Ashwath Rao B, Muralikrishna S N, Ashalatha Nayak

DEVELOPING A DEPENDENCY TREEBANK FOR KANNADA

Ashwath Rao B[*1], Muralikrishna S N[2], Ashalatha Nayak[3]

*[1]Computer Science & Engineering. Dept., Manipal Institute of Technology, Manipal, Karnataka State,India
ashwath.rao@manipal.edu[1]
[2] Computer Science & Engineering. Dept., Manipal Institute of Technology, Manipal, Karnataka State,India
murali.sn@manipal.edu[2]
[3] Computer Science & Engineering. Dept., Manipal Institute of Technology, Manipal, Karnataka State,India
asha.nayak@manipal.edu[3]

Abstract: Language syntax and semantics can be recorded in various forms like grammar rules, dictionary etc. An alternate to this approach is annotated corpus. There have not been major effort in this direction for Kannada. As part of this project we are carrying out annotation of about 200K words for POS, Morph and dependency annotation of sentences. The broad guidelines framed for this task is explained in this paper.

INTRODUCTION

Treebank is a major linguistic resource for various NLP tasks. The treebanks have proved to be a crucial resource for higher level NLP research and developing solutions for various socially relevant NLP applications. A major bottleneck in developing various natural language applications for Indian languages is the unavailability of appropriate language resources. For any NLP application, certain linguistic knowledge is required. This knowledge can be prepared in the form of dictionaries, grammars, word-formation rules etc. An alternative approach is to annotate linguistic knowledge in electronic texts. The annotated texts can be used for machine learning, developing these resources by extracting the knowledge etc. Penn Treebank for English (Marcus et al., 1993), Prague Dependency Tree bank for Czech (Hajicova, 1998) etc. are some of the efforts in this direction.

We are annotating sentences with the following objectives

(A) Monolingual General purpose : 100k words. Tourism : 35k words, Conversational : 25k

(B) Parallel Treebanks, apart from developing monolingual treebanks for the languages mentioned above, it is planned to also develop parallel treebanks of 40k words where the other langauage is Hindi. The domains for this treebanks would be administrative domain (circulars, letters etc).

We are annotating the Corpora for

1) POS
2) Chunk
3) Morph
4) Dependency relations across chunks
5) Sentence type
6) Voice type

Apart from the above, we are also developing

- Developing verb frames to facilitate dependency annotation
- Developing tools for quality checking
- Developing intra-chunk dependency annotators for automatic expansion of the chunks
- Data driven Parsers with a target labelled attachment accuracy of 40% - 50%.

POS-TAGSET FOR KANNADA

We are following Bureau of Indian Standards Part-of-Speech Tag set. The tag set follows a hierarchial structure. The tag set is shown below.

Category	Label	Annotation Convention
Noun	N	N
Common	NN	N__NN
Proper	NNP	N__NNP
Nloc	NST	N__NST
Pronoun	PR	PR
Personal	PRP	PR__PRP
Reflexive	PRF	PR__PRF
Relative	PRL	PR__PRL
Indefinite	PRI	PR__PRI
Reciprocal	PRC	PR__PRC
Wh-word	PRQ	PR__PRQ
Demonstrative	DM	DM
Deictic	DMD	DM__DMD
Relative	DMR	DM__DMR
Wh-word	DMQ	DM__DMQ
Indefinite	DMI	DM__DMI
Verb	V	V
Main	VM	V__VM

Finite	VF	V__VM__VF
Non-finite	VNF	V__VM__VNF
Infinitive	VINF	V__VM__VINF
Gerund	VNG	V__VM__VNG
Verbal Noun	NNV	N_NNV
Auxiliary	VAUX	V__VAUX
Nonfinite	VNF	V_VM_VNF
Infinite	VINF	V_VM_VINF
Adjective	JJ	JJ
Adverb	RB	RB
Postposition	PSP	PSP
Conjunction	CC	CC
Coordinator	CCD	CC__CCD
Subordinator	CCS	CC__CCS
Quotattive	UT	CC__CCS__UT
Particles	RP	RP
Default	RPD	RP__RPD
Classifier	CL	RP__CL
Interjection	INJ	RP__INJ
Intensifier	INTF	RP__INTF
Negation	NEG	RP__NEG
Quantifiers	QT	QT
General	QTF	QT__QTF
Cardinals	QTC	QT__QTC
Ordinals	QTO	QT__QTO
Residuals	RD	RD
Foreign word	RDF	RD__RDF
Symbol	SYM	RD__SYM
Punctuation	PUNC	RD__PUNC
Unknown	UNK	RD__UNK
Echo words	ECHO	RD__ECHO

Apart from the tags mentioned in BIS tag set, we have considered Indefinite Pronouns(PR__PRI) and Indefinite Demonstratives(DM__DMI).

MORPHOLOGICAL ANALYZER

For each token in the corpus, the Morphological features of the token is annotated. Morphological annotation is done after Sandhi splitting. The Morphological features annotated include, root, lexical category, gender, number, person, case direct/oblique, case marker & Tense, Aspect, Modality. An example of Morphological Analyzed sentence is shown below.

ಉಪಟಳಕ್ಕೆ <fs af='ಉಪಟಳ,n,,sg,3,o, ಕ್ಕೆ,kkeV'>

ತುತ್ತಾಗುತ್ತದೆ <fs af=' ತುತ್ತಾಗು,v,,sg,3,,ಉತ್ತ+ಅದೆ,uwwA+axeV'>

A. Issues

- Kannada language being morphologically rich, arriving at common case and TAM marking among annotators.
- If comma appears as a symbol, the root of the symbol cannot be put as comma as comma is used to separate features. To solve this, a word COMMA instead of symbol comma is used.

B. Tools Developed

Quality of corpus in terms of correct annotation is important. In this regard several in-house tools are developed. Few tools already developed are,

1) A tool for verification of missing morph data for any token
2) A tool for verification of correct number of features
3) A tool for verification of wx feature to Unicode feature mapping for Case / TAM

DEPENDENCY ANNOTATION

The theoretical model that has been adopted for the sentence analysis is Panini's grammatical model which provides a level of syntacticosemantic analysis. The model, not only offers a mechanism for SYNTACTIC analysis, but also incorporates the SEMANTIC information (dependency analysis). Indian languages have a relatively free word order, hence a dependency grammar based approach would be better suited for sentence analysis. The meaning in a sentence is encoded, not only in the words
(lexical items), but also in the relations between words. Thus every word in a sentence has a twofold role towards composing the larger meaning;
1) the concept it represents and
2) the participatory role it plays in the sentence in relation to the other words.

The latter (ii) is, most often, expressed through some explicit markers such as nominal inflections, verbal inflections etc. This implies that certain linguistic cues are explicitly available in a sentence using which one can extract the meaning from a sentence. Morphologically rich languages such as Kannada,Sanskrit(a classical Indian language), Telugu, Tamil etc(some of the modern Indian languages) mark the grammatical information in the words themselves (through affixes).The grammatical relations which
have been considered here are of two types:

1) kaaraka, and
2) Relations other than kaarakas.

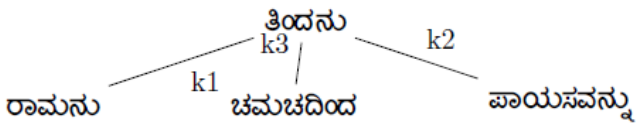
Kaaraka, according to Patanjali, is the one which performs an action (karotiiti kaarakam). A number of direct participants are needed for an action to be completed successfully. Doer of an action, time when the action is carried out, receipient of an

action which requires transfer of some sort, source of an action which denotes a point of departure etc are some examples of the direct participants(kaarakas) of an action. There could also be other players when an action is being carried out. These players may not have any direct role in the action though. Reason and purpose are two examples of such players. 'kaarakas' are the roles of various direct participants in an action. An action in a sentence is normally denoted through a verb. Hence, a verb becomes the primary modified (root node of a dependency tree) in a sentence. Panini has spelled out six kaarakas (Bharati et al., 1995).

The sentence may contain a number of relations between words which are not 'kaaraka' relations. The scheme adopted for annotating dependency relations in this treebank refers to these relations as 'other than kaaraka' relations.Purpose, reason,
genitive etc. would fall under the second type of relations within the Paninian framework. The six **kaarakas** given by Panini are '**kartaa**' (doer of an actions), **karma** (locus of the result of theaction), **karana** (instrument), **ampradaana**(receipient/beneficiary), **apaadaana** (source) and **adhikarana** (location).
An example of a dependency annotation is shown below.

ರಾಮನು ಚಮಚದಿಂದ ಪಾಯಸವನ್ನು ತಿಂದನು

rAmanu camacaxiMxa pAyasavannu wiMxanu
Ram erg sppon with rice-pudding ate
'Ram ate the rice-pudding with a spoon'.



The complete list of dependency annotation tags are shown below.

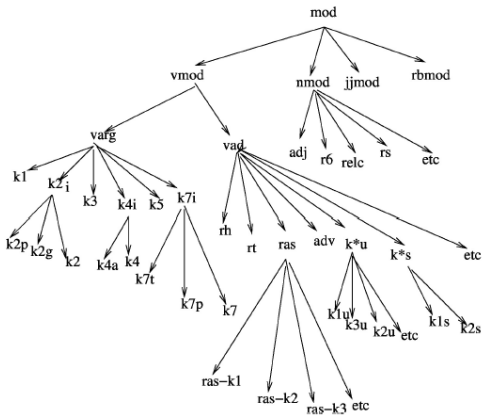


Figure 1: Dependency Relation Types

A. Tools to be developed
An extensive quality check tools need to be in place. The various tools we are planning to develop
are,
1) A tool for checking missing dependency relation marking
2) A tool for checking improper tag/chunk marking
3) A tool for checking invalid tag like multiple karta etc. under a single tree branch.

REFERENCES

[1] L. van der Beek, G. Bouma, R. Malouf, and G. van Noord.2002. *The Alpino dependency treebank*. Computational Linguistics in the Netherlands.
[2] A. Bharati, V. Chaitanya, R. Sangal. 1995. *Natural Language Processing: A Paninian Perspective*. Prentice-Hall of India, New Delhi.
[3] Cristina Bosco and V. Lombardo. 2004. *Dependency and relational structure in treebank annotation*. In Proceedings of Workshop on Recent Advances in Dependency Grammar at COLING'04.
[4] S. Brants, S. Dipper, S. Hansen, W. Lezius and G. Smith.2002. *The TIGER Treebank*. In Proceedings of the Workshop
on Treebanks and Linguistic Theories.
[5] E. Hajicova. 1998. *Prague Dependency Treebank: From Analytic to Tectogrammatical Annotation*. In Proc. TSD'98.
[6] P. Kingsbury and M. Palmer. 2002. *From Treebank toPropBank*. In Proceedings of the 3rd LREC, Las Palmas, Canary Islands, Spain.
[7] P. Kiparsky and J. F. Staal. 1969. *'Syntactic and Relations in Panini'*. Foundations of Language 5, 84-117.
[8] M. Marcus, B. Santorini, M.A. Marcinkiewicz. 1993. *Building a large annotated corpus of English: The Penn Treebank*.
In Computational Linguistics.
[9] M. Marcus, G. Kim, M. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, B. Schasberger. 1994. *The Penn treebank: Annotating predicate argument structure*. In Proceedings of the ARPA Human Language Technology Workshop.
[10] O. Rambow, B. Dorr, I. Kucerova and M. Palmer. 2003. *Automatically Deriving Tectogrammatical Labels from other resources- A comparison of Semantic labels across frameworks*. The Prague Bulletin of Mathematical Linguistics 79-80, 23-35 (2003)
[11] O. Rambow, C. Creswell, R. Szekely, H. Taber, and M. Walker. 2002. *A dependency treebank for English*. In Proceedings
of the 3rd International Conference on Language Resources and Evaluation.
[12] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, B. Webber. 2008. *The Penn Discourse Treebank 2.0*. In Proceedings of the 6th LREC.
[13] C. Shastri. 1973. *Vyakarana Chandrodya* (Vol. 1 to 5). Delhi: Motilal Banarsidass. (In Hindi)