IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 1, January 2015.
www.ijiset.com
ISSN 2348 – 7968

An Annotation Scheme for English Language using Paninian Framework

Amita[1], Ajay Jangra[1]

[1] Computer Science & Engineering Department, University Institute of Engineering and Technology, Kurukshetra, Haryana, India

Abstract

This paper presents a comprehensive study about the Panini's karaka relations for English. Paninian framework is suitable to all Indian language but some issues occur when applied to English languages. This paper discuss what are these issues and different approaches that were used in past.

Keywords: *NLP (Natural Language Processing), HyDT(hyderabad dependency treebank), POS(part of speech tagging)*

1. Introduction

In NLP (natural language processing), models of natural languages are build for its generation and analysis. First, to know about human communication process using natural language, there is cognitive and linguistic motivation. Second, to build technological advance intelligent computer system such as natural language interface to databases, speech recognition systems, computer aided design systems, text analysis, understanding system, man-machine interfaces, machine translation systems, and system that read or understand printed or handwritten text. Paninian Grammar Model revolves around "Information". Information is regarded as the core of Paninian Grammar Model for studying the language. This can be explained by the process of communication between two persons in which one person is giving information to another. The former person can be termed as person who encoded the information in a language string and the latter can be termed as person who decoded the information in the language string. There are two levels of representation in language use. One level is the sentence (written or oral) and another is "What the person has in his mind". Oral or written sentence is regarded as the surface level or base level. Karaka Level and Vibhakti level is also two another important levels of Paninian Grammar Model.

There are verb & verb relations and Karaka relations at the Karaka Level. These relations are generally between verb and other constituents in a sentence like noun. " What speaker has in his mind " is considered as topmost level. Karaka Level lies between this topmost level and vibhakti level and includes karaka relations. So there are several levels between the karaka level and the topmost level.

The information about TAM (tense, aspect and modality) is given by the vibhakti for a verb. For verbs vibhakti includes the verb and the auxiliary verb. Tense, aspect and modality are determined by the verb and auxiliary verbs. They are purely syntactic. At vibhakti level there are word group based on case endings, proposition markers or preposition.

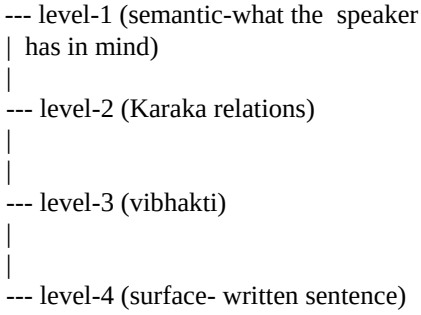

Fig. 1 Levels in the Paninian model

Karaka vibhakti mapping depends on verb and tense, aspect & modality factors. Paninian Grammar Model is well suited to Indian languages because Indian languages have free word order.

2. Literature Review

In Paninian Grammar model, sentence is treated as a series of relations. These relations are called modifier-modified. Thus, a sentence contains modifier and modified which is root of dependency tree. The modifiers of the verb take part in the action specified by the verb. The modifier relations with the verb are '*karaka*'. The name given to the relation subsisting between noun and verb in a sentence is

ISSN 2348 – 7968

'Karaka'. Panini described six types of karaka relations (Kiparsky and Stall, 1969[1]): adhikarana as 'location', apaadaan as 'source', sampradaan as 'recipient', karana as 'instrument', karma as 'theme', karta as 'agent' respectively. Karaka relations are not equivalent to theta roles (although they are mapped sometimes, for the sake of elucidation). In 1995 A. bharti et al. [2] represents a paninian's prespective to Indian languages. While thematic roles are purely semantic in nature karaka relations are syntactico-semantic.PG is very well suited for languages that have a relatively free word order. (Bharti et al. 1993)[3].

Paninian Grammar Model's previous application to English by Bharati et al. (1997)[4] proved that free word order languages and positional languages can be explained using PG model. (Begum et al., 2008) [5]In PG the task of case endings or markers (post-positions, verbal inflections) is emphasized. Positions or word order are considered only when necessary, since they contain only secondary information in free word order languages. Further, Begum et al. opine that if the dependency are chosen wisely then dependency framework is more closer to semantic than phrase structure grammar dependency framework is closer to semantics than phrase structure grammar. Given this, more and more research groups have been shifting to dependency analysis, of late. Thus, extending the annotation scheme based on the CPG model to English, helps capture semantic information along with providing a syntactic analysis. The semantics level reflect the surface form of the sentences, and is important syntactically. Such a level of annotation makes available a syntactico-semantic interface that can be easy to exploit computationally, for linguistic investigations and experimentation. This includes facilitating mappings between semantic arguments and syntactic dependents. In their preliminary work, where they proposed an annotation scheme for English language. This annotation scheme is based on computational Paninian Grammar framework. Vaidya et. al (2009)[6] described some level of local semantic of a verb in a sentence are captured by karaka relation and taking cues such as vibhakti from the morph-syntactic surface level.

In the direction application of the CPG model to English a step ahead is taken by Himani Chaudhary et al. work, taking forward the preliminary task carried out by (Vaidya et al.,2009)[6] and use these basic karaka definitions:

k1: *karta*: central to the action of the verb
k2:*karma*: the one most desired by the karta
k3: *karna:* instrument which is essential for action for take place
k4:*sampradaan:* recipient of the action
k5:*apaadaan:* movement away from source
k7: *adhikarana:* location of action

This work elaborates an annotation scheme with the help of some syntactic construction in English and also described some type of sentences in which construction fails [4].

For example, in figure 2 'It' is a dummy element in the sentence to fill the empty subject position.

We mark it with a special relation 'dummy', which reflects the fact that 'It' is semantically vacuous. We also add the information about the expletive construction to the feature structure of the head. The semantically vacuous 'It' will fail the test for k1 although it is in the subject position and agrees with the verb.

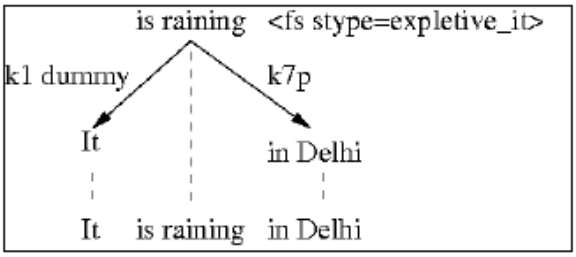

Fig. 2 Expletive sentence

They have adopted the HyDT annotation scheme and adapted it to suit the needs of the language wherever required. They discuss the task in its various aspects. In 2011, Himani et al.,[10] explained the annotation scheme and computational Paninian grammar used for this work. They used a corpus of 25000 words and Sanchay annotation interface. They talked how the application of annotation scheme in English varies from Hindi. In 2013, Himani et al., [11] introduced what are the Divergences in English-Hindi Parallel Dependency Treebank. This paper presented the divergence between the treebanks of English and Hindi[6]. The two treebanks diverge mainly from two aspects:

A. Stylistic
B. Structural

The two treebanks were considered 'divergent' if the parallel trees fell under any of the following:

- Differences in the construction (structure)
- Difference in relations marked (on the parallel sentences)
- Difference in tree depth
- Difference in the frequency of annotation labels

Changes in lexical category of a word of one language and its counterpart in the other, lead to Categorical divergence visible in the data. 'It suffices.' would be translated in Hindi as 'yaha kAfI hE.' (It sufficient is). While the word 'suffices' is realized as the main verb in English it is an adjectival modifier 'kAfI' (sufficient) in the phrase 'kAfI

ISSN 2348 – 7968

hE', in Hindi. Figure 3 shows the divergent trees for the sentence pair.

Hindi: 'yaha kAfI hE.'
Yaha Kaafii hE
It sufficient is
English: 'It suffices'

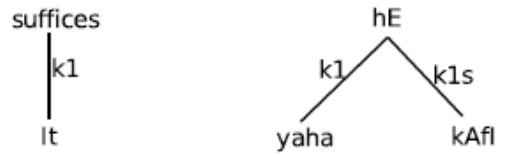

Fig. 3 Example showing categorical divergence.

D.M. Sharma et al. in 2007[9], used the Paninian Grammar models to give the annotating guidelines for Treebank for Indian Languages. Details of the tagging scheme and description of the grammatical model are contained in dependency annotation of Hindi. It also contain analysis of Paninian dependency model and examples of certain typical construction of Hindi. Developing dependency Treebank for Hindi is their task.
Mridul Gupta et al. 2008[10], describe an approach using Paninian dependency framework to automatically annotate Hindi Treebank. The annotator used there is a rule based system and certain syntactic cues available in a sentence are used by the rules. Their purpose is to facilitate the process of annotation and to fine grain and correct the output of broad coverage of constraint based parser. Certain syntactic cues such as post position , POS labels etc. to formulate the rules on which the system works.
Nidhi Mishra and Amit Mishra 2011[11], described POS tagging for Hindi corpus. In their method sentences and words are extracted by scanning the Hindi corpus by the system. The tag pattern is searched by the system from the database and the tag is displayed for each Hindi word like word tag, number tag, etc.

3. Issues and Challenges

Issues and challenges for Panini's framework for English are mentioned below:

3.1 Not exact mapping

Stanford parser parses the English sentences and output the dependencies between tokens. We cannot map these English dependencies to Hindi dependencies directly because of free word order of Hindi language.

3.2 Copula verbs

Copula verbs show the relation between the copular verb and the complement of a copular verb [1]. (We normally take a copula as a dependent of its complement)

"Bill is big"	cop(big, is)
"Bill is an honest man"	cop(man, is)

But there is no concept of copula verbs in Hindi language.

3.3 Control verbs

For English, the control verbs such as promise or persuade are not analyzed as cases with an empty PRO. Instead, the analysis shows a difference in the verb semantics of promise and persuade, which again amounts to making the lexicon richer. In (Fig. 4), the tree does not show a missing element but analyses the verb semantics of persuade differently from the semantics of a verb like promise (a subject-control verb) in (Fig. 5) Persuade is shown to take a Karta (k1), a karma (k2) and tadarthya (rt or purpose) labels. Promise on the other hand takes karta(k1), karma(k2) and sampradaan (recipient of the action -k4) labels.

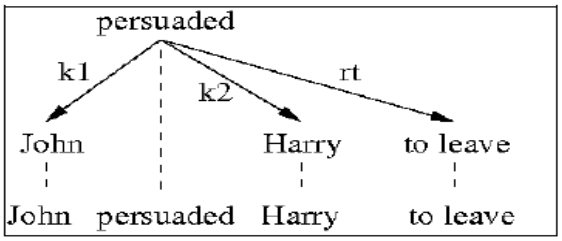

Fig. 4 Object control verb

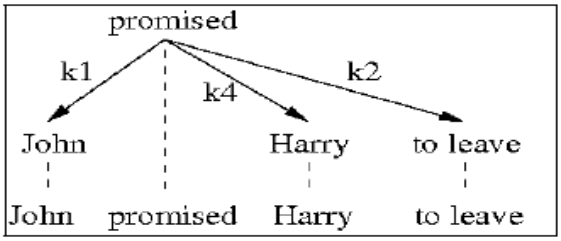

Fig. 5 Subject control verb

3.4 Structural differences

Structural differences between Hindi and English treebanks.

ISSN 2348 – 7968

4. Conclusion

As discussed earlier that Paninian framework is suitable to all Indian language and can be applied to English language. While mapping karaka relation to English language, there are many issues that occurred, this is due the structural difference between the English and Indian languages. A new approach for mapping these Panini's karaka relation to English language can be developed that can be helpful to translation of English to Indian languages.

References

[1] Kiparsky and J. F. Staal. 1969. Syntactic and Semantic Relations in Panini. Foundations of Language 5, 84-117.

[2] A. Bharati, V. Chaitanya, and R. Sangal, Natural Language Processing: A Paninian Perspective, Prentice-Hall, New Delhi, 1995.

[3] A. Bharati , R. Sangal. 1993. Parsing Free Word Order Languages in the Paninian Framework. ACL93:Proc. of Annual Meeting of Association for Computational Linguistics.

[4] A. Bharati, M. Bhatia, V. Chaitanya, and R. Sangal, "Paninian grammar framework applied to English," South Asian Language Review, 1997.

[5] Begum, S. Husain, A. Dhwaj, D.M. Sharma, L. Bai, and R. Sangal " Dependency annotation schemefor Indian languages" , in *Proceedings of IJCNLP-2008*.

[6] A. Vaidya, S. Husain, P. Mannem and D. M. Sharma, "A karaka based annotation scheme for english", in Computational Linguistics and Intelligent Text Processing 2009, Springer,pages 41–52.

[7] H. Chaudhary and D.M. Sharma, "Annotation and Issues in Building an english dependency treebank". In proceedings of ICON-2011: 9th International Conference on Natural Language Processing. Chennai 2011

[8] H. Chaudhary, H. Sharma and D.M. Sharma, "Divergences in English-Hindi Parallel Dependency Treebank",in proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), prague, pages 33–40, August 27–30, 2013.

[9] D. M. Sharma, R. Sangal, L. Bai, R. Begam, and K.V. Ramakrishnamacharyulu. 2007. AnnCorra :TreeBanks for Indian Languages, Annotation Guidelines (manuscript), IIIT, Hyderabad, India.

[10] M. Gupta, A. Bharti, V. Yadav, K. Gali, D. M. Sharma "Simple parser for Indian languages in a dependency framework" in proceedings of Third linguistics annotation workshop, ACL – IJCNLP 2009, pages 162-165.

[11] N. Mishra and A. M. Buddha, "Part of Speech Tagging for Hindi Corpus", International Conference on Communication Systems and Network Technologies, 2011.