Various Parsers Available for Indian and Foreign Languages: A Survey

Ishan Kumar
Ph.D. Research Scholar
NIT Jalandhar

Renu Dhir
Associate Professor
NIT Jalandhar

Sanjeev Kumar Sharma
Assistant Professor
DAV University Jalandhar

ABSTRACT

Parsing is one of the essential activity to analyze the sentence structure. A sentence will be parsed if it is grammatically correct i.e. follow the grammar rule of that particular language for which it has been designed. In many natural language processing applications such as grammar checker, question answering, summarization and machine translation etc., parser play an important role. In this research work author has provided a survey regarding various parsers developed for Indian and foreign languages. This paper addresses the various developments in parsers for Indian language, which is very essential computational linguistic tool needed for many natural language processing (NLP) applications like grammar checker and sentence structure analysis.

Keywords

Parser, grammar, sentence structure, parser for Indian languages.

1. INTRODUCTION

Parsing is the process of analyzing a sentence to determine its grammatical structure using the grammar rules of the underlying language. A parser applies the grammar rules on the input sentence to determine its structure and establish relationships between various syntactical components like phrases, clauses etc. of that sentence. A sentence is termed as ungrammatical if that fails to get parsed completely. In order to discern the error(s) present in the structure of a sentence,a typical parser can relax some constraints to parse ungrammatical sentence. This approach is called constraint relaxation.

2. AVAILABLE PARSER FOR INDIAN LANGUAGE

As compare to foreign languages, a very little work has been done in the natural language processing for Indian languages. Various Parsers for Indian Languages like Hindi, Bengali, Telugu, Marathi, Kannada and Assamese are available but further parsers are being developed for other Indian languages. Accuracy of the parser depends upon the approach used for developing the parser hence techniques used is also provided. A list of parsers developed for Indian languages is provide in table 1.

Table 1. Parsers developed for Indian languages

Sr. No.	Language	Author name	Year
1	Hindi, Bangla, Telugu	JoakinNirve	2009
2	Hindi	Akshar Bharti et. al.	2009
3	Assamese	Rahman, Mirzanur et. al.	2009
4	Bengali	Anirudh Ghosh et. al.	2009
5	Kannad	Antony P J	2010
6	Bangla	Sankar et. al.	2009
7	Marathi	Dhanshree Kulkarni et. a.	2014
8	Tamil	SelvamM et. al.	2008

3. AVAILABLE PARSERS FOR FOREIGN LANGUAGES

A lot of work has been done for foreign languages. Various techniques used for developing the parser for foreign language includes Constraint grammar framework approach, METAL grammar formalism approach, Two level morphological model based approach, Transformation based approach, Stochastic approach, Finite state transducers, Memory based learning approach, Maximum entropy based approach and Lexical functional grammar (LFG).Given below are some of the existing parsing systems:

- Karlsson (1990) used constraint grammar (CG) framework for English. This parsing system consists of various modules like morphological analyzer, morphological disambiguator, morphosyntactic mapping, context-dependent morphological disambiguation, clause boundary detection, and disambiguation of surface syntactic functions. Various constraints are applied to reduce morphological and syntactic ambiguities present due to lexical information or morphosyntactic mapping.
- Voutilainen and Heikkila (1993) provided English Constraint Grammar (**ENGCG**), a parsing system for English. Morphological analyzer is based on the two-level morphology model. This parsing system is based on constraint grammar framework. The morphologically analyzed and disambiguated output is given to the syntactic parsing module that assigns all the surface-syntactic functions to the input word forms and then there is a disambiguation module that performs similar to morphological disambiguation module to discard the unnecessary syntactic tags based on the context information.
- Xia and Wu (1996) provided a parsing system for Chinese based on Context Free Grammar (CFG) framework. The authors introduced two extensions to standard CFG, right hand side contexts and non-terminal functions. Right hand side contexts are introduced to restrict the compounding of noun phrases and non-terminal functions are introduced to formulate conditions to reduce ambiguities. The non-terminal functions can apply equally to both terminal and non-terminal symbols. The accuracy results are reported to be within the range of 78-85%, depending upon the sentence length. The system is more accurate in parsing and bracketing shorter sentences. The authors mentioned that the system may produce more than one parse tree, if possible for an input sentence.
- Tapanainen and Järvinen (1997) described a parsing system for English. The main idea behind this parser is to show unrestricted dependencies. The parser links the headwords and their modifiers or dependents, and marks the links with the syntactic function or relationship like subject, object, determiner etc. The authors found this approach promising when compared with other such existing parsing techniques like ENGCG. This work is partly based on the work presented by Karlsson (1990) for parsing using constraint grammar.
- Collins et al. (1999) provided a statistics based parser for Czech. The parser uses a lexicalized grammar with each non-terminal having a headword and part-of-speech. This system is reported to have achieved an accuracy of 80%.
- Kuboň (1999) presented a robust parser for Czech. This parser uses the grammar designed to cover a wide range of Czech sentences. The input sentences may contain grammatical errors. The author hoped that this system can help in designing similar systems for other Slavic languages. The grammar of this parser covers most frequent syntactic structures for simple clauses and certain types of complex sentences.
- Daelemans et al. (1999) presented a memory based learning approach to shallow parsing of English. This framework has different memory-based modules for part-of-speech tagging, chunking, and identification of syntactic functions.
- Bangalore and Joshi (1999) presented **Supertagging**, a novel approach for robust parsing of English. This approach is presented in the context of Lexicalized Tree Adjoining Grammar (LTAG) formalism but the authors noted that it is equally applicable to other lexicalized grammar formalisms. In simple terms, this parser takes POS tagged and disambiguated text as input and then assigns supertags to the lexical items in that input. Each lexical item is assigned as many supertags as the number of different syntactic contexts in which that item can appear.
- Charniak (2000) presented a parser for English that uses maximum entropy model for conditioning and smoothing the probabilities required by this generative parser. This model is an improvement of the probabilistic model presented by the author in 1997.
- Joshi and Hopely (1996) described a parsing system for English that was based on finite state transducers (FST). This program works by performing lexicon look up to assign all possible tags to the words in the input sentence. Then it replaces some ‘grammatical idioms’ with single part-of-speech like ‘of course’ with adverb etc. After that rule based system is applied for POS disambiguation.

- Aldezabal et al. (2000) presented a parsing system for unrestricted Basque text. This parser consists of two sequential modules, one using unification based grammar and other based on finite-state models. Unification based parser builds basic syntactic units in the sentence. This parser is based on a grammar containing 120 rules and performs bottom-up parsing to build a chart as output. The output thus produced is still ambiguous; to resolve this ambiguity finite-state based parsing is used. This finite-state parsing module performs syntactic disambiguation and filters the results produced by unification based parsing module. The authors noted that finite-state networks are not able to handle complex agreement and free constituent order nature of Basque, therefore, unification-based parser is required for this.
- Knutsson et al. (2003) described a shallow parsing system for Swedish known as Granska Text Analyzer **(GTA)**. The parser does not build complete tree structures but identifies internal structure of phrases and detects clause

boundaries. The parser is robust in terms that it can handle noisy and ill-formed input text. The parser works on POS tagged and disambiguated Swedish text. This is a rule-based parser using handwritten rules augmented with features. It uses 200 rules for identifying phrase structures, 40 heuristic rules for disambiguating ambiguous phrase identifications, and 20 rules are for clause boundary identification.

- Dubey and Keller (2003) presented a probabilistic parsing system for German. This model uses sister-head dependencies unlike head-head dependencies found successful for English. The authors found that the model followed for English fails to outperform even the unlexicalized baseline model for German. This new model was trained on Negra, a syntactically annotated corpus for German.
- Taskar et al. (2004) presented a discriminative approach for parsing of English. This approach called maximum-margin is inspired by the large-margin criterion underlying support vector machines. This framework permits the use of a number of input features. This approach is reported to perform well when compared with other similar approaches. However, the authors noted that this model being discriminative requires more computational time for training.
- Haque and Khan (2005) proposed a parsing system for Bangla language. This parser is based on Lexical Functional Grammar (LFG) formalism. The authors showed the parsing results using a toy grammar of Bangla for simple sentences only.
- Hettige and Karunananda (2006b) presented a parsing system for Sinhala to be used as a part of machine translation from English to Sinhala. The parser uses grammar of the Sinhala language and works on the output produced by morphological analyzer of Sinhala. The parser produces a parse tree of the input sentence, in which each word is marked with its grammatical information and grammatical relations between various words are marked as subject, object etc. The authors reported that this parser can handle simple and complex sentences of Sinhala.
- A transformation based approach to parsing of free English text has been provided by Brill (1993). This system starts with the knowledge about basic phrase structure and then learns simple structural transformations by comparing its output with the training corpus. These learnt transformations help in error reduction for future parses. The transformations are learnt by repeated comparison of bracketing in the current state with the proper bracketing present in training corpus.
- Loftsson&Rögnvaldsson (2007) and Loftsson (2007) presented details of a finite-state parsing system for Icelandic language. This is a shallow parsing system. It takes POS tagged text as input and produces a shallow syntactically annotated output. Its POS tagset consists of 660 tags. The parser consists of a series of finite state transducers to annotate text with phrase tags and syntactic functions. The parser consists of two separate modules, one for marking phrases and other for denoting syntactic functions of those phrases. The former has 14 transducers and the latter works on using 8 transducers. The authors planned to use this parser for grammar checking application so due to this they have relaxed feature agreements while marking up phrases.
- Thurmair (1990) explained a parsing system for use in grammar and style checking applications. As for grammar checking, the text can be ill-formed, so parser needs to deal with that. This parser uses METAL grammar formalism. This parser uses fallback rules, in its grammar, similar to constraint relaxation approach to parse ill-formed text and mark the sentence as ungrammatical as soon as a fallback rule is fired.

List of the parsers developed for foreign languages is provided in table 2.

Table 2.list of parsers developed for foreign languages

Sr. No.	Language	Author name	Technique used	Year
1	English	Karlsson	Constraint grammar framework	1990
2	German	Thurmair	METAL grammar formalism	1990
3	English	Voutilainen and Heikkila	Two level morphological model	1993
4	English	Brill	Transformation based	1993
5	English	Schabes et al.	Stochastic approach	1993
6	English	Joshi and Hopely	Finite state transducers	1996
7	Chinese	Xia and Wu	Context free grammar	1996
8	English	Tapanainen and Järvinen	Constraint grammar	1997
9.	Czech	Kuboň	Grammar based	1999
10.	Czech	Collin et. al	Lexicalized grammar	1999
11.	English	Daelemans et al.	Memory based learning approach	1999
12.	English	Bangalore and Joshi	Lexicalized tree grammar formalism	1999
13.	English	Charniak	Maximum entropy	2000
14.	Basque	Aldezabal et al.	Finite state transducers	2000
15.	Swedish	Knutsson et al.	Rule based	2003
16.	German	Dubey and Keller	Probabilistic	2003

17.	English	Taskar et al.	Maximum margin approach	2004
18.	Bangla	Haque and Khan	Lexical functional grammar (LFG)	2005
19.	Sinhala	Hettige and Karunananda	grammar	2006
20.	Iceland	Loftsson&Rögnvaldsson	Finite state transducers.	2007

4. CONCLUSION

From this survey, it is concluded that parsing is an essential component for many natural language processing applications. Parser for most of the foreign languages has been developed but a lot of work yet to be done for Indian languages. Moreover, the accuracy of parser also depends upon the technique used for its development. The technique used depends upon the type of language and morphological features of the language. Therefore, efforts must be done to develop the parser for morphologically rich Indian languages.

REFERENCES

[1]. Aldezabal, Izaskun, Koldo Gojenola, and Kepa Sarasola. 2000. A Bootstrapping Approach to Parser Development. In *Proceedings of the Sixth International Workshop on Parsing Technologies (IWPT-2000)*, Trento, Italy.
[2]. Charniak, Eugene. 2000. A Maximum-Entropy-Inspired Parser. In *Proceedings of NAACL*, pages 132-139, Seattle, WA.
[3]. Church, Kenneth Ward. 1988. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of Second Conference on Applied Natural Language Processing*, pages 136-143, Austin, Texas.
[4]. Collins, Michael, Jan Hajič, Lance Ramshaw, and Christoph Tillmann. 1999. A Statistical Parser for Czech. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 505-512, University of Maryland, College Park, MD.
[5]. Freeman, Andrew. 2001. Brill's POS Tagger and a Morphology Parser for Arabic. In *Proceedings of the ACL/EACL – 2001, Workshop on Arabic Language Processing: Status and Prospects*, Toulouse, France.
[6]. Hettige, Budditha and Asoka S. Karunananda. 2006b. A Parser for Sinhala Language – First Step Towards English to Sinhala Machine Translation. In *Proceedings of the First International Conference on Industrial and Information Systems (ICIIS)*, pages 583-587, Sri Lanka.
[7]. Joshi, Aravind K. and Phil Hopely. 1996. A Parser from Antiquity. *Natural Language Engineering*, 2(4):291-294.
[8]. Knutsson, Ola, Johnny Bigert, and ViggoKann. 2003. A Robust Shallow Parser for Swedish. In *Proceedings of the 14th Nordic Conference on Computational Linguistics (NODALIDA-2003)*, Reykjavik, Iceland.