IMPACT: International Journal of Research in Engineering & Technology (IMPACT: IJRET)
ISSN(E): 2321-8843; ISSN(P): 2347-4599
Vol. 3, Issue 5, May 2015, 17-22
© Impact Journals

MULTIWORD EXPRESSION EXTRACTION FROM NOISY TEXT USING LINGUISTIC RULES

VINEET KUMAR BIRLA

Research Scholar, Department of CSE & MEWAR University, Southern Rajasthan, India

ABSTRACT

Language Technology units such as Machine Translations require dictionaries. But available dictionaries are simple set of word pair [3]. Since the text is collection of inter-related sentences and in which group of words may mean differently than the meaning of individual words, dictionary proves insufficient to provide requisite knowledge to language technology units. To enable Language Technology units with requisite information, therefore, multiword expressions are required. While syntactic multiword extraction is simpler, that of semantic Multiword expression is difficult for process automation, since the identification itself is difficult. This papers presents algorithm for extraction of Multiword Expression from a given English text.

KEYWORDS: Corpora, Noisy Text, Charniak Parser, Multiword, Unique Words, Named Entities, Extractor

INTRODUCTION

Multiword Expressions are phrases that are not *entirely* predictable on the basis of standard grammar rules and lexical entries. Multiword expressions are used in various areas of academics as well as research. Some of them are Machine translation, Data Mining, Knowledge engineering, query refinement for search system and search engine indices etc.

The mechanism suits Multiword expression extraction from specific domain if domain specific dictionary is attached to it. Although for correctness and completeness manual verification is suggested, the overall effort in manual reduces considerably through the process. This mechanism has also been used for extraction of Multiword expression, unique words and named entities for an ongoing project titled English to Urdu and Punjabi Machine Translation based on Anglabharati methodology.

In rule based Machine Translation such as Anglabharati, there is a need of lexicon which contains root words of source language with their information like parts of speech and target language meaning etc. In order to have collection of words in the lexicon there is a need to find the words. To find out the appropriate words for the lexical database we use the extraction process for multiword expression and unique words etc. The process is comprised of multiple algorithms used at different phases of the process. The approach is only meant for the source language which is English.

NOISY TEXT

Noisy text is often the result of an end user's excessive use of idiomatic expressions, abbreviations, acronyms and business-specific lingo. It can also be caused by poor spelling, lack of punctuation, poor translations from optical (OCR) and speech recognition programs or typographical errors. [4] Noisy text is particularly prevalent in the unstructured text found in blog posts, chat conversations, discussion threads and SMS text messages. Since corpora is collected through various sources such as web in form of html pages, e-books in PDF format, OCR and many more. In order to extract information from noisy text, the text should be clean by removing errors like boundary of sentence should not be end at abbreviations like Mr., Dr. Etc. Removal of Headers, footers and link of web addresses, partial sentences, poor spell words, fill in the blanks and some hidden or control characters are also removed. The approach is semi-automatic. So there is a need of some manual observations also. In rule based Machine Translation such as Anglabharati, there is a need of lexicon which contains root words of source language with their information like parts of speech, category code, semantic, meaning and gender information etc. Language technology units such as Machine Translation system require dictionaries. But available dictionary are simple set of words [3]. Since text is a collection of inter-related sentences in which various information like Multiword Expressions. Although for correctness and completeness manual verification by language expert is suggested, the overall effort reduces considerably through the process.

EXTRACTION METHODOLOGY USED

It is a process comprising of multiple algorithms operated either in sequence or in parallel. The Multiword Expression methodology consists of five steps:

- Text Corpora Cleaning
- Tagging
- Parsing
- File processing
- Extractor

Text Corpora Cleaning

Corpora are collection of unified text. While grabbing the text from various resources often may contain unwanted characters, characters used for phonetic expressions, those used for emphasizing or headings, sub headings etc. Partial sentences also appear as a junk in view of unification of corpora. These unwanted words or sentences needs to be cleaned.

Algorithm is designed to identify sentence boundary through "." or "?" avoiding full stops appearing after certain terms such as Mr., Dr. and Prof. Etc. The identified sentences are checked with a lite version of morphological analyzer for its structure e.g. SOV (Subject Object Verb) form of English sentence, conjunct formation through "and" for noun and noun or parts of sentence and part of sentence. After running through this algorithm, raw text gets converted to clean corpora.

Tagging

Even the cleaned corpora may have anomalies. To identify, delete, modify or remove such anomalies before tagging a sentence, sentence boundary algorithm again is used. It is semi automated tool which expects 10 to 20 % of manual observations and redirections. This simple algorithm attaches <s> and </s> tag at the start and end of sentence respectively. Running through the algorithm cleaned text file gets converted to tagged text file.

3) Parsing

Charniak parser is an open source tool used to retrieve parsed information of text corpora. It takes tagged text file as an input and result parsed file as a output. The parsed file contains parsed text which is in form of part of speech is associated to each sentences and each word of sentence using lisp format. Charniak parser convert bounded English text to parts of speech based parsed English text. Following are some notations used to mark POS information for words

- CC Coordinating conjunction
- CD Cardinal number
- JJ Adjective
- NN Noun, singular or mass
- NP Proper noun, singular
- NPS Proper noun, plural
- PP Personal pronoun
- RB Adverb
- RP Particle
- VB Verb, base form

File Processing

The parsed file is given to algorithm designed exclusively to extract entities such as Multword expressions and unique words are required for AnglaBharati engine. The algorithm uses linguistic rules given below [1].

- **Noun Phrase (NP): Chicken in Red Wine Sauce**

  Noun Phrase can be seen as combination of following POS:

  NNs + JJ + NNs

  JJ + NN + In + NN + NN (red chicken in wine sauce)

  NN + NN (horse riding)

  NNP + NNP (Kanak Vrindavan)

  DT + JJ + NNP (The Pink City)

- **Verb Phrase (VP): Support Healthy Eating**

  Verb Phrase can be seen as combination of following POS:

  VB + NP

- **Adverb Phrase (ADVP): as Far as Possible**

  Adverbial Phrase can be seen as combination of following POS:

RB + RB + IN + JJ (as soon as possible)

- Verb + Preposition: **dealt in**
- Preposition + Noun: **down to earth**
- Capital letter + Preposition + Capital Letter: **Government of India**

Extractor

At the final stage of generating the content for the lexical database, some of the pre-final processes have to take place. It includes garbage removal, duplicates removal strategy, pattern discard technique and manual verification.

The process starts by reading text sentence by sentence. By taking count of rules, the search for each of the rules is made on the sentences. The basis of algorithm depends on the structure of parsed text. In parsed text, the set of words are collected on the basis of predetermined rules with number of left and right braces to be equal. As per the occurrence of rules in the sentences, set of words are collected in the appropriate flat file which is further distributed for manual verifications [2]. The flow chart of Multiword Expression extraction is given in figure 1.

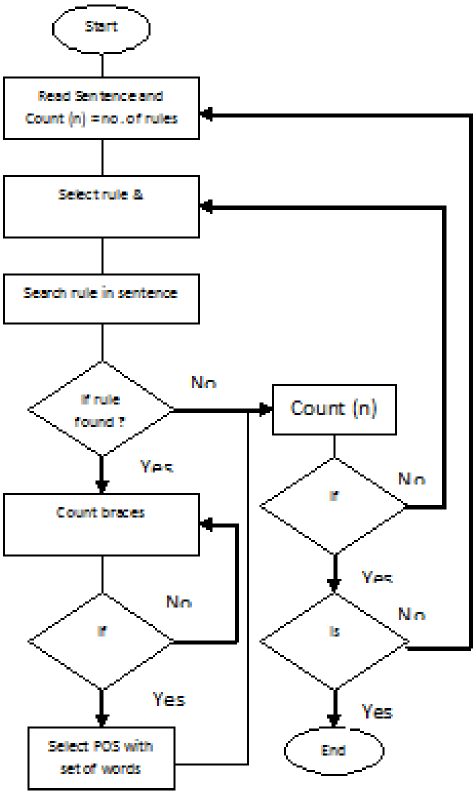

Figure 1: Flowchart for Extractor Process

IMPLEMENTATION AND RESULTS

The methodology has been implemented for Anglabharati machine translation system. The sample corpus used has about 1, 41,130 words out of which 9281 expressions are extracted for a single rule using Extractor and further 2668 Multiword Expression is extracted after analyzing by language expert. The Multiword expression extraction process is given in figure 2.

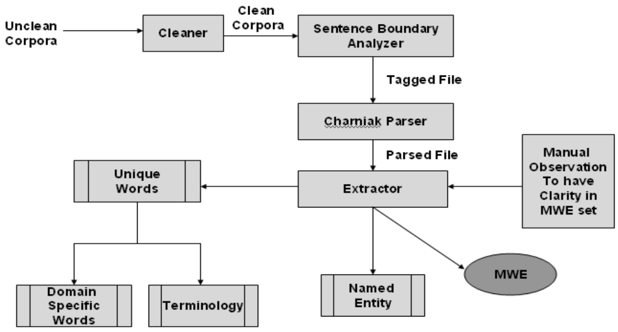

Figure 2: Steps Used in MWE Extraction

CONCLUSIONS

In this paper we tried to show that algorithm designed to retrieve MWEs carried out throughout the noisy text. Methodology is far more diverse and interesting than is normally appreciated. Like the issue of disambiguation, MWEs constitute a key problem that must be resolved in order for linguistically precision. The goal here has been primarily to illustrate the reduction of manual work which has to be done by linguists in the course of extraction of MWE, unique words and Named entities. Since the extraction of MWE is a complex phenomenon however, the methodology converge the process to semi-automated approach. It has proved its simplicity in the extraction process of Multiword expressions and unique words with practical application. Since this methodology is used for extraction of MWE in English language. Further it would be implemented for other Indian regional languages.

ACKNOWLEDGEMENTS

First, a special thank is given to Prof. R.M.K. Sinha, Ex. HOD of Dept. of Computer Science, IIT Kanpur, India for his continuous support and ideas for this research paper. He was always there to listen and to give advice. He is responsible for involvement in the process at the first place. He showed different ways to approach a research problem and the need to be persistent to accomplish any goal. Last, but not least, we thank to Prof. K. K. Goswami senior language expert for giving his precious time in analyzing the results of the methodology and leads to completion of the process.

REFERENCES

1. Spela Vintar, Darja Fiser, “Harvesting Multi-Word Expressions from Parallel Corpora”, Dept. of Translation, Faculty of Arts, University of Ljubljana, Slovenia, [15.03.2008].
2. Akshar Bharati Rajeev Sangal, Deepti Mishra, Sriram V.,Papi Reddy T., “Handling Multi-Word Expressions without Explicit Linguistic Rules in an MT System”, International Institute of Information Technology – Hyderabad, [1–10, 2004].
3. Eneko Agirre and Izaskun Aldezabal and Eli Pociello, “Lixicalization and Multiword Expressions in the BasqueWordNet”, IXA NLP Group University of Basque Country, Donostia Basque Country, [2005].
4. Vineet Kumar Birla, V. N. Shukla “Information Retreival from Noisy text”, Role of Translation in Nation Building, Nationalism and Supra-Nationalism [December 16-19, 2010].