Designing Contextual Part-of-Speech Tagger for Garden-path Sentences

Rudranarayan Mohapatra, Bishnupriya Otta

Abstract:

This paper will look at the process of creating a contextual Part-of-Speech tagging modular solution for English language that would be suitable for better clause boundary detection at the time of English-to-Indian Languages machine translation process. Though a many researchers have experimented in this area considering English as the source language, however capturing the contextual morph-syntactic information in English garden-path sentences POS string itself and its further use to capture the same morph-syntactic information in target languages side very specific to Agglutinative nature languages like Odia, Bangla itself a challenge task. Looking this nature of scope, this paper is trying to explain the hybrid approach with a super tag set for garden-path sentences and its scope, constraints and result.

Key Words- Machine Translation (MT), Garden-path (GP), Contextual part-of-Speech (CPOS), human sentence parsing mechanism (HSPM), Background Information (BI), Foreground Information (FI) and Constraint elements (CE).

I. INTRODUCTION:

One of the central goals of Natural Language Processing is to provide a systematic account of how, the system interprets structurally ambiguous sentences. The need to develop a useful and satisfactory Machine Translation system is increasingly appreciated in contemporary e-society and globalization of economy development. Many approaches, stochastic, rule-based and /or hybrid have been experimented to overcome technical barriers through the course of time. However it is always have a challenge to fine grain and extract the relevant contextual information from the source sentence itself. This paper mainly discusses to capture the contextual Part-of-Speech tag string of Garden-Path sentences especially in English Corpus with an aim to translation them to Indian languages.

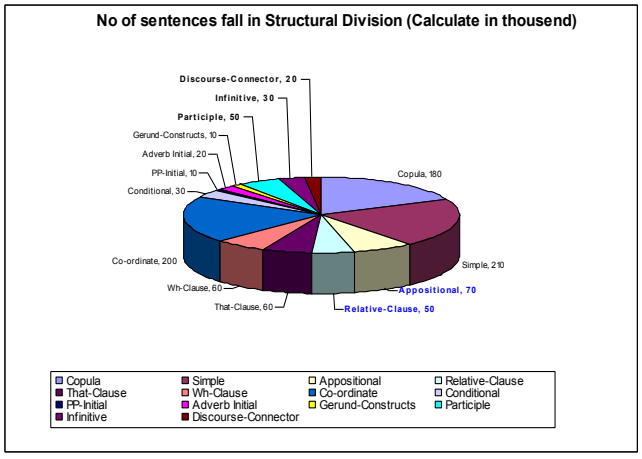
Garden-path Sentences:

Garden-path (GP) sentence is a special linguistic phenomenon in which the original clausal processing breaks down when the backtracking occurs. So that it extends human sentence parsing mechanism (HSPM) and exert the lingering effects on cognition. Crain and Coker (1979) note: "garden path sentences result from the omission of all syntactic markers which signal that one is parsing a Complex NP".[1] The existence of garden path effects and semantic anomalies is evidence that the human sentence parsing mechanism (HSPM) makes rapid decisions about the alternative or grammatically well formed structural representation to adopt when the input is itself ambiguous. The factors that influence the decision-making of the parser and the manner in which the parser is influenced are matters of controversy. However, on account of ambiguity resolution is known as the garden path model (Frazier, 1979).[2] In this paper, the below described rule based process is a small step to resolve the ambiguities at the time of machine translation from English to Indian languages.

Corpus Analysis Report for Garden-Path Structures:

To make the live corpus analysis we have considered random 1000 sentences for initial analysis followed by another 1000 sentences are used for evaluation. In analysis 1000 corpus by manual preprocess and structure identification step we have considered some structural pattern as so-called possible garden path structures. So irrespective of multiple grammatical information in the analysis sentences based upon their garden-path priority, we have assigned only one category to one sentence rather than the multiple. We have considered the below mentioned five categories in this area at an instant manual analysis process.

Figure No. 1



Syntactic structures considers under G-Path structure are: (1) Infinitive Clausal sentences (2) participle structures, (3)

[1] Predicting Garden Path Sentences*, COGNITIVE SCIENCE 6, 349-373 (1982), ROBERT WILLIAM MILNE, Department of Electrical Engineering Air Force Institute of Technology, Wright-Patterson Air Force Base Ohio 45433

[2] Frazier, L. (1979). On comprehending sentences Syntactic parsing strategies, unpublished doctoral dissertation, University of Connecticut, Distributed by the Indiana University Linguistics Club, Bloomington. Taken from, "Sidestepping Garden Paths: Assessing the Contributions of Syntax, Semantics and Plausibility in Resolving Ambiguities*, Weijia Ni, Stephen Crain and Donald Shankweiler, Haskins LAboratories Status Report on Speech Research 1994-1995, SR-119/120, 139-173

Embedded Relative clause verb sentences, (4) Appositional & (5) Discourse Connector Sentences. According to analysis this pattern captures average **22%** of whole corpus. GP sentence usually is created by punctuation absence, Conjunction, Compliment or Relative Pronoun marker absence in between the clauses, the deviation of semantic category and shift of lexical meaning. Though it always difficult to exactly determine the garden-path sentences, we have consider the mentioned structural pattern in this broader area with an argument that it back tracks the reader's process to identify its behavioral cognition.

The Appositional and Discourse connector structural pattern sentences most cases have different syntactic behavior. For example: "*The Modhera Temple, built by the Solanki kings in 1025 AD was dedicated to Surya the Sun God."* In this sentence the garden part clause 'built by the Solanki kings in 1025 AD' , is the modifier to the subject of 'The Modhera Temple was dedicated to Surya the Sun God' and 'the verb 'built' is behaves as the embedded relative clause verb to drive the whole sentence as garden part clause nested in the main clause.

Exemplary sentences and their sentential structure pattern:

The below mentioned Table No.1 (a) & (b) has the exemplary Garden-path sentences that appears in live corpus analysis of tourism domain with their syntactic POS tagged output received from availed Stanford POS Tagger version 3.0.

Table No. 1 (a)

G-path Category	Examples
Having Infinitive clausal Infinitive marker	The annual trade event of Tourism Australia – Australian Tourism Exchange (ATE) is going to be held at Melbourne
Having Embedded Relative Clause verb	Akshyaya Trutiya is exclusively an agricultural festival held on the third day of the Hindu year.
Having embedded Progressive participle Verb	Then they sow seeds ceremonially praying the Goddess for a rich bumper crop.
Having VP-NP-VP	The room that houses the actual clock mechanism is full of Victorian mechanical wonders

Table No. 1 (b)

Output of Stanford Tagged POS
The@@@@DT annual@@@@JJ trade@@@@NN event@@@@NN of@@@@IN Tourism@@@@NNP Australia@@@@NNP -@@@@: Australian@@@@JJ Tourism@@@@NNP Exchange@@@@NNP <@@@@NN ATE@@@@NNP >@@@@NN is@@@@VBZ going@@@@VBG to@@@@TO be@@@@VB held@@@@VBN at@@@@IN Melbourne@@@@NNP
Akshyaya@@@@NNP Trutiya@@@@NNP is@@@@VBZ exclusively@@@@RB an@@@@DT agricultural@@@@JJ festival@@@@NN held@@@@VBN on@@@@IN the@@@@DT third@@@@JJ day@@@@NN of@@@@IN the@@@@DT Hindu@@@@NNP year@@@@NN
Then@@@@RB they@@@@PRP sow@@@@VBP seeds@@@@NNS ceremonially@@@@RB praying@@@@VBG the@@@@DT Goddess@@@@NNP for@@@@IN a@@@@DT rich@@@@JJ bumper@@@@NN crop@@@@NN
The@@@@DT room@@@@NN that@@@@WDT houses@@@@VBZ the@@@@DT actual@@@@JJ clock@@@@NN mechanism@@@@NN is@@@@VBZ full@@@@JJ of@@@@IN Victorian@@@@JJ mechanical@@@@JJ wonders@@@@NNS

English Contextual POS Tagger:

It is a modular solution comprising different grammatical sub functionalities of English Language to get a compressive and Contextual TAG set in order to lead the system introduction and increasing output accuracy. The Contextual POS Tag set digest the standard lexical tag set and add the new Contextual Tag Sets viz. Conditional, relative clause verb, interrogative marker, gerundial and participle markers etc. as micro levels Part-of-Speech Tag Sets. For this we have followed a two step approach of combination of stochastic Stanford POS tagger followed by heuristically & hierarchical rules created manually observed from live corpus structure pattern analysis. Table : 2

Division of Verbs with respect to Garden-path direction:

Syntactic POS Nomenclature	Stanford Tagger used POS	Used Contextual Tag set	Contextual POS Nomenclature
Copulative Verbs	VB	VX	Only Copulative Verbs
VERB Base Singular	VBZ	TXS	Verb Base singular except copulative
VERB Base Plural	VB	TX	Verb Base Plural
VERB Progressive	VBG	TXG	Verb Progressive
		ADJG	Verbal Progressive-participe dérive Adjectives
		GER	Gerundial Noun marker
		PrPrt	Progressive participle clause Verb
VERB past	VBD	TXD	Verb Past
VERB Participle	VBN	TXN	Verbal Participle
		ADJN	Verbal Past-participle derive Adjectives
		PtPrt	Past participle clause Verb
		ReClv	Embedded Relative Clause verb

Part-of-Speech Hierarchical division in order to fix them in rule based Approach: The resolving of part of speech for G-path sentences, we have made them classification of Pos in two steps. In first step we are trying to solve the 'TXS, TX, TXG, ADJG, GER, TXD, TXN, ADJN' pos in rule based approach. And in second step we are applying super TAG set to demarcate the INF, PrPart, ReClv etc according to their complexity hierarchy in Semantics and Contextual nature of behavior.

Applied rules for embedded Relative Clause verb:

"One characteristic of garden-path sentences that seems to strongly influence ease of reanalysis concerns the syntactic relationship between the error signal and the head of the phrase that has been misanalyzed (Ferreira & Henderson, 1991, 1998)."[3] The rule for garden path effects can either be instigated or deterred by substituting one pre-nominal modifier for another. The rule in between the function triggered by the focus operator i.e. the token (Verbal modifier) itself to determine is it tends to garden path effects or not. And to control the focus operator we have considers three other representatives of that sentences. Those are: Background Information (BI), Foreground Information (FI) and Constraint elements (CE). Where the background information represents the previous syntactic and morphological Information of the 'Focus Operator', the Foreground Information leads the Focus Operator based upon post syntactic and morphological Information. The Constraint elements are not mentioned explicitly in the sentence; instead, it is presupposed to exist or not.

Under the pseudo code of determination of Reduced Relative clause Verb marker

Assign ReClv category to (TXN) if it matches to specific pattern

If,

DET/NOUN/ADJ + AUXILARY (minus) + ADV (optionnel) + TXN + ... (-Conjonction/Ponctuation/Relative-prônons/complime nt 'that') + VERB/ AUXILARY

Then

TXN = ReClv

However if any bracketed or quoted part comes out then that whole part will be considered as single part & will again re-loop the Embedded Relative Clause verb Rules.

Where 'DET' (Determinants), NOUN includes ('NP, NN, EX, FW, LS, PP, NPS, NNS, NNP, POS, PRPDOLLOR, PRP and NNPS') ADJ (JJ, JJR, JJS, & VBG changes to Adjectives by hierarchical rules) and ADV represents Stanford tag set 'RB, RBR & RBS'.

[3] Misinterpretations of Garden-Path Sentences: Implications for Models of Sentence Processing and Reanalysis, Fernanda Ferreira,1,4 Kiel Christianson,1,3 and Andrew Hollingworth2,3, Journal of Psycholinguistic Research, Vol. 30, No. 1, 2001

Reduced Relative clause Verb marker rule, TXN is the Focus Operator (FO) Where Punctuations, Conjunctions, Relative Pronouns and Compliments are Constraint Elements. And all others are treated as BI or FI depending upon their positions with respect to FO. As the Syntactic rules are superseded by a strict functional behavior so for a particular set of Focus Operator (FO) its other previous rules Focus operators may be BI or FI but not CE if not mentioned particularly.

After Conversion of Lexicalized Part-of-Speech tag set to Contextual Tag Set and the above mentioned sentences Contextual tag string will be as follows:

Examples	Tagged POS
The annual trade event of Tourism Australia – Australian Tourism Exchange (ATE) is going to be held at Melbourne	The@@DET annual@@JJ trade@@NN event@@NN of@@IN Tourism@@NNP Australia@@NNP -@@: Australian@@JJ Tourism@@NNP Exchange@@NNP <@@OBrac ATE@@NNP >@@CBrac is@@AUX going@@TXG to-be-held@@INF at@@IN Melbourne@@NNP
Akshyaya Trutiya is exclusively an agricultural festival held on the third day of the Hindu year.	Akshyaya@@NNP Trutiya@@NNP is@@VX exclusively@@RB an@@DT agricultural@@JJ festival@@NN held@@ReClv on@@IN the@@DT third@@JJ day@@NN of@@IN the@@DT Hindu@@NNP year@@NN□□
Then they sow seeds ceremonially praying the Goddess for a rich bumper crop.	Then@@RB they@@PRP sow@@TXP seeds@@NNS ceremonially@@RB praying@@PrPART the@@DET Goddess@@NNP for@@IN a@@DET rich@@JJ bumper@@NN crop@@NN□
The room that houses the actual clock mechanism is full of Victorian mechanical wonders	The@@DT room@@NN that@@COMPLT houses@@TXS the@@DT actual@@JJ clock@@NN mechanism@@NN is@@VX full@@JJ of@@IN Victorian@@JJ mechanical@@JJ wonders@@NNS

Evaluation:

Experiments and analysis were performed to examine the performance of the proposed approach. To evaluate the performance of the proposed method, analysis on the coverage of the extracted Contextual POS Tag strings for the source language was conducted. For this we have considered

a sample of other 1000 garden-path sentences of above mentioned categories having correct Stanford POS string for subjective and manual evaluation. On that basis, we have got an average of 759 sentences are providing the desired corrected contextual POS Tag String on an average of 75% accuracy.

However, in un-cleaned input sentences and not removing of incorrect Stanford POS string for the Overall evaluation of Contextual Part-of-Speech we have got a fairly less number of corrected contextual POS Tag String. On this basis we have got a series of evaluation from source sentences evaluation of preprocessing. On that case on an average of 5% sentences are found to be discarded as their incorrectness in input string itself. Out of correct input, further evaluating the incorrect Stanford Part-of-Speech tag string on the same way of 5% sentences are found to be discarded, we are reached to 712 sentences Contextual POS string having seems OK, i.e. 71%.

Constraints:

As this system a hybrid system in which linguist users introduce high level contextual rules those are inspired in the Constraint Grammars formalism those to be applied in combination with a tagger based on a Hidden Markov Model (HMM). The super tagging rules are highly fluctuated due to different source language constraints like: (1) Random use of punctuation & conjunction markers (2) Case Sensitive problem, (3) Statistical tagger learning dependencies, (4) Single string single POS etc with number of ifs & buts etc. As English is a case sensitive language and in that direction in some cases the stochastic POS tagger itself provides the output of the same token as NOUN group instead of verb. In this condition it is hard to probable the exact contextual POS of the same token in rule base approach. For instance: For the sentence 'Visit this place', the stochastic POS tag string is '[Visit@@@@NN, this@@@@DT, place@@@@NN]' due to sentential case sensitive, Where it is hard to probable the contextual tag set of 'Visit' as 'Verb'.

Conclusion

Our results demonstrate that by incorporating more probabilistic rules and enhancing the stochastic learning, we can achieve impressive performance gains across a range of scenarios. In this paper, we showed that performance improves as the number of rules increases. We were limited by our corpus, but the future work can be increased by addition of massively parallel corpora that to involving as well as learning from languages. Our experiments also have makes a clear picture that performance can vary greatly depending on the choice of hierarchical rules and addition of conditions.

References:

[1] Predicting Garden Path Sentences*, COGNITIVE SCIENCE 6, 349-373 (1982), ROBERT WILLIAM MILNE, Department of Electrical Engineering Air Force Institute of Technology, Wright-Patterson Air Force Base Ohio 45433

[2] Frazier, L. (1979). On comprehending sentences Syntactic parsing strategies, unpublished doctoral dissertation, University of Connecticut, Distributed by the Indiana University Linguistics Club, Bloomington. Taken from, "Sidestepping Garden Paths: Assessing the Contributions of Syntax, Semantics and Plausibility in Resolving Ambiguities*, Weijia Ni, Stephen Crain and Donald Shankweiler, Haskins LAboratories Status Report on Speech Research 1994-1995, SR-119/120, 139-173

[3] Misinterpretations of Garden-Path Sentences: Implications for Models of Sentence Processing and Reanalysis, Fernanda Ferreira,1,4 Kiel Christianson,1,3 and Andrew Hollingworth2,3, Journal of Psycholinguistic Research, Vol. 30, No. 1, 2001

[4] Huet, G´erard 2003a. Zen and the Art of Symbolic Computing: Light and Fast Applicative Algorithms for Computational Linguistics, in Practical Aspects of Declarative Languages (PADL) symposium. pauillac.inria.fr/~huet/PUBLIC/padl.pdf.

[5] Mitchell P. Marcus, B. Santorini, M.N. Marcinkiewicz, 'Building a Large Annotated Corpus of English: The Penn Treebank', University of Pennsylvania; http://acl.ldc.upenn.edu/J/J93/J93-2004.pdf

First Author: Dr. Rudranarayan Mohapatra presently working as a Senior Technical Officer at C-DAC, Pune, India. He has a number of publications in Computational Linguistics field and NLP area in international journals and conferences.

Second Author: Dr. Bishnupriya Otta, presently in the position of Reader at Utkal University, India. She has expertise very specific to Odia language and having number of publications in this area.

