Textual Analysis of Big Data with Object Knowledge Model: A Knowledge Representation framework

Dr. S Padmaja, Mr. Sasidhar Bandu and Prof.S Sameen Fatima

Abstract — This study aims to extract greater semantics from the English language sentences of big data by proposing a frame-based knowledge representation framework called Object Knowledge Model (OKM). This model has only three prime non-terminals, i.e., set of descriptors d, noun n, and verb v, which converts a sentence to attributed sentence for enabling semantic/knowledge/context- based search on the contents of big data. This approach can be applied in various knowledge representation domains like big data analytics, semantic web, cross lingual and multilingual search and also in information extraction of named entities.

Index Terms — Textual analysis of big data, Natural Language Processing, greater semantics, context- based search

1 Introduction

A novel approach to extract the greater semantics from textual big data has been proposed in this paper. It can be performed in any language with OKM, a frame-based knowledge representation framework, for enabling textual analysis of semantic knowledge/content-based search in big data.

A. The limitations of the existing semantics in big data and motivation for OKM.

Currently the semantics from textual big data using web search engines depend on searching RDF/RDFS/OWL documents which contain metadata about the web pages, but do not contain any semantic information on the content of the relevant web pages. In reality, the user of the semantic web will be more interested in making a semantic search on the content of the semantic web documents, instead of a search on only the metadata of those documents (as is being done today), as carried in the RDF/RDFS metadata descriptions of these documents. This is due to the web surfing user having a particular meaningful or semantic query in his/her mind while searching. The query may be expecting answers so as to find more of the content, given a few keywords representing the entities or relationship comprising the content.

RDF, RDFS, and OWL can describe the identification, description and classification of objects which are Web Resources only, i.e., web pages, Authors, Publishers, etc. All possible metadata about the web pages can be captured in RDF and RDFS or OWL. They do not address the internal subject content of the individual web pages.

It's peak time for researchers to work on Sir Tim Berners Lee dream to make semantic web the next version of WWW and implement it into practice [1]. This study aims to enhance the semantic/content/knowledge-based search of not only web pages, but also research papers, equity reports etc. [4], [5]. This work was implemented using Python Natural Language Processing libraries NLTK and Goslate.

The heart of the OKM is its grammar, proposed in this paper, which will parse any given English sentence. It is written in such a way that given any sentence it will generate the parse tree with non-leaf elements as a set of descriptors (d), nouns (n) and verbs (v), and leaf elements as attributed words of the sentence ex: Ram (111).

B. Related Work

Research in this area was conducted by many re-searchers, but most of the prominent work was done by Stanford by developing the pos tagging [2] grammar to parse a sentence and generate dependencies in English which was further enhanced to include other languages ex: Arabic, Chinese, and Spanish etc. [3].

C. Knowledge Search using OKM

OKM enables frame-based Knowledge Representation. The knowledge content of a web page in any given Indian language can be converted to the Object Knowledge Model (OKM) based Knowledge Representation. For each sentence, there will be an equivalent frame of knowledge created. For the whole page, a complete knowledge base will be created. Once the knowledge base is created by using this approach, it can be searched by using the key words denoting the objects or their relationships. Given a keyword or input, it has to be used as an input for searching the knowledge base equivalent of a web page. If the input keyword is indicating an object, then all the objects related to the given input object can be located very easily in the frame of the knowledge base. Each knowledge frame can be searched with the given input literal (for object) for string (i.e., keyword) matching. If a match is found, then

all the relationships of that string or keyword (object) with other objects and the nature of their relationships can be extracted directly and displayed as:
Input object relationship other object(s) object.
Multiple matches can be shown in multiple displays.

2. Implementation

A. Simple English sentence

First, let us see the structure of a simple sentence in English grammar consisting of a Set of descriptors (Adjective, Adverb, Articles) + Noun phrase + Verb phrase.

Ex: Ram went to school by bus. Brother, a letter for you.

B. Word Attributes

The attributes which are assigned are divided into two parts:

- Noun Attributes
- Verb Attributes

Noun Attributes: Given a sentence, all the nouns and pronouns in the sentence are attributed by some features such as, gender g (1: male, 2: female, 3: neutral), number n (1: singular, 2: dual, 3: plural) and case c (1-8) ordered as per Sanskrit Grammar cases.
Cases in order are as follows:

1) Subjective
2) Objective
3) Instrumental
4) Dative (indirect object)
5) Ablative
6) Genitive (possessive)
7) Locative
8) Vocative

Attributed noun word will look like:

Noun word (g n c) ex: Ram (1 1 1).

Verb Attributes: Given a sentence, all the verbs in the sentence are attributed by some features such as, person p (1: first person, 2: second person, 3: third person), number n (1: singular, 2: dual, 3: plural) and tense (-1: past tense, 0: present tense, 1: future tense). Attributed verb word will look like: Verb word (p n t) ex: went (3 1 -1).

Now, given a sentence, let us say: Ram went to school by bus. When parsed through OKM converted into

(((Ram 111) (went 31-1)) ((school 312)) ((bus 313))). Prepositions are neglected in OKM framework.

The parentheses which are assigned to the sentence are nothing but segmenting a sentence using OKM grammar.

C. OKM Grammar

The OKM grammar is given below:

1) k ::= (d o)*
2) d ::= (D)*
3) D ::= set of descriptors (adjectives, adverbs)
4) o ::= on * ov *
5) on ::= nr na
6) na ::= (g n c) *
7) g ::= 1..3 ;; gender
8) n ::= 1..3 ;; number (singular, dual, plural)
9) c ::= 1..8 ;; case
10) nr :: set of noun roots
11) ov ::= vr va
12) vr ::= set of verb roots/procedure/process/web service
(with Oni as input and Oni+1 as output)
13) va ::= (p n t) *
14) p ::= 1..3 ;; person (first, second, third)
15) n ::= 1..3 ;; number (singular, dual, plural)
16) t ::= -1,0,1 ;; tense (past, present, future) and mood An asterisk (*) denotes zero or more repetitions of the preceding element.

3. Results

Parsed sentence was extracted with gender g and number n attributes of noun words and number n and tense t attributes of verb words. But for case c, which is noun attribute, and person p, which is verb attribute, the dependencies (noun and verb) were considered by sentence segmentation using OKM grammar and Stanford universal dependencies. The sentence was parsed and the parse tree generated using the OKM framework. In the tree, the non-leaf element as descriptor d, noun n and verb v, and leaf elements as attributed words of the sentence are compared with the parse tree generated by the grammar of Stanford. The number 999 is the default value given to D in all the figures. Further this work can be extended for any language. The output generated is shown in the figure below:

e.g. 1: Ram went to school by bus.

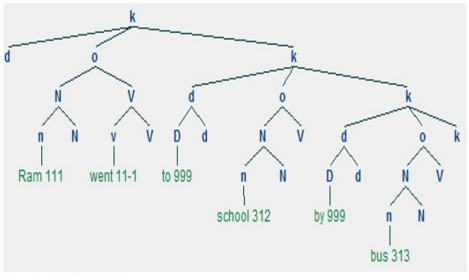

Fig 1: (a) OKM

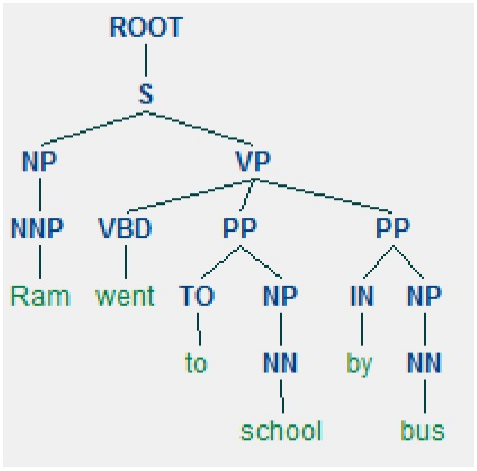

Fig 1(b): Stanford

e.g. 2: Brother, a letter for you.

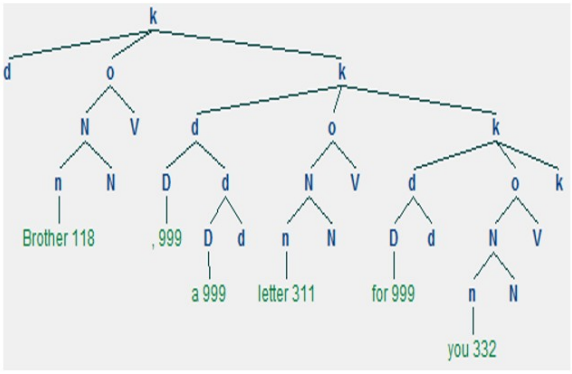

Fig 2(a): OKM

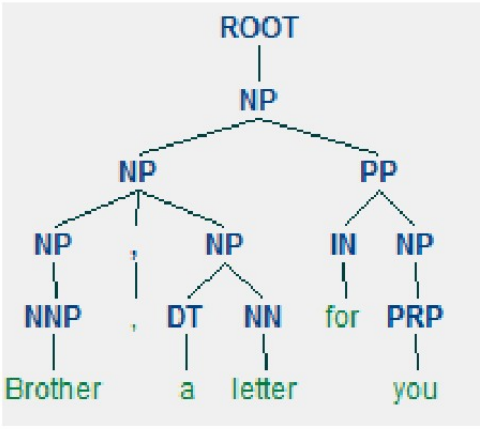

Fig 2(b): Stanford

e.g.3: Good Morning

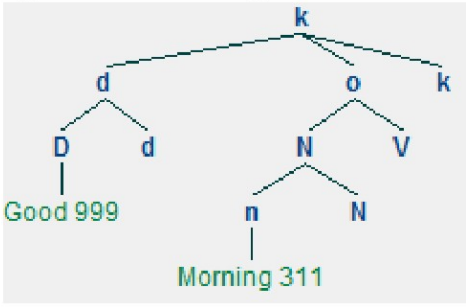

Fig 3(a): OKM

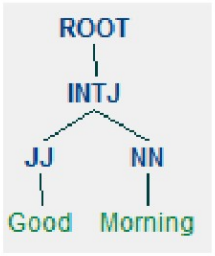

Fig 3(b): Stanford

e.g.4: Here my pen is.

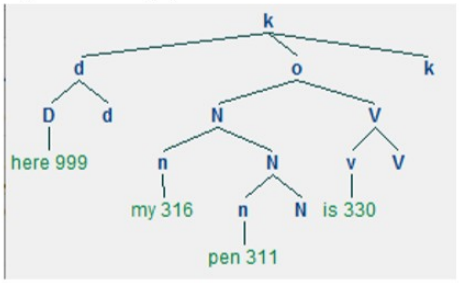

Fig 4(a): OKM

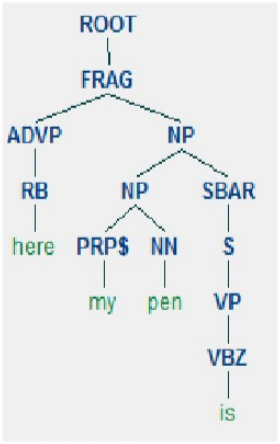

Fig 4(b): Stanford

e.g.5: She is Rohit's friend.

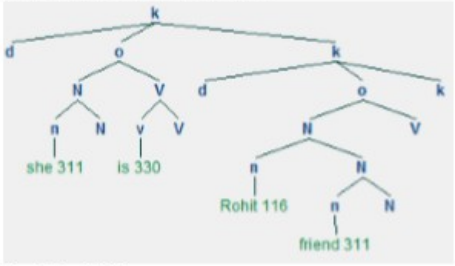

Fig 5(a): OKM

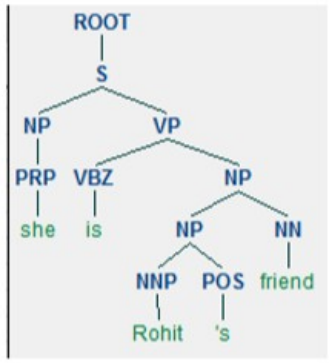

Fig 5(b): Stanford

e.g.6: Jai Mata Di.

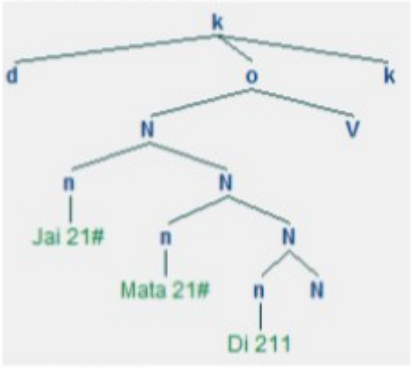

Fig 6(a): OKM

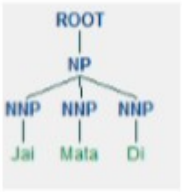

Fig 6(b): Stanford

e.g. 7: The leaf is falling from the tree.

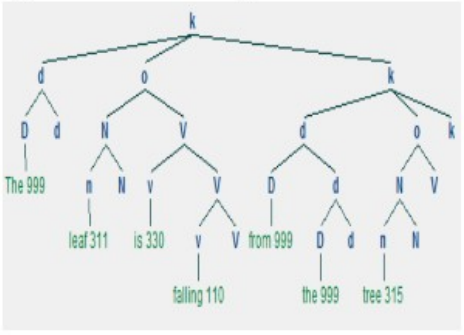

Fig 7(a): OKM

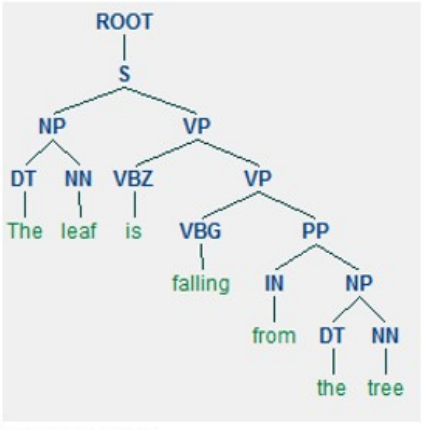

Fig 7(b): Stanford

4. Conclusion

Given any sentence, OKM grammar is able to parse a sentence and segment the sentence into features, but assigning the attributes to the word depends on the accuracy of the Stanford POS Tagging and Universal dependencies. There are certain conclusions which were drawn from the results.

1) The OKM is easy to understand, as can be seen in the Fig 3(b), there are some nonterminals used which were not used in any of the Figures (example sentences above) of Stanford and there are many nonterminals to remember in Stanford grammar when compared to OKM grammar. This is due to OKM having only three prime non-terminals, i.e., set of descriptors d, noun n, and verb v, whereas Stanford has many nonterminals which are hard to remember.
2) There are certain cases where the sentence segmentation is same in both OKM and Stanford as seen in examples in Fig 4 and Fig 6.

   The sentences, here my pen is and Jai Mata Di, are segmented as (here) (my pen is) and (Jai Mata Di) respectively. Whereas in OKM each and every noun and verb are attributed with the semantic knowledge which is not present in Stanford.
3) OKM is more powerful than Stanford in sentence segmenting which can be seen in some of the cases, for example, in Fig 2, Fig 5 and Fig 7.

- For the sentence in Fig 2, the sentence is segmented as: Stanford: (Brother, a letter) (for you) OKM: (Brother) (, a letter for you)
- For the sentence, She is Rohits friend in Fig 5, the sentence is segmented as: Stanford: (She) (is Rohits friend). OKM: (She is) (Rohits friend).
- For the sentence, The leaf is falling from the tree in Fig 7, the sentence is segmented as: Stanford: (The leaf) (is falling from the tree) OKM: (The) (leaf is falling) (from the tree).

  In the above three examples one can clearly see the difference in sentence segmenting where OKM is more intelligent when com- pared to Stanford grammar.

5. Future Work And Applications

- OKM can be applied in Semantic Web as a tool for better knowledge representation and better semantic search of the Web.
- It can be applied in Search Engine technology for cross lingual, multilingual search.
- It can be applied in text analysis and information extraction for machine learning classification of the Named Entities in a given text.
- It can be applied in language-based learning, by deploying OKM as a Knowledge Representation methodology with learning as input and modification/extension of the knowledge represented.

REFERENCES

[1] T Berners-Lee, J Hendler, and O Lassila. Scientific american: Feature article the semantic web, 2001.
[2] Marie-Catherine De Marneffe and Christopher D Manning. Stanford typed dependencies manual. Technical report, Technical report, Stanford University, 2008.
[3] David D Lewis. Feature selection and feature extraction for text categorization. In Proceedings of the workshop on Speech and Natural Language, pages 212–217. Association for Computational Linguistics, 1992.
[4] Kirti D Pakhale and SS Pawar. Focused retrieval of e-books using text learning and semantic search. International Journal of Innovative Research and Development, 3(7), 2014.
[5] Jane Zhang. Ontology and the semantic web. 2007

Dr. Padmaja S received her PhD in computer science from Osmania University. She is an Associate Professor at KMIT, Hyderabad. She regularly contributes to scholarly journals, conferences and is also reviewer. She is a resource person for various courses offered at research and academic institutions of repute. Natural Language Processing, Machine Learning, Big data Analytics are few of her areas of research interest. She can be reached at bandupadmaja@gmail.com for research collaboration.

Mr. Sasidhar B is an EFL Lecturer and heads Professional Development Unit at Prince Sattam Bin Abdulaziz University, Saudi Arabia. He is a resource person for various institutions and corporate houses in India on ELT, Teacher Training, CALL and Educational Technology. He can be reached at sasibandu@yahoo.com.

Prof Syeda Sameen Fatima has over 33 years of experience in teaching, research and administration in India, USA and UAE. She took over as Principal in July 2016, and holds the distinction of being the first lady Principal, in the history of the College of Engineering, Osmania University. Currently she is a Professor at the Department of Computer Science and also the Director, Centre for Women's Studies at Osmania University. She has published several papers in national and international journals and conferences. Her areas of interest include Machine Learning, Text Mining and Information Retrieval Systems. She is a life member of the Computer Society of India. She received the "Best Teacher Award" by The Government of Telangana, India in the year 2017. She can be reached at sameenf@gmail.com .