Chapter 2

Processing of Compound Nouns in NLP

2.1 Introduction

Conceptual combination by which a novel compound is formed is a complex cognitive process and has been a major research topic in Cognitive Psychology and Cognitive Linguistics in particular and Cognitive Science in general. The compound nouns are very productive in many languages including English and generally speakers do not have any problem in the interpretation of meaning of these nouns when they are used in the discourse. What processes are used by human beings to understand the relations between the constituent words of the compound nouns had remained a major area of research in 1990s and early of the 21st century. The diversity of this process of concept combination is reflected in different theories F. J. Costello (2002); F. Costello et al. (2004); F. J. Costello et al. (n.d.); F. Costello & Keane (2000); F. J. Costello & Keane (2001a); F. J. Costello (1996); F. Costello & Keane (1997); F. J. Costello & Keane (2000, 2001b); Gagné (2000, 2002); Gagné (2000, 2001); Wisniewski & Love (1998); Wisniewski

(1997). In Cognitive Linguistics also, conceptual blending theory attempts to analyse how two concepts from two different mental spaces are blended to make a new compound in a new blended mental space Fauconnier & Turner (2003).

In NLP and applied Artificial Intelligence research some of the insights of these theories have been taken and implemented to solve the problem of compound noun interpretation. This chapter will mainly review the works done in NLP for the automatic interpretation of compound nouns. Most of these works are actually in English with some exceptions in some other languages. The work in Indian languages on automatic interpretation of compound nouns is scarce especially using a data set of an Indian language. This motivates the present research as well as establishes its relevance in the context of Indian language NLP.

2.2 Compound Nouns as a Multi-Word Expression

Multiword Expression is a bottle-neck in all NLP tasks. Multiword expressions (MWEs) are expressions which are made up of at least two words, and their component words cannot predict their meaning. According to Sag et al. (2002), Multiword expressions (MWEs) are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and statistical idiomaticity. Moreover, they act as a single unit at some level of linguistic analysis. A Noun compound is a primary type of Nominal MWE. In a Noun compound, there are two or more than two noun sequences combining to form a noun, such as *traffic light, golf club, girlfriend* etc. In this type of series, there is a relation like modifier and modified. Noun compounds are ambiguous or do not have enough details about semantics which makes it an essential problem in NLP research. Our study focuses on identifying the syntactic and semantic relations between the constituents of Hindi noun compounds.

J.-H. Kim et al. (2001) present the problems and currently available techniques for MWE analysis in NLP. They reported there are two main problems in general for treating MWE. The first is overgeneralization: for example, a generation system that is uninformed about both the patterns of compounding and the particular collocational frequency of the relevant dialect would correctly generate *telephone booth (American) or telephone box (British/ Australian)*, but might also generate such perfectly compositional, but unacceptable examples as telephone cabinet, telephone closet, etc. A second problem is referred to as the idiomaticity problem: how to predict, for example, that an expression like kick the bucket, which appears to conform to the grammar of English VPs, has a meaning unrelated to the meanings of kick, the, and bucket. Syntactically-idiomatic MWEs can also lead to parsing problems. MWES are classified as the institutionalised phrases and lexicalised phrases. Compound nouns come under lexicalised phrases (semi-fixed expressions).

2.3 Defining Compound Nouns in NLP

Compound nouns in NLP do not follow all the theoretical criterion of compound noun definition. Characterisation of compound nouns in theoretical linguistics follows several criteria and applying all those rules in NLP will imply an intensive annotation framework and the processing of the compound will be more complex. Spärck Jones (1983) argues that viewing compound interpretation from a computational perspective adds new problems that are otherwise not encountered in linguistic studies of compounds. For example, identifying whether or not a sequence of words is a sequence of nouns (i.e. PoS tagging) and consequently determining if this sequence of nouns is a valid compound itself.

According to Levi (1978), Complex nominals are of three types : nominal compounds (e.g. *pineapple cake*), nominalizations (e.g. *movie producer*; the noun producer is derived from the verb to produce) and noun phrases with non-predicating

adjectives (e.g. *electrical engineer*; the adjective is considered non-predicating because the construction cannot be paraphrased as *an engineer who is electrical). Most scholars in NLP have taken this categorisation into consideration and defined the compound nouns as a sequence of two or more than two nouns which are mostly in modifier-head relations. We have also used this definition in our study.

2.4 On the Compound Noun Interpretation

In this section, we will discuss the major approaches and theories used for processing of compound nouns in Cognitive Psychology and NLP. The present study is focused on the automatic identification of the senses of the compound nouns of Hindi and falls under NLP. However, we are briefly discussing the approaches used for explaining the computation of the compound words by human minds as discussed in Cognitive Psychology as we believe those studies are the foundations of any study on compound nouns within NLP. Both the works broadly fall under the umbrella domain of Cognitive Science. Any development in explaining the working of the human mind leads to a successive development in the domain of applied Artificial Intelligence.

2.4.1 Major theories in Cognitive Psychology

The interpretation of the novel compounds had been a favourite research topic of the cognitive scientists in the last decade of the 20th century and first decade of the 21st century. From Cognitive Psychology, Wisniewski (1996, 1997); Wisniewski & Love (1998); Wisniewski & Middleton (2002); Gagné & Shoben (1997); F. J. Costello (1996); F. Costello & Keane (2000); F. J. Costello & Keane (2001a); Lynott & Keane (2002) are few important names to be mentioned. As a result of these works different theories evolved to understand conceptual

combination. The most important of them are the dual-process theory by Storms & Wisniewski (2005) and the constraint theory by F. J. Costello et al. (n.d.).

The Dual-process theory explains the comprehension of a noun-noun compound, which is in a modifier-modified relation, in terms of three types of interpretation: relational interpretation, property interpretation and conjunctive interpretation. In Dual-process theory there are three different approaches to interpretation. One is based on relational combination between the constituents. For instance, a street dog is interpreted as a dog which lives on the street. The other is property based when some common properties of the two constituents are taken for the interpretation of the meaning of the new compound. For instance, the cactus fish is a fish which has some prickly features like the cactus. In a third kind of conjunctive interpretation, the interpreter produces an interpretation combining both the concepts. The typical example is *'pet bird'* which is a pet as well as a bird. Surveys on how people interpret a new compound have shown that the first two types of interpretations are more frequently used. Wisniewski (1996), Murphy & Wisniewski (2006) argued for primacy and preference for relation interpretation. However, in another paper Wisniewski & Love (1998) showed as a result of some experiments that even property relation is as important as the relation interpretation and any model of conceptual combination must include both. The relational interpretations are produced by a scenario construction, whereas both the property and the conjunctive interpretations are produced by 'comparison, alignment and property re-instantiation' F. J. Costello & Keane (2001a).

Constraint theory describes conceptual combination as a process where people select from some possible meanings using three constraints of diagnosticity, plausibility and informativeness. Diagnostic theory says that when an interpretation is selected by human minds, some diagnostic property of the modifier constituent is selected and used for the modified head when a compound is formed. The diagnostic property is the prototypical property associated with a word like *pricklyness* with cactus and

it is selected for modifying the head in the compound *cactus fish*. These processing theories guide us to understand how the compound noun interpretation happens and accordingly identify the correct relations between the constituents.

2.4.2 Representation of Compound Nouns in Linguistics

In the 20th century the compound noun semantic works of Hatcher (1960); Jackendoff (1988); Noreen (1904); Jespersen (1961) focus on the description and semantic relation inventory based approach on the identifying variety of semantic relations observed in the compound nouns. From then the compound noun semantic problem was treated as a semantic relation identification and extraction problem. The rise of generative grammar in the 1950s and 1960s led to a greater concern with matters of representation. In a transformational (or at least multistrata) framework it made sense to analyse compound nouns as derived from a fuller structure at a deeper representational level. Lees (1970) describes an underlying representation where the constituents of a compound fill the thematic role slots of one of a small number of generalised verbs. For example, air rifle, motor car and water wheel are all derived from triples V-Object-Instrument, where V is a generalised verb with the meaning shared by the English verbs energise, drive, power, actuate, propel, impel,. . ., the Object role is assigned to the modifier noun and the Instrument role is assigned to the head. Other works include Li (1971); Warren (1978, 2003); Levi (1978).

Compounding is a highly productive word formation process and frequently used in natural languages to form new words. 2-4% of the tokens in various corpora are part of noun compounds Baldwin & Tanaka (2004). It is productive in the sense that it has an open -ended ability to form new compounds and frequent that we can see it a sheer amount of time in natural languages. In written and technical domains the frequency of the compound nouns is quite more compared to the spoken language. Levi characterises compound nouns as "naming devices that pick out the relevant

categories of the speaker's experience" . The "naming device" perspective explains the productivity of compounds and also shows how they become institutionalised. The term institutionalised is used here in the sense defined by Bauer (2001) to mean 'compounds that have been accepted by the community as a conventionalized name for a particular concept'. In every technical domain naming of any concept is usually done by forming a compound like colon cancer treatment. We have used Ayurveda domain compounds in this thesis and many of the concepts in Ayurveda are represented as compound nouns . The ability to represent a lot more concepts with a small number of words is the main serving of compound and this makes it a highly crucial part in the word formation process. Compounds are used to compress the syntactic and semantic information in a compact form Levi (1978). For example as Li (1971) gave, the noun–noun compound cradle song roughly packs the sentential structure "a song to lull a child in the cradle to sleep".and also the compounds are used as a naming device to refer to entities that have no name, and in this context Downing (1977)(p. 824) aptly states that "compounding thus serves as a backdoor into the lexicon".

Therefore, the syntax and semantics of complex nominals is studied extensively in theoretical linguistics. Complex nominals as a linguistic construction—have received a fair amount of attention and scrutiny in the functional and generative schools of linguistics Li (1971); Aerts & Gabora (2005); Downing (1977); Warren (2003); Pustejovsky & Bouillon (1995).

2.5 English Compound Noun Interpretation: A brief review

Major works in NLP on English compound noun interpretation are of Lauer (1995); Lapata & Keller (2004); Girju et al. (2004); Vanderwende (1994); Girju et al. (2005, 2007); Girju (2006); Tratz & Hovy (2010); Séaghdha & Copestake

(2007, 2008, 2009); Séaghdha (2007) . As Lauer (1995) said "The likelihood of an NLP system encountering a noun compound is high and the inability to effectively interpret noun compounds will degrade the performance of many practical NLP applications' '.

In this section we have selected the major studies on compound noun interpretation tasks and presented the review of those studies. Most of the studies are motivated by the relevance of our own work as well as some studies also provide a brief account of some other problems related to compound noun analysis tasks. Our main focus of this review is on the compound noun interpretation task as a classification problem using different computational methods over a finite set of relations.

2.5.1 The theoretical work that influenced NLP research

Levi (1978) presents one of the earliest and influential complex nominal studies from theoretical perspectives. The theory proposed in the work has influenced all the works related to complex nominals analysis in NLP studies. Levi worked on the highly detailed semantic analysis of English complex nominals and compared it with the Hebrew compound nominals. According to Levi, Complex nominals are of three types : nominal compounds *(e.g. pineapple cake)*, nominalizations emph(e.g. movie producer; the noun producer is derived from the verb to produce) and noun phrases with non-predicating adjectives (e.g. electrical engineer; the adjective is considered non-predicating because the construction cannot be paraphrased as *an engineer who is electrical). The noun phrases with non predicating adjectives are included due to the similarity between the compound nouns and NP phrases with non predicating adjectives. Her claim is that non-predicating adjectives are derived from underlying nouns, and therefore behave like a noun when they are part of complex nominals.

Levi argues that complex nominals show syntactic and semantic regularities which help in the interpretation of new compound formation. She claims not to treat complex nominals as strictly idiosyncratic constructions which need to be included in lexicon except for some compound constructions like idiomatic CN, lexicalized and metaphorical CNs.

The semantics of compound nouns in the work is captured by nine Recoverably Deletable Predicates (RDPs). The description of these predicates are as follows:

USE steam iron CAUSE1 flu virus

CAUSE2 birth pains HAVE1 school town

HAVE2 company liabilities MAKE1 honey bee

MAKE2 stone wall BE chocolate bar

IN mountain lodge FOR headache pills

FROM bacon grease ABOUT adventure store

The steam iron is an iron using steam to function. Here the USE refers to an instrument , flu virus is a virus that causes flu, headache pills are pills for the headache that intend the meaning of purpose and so on. Levi also proposes a new compound: nominalisation type As Séaghdha (2008) presents, according to Levi (1978)'s annotation principles history professor and history teacher are represented by different categories because teacher is derived from the verb to teach. So similar concepts sometimes refer to different semantic categories. The theory described that a given complex nominals derived from RDP deletion potentially can be twelve way ambiguous depending on the actual discourse by lexical, semantic and pragmatic factors. Levi presents that semantic clues provided by the surface nouns enable us to derive marginal notes different from the presidential notes, the world's knowledge of the semantic property of the constituent nouns suggest the locative and subjective

interpretation in both the compound nouns. Similarly the pragmatic factors (the knowledge about present day technology) influences the interpretation. For instance, a musical clock is a clock producing music rather than a clock powered by music, like the electric clocks which are powered by electricity. Another ambiguous example provided by Levi is that in using the construction musical talent the speaker means 'talent in music' (in deletion) whereas the listener means 'talent for music' (for deletion), but this constructions do not posit much restriction in the communication between the speaker and hearer and is an example of slight misinterpretation. Levi's theory suggests that a compound is not restricted to be derived from only one predicate. A compound noun can be derived from two or more than two predicate deletions as in the example Chocolate bunny. The deletion can be of BE or MAKE2 as the interpretation can be bunny is chocolate or bunny made from chocolate. We have also observed this kind of ambiguous compound in our Hindi Compound Nouns which we have presented in detail in the following chapters.

The proposed predicates are too general to use in fine grained distinction of compound nouns for computational processing. It is the most influential study on compound noun interpretation. All the other works after this study followed Levi's work to some extent.

2.5.2 First Corpus based study of Compound nouns in NLP

Warren (1978, 2003) **Warren's** study uses an empirical investigation of 4566 Compound nouns extracted from Brown Corpus. Warren's study includes all the naturally occurring compounds and the analysis is context dependent i.e. compound is classified after analysing the compound in its context of use. The present study also uses this approach in analysing the Hindi compound nouns.

Warren identifies six semantic relations:

1. A is something that wholly constitutes B, or vice-versa: source-result (e.g. leather shoe), result-source (e.g. paste wax), copula (e.g. girl friend).

2. A is something of which B is a part or feature or vice-versa: whole-part (e.g. eggshell), part-whole (e.g. wheelchair), size-whole (e.g. 22-inch board).

3. A is the location or origin of B in time or space: place-obj (e.g. city lights), time-obj (e.g. morning train), origin-obj (e.g. seafood). 4. A indicates the purpose of B: purpose (e.g. coffee cup), goal-obj (e.g. moon rocket).

5. A indicates the activity or interest which B is habitually concerned with: activity actor (e.g. sportsman).

6. A indicates something that B resembles: comparant-compared (e.g. clubfoot).

Warren assigns semantic features to the constituents of compound nouns to delimit the differences in different categories. For example, in distinguishing material artifact compounds like clay bird, the first noun is required to have the +Material feature, while the result should have the +Man-made and +Concrete features. Other features that are used in the classification are +Abstract, +Material, +Artifact, +Shape, +Group, +Animate, +Inanimate, +Human, +Body Part, +Animal, +Building, +Plant, +Area, +Time, +Place, +Natural, +Event, +Phenomenon, +Organization, etc. Warren also proposes prepositional paraphrases to mark the internal semantic relations between the constituents of compound nouns. She is one of the first to use prepositional paraphrases in interpretation of compound nouns. She lists the prepositional paraphrases that are usually associated with the proposed participant roles, e.g. source-result - of : student group - group of students; part-whole - of, with: clay soil - soil with clay; place-obj and time- obj: in, at, on; origin-obj - from; goal-instrument - for; comparant-compared - like. The present work (section 4.3.1) also takes inspiration from Warren's work on the semantic features of the constituents in developing the dataset for Hindi Compound noun semantics. Warren's Study serves as inspiration for many further studies in compound noun

interpretation in NLP using prepositional paraphrases. Lapata (2002); Nakov & Kim (2011); Bos & Nissim (2015)(inter alios.)

2.5.3 The other major works in NLP

T. Finin (1980, 1986); T. W. Finin (1980) is one of the firsts in the study of computation of compound noun semantics. Finin proposes a rule based approach for automatically deriving compound noun interpretation of English compound nouns. The theory proposes the concept of frames and slots ; the slots further can have multiple facets. The Finin approach assumes that the one noun in the compound denotes an event and hence has an event structure and the other noun fills a role in that structure. Finin defines several classes of rules; idiomatic rules, applied mostly on the exocentric compounds and on those compounds where the meaning cannot be easily identified based on the interaction patterns of the constituents. productive rules, define a general pattern with many possible instantiations. The third rule is structural rules, most of the work is based on these rules. The rules create a structural relationship between the modifier and the head ;they are particularly useful for analysing compounds containing nominalized verbs. For example , Iron wheel, where the concept iron is considered to fill the raw-material slot of the concept wheel; March flight, where March is taken to fill the time slot of the concept to-fly Finin system takes as input all the concept representation with all frames, slots and frames then calculates the probability score of the output with all the possible semantic interpretation. The most probable interpretation will be the highest score interpretation. Finin work is the first one to take idiomatic compound noun interpretation work.

Lauer (1995) work is one of the first to use a statistical modelling of the noun compounds, by leveraging information obtained from large amounts of unannotated text. He uses a prepositional paraphrasing approach for the semantic analysis of noun compounds. The compound is annotated with 8 sets of prepositions: of, for, in,

about, with, from, on and at. The test set consists of a random sample of 400 noun-noun compounds from the Grolier corpus, which he manually annotates, taking the actual sentential context into account. For each compound and preposition model calculate two probabilities depending on the meaning attached to the head or the object of the preposition(modifier). The model assumption was that the modifier and the head have independent preference for choosing a preposition and the most likely preference will be the one with the highest probability. The model achieves an accuracy of 47%, where the most frequent preposition baseline (of) is at 33%.

Rosario et al. (2002); Rosario & Hearst (2001) : One of the influential works on the classification of semantic compounds is the work by Barabara Rosario. She worked using empirical methods. She presented an automatic corpus-based technique to classify the semantic relations using a classification algorithm of machine learning. She has performed the task on biomedical text. To create a collection of noun compounds; she has performed searches from MedLine, which contains references and abstracts from 4300 biomedical journals. On these titles and abstracts, she ran a part-of-speech tagger Cutting et al. (1992) and a program that extracts only sequences of units tagged as nouns. In theoretical linguistics, there are contradictory views regarding the semantic properties of noun compounds (NCs). Levi (1978) argues that there exists a small set of semantic relationships that NCs may imply. Downing (1977) claims that the semantics of NCs cannot be exhausted by any finite listing of relationships. Between these two extremes lies Warren (1978) taxonomy of six significant semantic relations organised into a hierarchical structure. Rosario identified 38 relations and compared it with Levi's and Warren's classification. She wrote that while trying to reproduce Levi's classes she realised that they are too general for her purpose. Warren's classification is also very detailed. Her classification is mainly based on the relation of the constituent nouns and not on the semantics of the head nouns. Rosario & Hearst (2001) study reports that the compound meanings were associated with pairs of concepts in a domain-specific hierarchical lexical ontology. The constituents of each compound

were mapped onto the ontology's top level and then specialised by moving down in the hierarchy to remove relational ambiguities. For example, heart arteries, heel capillary and limb vein were all mapped onto the same pair of lexical concepts (Body Regions-Cardiovascular System) and are judged to express the same relation. This method generates thousands of concept pairs of varying frequency and specificity, and is shown to accurately generalise to unseen concept pairs. In this task of multi-class classification (with 18 classes) she has achieved an accuracy of about 60%.

Girju et al. (2005) have used paraphrasing and abstract semantic relations methods with the statistical methods for interpretation of Noun Compounds. In the work they have taken both the two word and three word sequences of noun compounds. Girju annotated the compounds with the 8 prepositional paraphrases(developed by Lauer (1995) and their own 35 semantic relations. The most frequent relations are part-whole(*boy mouth*), attribute-holder(*quality base*), purpose(*headache medicine*), location(*Texas university*), topic(*craft museum*) and theme(*car salesman*). Girju used semantic relation tasks as a classification task in machine learning. The best result is obtained using binary SVM classifiers. Semantic scattering and SVM method was used to classify the semantic relations. Wordnet based features are used for feature identification of head and modifier nouns. The supervised model performed better with paraphrasing methods than abstract semantic relations. According to Girju the word sense disambiguation affected the models performance.

Girju (2006) analysed compound nouns from multilingual perspective and translated the English compound nouns into four languages: French, Italian, Spanish and Romanian. 22 semantic relations were used with the prepositions. The result showed that adding the preposition from translation as a feature improved the accuracy.

S. N. Kim & Baldwin (2005) studied noun compounds in detail along with their study of multiword expressions. They used lexical similarity based methods using the wordnet similarity of compound nouns and also worked on the bracketing of noun compounds. The dataset contains a total of 2,169 binary compounds and 1,571

three-word compounds. They used 20 semantic relations defined by Barker & Szpakowicz (1998). The Girju data was also used to check the model's performance on this data.

Séaghdha (2008); Séaghdha & Copestake (2008, 2009) presented a comprehensive analysis of kernel methods for automatic interpretation of noun compounds. Séaghdha (2007) used Levi' s 8 RDP for its analysis. The assumption of the work is that compound interpretation is an analogical process, i.e. the relational meaning of one compound can be derived or predicted from 'similar' compounds. Seagadha used BNC(British National Corpus and wikipedia dumps to extract the relation of similarity between the compounds.

Nakov (2007); Nakov & Hearst (2008) presented a brief overview of syntax and semantics of noun compounds, from a linguistic and computational point of view. They have presented Levis theory, adjacency model, dependency model for the syntactic structure of compound nouns, relational approach and attributional approach for the semantics of noun compound. Nakov's works also includes more than two noun sequences and interpretation of NCS in context. They have also shown how noun compound understanding solves textual entailment problems.

Tratz & Hovy (2010) presented a novel taxonomy of 43 noun compound relations, a dataset and automatic classification method for noun compound interpretation. The taxonomy created by Tratz et al. used Amazon Mechanical Turk and inter-annotator agreement study to test relation sets. Categories are defined by definition as for location category definition is as n1 is the location where n2 is at, near or around. Dataset is extracted from a large corpus using POS Tagging and the Wall Street Journal section of Penn Treebank. They used a Maximum Entropy classifier with boolean features as WordNet-based features, Roget's Thesaurus based feature and Web 1T N-gram features. The Tratz data is the most used data for compound noun interpretation tasks. The Tratz model used a supervised learning method backed up with linguistics resources. The compound noun set was around 17000

words and the inter annotator agreement was also calculated . The agreement result was also better than previously reported work. Girju et al. (2004); Séaghdha & Copestake (2008)). The bracketing of compounds having more than two sequences was also included in this work.

Dima (2016) have used the deep neural network and word embedding for automatic interpretation of noun compounds. In this paper, they have presented a Deep Neural Network classifier approach for the task of automatic noun compound interpretation for English. This approach achieves comparable results to the state of the art system trained on a closely-related dataset and significantly outperforms this earlier system when confronted with unseen compounds. The work used the distributional semantic hypothesis for analysing compound nouns by making the embedding represntation of compound nouns. Dima used Tratz & Hovy (2010) sets and the best result was 77%. This was the first work using embeddings.

Shwartz & Waterson (2018) analysed compound nouns by proposing the paraphrase method on Tratz and Hovy dataset. The model was based on the path embeddings. The embedding was learned using dependency paths between the constituents of a compound noun by training the English Gigaword Corpus Parker et al. (2011) and a Wikipedia dump. Their models did not perform better than Dima & Hinrichs (2015), yet a new approach was provided for compound noun interpretation.

Fares (2019) studied the noun-noun compound interpretation in a general purpose meaning representation framework. He created a noun-noun compound dataset with bracketing and semantic annotation. He used NomBank and PCEDT features to annotate the compound nouns. The fares is the first to treat compound noun problems as a downstream task and used transfer learning and multi task learning for interpretation of compound nouns.

Ponkiya et al. (2018) work includes automatic interpretation of English compound nouns having two sequences. Ponkia used Framenet to annotate the Tratz dataset and then used LSTM and other neural networks to interpret the compound nouns. Ponkia also used Girju paraphrase to use with the BERT model to understand the BERT model performance on Compound noun interpretation. The study used a masked method to train the BERT model and evaluated it based on the most probable preposition masked by the BERT model.

2.6 Compound noun extraction and Interpretation works in Indian languages

For Indian Languages, Several attempts have been made for extraction and identification of multiword expressions from the corpus using probabilistic methods.

The first work as far as we know was done by Sinha (2011) on stepwise multiword expression extraction of Hindi. He has examined MWE from a machine translation point of view. They reviewed different types of multiword expressions found in the literature and then worked on those MWEs which are found in Hindi Text Corpus. The corpus is collected from different news articles, grammar books and iit Bombay corpus. For identifying MWEs a monolingual corpus with a dictionary and Hindi WordNet is used for a semi automatic method.

The process used for the identification is as follows: sentence boundary identification; POS tagging; morphological analysis; identification of acronym and abbreviation with dots; Hindi chunker and verb-phrase form separation; identification of replicating class; identification of doublet class; identification of *vaalaa* morpheme construct class; complex predicates and compound verb identification; identification of acronym (with no dots); and identification of named-entities. The table represents the experiment result.

MWE Type	F-score
acronym and abbreviation with dots	92.2%
replicating class	97.4%
doublet class	73.6%
'vaala' construct class	90.7%
Complex predicates and compound verbs	77.2%
acronym (with no dots) and named entity	27.5%

Table 2.1: Sinha's multiword expressions list

Gayen & Sarkar (2013) worked on the identification of Bengali Noun Noun compounds from the corpus using Random forest classifier. The work classified the compound into non multi word bigram CN and multiword bigram CN. The compound noun candidates are extracted using chunk information and applying heuristic rules on the corpus. Different association measures, syntactic clues and linguistic features and WordNet-based similarity features have been used for this work. 8546 compound nouns identified automatically using chunker and heuristic rules. The result of the Random forest classifier was a 0.8 F1 score. Gayen model is used to extract the CN MWE from Bengali running texts. This work did not treat the interpretation of compound nouns.

Kunchukuttan & Damani (2008) model worked for the automatic extraction of Hindi compound nouns from the corpus. The model gives a recall of 80% and precision of 23% at rank 1000. The study identified major categories of compound noun MWEs, based on linguistic and psycholinguistic principles. The extraction methods used various statistical co-occurrence measures to extract compound nouns MWEs. This work also addressed the extraction of reduplicative expressions using lexical, semantic and phonetic knowledge. As Damani states the limitation of their work was use of a limited size of corpus. The Named Entity Recognizer was also not used for this work as many compound can be a named entity.

Kumar et al. (2010) worked on the analysis of sanskrit compounds and developed a tool which segments a sanskrit compound in its several parts and generates the

paraphrase of the compound. The semantic analysis of Sanskrit involves four steps; segmentation of the constituents, deciding the correct bracketing , identification of type of compounds and lastly the paraphrase generation for understanding the meaning of the compound. The tool is based on Paninian grammar. The Panini sutra was used to create rules and then using these rules the tool was developed. The model result as given by Kumar, for 200 simple compounds and 100 nested compounds the paraphrase generated by the model has 89% accuracy for simple compounds and 80% for nested compounds.

In a 800k size corpus compounds are tagged in context by the Sanskrit Consortium. Total 92K instances of compound words was there. The relations were classified into four categories depending on the head of the compound. The distribution of the frequent compounds is given in Table.

Type	Percentage
Endocentric	58.70%
Karmadharaya (IS_A)	18.11%
Exocentric	11.04%
copulative class	5.67%

Table 2.2: Distribution of Sanskrit Compounds Kulkarni et al (2010)

Kulkarni et al. (2012) examined the sanskrit compounding process and the insight gained from sanskrit grammatical tradition is used for the analysing semantics of Hindi and Marathi compound nouns. The study is based on the work of Kumar et al. (2010)kumar who worked on in depth analysis of Sanskrit compound processing and developed a compound processor. Based on Kumar's study Kulkarni et al gained the insight that for understanding the meaning of a compound first one needs to understand the underlying constituency structure of that compound . The example given was of South Indian Food Plaza:

South Indian Food Plaza = «<South-Indian>-Food>-Plaza>

And another is to identify different semantic relations between the constituents of the compounds.

Kulkarni's work based on the semantic analysis of Hindi and Marathi compound nouns using the corpus of compound nouns. The aim was to prepare a tagset of underlying relations found in the compound nouns. The work classified the Hindi and Marathi compound nouns based on paraphrasing methods and observed the result as given in table 2.3.

Type	No of Instances	Percentage
Genitive	270	47.12
Paraphrasing with Vibhakti alone	80	13.96
Hyponymic Relation	68	11.86
Paraphrasing with verb + Vibhakti	40	6.98
Copulative	20	3.49
Reduplications	15	2.62
Other kinds of Paraphrases	14	2.44
Difficulty in annotation	66	11.51

TABLE 2.3: Analysis of Hindi data for various types of paraphrases

Kulkarni observed that the Hindi compound deviates from Sanskrit in two aspects:

1. The dataset does not have any instances of Bahuvrıhi (exo-centric) compound.

2. Hindi data has many cases where a lot of compounds require a verb as well as vibhakti (a case marker) for its paraphrasing.

3. The genitive was the most dominating type of compound in both the languages . There are some instances of genitive which can not be analyzed using only genitive marker.

Kulkarni work attempted to classify the underlying relations posited by the Hindi and Marathi compound nouns. As the work suggests that most of the cmpound

of genitive needs other paraphraing for understanding the meaning. We have observed such type of instances in our dataset. For deep understanding of meaning of compound nouns we need other approaches for analysis of the compound nouns.

Rallapalli & Paul (2012) used a hybrid approach for compound noun interpretation problems. Ralapalli used a PURPOSE Net and rules to extract the relation from the purpose net and interpret the compound nouns. The ontology based approach used a lookup technique which searches the words on indexed and preprocessed PurposeNet to derive the semantic relation between the constituents in a noun compound. The work also compared this ontology with wordNet and Concept Net. As Ralapalli states in the study, comparison of results of all the three ontologies indicates that the best performance was on PurposeNet. She also used some Hindi compounds translated into English using a dictionary to check the performance of the approach on other languages. The model result was 60% for English and 30% for Hindi Compound Nouns. The result was not good as the translation affected the result. The work was performed on a limited number of compound nouns which has its entry in PurposeNet.

2.7 Summary and discussion

In the section, we summarise the studies in NLP we reviewed from two points of views: (1) Representation Schema (2) The probabilistic models and features used. The graph representation of the literature review is illustrated in the figure 2.1.

Representation Schema: Lauer (1995) approximates the semantics of noun–noun compounds using eight prepositional paraphrases on a dataset of 282 two-word and 244 three-word compounds. Rosario et al. (2002) has used 18 relations for compound nouns of biomedical text. Girju et al. (2005) annotated a little over 4, 500 binary compounds and 484 three-word compounds starting with a taxonomy

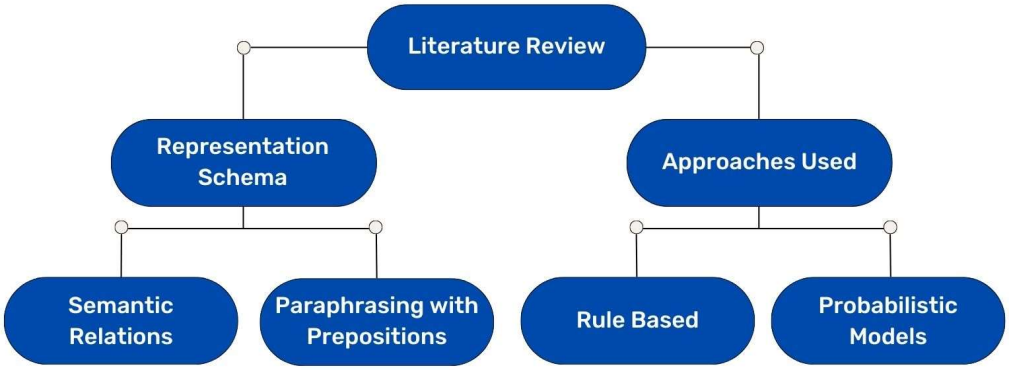

FIGURE 2.1: Classification of Literature Review

of 35 semantic relations, but they report that only 21 of these relations occur in the dataset. SÉAGHDHA & COPESTAKE (2007) relied on the relations introduced by LEVI (1978) to create a coarse-grained inventory of six relations (with two additional levels of granularity), which they then used to annotate 1, 443 two-word compounds extracted from the BNC. S. N. KIM & BALDWIN (2006) used the set of 20 semantic relations defined by BARKER & SZPAKOWICZ (1998) to annotate 2, 169 two-word compounds extracted from the WSJ segment of the PTB. S. N. KIM ET AL. (2013) report bracketing and annotating 1,571 three-word compounds, but only the first dataset, i.e. S. N. KIM & BALDWIN (2008), is available on-line. Finally, TRATZ & HOVY (2010) annotated a relatively large dataset of 17, 509 binary compounds using a taxonomy of 43 semantic relations. Recent work of FARES (2019) used NomBank and PCEDT for generating the noun compound dataset and the relations. Ponkia has used FrameNet based annotation relation and Lauer used 8 prepositions.

Statistical Models and Features used : We showed that noun–noun compound interpretation is, by and large, approached as an automatic classification problem . Several machine learning models have been already applied to learn compound

interpretation, including rule based learning algorithms and decision trees like C5.0 Nastase & Szpakowicz (2003); Nastase et al. (2006) nearest neighbour classifiers using semantic similarity based on lexical resources (Kim & Baldwin, 2013), kernel-based methods like SVMs using lexical and relational features Girju et al. (2005); Séaghdha & Copestake (2013), maximum entropy with a relatively large selection of lexical and surface form features such as synonyms and affixes Tratz & Hovy (2010) and, most recently, neural networks solely relying on word embeddings to represent the compound's head and modifier nouns Dima & Hinrichs (2015); Shwartz & Dagan (2019). It was clear, throughout the literature review, that past work depended heavily on features extracted from lexical resources such as WordNet Girju et al. (2005); Tratz & Hovy (2010); S. N. Kim & Baldwin (2007). Interestingly, though, the earlier studies by Lauer (1995); Lapata (2000) used statistical methods that do not require lexical resources, but the nature of relations they used made this possible; that is, prepositional paraphrases. More recent approaches, however, make use of co-occurrence distributions Séaghdha & Copestake (2009) and distributional semantic models such as word embeddings Dima (2016); Shwartz & Dagan (2019) and transfer learning models Fares (2019); Ponkiya et al. (2020) pretrained language models.

A summary table of the previous major works for compound noun interpretation in NLP is given in table 2.4

Author	Dataset Used	Semantic Relation/Paraphrasing	Model used	Result
Lauer(1995)	244 and 282 English	8 Preposition Paraphrase	Co occurance probabilities	47%
Rosario(2001)	1660 Biomedical	18 relation	ML algorithm and a database	60

Girju(2005)	4500	8 Pre, 35relation	SVM,WordNet based	45
Kim etal(2005)	3500	20 relations	semantic scattering, Decision tree and naive bayes	43
Seagadha(2007)	1443	Levi RDP	SVM with kernels	60%
Tratz(2010)	17200	43 relations	MEntropy, lexical resources	79
Dima(2015)	Tratz Data	Tratz relations	word embeddings and neural Networks	77
Kulkarni(2010)	Hindi Marathi	paraphrase based on sanskrit	Corpus analysis	
Kumar (2012)	Sanskrit	Sanskrit grammar	Rule based	
Damani(2008)	Hindi CN extraction		Statistical methods	79
Gayen(2013)	Bengali CN extraction		Random forest	85
Ralapalli(2018)	English ,Hindi CN translated into English	Girju relations	Ontology based using rules	76
Fares(2019)	English CNs	Embedding, and relations	TL and MTL	72

Table 2.4: Summary table of major works

Tratz & Hovy (2010) argue that "no set of relations proposed to date has been accepted as complete and appropriate for general-purpose text (p. 678). Each proposed set has its own advantages and disadvantages Séaghdha & Copestake (2007). Levi (1978) said that the degree of ambiguity afforded by the RDPs is sufficiently restricted for the hearer to interpret the relation intended by the speaker while still allowing semantic flexibility. Downing (1977) argued that noun compounds can encode an infinite set of semantic relations. There is considerable overlap between these sets of relation and the use of sets of relation is based on application. Nakov (2008) said that using abstract relations, rather than paraphrases, can be problematic because it is unclear which relation inventory is best. Some previous works were also based on the relational similarity between the constituents of a compound noun and the semantic similarity between the constituents of compound nouns.

Most of the models failed to handle these three challenges in compound noun interpretation task.

1. Data sparsity: compound noun formation is a very productive word formation and as stated by Baldwin & Tanaka (2004) 60.3% of compound nouns occur only once in BNC corpus. Data sparsity posit difficulty in applying statistical approach.

2. Knowledge intensity : semantic analysis depends on several factors. The structure of the word, its lexical information and meaning, its nature in the context and also the individual's understanding level, speaker and hearer involvement. Therefore for understanding the meaning of a compound, a deep semantic knowledge is required which is not possible with machine learning algorithms in the current scenario.

Tratz & Hovy (2010) model performance and dataset shows that machine learning models give better performance and result when backed up with rich linguistics resources.

3. Context-Dependent Analysis: All the previous work included context independent analysis, context dependent analysis requires more deep world knowledge.

2.8 Conclusion

In this chapter we have introduced the compound noun as a type of multiword expressions and defined the compound nouns for our study. As we have shown there is little disagreement in defining compound nouns for study in NLP, most of the scholars have treated noun noun sequence as a compound noun. We also presented an overview of the different tasks related to compound noun analysis and what different challenges occur in the study of compound noun analysis. We also demonstrated by providing the example of some NLP application for which compound noun analysis is a pivotal task. Then we gave a brief overview of the most crucial task in compound noun analysis and the central task on which our whole thesis revolves: compound noun semantics; the task of semantic relation identification and classification between the constituents of a compound noun.

The second part of the chapter we presented the literature review on compound noun interpretation from the perspectives of Cognitive Psychology , Theoretical Linguistics and NLP. We provided an in depth summary of some influential previous studies done in NLP for automatic interpretation of compound nouns. We provided the summary of the reviewed works. We reviewed some works on Indian languages for compound noun analysis.

The concept of semantic relations for semantic interpretation of compound nouns is fundamental in all the areas of language related research. Several NLP tasks have used fine grained to coarse grained semantic relations for the interpretation of compound nouns depending on the application, model and datasets. This work also uses the same approach for the Hindi Compound noun interpretation task. Earlier

works on compound nouns for Indian languages particularly Hindi have not taken into account the probabilistic approach. Our work is an attempt towards analysing Hindi compound noun semantics using machine learning models.

Following chapters will talk about the Hindi Compound noun semantics analysis, automatic classification of semantic relations and knowledge base development using generative lexicon for use as knowledge resource with probabilistic models.