Automatic Generation of Synsets for Wordnet of Hindi Language

Priyank Pandey, Department of Computer Science, Graphic Era University, Dehradun, India
Manju Khari, Computer Science and Engineering Department, AIACTR, Delhi, India
Raghavendra Kumar, Computer Science and Engineering Department, LNCT College, Indore, India
Dac-Nhuong Le, Faculty of Information Technology, Haiphong University, Haiphong, Vietnam

ABSTRACT

India is a land of 122 languages and numerous dialects. Lack of competent lexical resources for Indian languages is a ubiquitous fact, which negatively affects the development of tools for NLP of Indian languages. Recent advancements like the Indo WordNet project has significantly contributed to dealing with the scarcity of lexicons, but the progress and coverage is a matter of dispute. The bottlenecks, cost, time, and skilled lexicographers further slackens the progress. In this article, the authors propose a technique to automate the generation of lexical entries using a machine learning approach which visibly expedites the process of lexicon generation like WordNet. The reluctance to adopt an automated approach is majorly credited to a lack of accuracy, the inability to capture a regional touch of a language, incorrect back-translation, etc. To overcome this issue, the author will use Wikipedia to validate the synsets.

KEYWORDS
Indian Languages, Lexical Database, Natural Language Processing, Support Vectors, WordNet

1. INTRODUCTION

Natural Language Processing (NLP) plays an important role in performing interaction between humans \textit{(or natural)} and computers. It is a tedious task for the computers to understand the meaning and the context of the word used by the human. Because, there could be multiple meanings and translations for one word used by the humans. Second, the scenario can exist where there are two languages involved. There can be multiple translations of one word, and which meaning is considered as valid meaning is the primitive task for a computer. In these scenarios, NLP comes into play. The main motive of the paper is to create a new WordNet that has not yet been created and perform successful mapping of senses between two words of different languages using existing wordnet. Total 7000 living languages are present, out of them only 50 version of wordnets had been produced. In addition, many wordnet of different languages are still in early stage (De Melo & Weikum, 2012).

Lexical database (i.e. WordNet) used in the natural language processing tasks, is a repository that stores the synonym of a particular word (called as “synsets”) with the short meanings and

DOI: 10.4018/IJNCR.2018040103

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
usage examples and records the number of relations among these synonyms. WordNet includes all crucial information about the word, and to check whether it is noun, adjective, or adverb but ignores prepositions, determiners, and any other function. It shows all the possible senses of a specific word. First WordNet was created in the Cognitive Science Laboratory of Princeton University in 1985.

The famous lexical database used for English retrieves crucial information regarding the relationship of words and their senses, which is Princeton WordNet. This WordNet stores the word’s sense and diagnose synsets. (Synonyms set consist of senses that are similar in denotation as its semantic groups. The arrangement of term (or word) and synsets are in the form of network of nodes that is being held by numerous semantic relations which are lexical units. One relation, hyponymy is said to be as the one that “holds between a more specific, or subordinate, lexeme and a more general, or super ordinate, lexeme”. For E.g., it can be viewed as pairs like ‘red’: ‘colour’, ‘car’: ‘vehicle’. However, Hypernymy is an inverse relation to the hyponymy relation. The advantage of Princeton English wordnet is not confined to this wordnet only; however, it is also fruitful for natural language processing tasks in case of deploying new wordnets.

The remainder of the paper is organized as follows. Section 2 discusses the conventional way of automatic generation of wordnets in the form of literature survey covering most of the papers. Section 3 highlights the wordnet generation methodology, which includes basic terminologies, and the implementation. Section 4 focuses on the comparison of the accuracy with validation scheme. Section 5 highlights result and their analysis of accuracy, precision and recall along with the comparison of author’s work with the previous work. Section 7 concludes the entire paper along with the future scope.

2. CONVENTIONAL WAY OF AUTOMATIC WORDNET GENERATION

Before heading towards our approach in constructing wordnets and translating the senses of each term for multiple languages, this section will introduce the conventional approaches or models in creating wordnets that was used by early researchers.

First approach is merge model, in which conversion of thesaurus into a wordnet is mentioned. Then, this wordnet is linked semi-automatically to another wordnet. The disadvantage is this model is only confined to small number of languages. Unless there are few available word-net for every language, it is not applicable to a wide span of languages.

Second approach is the expand model, is better than the previous one as it consumes less pre-existing resources. This uses a strategy as follows:

1. Consider an available wordnet for some language $L_i$; generally English Princeton WordNet;
2. In every sense s mentioned in the wordnet, then, using translation dictionary, translation of the elements linked through s from $L_i$ to destination language $L_n$ is taken place;
3. In addition, preserve all related relations concerned with semantics among senses from t already available wordnet so that it can arrive for a new wordnet i.e. $L_n$.

Main issue is about relevancy in between translations and their senses. The translation done for a term is not always applicable to all of the senses for that translated word. For acknowledging this type of issue, various researchers highlighted their approaches. Author represents those ideologies and methodologies in a tabular format in Table 1. Table 2 represented the comparative analysis with the advantages and limitations.

3. WORDNET GENERATION METHODOLOGY

The automatic generation (De Melo & Weikum, 2012) to build wordnet requires a desired approach. The approach is like this: Let the language of new word net, which has to be built is denoted by $L_n$,
MP Modelling of Glucose-Insulin Interactions in the Intravenous Glucose Tolerance Test
www.igi-global.com/article/modelling-glucose-insulin-interactions-intravenous/58063?camid=4v1a

Box-Counting Dimension of Fractal Urban Form: Stability Issues and Measurement Design
Shiguo Jiang and Desheng Liu (2012). *International Journal of Artificial Life Research* (pp. 41-63).
www.igi-global.com/article/box-counting-dimension-of-fractal-urban-form/81213?camid=4v1a