Certificate

It is certified that the work contained in the thesis titled **Hindi Compound Noun Semantics using Machine Learning and Generative Lexicon** by **Vandana Dwivedi** (Roll No.: 16191501) has been carried out under my/our supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements of Comprehensive Examination, Candidacy, and SOTA for the award of the degree of **Doctor of Philosophy** to the Indian Institute of Technology (Banaras Hindu University) Varanasi.

Sanjukta 26/12/2022

Signature of Supervisor

Dr. Sanjukta Ghosh

Associate Professor

Department of Humanistic Studies

Indian Institute of Technology

(BHU)

Varanasi - 221005, India

सहयुक्त आचार्य/Associate Professor
मानवतावादी अध्ययन विभाग/Department of Humanistic Studies
भारतीय प्रौद्योगिकी संस्थान/Indian Institute of Technology
(काशी हिन्दू विश्वविद्यालय)/(Banaras Hindu University)
वाराणसी-२२१००५ (उ०प्र०)/Varanasi-221005 (U.P.)

Declaration

I, **Vandana Dwivedi**, certify that the work embodied in this thesis is my bonafide work and carried out by me under the supervision of **Dr. Sanjukta Ghosh (Department of Humanistic Studies)** from **January 2017** to **December 2022**, at the **Department of Humanistic**, Indian Institute of Technology (BHU) Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, *etc.*, reported in journals, books, magazines, reports dissertations, theses, *etc.*, or available at websites and have not included them in this thesis and have not cited as my work.

Vandana Dwivedi

Date: 26/12/2022
Place: Varanasi

Signature of Student
(Vandana Dwivedi)

Certificate by the Supervisor

It is certified that the above statement made by the student is correct to the best of my/our knowledge.

Sanjukta 26/12/2022

Signature of Supervisor

सहयुक्त आचार्य/Associate Professor
मानवतावादी अध्ययन विभाग/Department of Humanistic Studies
भारतीय प्रौद्योगिकी संस्थान/Indian Institute of Technology
(काशी हिन्दू विश्वविद्यालय)/(Banaras Hindu University)
वाराणसी-२२१००५ (उ०प्र०)/Varanasi-221005 (U.P.)

28/12/2022

Signature of Head of Department
(Dr. Ajit Kumar Mishra)

Head
Department of Humanistic Studies
Indian Institute of Technology
(Banaras Hindu University)
VARANASI-221005 (U.P.)

Copyright Transfer Certificate

Title of the Thesis: **Hindi Compound Noun Semantics using Machine Learning and Generative Lexicon**
Name of Student: **Vandana Dwivedi**

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University) Varanasi, all rights under copyright that may exist in and for the above thesis submitted for the award of the Doctor of Philosophy.

Vandana Dwivedi

Date: 26/12/2022
Place: Varanasi

Signature of Student
(Vandana Dwivedi)

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

Acknowledgments

All these PhD years have been a never-boring mixture of learning and relearning, experiments that worked in the end, experiments that didn't, stress, relief, more stress, some croquet and an awful lot of running. And at the end a thesis was produced. I want to take this opportunity to express my deep sense of gratitude to all who helped me directly or indirectly during this thesis work.A good number of people have contributed to getting me there, by offering advice, suggestion and feedback.

First I would like to thank My supervisor Dr. Sanjukta Ghosh, who has been an unfailing source of support, from the beginning to the end of my thesis for her continuous and immense support, guidance and encouragement. She has influenced many aspects of my writing and presentation for the better. I am grateful to my RPEC examiners, Dr. Anil Thakur and Dr. Sukomal Pal for their constructive comments, suggestions and good-humoured criticism. Their inspiration is the cause behind the successful completion of this Thesis work.

I sincerely thank Prof. R.K Mishra, Department of Electrical engineering, and Ex. Head, Department of Humanistic Studies, and Dr. A. K. Mishra , Head,Department of Humanistic Studies for providing continuous support, encouragement, and advice. I sincerely thank all the Professors, DPGC Convenors, Deans, and office staff of the Department of Humanistic Studies, India Institute of Technology (BHU) Varanasi, India. I express my gratitude to the Director, Registrars, Deans, Heads, and Student Alumni of the Indian Institute of Technology (BHU) Varanasi for this life-changing opportunity.

I would like to thanks all my friends, colleagues, friends and seniors, who constantly supported and helped me in my PhD journey.

Words cannot express my gratitude to my family for their constant support. I would like to express my deepest gratitude to my parents,grandmother and siblings for their patience and making this journey a rewarding and exciting one.

This thesis is dedicated to my late grandfather Acharya Sudarshan Dwivedi who introduced me to this field and for giving me a good head start.

I would like to end by paying my heartfelt thanks and prayers to the Almighty for his unbound love and grace.

- Vandana Dwivedi

Preface

This thesis investigates the machine learning based approach for analyzing the Hindi compound noun semantics. Noun compounds are interesting constructions from the perspective of theoretical linguistics and computational linguistics for the complex semantic relations between their constituents. The meaning of Compound Noun is composed of the meanings of the individual constituents and the way they are semantically related. Noun compound interpretation is the task of detecting this underlying semantic relation. For instance, a kitchen knife is a knife which is used in the kitchen (used in or purpose relation) whereas a steel knife is a knife made of steel (made of or constituent relation). This work explored different machine learning techniques as well as a knowledge base method which can predict the semantic relations between the constituents of compound nouns.

We developed a semantic relation set to annotate the semantic relations between the constituents of compound nouns,describing in detail the motivation for the relation set and the development process. Using the semantic relation set an annotated dataset is created to use as an experiment dataset with different machine learning algorithms. The inter-annotator agreements result indicate the reliability of the semantic relation and dataset for Hindi Compound Nouns analysis.

We treated compound noun interpretation problems as a classification problem for ML tasks and experimented with SVM and Random forest. Hindi WordNet is used as a linguistic resource and has got a significant output. We also experimented with Embeddings, since embedding captures more semantic features in language models and gives better results.

In this work, we also attempted to create a knowledge base in the domain of Ayurveda using the Generative Lexicon framework of lexical knowledge representation. Ayurveda text has a very limited corpus. Therefore, using a probabilistic model does not work. We developed a lexical knowledge representation database of some selected frequent nouns of Ayurveda corpus using the Generative Lexicon framework. The GL model is able to represent the natural polysemy of the word in the lexicon which is used for disambiguation tasks. Language models with rich knowledge encoded resources can be beneficial for developing linguistically precise probabilistic models.

Contents

List of Figures

List of Tables

Abbreviations

AI	**A**rtificial **I**ntelligence
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
RF	**R**andom **F**orest
NLP	**N**atural **L**anguage **P**rocessing
SVM	**S**upport **V**ector **M**achine
GL	**G**enerative **L**exicon
ML	**M**achine **L**earning
TP	**T**rue **P**ositive
TN	**T**rue **N**egative
FP	**F**alse **P**ositive
FN	**F**alse **N**egative
CN	**C**ompound **N**oun
MWE	**M**ulti **W**ord **E**xpression

Symbols

w	word
N	Noun
k	Kohen Kappa

ITRANS Transliteration Scheme

Vowels

ITRANS	DEVANAGARI
a	अ
aa	आ
i	इ
ii	ई
u	उ
uu	ऊ
RRi	ऋ
RRI	ॠ
e	ए
ai	ऐ
o	ओ
au	औ
aM	अं
aH	अः

Consonants

ITRANS	DEVANAGARI
k	क
kh	ख
g	ग
gh	घ
~N	ङ
ch	च
Ch	छ
j	ज
jh	झ
~n	ञ
T	ट
Th	ठ
D	ड
Dh	ढ
N	ण
t	त
th	थ
d	द
dh	ध
n	न
p	प
ph	फ
b	ब
bh	भ
m	म
y	य
r	र
l	ल
v	व
sh	श
Sh	ष
s	स
h	ह
x	क्ष
GY	ज्ञ
shr	श्र

FIGURE 1: ITRANS SCHEME