Dependency Parser for Tamil classical literature- *Kurunthokai*

Dr.V.Dhanalakshmi
Head of the Department, Department of Tamil, FSH, SRM University, Kattankulathur-603203
M.Anand Kumar
Assistant Professor, CEN, Amrita Vishwa Vidyapeetham, Coimbatore.
C.Murugesan
Research Associate, Department of Tamil, FSH, SRM University, Kattankulathur-603203

Introduction

Dependency parsing is a **form of syntactic parsing** of natural language based on the theoretical tradition of dependency grammar [4]. It is a process of analyzing the dependency structure of a sentence. Tamil is always head-final language. The verb comes at the end of the clause with a typical word order of Subject Object Verb (SOV) . However, Tamil language allows word order to be changed, making it a relatively word order free language. In Sangam literature poems the syntactic structure are more complex due to this relatively free word order nature of Tamil language. Dependency grammar is better suited for languages with free or flexible word order. Dependency parser for the contemporary Tamil language has been already developed [5]. This paper proposes a dependency parser for Tamil classical literature – kurunthokai.

Kurunthokai

The Sangam literature contains about 50, 000 lines of poetry contained in 2381 poems attributed to 473 poets including many women poets. '**Kurunthokai**' is one of the eight anthologies (Ettuth thokai), which belongs to the Sangam age. It has 401 short poems of 6 to 8 lines, and written by 205 poets. The syntactic structure of the poem is very complex since many clauses are embedded in it. So, developing an automatic dependency parser for the Tamil classical literature- Kurunthokai will be a most wanted work of the modern era.

Architecture of the Proposed System:

The architecture of the proposed system is given below:

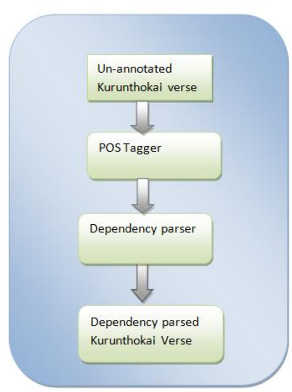

Fig:1 Architecture of the Proposed System

The proposed system contains three modules:

1. POS Tagger
2. Dependency Parser

Part-of-Speech Tagger: The Part of Speech (POS) tagging is done to obtain the labeled Part of Speech or other lexical class marker such as noun, verb, adjective, etc in the given verse .

Dependency parser:. Dependency parsing is done to obtain the grammatical relation of the words in the verse. This decides which relations (subject, object, location, etc.) they have with the main verb or the root word. This is also used to obtain the types of clauses present in the verse.

Methodology of the proposed system

The methodology adopted to develop the modules is the Machine learning approach. Machine learning deals with techniques that allow computers automatically learn to make accurate predictions based on past observations. The major focus of machine learning research is to extract information from annotated data automatically, by computational and statistical methods. So the annotated data plays a key role in supervised machine learning techniques.

There are two main tasks involved in supervised machine learning; 1) . learning or training, 2) . Prediction. The system is trained with a set of examples called training data. The primary goal is to automatically acquire an effective and accurate model from the training data. This is a typical task for supervised learning and is usually called concept learning or learning from examples. The accuracy of the model will be better if the training data size is larger. The second phase of machine learning is the prediction, wherein a set of inputs is mapped into the corresponding target values, based on the generated model. The main challenge of machine learning is to create a model with high prediction accuracy. Here the training data will be generated for all the verses in the Kurunthokai, so the testing data will be of gold standard data.

POS Tagger

MBT (Memory Based Tagger) is an approach to POS tagging based on Memory-based learning. It is an extension of the classical k-Nearest Neighbor (k-NN) approach to statistical pattern classification. Here all instances are fully stored in memory and classification involves a pass along all stored instances. The approach is based on the assumption that reasoning is based on direct reuse of stored experiences rather than on the application of knowledge (such as rules or decision trees) abstracted from experience. Hence the tagging accuracy for known words will be cent percentage. As our test data will be of gold standard data and we use MBT for POS tagging the kurunthokai verse.

Customized POS tagset

The customized tagset contains 23 tags. We are not considering the inflections. The customized tagset is developed with the perspective of Tamil grammar Tholkaappiyam [6]. The customized tagset with examples is given below Table (1) .

S.NO	**Main Category**	**Tag Category**	**Tag Set**	**Example**
1.	பெயர் (Noun)	பெயர்ச்சொல் (Common Noun)	<NN>	மரம்
2.		சிறப்பு பெயர் (Proper Noun)	<NNP>	அதியமான் முசிறி
3.		பதிலிடு பெயர் (Pronoun)	<PRN>	சுட்டுப்பெயர்கள் அவன், அது

4.		வினா பதிலிடு பெயர் (Interogative Pronoun)	<PNI>	யாவன், யாது, யார்
5.		பிரதி பதிலிடு பெயர் (Reflexive Pronoun)	<PNR>	தான்
6.	எண்ணுப் பெயர் (Numbers)	Cardinal	<CRD>	ஒன்று, 100, க
7.		Ordinal	<ORD>	ஒன்றாம், ஒன்றாவது
8.	வினை (Verb)	வினைமுற்று (Finite Verb)	<VF>	வந்தான்
9.		வினை எச்சம் (Adverbial Participle)	<AVP>	வந்து
10.		(Infinitive Participle)	<IVP>	வர
11.		நிபந்தனை எச்சம் (Conditional Participle)	<CVP>	வந்தால்
12.		பெயர் எச்சம் (Adjectival Participle)	<AJP>	வந்த
13.		தொழில் பெயர் (Gerundial Noun)	<GN>	வருதல்
14.		பெயரடைப் பெயர் (Adjectival Noun)	<AN>	நல்லவன்
15.		வினையாலணையும் பெயர் (Participle	<PN>	வந்தவன்

		Noun)		
16.	உரிச் சொல் Modifiers	பெயரடை (Adjective)	<ADJ>	நல்ல, அழகான
17.		வினையடை (Adverb)	<ADV>	அழகாக, நேற்று
18.		அளவடை (Quantifiers)	<QF>	பல, சில
19.	இடைச் சொல்	ஒட்டு (Clitic)	<CL>	உம், ஓ, ஏ, ஆ, தாம், தான்
20.		உவம உருபு (Comparision Word)	<CW>	ஒப்ப, உறழ
21.		வேற்றுமை உருபு (Case Marker)	<CM>	ஐ, ஆல், கு
22.		பின்னுருபு (Postpositional word)	<PPO>	போல, குறித்து
23.		அசை (Particle)	<PAR>	சின், மன், ஒரு

Table 1: Customized POS Tagset

Dependency Parser

Parsing is related to the automatic analysis of syntactic structure according to a grammar. It is usually performed after basic morpho-syntactic (POS) categories have been identified in a text. Dependency parsing is a form of syntactic parsing of natural language based on the theoretical tradition of dependency grammar. The dependency structure of the sentence is defined using dependency label and dependency head. Whenever two words are connected by a dependency relation, we say that one of them is the head and the other is the dependent, and that there is a link connecting them. In general, the dependent is in the form of modifier, object or complement. The head plays the larger role in determining the behavior of the pair.

In our proposed dependency parser system the representation the source of the edge represents the modifier and destination points to the head word. The customized dependency tagset contains 35 grammatical relations. MALT[1] parser tool is used to obtain the dependency parser for kunthokai verse. An example of the data set is given below.

S.no	Word	POS tag	Depn arc	Depn tag
1	யாயும்	<NN>	2	<CJ_SUB>
2	ஞாயும்	<NN>	3	<CL_SUB>
3	யார்	<PNI>	4	<COP>
4	ஆகியரோ	<VF>	23	<F_CL>
5	எந்தையும்	<NN>	6	<CJ_SUB>
6	நுந்தையும்	<NN>	9	<CL_SUB>
7	எம்	<PNI>	8	<ATT>
8	முறைக்	<NN>	9	<LOC_ATT>
9	கேளிர்	<NN>	23	<F_CL>
10	யானும்	<PN>	11	<ATT>
11	நீயும்	<PN>	14	<CL_SUB>
12	எவ்	<PNI>	13	<ATT>
13	வழி	<NN>	14	<LOC_ATT>
14	அறிதும்	<VF>	23	<F_CL>
15	செம்	<ADJ>	16	<ATT>
16	புலப்	<NN>	18	<ATT>
17	பெயல்	<VN>	18	<ATT>
18	நீர்	<NN>	19	<CL_SUB>
19	போல	<PPO>	23	<PP_CL>
20	அன்புடைய	<ADJ>	21	<ATT>
21	நெஞ்சம்	<NN>	22	<R_SUB>

[1]http://maltparser.org/userguide.html

22	தாம்	<PNR>	23	<NSUB>
23	கலந்தனவே	<VF>	0	<ROOT>

Conclusion

MALT parser tool is language independent tool used for dependency parsing which is implemented for several languages. Using this tool the dependency labels and position of head in Tamil language is obtained.

References

1. Akshar Bharati, Rajeev Sangal, Dipti Misra Sharma and Lakshmi Bai, AnnCorra: Annotating Corpora Guidelines for POS and ChunkAnnotation for Indian Languages, Technical Report, LanguageTechnologies Research Centre IIIT, Hyderabad, 2006.
2. Dhanalakshmi V, Anand kumar M, Vijaya M.S, Loganathan R, Soman K.P and Rajendran S. (2008), "Tamil Part-of-Speech tagger based on SVMTool", Proceedings of the COLIPS International Conference on natural language processing (IALP-2008), Chiang Mai, Thailand.
3. Dhanalakshmi.V, Anand Kumar M, Soman K P and Rajendran S (2011), "Shallow Parser for Tamil" Tamil internet Conference (2011), University of Pennsylvania, Philadelphia, USA.
4. Nivre J (2005), "Dependency grammar and dependency parsing", Technical Report MSI report 05133, Vxj University: School of Mathematics and Systems Engineering.
5. Lehmann Thomas (1983), "A Grammar of Modern Tamil", Pondicherry: Pondicherry Institute of Linguistics and Culture.
6. Tholkaappiyam, sollathikaaram, kazaka veLiyiidu (1994) .