदा अस्वपदविग्रहः। तथा च - समासघटकपदरहितं वाक्यम् अस्वपदविग्रहः इत्यर्थः। (अत्र एकं पदं समासघटकमेव प्रायो भवति।) -- समासः

[4]सङ्क्षेप-रामायणम्,Bālakāṇda, sloka-1

compound : '<' component '-' component '>' tag

component : word | compound

tag : A[1-7]

| Bs[2-7] | Bs[dgpsu] | Bsm[gn] | Bv[sSU] | B[bv]

| D[is] | K[1-5] | Km | T[1-7] | T[bgmnpk]

| Tds | [ESd] | U[1-5,7]

word : [a-zA-Z]+

Sanskrit compounds tagging syntax

Note here that the compound is binary except in the case of dvandva and bahupada bahuvrihi.

We define a compound formed with the leaf nodes of a binary tree as a simple compound. Compounds with at least one component as a compound (i.e. a non-leaf node) are termed as nested compounds.

7.2.1 Paraphrase Generation

In Sanskrit compounds, as mentioned earlier, only the last component contains a case suffix[5]. Hence the major task of the paraphrase generator is to decide the gender, number and case suffix of the last component of a compound. The paraphrase of a compound varies with the type of a compound. Appendix-A gives rules for generating paraphrases for different compound types. Here are some examples :-

Example-1 Input : <दशरथ-पुत्रः>T6

Output : दशरथस्य पुत्रः = दशरथपुत्रः

The general rule for generating the paraphrase of a T6 type

[5]with an exception of an aluk samāsaḥ

compound is

<W_1 - W_2>T6 => W_1'{6} - W_2'{1} = (W_1' + W_2'){1}

Where W_1' and W_2' are the प्रातिपदिकs (nominal stem) of the words W_1 and W_2 respectively, '{6}' and '{1}' indicate the vibhakti. W_1' + W_2' on the RHS is sandhied form of the two प्रातिपदिकs W_1' and W_2'.

Example-2

Input : <पीत-अम्बरः>Bs6

Output : पीतम् अम्बरम् यस्य सः = पीताम्बरः

The general rule for generating the paraphrase is

<W_1 - W_2>Bs6 => W_1'{g}{1}, W_2'{g}{1} yat {g'}{6} tat{g'}{1}

= (W_1' + W_2'){g'}{1}

Where g' is the gender of W_2 and g is the default gender of W_2. The word अम्बरः is used in masculine here and hence its gender is masculine, which is g' while the default gender of अम्बर is neuter which is denoted by 'g' above.

7.2.2 Paraphrase generation of simple compounds

The paraphrase generation involves two major steps. In the first step we analyse the components and in the second step we generate the required word forms and construct the paraphrase. So we now describe the algorithm.

Step-A

Input : <W_1 - W_2>T_n

a. Analyse W_1, Here W_1 is an 'iic' 'in init composite' i.e. a **समास-पूर्वपद**. So we are interseted only in that analysis of W_1 where it can occur as a **समासपूर्वपद**. In case there are more than one analysis possible then we get the one which is more probable. For example, for **दशरथ** - two analysis are possible

i) **दशरथ, पुं**, iic

ii) **दशरथ, नपुं**, iic

Here we choose **दशरथ,पुं** as it is more probable. Similarly in case of **राज**- two analysis are possible

i) **राज,पुं**,iic

ii) **राजन,पुं**,iic

Among these two **राजन्** is more probable.

To choose the more probable answer, we use a database of iics - **समासपूर्वपद**s. This has following 3 fields **पूर्वपद** form, **प्रातिपदिकम्** and **लिङ्गम्**.

Eg. **राज-,राजन,पुं**

दशरथ-, दशरथ,पुं

This database is generated by extracting the **समस्तपद** entries form Apte's dictionary.

Let the default **प्रातिपदिकम्** and gender for W_1 be W_1' and g_1 respectively.

b. Analyse W_2

The analysis of W_2 pose certain problems. Just as in the case of W_1, here also multiple analysis are possible. However only those ambiguities matter us where the analyser shows analysis

with more than one gender or vibhaktis while in the context corresponding to a certain reading only one gender or vibhakti is possible. Since our paraphrase generator does not look at the context, it produces the most probable answer and has a provision to produce other answer on demand.

Another problem with the analysis of 'ife' 'in fini compositi' समास-उत्तरपद is when the समस्तपद undergoes some operations related to क्लीबत्वं then the morphological analyser should handle it. Eg. consider सप्तगङ्गम्. Here the ife is गङ्गम् which appears in neuter gender while the default gender of the word is feminine with प्रातिपदिकम् - गङ्गा.

Similar problem of change in the gender one encounters in the analysis of ife is with the Bahuvrihi compounds where the उत्तरपद assumes the gender of the object it refers to. Eg. the defualt gender of अम्बर is neuter. But in a Bahuvrihi compound, say eg. पीताम्बरः, अम्बर is in masculine since the word पीताम्बरः refers to विष्णु. So to handle these cases, we need

i) a morphological analyser which analyses words even if they are declined in some other gender. Thus this morphological analyser should be able to analyse गङ्गम् and अम्बरः, in addition to regular forms गङ्गा and अम्बरम्.

ii) a database that gives the defualt gender of a प्रातिपदिकम्.

Step-B

In this step, a paraphrase is generated. The Paraphrase has two parts :The phrase explaining the meaning and a compound word

denoting this meaning. For instance for the compound पीताम्बरः (a person who is in yellow dress) the paraphrase is पीतम् अम्बरम् यस्य सः=पीताम्बरः. In this, the पीतम् अम्बरं यस्य सः is LHS and पीताम्बरः is the RHS. The अम्बरं on the LHS has its default gender while अम्बरः on the right hand side is the gender of the object the word refers to.

Step-C Finally, if the compound word is not in the प्रथमा-विभक्ति, then appropriate pronominal phrase is also generated as

tat{g}{vibh}{num}

W_2{g}{vibh}{num}

Where g, is the gender, vibh and num are the vibhakti and number of the ifc - समास-उत्तरपद.

7.2.3 Paraphrase generation of nested compounds

In case of nested compounds, one or both the components are compounds. So we apply the above procedure repeatedly starting with the innermost compound which is a simple compound and go on simplifying it till all the compounds are covered.

7.2.4 Problem cases and thier solutions

- अलुक्समासः:-

  Only the last component of a compound has a case suffix. However as noted earlier there are exceptions typically with certain compounds whose first and intermediate components also have Vibhaktis. Such compounds are called aluk samāsa. The tagset of the Sanskrit Consortium does not mark aluk samāsas. Since the aluk samāsas are few in number, we treat as exceptional cases and produce their

output just by table lookup.

- मध्यमपदलोपिसमासः :-

  This is a special type of compound in which some of the words in the paraphrase do not occur in its compound form. e.g. Devabrāhmaṇaḥ is a compound whose paraphrase is Devapujakaḥ brāhmaṇaḥ(a brāhmin who worships god). So to get the paraphrase of such compounds mere components are not sufficient. One should also know the context to supply the missing words. We again list out these compounds as exceptions. It is neccessary to study these compounds seperately and see if it is possible to provide some semantic criterion to provide the missing elements.

- **Special cases from** Gaṇapāṭa etc. :-

  Compounds with special paraphrases have been listed by Pāṇini separately in a list. Examples of such compounds are Mayuravaymsakaḥ, Kambojamuṇḍaḥ, Yavanamuṇḍaḥ etc. Each one of them have a special paraphrase. Readymade paraphrases of such compounds are provided.

- उपपदसमासः :-

  An upapada tatpuruṣa samāsaḥ has a verbal noun (kṛdanta) as a post component(e.g. jnaḥ and kāraḥ in Tattvajnaḥ and Kumbhakāraḥ respectively). These forms are special and occur only as bound forms in a compound. Hence, a special morphological analyser to handle these forms is built.

- **The requirement of a special morph** :-

  From generation point of view, determining the gender of

constituents is the most difficult one. For instance in **उपगङ्गम्** (Near to Ganges river), the second constituent of the compound viz **गङ्गम्** is in neuter gender, derived from the word **गङ्गा** (the Ganges river). The word **गङ्गा** is in feminine gender. Another instance from **बहुव्रीहि** compound is **पाचिकाभार्यः** (The person whose wife is a cook.). Here, the word **भार्यः** is the second constituent and it is in mesculine gender derived from the word **भार्या** (the wife). The word **भार्या** is in feminine gender. So the word **गङ्गा** and the word **भार्या** can be easily analysed by the morph. But when these words occur in compounds then the word **गङ्गा** becomes **गङ्गम्** and the word **भार्या** becomes **भार्यः** due to compound formation and this is place where our morphological analyser fails to analyse the words. Here, we require a special morphological analyser which can analyse these kind of words and can provide the correct stems, genders and the information of first and second components.

- **The requirement of Sandhi module** :-

  As Sandhi is mandatory in compounds, it becomes neccessary to have a Sandhi module which can join the constituents according to the Paninian theory. The Sandhi module comes in picture at the stage of paraphrase generation in RHS (Right hand side) of the paraphrase. For instance **सुमित्रानन्दः** (the happiness of Sumitra) and it is tagged as <**सुमित्रा-आनन्दः**>T6

  Here, x = **सुमित्रा**, y = **आनन्दः** and T_n = T6 (**षष्ठी-तत्पुरुषः**)

  सुमित्रायाः आनन्दः = **सुमित्रानन्दः** is the complete paraphrase of <**सुमित्रा-आनन्दः**>T6.

In paraphrase

LHS = सुमित्रायाः आनन्दः

and

RHS = सुमित्रानन्दः

Now, for generating LHS a module is not required but when it goes to RHS, it needs a Sandhi module to generate सुमित्रानन्दः.

- **Special treatment of Dvandva compounds**

The gender and number of द्वन्द्व compounds depend on number of components and sometimes even on the semantics. For instance बककाकौ(The Crane (bird) and the Crow) and बककाकाः (The Crane or the group of Cranes and the group of Crows). The instance बककाकौ contains only two components बक and काक and it is in dual number. So simply by looking at the number of components and the number of compound, the paraphrase can be बकः च काकः च = बककाकौ which gives the meaning that only two single birds are there. But बककाकाः is not the same case, although the compound contains only two components and has the same gender, the word is now in plural not in dual. It is becuase of the involved semantics. Here either of both the words बक and काक denote a जाति and not an individual. Hence the paraphrase in this case is बकाः च काकाः च = बककाकाः or बकः च काकाः च = बककाकाः. Now the question is how to treat such compounds ? Where is the information that the component refers to a जाति and not a व्यक्ति. One may argue retering to the सूत्र -"जातिरप्राणिनाम्" (2.4.6) that one may list such words and handle separately. But as we see above both <बक-काकाः> as well as <बक-काकौ> is possible. In one case it refers to the जाति and in another

case व्यक्ति. So one may then argue that the information is in the form itself. Yes it is true that the information is in the word form. But when there are 3 or more than 3 components or if such components are part of another compound then this information will not be available. In such cases we asume that the word refers to a व्यक्ति.

7.3 Evaluation

The evaluation here is simple one. It does not involve any precision or recall figures, but just the % cases that produce correct paraphrase. We tested 200 simple compounds and 100 nested compounds. 89% simple compounds and 80% nested compounds paraphrases were correct.