Improving Existing Punjabi Morphological Analyzer

Gagan Bansal[1], Satinder Pal Ahuja[2], Sanjeev Kumar Sharma[3]

[1]Indo Global College of Engg., Mohali

[2]Associate Professor, Indo Global College of Engg., Mohali

[3]Associate Professor, B.I.S College of Engineering and Technology, Moga

Abstract: *Morphological analysis is essential for natural language processes like generation of Treebanks, training of parsing models and parsing. Rule based approach is applicable to the languages which have well defined set of rules to accommodate most of the words with inflectional and derivational morphology. Rule based morphological analysis is very difficult and cannot accommodate all combinations through the rules due to inflections and exceptions especially in languages like Punjabi. Statistical methods are very important which in turn need large volume of electronic corpus and automated tools which are not available in Punjabi. Lexicon based morphological analyzer has been developed for Punjabi. This can be further improved by adding new words with their grammatical information.*

Introduction

Natural Language Processing (NLP) is a set of computational techniques for analyzing and representing text in Natural Language (NL) with linguistic analysis for achieving human-like language processing for a range of tasks or applications. It deals with interactions between computer and human (natural) languages. Computational linguistics (CL) is an interdisciplinary field dealing with statistical and/or rule based modeling of NL from a computational perspective .Major components driven by these broad areas are syntax, semantics and pragmatics. NLP applications

like Grammar Checker and Language Analyzer need a parser with an optional parsing model. Parsing is the process of analyzing the text automatically by assigning syntactic structure according to the grammar of NL. A parser is a computational system which processes input sentences and builds one or more constituent structures called parses or parse trees. For a simple parsing task considering languages with a limited vocabulary, parsers using rule based techniques are usually sufficient. For applications requiring a large vocabulary parsers based on more sophisticated parsing models are needed, for example models which use probability distributions over parses to accomplish the disambiguation task in ambiguous sentences. Adequate parsing models can be created by adding structural components in statistical methods which satisfy the constraints needed for the parsing process. In order to build a parsing model, large volume of Treebank is needed. Treebank is a corpus with linguistic annotation beyond word level. Part of Speech (POS) tagging and phrasing are essential for the development of Treebank. POS tags are generated through morphological analysis. POS tagging is used to assign or select correct POS tag to a word before syntactic analysis. Phrasing is the process of applying morpho-syntactic relations among words in the formation of constituents which in turn build parse trees. Collection of phrase structured sentences together constitutes a Treebank. Morphological analysis is the process of segmenting a given word into a sequence of morphemes. It is closely related to POS tagging but word segmentation is required for natural languages because morpheme boundaries are not indicated. Inflectional morphology gives different forms added to a root word whereas derivational morphology derives new words by inclusion of affixes. Lexical and surface levels of words are studied through morphological analysis. Based on that, POS tags are suggested to words in a sentence.

Overview of Morphological Analyzer

Morphological Analyzer is a device which gives analysis of a given word. A 'word' here refers to a string of characters separated by a space. This process can be described as similar to the understanding process that takes place in the human mind. A Morphological Analyzer, MA, is a program or algorithm which determines the morpheme(s) 1 of a given inflected or derived word form including the analysis of the bound morphemes in it's grammatical form. For the MA to function, a stemmer should be priory present

in the system. Hence, wherever required, we would also be talking about the stemmer throughout the document. Stemmers are used to find the root of the inflected or derived word form. MA uses the information of the stemmer and keeps track of the bound morpheme(s) present in the original inflected or derived word and in addition to this provides grammatical information. MA is generally used for information retrieval and other Natural Language Processing Applications like the Parts-Of-Speech (POS) Tagging, Machine Translation etc. MA have been designed and implemented for several languages of the world. Complexities in design and development, however, prevail for inflectionally and derivationally rich languages like Nepali. The stemmer and consequently the MA being developed for Nepali currently does not handle compound words formed as a result of the concatenation of individual words. Morphological analyzer and morphological generator are two essential and basic tools for building any language processing application for a natural language. Morphological analysis means to study the internal structure of the words of a language. A Morphological analyzer gives the morph analysis of a word i.e. for a given word a morphological analyzer will return its root word and word class along with other grammatical information depending upon its word class. Like for nouns it will provide gender, number, and case information and for verbs it will provide tense, phase etc. in addition to this. Morphological generator does exactly the reverse of it, i.e. given a root word and grammatical information it will generate the word form of that root word.

Punjabi World Classes and Inflection

Inflection is usually a suffix, which expresses grammatical relationships such as number, person, tense etc. Punjabi words may be inflected or uninflected. The word classes of Punjabi are as follow:

Noun: - Nouns have assigned gender, either masculine or feminine, though some nouns may be used in both genders. These inflect for number and case, grammatical categories. There are two numbers, singular and plural. There are six cases.

Pronoun:-Pronouns inflect for number and case, except the genitive case forms of all the pronouns, which show inflection for gender also in addition

to number and case. Except first and second person personal pronouns, all the other pronouns are in third person and may function as adjectives in sentences.

Adjective:-An adjective modifies a noun or pronoun i.e. highlights some of the properties of a noun or pronoun. In Punjabi, adjectives usually precede the nouns but follow the pronouns.

Cardinal:-Cardinals can equally be used for both the genders and change forms for case – direct and oblique. Except 'one', which is in singular number, all the remaining cardinals are in plural number.

Ordinal:-The ordinals inflect for gender – masculine and feminine, and case – direct and oblique. All the ordinals are used in singular number, generally.

Main verb:-Verbs inflect for gender (masculine and feminine), number (singular and plural), person (first, second, and third), phase (perfect and non-perfect), and tense (future). Inflectional root for verbs have assigned transitivity (transitive and intransitive) and causality (none, simple, and double). Transitive verbs are those that require an object in a sentence unlike

Auxiliary verb:-Auxiliary verb roots are for two tenses – present and past. ਹੈ hai (for present tense) and ਸੀ sī (for past tense) inflect for number – singular and plural, and person – first, second, and third. These forms can equally be used for both the genders.

Adverb:-Word acting as a modifier for verb in a Punjabi sentence is termed as adverb. Adverbs can indicate manner, time, place, cause etc. Unlike adjectives, adverbs can virtually be used at any place in Punjabi sentence. The adverbs have been classified into two categories, inflected and uninflected.

Postposition:-Postpositions are similar to prepositions in English. These link noun, pronoun, and phrases to other parts of the sentence. Some In Punjabi, postpositions follow the noun or pronoun unlike English, where these precede the noun or pronoun, and thus termed prepositions.

Conjunction:-The words of this word class are used to combine more than one clause in a sentence to form complex or compound sentences.

These words are uninflected and are categorized based on the type – co-ordinate and subordinate.

Interjection:-The words of this word class are uninflected. It is never a part of the sentence, so no agreement checking in terms of any grammatical category is required.

Particle:-All the words of this word class are uninflected. These have been classified based on their type like honorific, negative and emphatic. There are few particles which do not fall into any of the above three categories, so these can be grouped into one type – undecided.

Vocative particle:-Vocative particles are inflected for gender – masculine and feminine, and number – singular and plural,.

Verb-part:-The words of this word class are uninflected. It differs from other similar word classes in the sense that the words of this word class can only be used as a part of a verb phrase in a sentence.

Existing System

The existing approach followed for morphological analysis using full-form lexicon is just database look up. Each and every unmarked token in the input text is searched for in the full-form lexicon. If found in the lexicon then root and POS tag is assigned to that token. If not found then the token is looked in the similar words database as a wrong word. If it is found there then its correct form is taken to look for in full-form lexicon. If a word is not found in both full-form lexicon and similar words database, then it is marked as unknown.

Drawbacks of Existing System

The existing lexicon lacks the no of words and because of this a no of words comes out as unknown words. Even many common words like name of the city (.K>,.K9>2@) and villages is not present in the lexicon. Also the similar word database needs to be modified and the logic for checking the similar words needs to be improved. For example if we search for a word ਅੰਗਰੇਜ਼ੀ the existing morphological analyzer will give the complete description of this word with appropriate tag but if someone enters the word ਅਗਰੇਜ਼ੀ

that is he miss the *tippi* then the existing morphological analyzer will give it as unknown word instead of searching for a similar word ਅੰਗਰੇਜ਼ੀ . Also the existing system uses two different databases one containing the words and the other containing the similar words and it does not check the unknown word for similar word in the first database.

Proposed Method of Improvement

There is need of improvement at two major points. First the lexicon needs to be improved. This can be improved by adding more words. But as there is no limit of the words so we can add as many words as possible. Secondly we can implement some logic to find the similar words that have a difference of *tippi , adak* and *bibdi.* This is very challenging work as to add the word in the database first of all we will need to know the words that are not present in the database. For this we will use the following steps:-

Collection of corpus:- More the corpus available more will be the improvement. Another thing that should be kept in mind is that the corpus should be accurate and it should be from different domains. So we can start our work with the collection of accurate corpus. While collecting the corpus we kept the following points in our mind:-

The corpus should be in Unicode.

The corpus should be accurate i.e. it should have minimum no of spelling mistake.

The corpus should not be domain specific.

The corpus should contain as many different words as possible.

The corpus should contain maximum no of unknown words.

The main sources of our corpus are:-

- http://punjabikhabar.blogspot.com
- http://www.quamiekta.com
- http://www.europediawaz.com
- http://www.charhdikala.com
- http://punjabitribuneonline.com
- http://www.sadachannel.com
- www.veerpunjab.com

- www.punjabinfoline.com

Annotation of the corpus: Annotation of the corpus means giving a tag to the every individual word. The next step that we will perform after the collection of corpus will be annotation of the corpus. We will annotate the corpus by using a tool named TAGGER. This tool is developed by us. This tool has been developed from a pre existing Rule based POS Tagger. We made some alteration in that pre existing tool and used it for the annotation of our corpus. After passing the corpus from this toll we will get the list of Unknown words that are not present in the existing database (Lexicon).

Extracting the unknown words from the annotated corpus:- From the annotated corpus developed in the above step we will extract the unknown words. This can be done by just using a search module developed in c# language. This module will scan the whole annotated corpus and will separate the unknown words and will write them on a separate file.

Analysis of Unknown Words

The unknown words extracted in above stem will be analyzed and will be divided in to the following categories:

Words with spelling mistake

Words from other languages

Words with spelling mistakes will be further analyzed for those having spelling mistake due to *tippi, adak* and bindi. These words will be used for testing of our logic to find similar words having mistake due to tippi, adak and bindi.The words from other languages and the words that are not present in the database will be added to the database with their basic information that is their root word and Tag.

Adding the New Words to Database

The new words extracted and identified in above steps will be added to database. The database will be created in xml language and the programming will be done in c#.net.

References

1. Bharati, Akshar, Vineet Chaitanya and Rajeev Sangal. (1995). Natural Language Processing: A Paninian Perspective, Prentice-Hall of India, New Delhi.
2. Bharati, Akshar, Amba P. Kulkarni, Vineet Chaitanya. (1998a). Challenges in Developing Word Analyzers for Indian Languages, Presented at Workshop on Morphology, CIEFL, Hyderabad, July 1998.
3. Bharati, Akshar, Rajeev Sangal and S.M. Bendre (1998b). Some Observations on Corpora of Some Indian Languages. Knowledge Based Computing Systems, Tata McGraw-Hill.
4. Goldsmith, John. (2001). Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics, Vol 27, No. 2, pp 153-198.
5. Daniel Jurafsky, James H. Martin.Speech and Language Processing: An introduction to speech recognition, natural language processing, and computational linguistics.
6. LTRC, IIIT Hyderabad http://ltrc.iiit.ac.in
7. Gill Mandeep Singh, Lehal Gurpreet Singh, Joshi S.S., A full form lexicon based Morphological Analysis and generation tool for Punjabi,International Journal of Cybernatics and Informatics, Hyderabad,India,October 2007, pp 38-47
8. Aniket Dalal, Kumar Nagaraj, Uma Sawant and Sandeep Shelke, "Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach", In Proceeding of the NLPAI Machine Learning Competition, 2006.
9. Antony P.J, Santhanu P Mohan, Soman K.P,"SVM Based Part of Speech Tagger for Malayalam", IEEE International Conference on Recent Trends in Information, Telecommunication and Computing, pp. 339-341, 2010
10. Agarwal Himashu, Amni Anirudh," Part of Speech Tagging and Chunking with Conditional Random Fields" in the proceedings of NLPAI Contest, 2006
11. Bird, S., E. Klein and E. Loper, 2007. Natural language processing in python. University of Pennsylvania, Nltk. Sourceforge. http:// mail.python.org/pipermail/python-list/2007-May/442489.html.

12. Blunsom, P., 2004. Hidden Markov models. Technical Report.
13. Brants, TnT – A statistical part-of-speech tagger. In Proc. Of the 6th Applied NLP Conference, pp. 224-231, 2000
14. Cutting, J. Kupiec, J. Pederson and P. Sibun, A practical part of-speech tagger. In Proc. of the 3rd Conference on Applied
15. NLP, pp. 133-140, 1992
16. Dermatas and K. George, Automatic stochastic tagging of natural language texts. Computational Linguistics, 21(2): 137-163, 1995
17. Ekbal, Asif, and S. Bandyopadhyay,"Lexicon Development and POS tagging using a Tagged Bengali News Corpus", In Proc. of FLAIRS-2007, Florida, 261-263, 2007
18. Ekbal, Asif, Haque, R. and S. Bandyopadhyay, "Named Entity Recognition in Bengali: A Conditional Random Field Approach", In Proc. of 3rd IJCNLP, 51-55, 2008
19. Ekbal, A. Bandyopadhyay, S., "Part of Speech Tagging in Bengali Using Support Vector Machine", ICIT- 08, IEEE International Conference on Information Technology, pp. 106-111, 2008
20. E. Dermatas and K. George, Automatic stochastic tagging of Natural language texts, Computational Linguistics, 21(2): 137-163, 1995
21. Ekbal Asif, et.al, "Bengali Part of Speech Tagging using Conditional Random Field" in Proceedings of the 7th International Symposium of Natural Language Processing
22. (SNLP-2007), Pattaya, Thailand, 15 December 2007, pp.131-136
23. Gurpreet Singh, "Development of Punjabi Grammar Checker, Phd. Dissertation, 2008
24. Jurafsky D and Marting J H, Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Pearson Education Series 2002
25. James Allen, Natural Language Understanding, Benjamin/ Cummings Publishing Company, 1995.

* * * * *