HINDI COMPOUND NOUN SEMANTICS USING MACHINE LEARNING AND GENERATIVE LEXICON

A thesis submitted in partial fulfillment of the

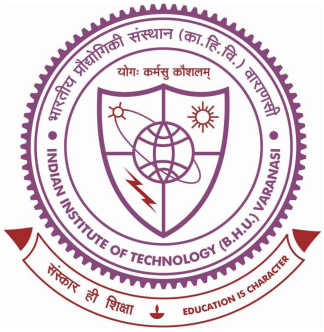
requirements for the award of the degree of

DOCTOR OF PHILOSOPHY

by

VANDANA DWIVEDI

ROLL NO.: 16191501



DEPARTMENT OF HUMANISTIC STUDIES

INDIAN INSTITUTE OF TECHNOLOGY

(BANARAS HINDU UNIVERSITY),

VARANASI-221 005

DECEMBER 2022

Chapter 6

Conclusion and Future Work

This chapter will summarize the main contributions of the thesis in different chapters. It will point out the limitations of the present work and future research directions that can be taken from this work.

6.1 Introduction

To our knowledge, this work is the first attempt to identify and classify semantic relations between the constituents of a compound noun of Hindi using a statistical approach. So far, no work exists in the automatic interpretation of compound nouns in Hindi using only the Hindi dataset. This is also a first attempt to build a Generative Lexicon based Knowledge Representation of the most frequent headwords of the compounds found in the Ayurveda domain. The Generative Lexicon has not been explored as a lexical knowledge base to support the computational works for any Indian language. The work makes a small beginning towards that kind of work in the future.

6.2 Summarization of the thesis

This work is divided into six chapters.

Chapter one provides a basic introduction of the theoretical concept of the compound noun. Compound nouns have played an important role in the research in three different fields: firstly in theoretical linguistics especially in morphology and semantics, secondly in Cognitive Psychology and Psycholinguistics for issues related to its representation and processing in human minds and finally in Natural Language Processing as an identification and classification problem. The chapter briefly mentions these to establish the relevance of the topic in current linguistic and computational linguistic scenarios. After that, it discusses the main theories on the classification of compound nouns in Sanskrit and Western grammatical traditions. In the next sections, the main objectives and research issues discussed in the thesis as well as the methodology and theoretical frameworks used for achieving that are explained. At the end, a brief outline of the other chapters is provided.

Chapter two opens up with setting up the compound noun interpretation problem in the broad research area of Multi Word Expressions (MWEs) in NLP. It discusses the major theoretical works that influenced the researchers in NLP for handling this problem as well as provides a brief review of literature on the computational analysis of English compound nouns. Apart from English, the major works in other non-Indian as well as Indian languages have also been reviewed in this chapter.

Chapter three discusses the first contribution of the thesis in making a dataset of compound nouns from the corpus of Hindi health domain available at the TDIL website. The relation set of 20 semantic relations which we developed for annotation of the noun compounds of Hindi is also discussed in this chapter. A domain independent data set of compound nouns has also been made from the MWE list available in the website of CFLIT, IIT Bombay. The reliability of the annotated data of compounds by four respondents has been checked with an inter-annotator

agreement score. Thus, this chapter prepares a gold standard dataset ready to be used for machine learning in the next chapter.

The fourth chapter describes three different experiments done on the data of compound nouns to check how far the machine learning algorithm is able to predict correct relations between the constituents of the compounds of Hindi. We experimented with three different features viz; individual noun features, word2vec embedding and BERT embedding and three classification algorithms; Support Vector Machines, Random Forest and Bert classifier. The results of the experiments are also presented in the chapter. We conclude from the experiments that for the classification task of semantic relations SVM classifier with RBF kernel outperforms all other classifiers for binary class classification. We also found that word embedding features perform better for understanding the nature of the constituents and the relationship between them. BERT based embedding cannot perform better than word2vec embedding due to the difference between the datasets for pretraining and semantic relations classification tasks. From the results of the analysis, we also concluded that for the semantic analysis tasks a very rich semantically encoded linguistics resource is required to use with any probabilistic model.

Chapter five discusses the theoretical framework of Generative Lexicon that has been used to build a lexical knowledge base for the compound nouns of Ayurveda domain. A third data set of compound nouns has been made from the Ayurveda domain to increase the number of data in the general health domain as well to build a lexical resource for analyzing Ayurveda compounds. The most frequent noun found in this domain in the compound word was rasaayana 'rejuvenation' that occurs 35 times in the corpus built. Therefore, we made the qualia representation of this head word and showed how senses of different compounds made with this head can be analyzed using Generative Lexicon representation. For total 25 frequent head words found in this corpus, qualia representation is made and added at the end in the appendix.

Chapter six concludes with a summary, limitation and future direction of the work.

6.3 Limitations of the work

The thesis is an attempt towards the automatic interpretation of Hindi compound nouns. This work is the first of our knowledge to study Hindi compound semantics from an NLP perspective and automatize the Interpretation of Hindi Compound nouns using machine learning and knowledge base. Despite several contributions, there are some limitations of the proposed work. The thesis presented a machine learning approach for automatic interpretation of Hindi compound nouns. The machine learning model used in the thesis achieved significant results for only four classes in a multi-class classification task. For remaining classes we need more data and more features.

The dataset taken for the experiment has compound nouns with only two constituents. Compounds having three or more constituents were excluded from the study in the beginning.

For experiment with BERT classifier we have used a pre-trained language model trained on a limited Hindi corpus. The result could have been better if we would have developed our own training model. In this study, we did not consider the pragmatic factor influencing the interpretation of the compound noun. We also have not studied the compound nouns used as metaphors.

The Knowledge Base developed for compound noun interpretation consists of a small number of nouns. Corpus collected is also not a large corpus. The copus does not contain any instance of a verbal noun as the head. To make a generalization about compound noun semantics using qualia relation and structure we need to analyze every instance of compound nouns.

One important limitation of the work is that compound noun meaning can be ambiguous as a whole. Thus, the semantic relations can be ambiguous too. One compound noun can fall in one category in one context or another in any other context. Ambiguity is due to two main reasons: one is due to the nature of the constituent words being ambiguous individually and other the word meaning can be the same but the relation can be ambiguous due to context. Contextual factors also influence the compound noun semantics. This problem in NLP is called multilabel class classification. Our work has not taken account of this problem.

6.4 Future direction of research

In order to get a more accurate result of automatic interpretation, the future work will focus on the development of a balanced dataset inclusive of all the semantic relations for further machine learning experiments with other algorithms and features. We also want to combine different features together (WordNet Features and Embedding Features) and then use different machine learning algorithms to check the model performance. The efficacy of the automatic interpretation of the work can also be checked by using different neural networks for the Hindi compound noun interpretation task.

The thesis used compound noun interpretation problems as a classification task in machine learning. We would like to work on developing a semantic model which can predict the underlying semantic relationship between the constituents automatically and provide the most appropriate meaning to the compound nouns.

The knowledge base developed for the Ayurveda domain is based on a limited corpus and is used only for compound noun interpretation. In future, we will add more nouns to the database which can be used effectively as a lexical knowledge base for other semantic interpretation tasks in the Ayurveda domain. We also plan to

develop a computational model using embedding trained on the generative lexicon based ontology and a pretrained model trained on Ayurveda domain for domain specific compound noun interpretation.

In the future, we would like to extend this work with the noun compounds having more than two noun constituents and other compound constructions such as noun-adjective, noun-verb, and reduplicative compounds.

The future study may also include the metaphorical and idiomatic compound nouns as well as the pragmatic factors influencing the compound noun interpretation.