CHAPTER 2

LITERATURE SURVEY

In this chapter, we discuss the work carried out for various NLP tasks using rule-based and machine learning approaches. The main focus of this thesis is the application of minimally supervised pattern based bootstrapping approaches for handling a number of NLP tasks. First, we discuss the existing rule-based approaches to the UNL semantic relation extraction. Then, we discuss the use of bootstrapping techniques for different NLP oriented tasks across different languages.

We begin by exploring some of the canonical work of Hearst and Riloff, which are very influential in the development of many sophisticated and effective bootstrapping methods. This chapter then goes on to discuss different methods used for UNL semantic relation extraction. We also discuss the unsupervised approaches to relation extraction. We then describe some existing systems that use pattern based bootstrapping techniques for semantic relation extraction right from the simple exact word matching by DIPRE to the sophisticated system Expresso that learns new patterns and performs statistical scoring. We then go on to discuss some bootstrapping approaches to NLP tasks such as Word Sense Disambiguation, Nested graph extraction, Anaphora and coreference resolution, Multiword Expression identification and finally graph operations for single and multi-document summarization.
2.1 APPROACHES TO HANDLE NLP TASKS

The approaches to handle various NLP tasks are classified into rule-based and machine learning techniques. In Indian languages, most of the NLP work has been done in Hindi, Tamil, Bengali, Telugu, Malayalam and Marathi. Handling NLP tasks are quite difficult for morphologically rich languages compared with English and other European languages due to the partial free-word order nature and the availability of less resources and the gold standard annotated dataset. However, less-resourced languages also often lack human resources, but this aspect seldom receives the necessary attention in research or discussions. It is often hard to find computational linguists working on these languages and therefore the task of annotating training data and training a machine-learning algorithm often falls to linguists or mother-tongue speakers who are computer literate. Unfortunately, these individuals often lack the necessary skills to collaborate in the computerization of less-resourced languages.

Antony & Soman (2011) described the works carried out on part-of-speech tagger for various Indian languages. Due to the morphological richness of Indian languages, complex linguistic rules are needed and the rule based approach did not produce good results in many cases. In contrast, POS taggers developed using stochastic and other approaches gives better results. However, stochastic methods need very large corpora to obtain the correct tagging.

Saha et al (2008) described a Maximum Entropy based NER system for Hindi. They collected some gazetteer lists from the web pertaining to the English language and incorporated them in the system to increase its performance, apply these English lists to the Hindi NER task, they proposed a two-phase transliteration methodology.
Agarwal et al (2004) describe an algorithm for automatic extraction of MWE in Bengali with the use of morphological analyzer and a root lexicon. Each NL word is morphologically analyzed and provided with possible parts-of-speech tags. With the help of the possible POS tags, plausible candidates are extracted from the corpus with the corresponding frequencies. Finally, the candidates are assigned a significance value based on a statistical measure.

Chatterjee & Misra (2009) presented a trainable model for Word Sense Disambiguation (WSD) for resolving this ambiguity. The proposed model applies concepts of information theory to find the appropriate sense of a word when the context is known. Given a training text tagged with the correct senses of a particular word, their model learns to classify each occurrence of the target word with its correct sense in the unseen text.

Chatterji et al (2011) proposed a two-stage pronoun reference resolution system for three languages namely, Bengali, Hindi and Tamil. In the first stage, the Bengali Markable annotation data were used as a training data and Conditional Random Field is used for segmenting and labeling sequential markable data. In the second stage, they used a Markable annotated Training data of both the three languages for training three models.

In general, the Machine Learning (ML) techniques make use of a large amount of annotated data to acquire high-level language knowledge. ML based techniques facilitate the development of recognizers in a very short time. Several ML techniques have been successfully used for various NLP tasks. We have discussed a few tasks that have used ML techniques.

From the above discussions, it is clear that statistical and supervised approaches require a large amount of labeled resources as training datasets. In contradistinction to English, morphologically rich languages have neither any
semantically tagged corpus to aid machine learning approaches for morphologically rich language texts, nor any suitable parallel corpora. To overcome these difficulties, semi-supervised learning techniques seems to be promising approaches to handle the NLP tasks of resource scarce languages. Section 2.2 focuses on the bootstrapping approaches carried out to handle various NLP tasks.

2.2 **BOOTSTRAPPING APPROACHES**

In general, bootstrapping is a semi-supervised learning procedure that starts with a small set of training data and a large set of test data. From the training data, the example patterns are extracted for the labeled task. With the use of patterns represented as a sequence of tuples, bootstrapping is performed by iteratively, matching similar patterns in and extracting them from the test corpus. This minimally supervised bootstrapping learning algorithm is applied to various NLP tasks.

Hearst (1992) described the first approach for the semantic relation extraction task using patterns. He exploited manually defined lexico-syntactic patterns for extracting hyponym (is-a) relation in a text. He identified six generic lexico-syntactic patterns, which can extract hypernym and hyponym relations from a text. Based on the work of Hearst (1992), Berland & Charniak (1999) devised five lexico-syntactic patterns for extracting part-whole relationships from text.

Yarowsky (1995) presented an unsupervised Word Sense Disambiguation (WSD) system which competed with supervised techniques. He exploited “One sense per collocation” and “one sense per discourse” properties of natural language and incrementally learnt the decision list with a few seed instances. Collins & Singer (1999) presented a new algorithm which combines Yarowsky and co-training algorithm and induced a decision list for
the named entity classification starting from a few seed instances by iteratively learning a spelling decision list and contextual rules. Thelen & Riloff (2002) presented a framework which extracts semantic lexicons for multiple categories. They started with a small set of seed words and find all patterns that matched these seed words in the corpus. Ravichandran & Hovy (2002) proposed a bootstrapping approach to extract surface text patterns for a question answering system. They used the sequence, word orderings and suffix trees to identify the optimal answers for a given question query.

Abney (2004) extended the original Yarowsky algorithm by defining a number of variants to optimize maximum likelihood criterion, the formal measure of performance. Li & Li (2004) proposed a bilingual bootstrapping procedure that makes use of a translation dictionary and a comparable corpus to disambiguate word senses in the source language, by exploiting the asymmetric many-to-many sense mapping relationship between words in two languages.

Curron et al (2007) presented a mutual exclusion bootstrapping which minimizes semantic drift using mutual exclusion between semantic classes of learned instances. Now, we go on to discuss the use of bootstrapping approaches in detail for various NLP tasks. First, we will discuss the approaches employed for the semantic relation extraction task.

2.3 SEMANTIC RELATION EXTRACTION

This section begins with the description and analysis of the extraction of generic UNL semantic relations or the so called UNL enconversion and follows it up with the rule-based and machine learning approaches.
2.3.1 Universal Networking Language (UNL) Enconversion

The conversion of a natural language text to the Universal Networking Language (UNL) representation (commonly called Enconversion or UNLization (Enconverter Specifications 2002)) and vice versa (Deconversion or NLization (Deconverter Specifications 2002)) has been attempted in different languages. Conversion of natural language text to UNL representation has been actively studied in Indian Languages such as Hindi (Surve et al 2004), Punjabi (Bhatia & Sharma 2009), etc., and non-Indian languages such as English (Jain & Damani 2008), Arabic (Adly & Alansary 2009), French (Blanc 2005), Russian (Bogusvolsky et al 2003; Dikonov 2011), etc. The commonly used approaches for UNLization and NLization are rule-based, framework based and statistical based approaches.

Jain & Damani (2008) presented an Enconverter for English. Given an English sentence, this system uses a lexicalized probabilistic parser to obtain the typed dependency tree and the phrase structure tree. The dependency relations are then converted into UNL relations and attributes based on the POS tags of the words and a set of enconversion rules.

Bangla to UNL Enconversion has been actively studied and various works have been carried out. Mridha et al (2010) designed the rules based on the Bangla root, verbal suffix and primary suffix for Bangla to UNL representation. At the same time, Ali et al (2010) has presented the case structure based theoretical rules for conversion of Bangla to UNL expression.

Bhatia & Sharma (2009) designed a Punjabi Enconverter consisting of two phases for the generation of UNL attributes and the disambiguation of UNL relations. The Punjabi morphology and case markers are used for attributes generation and relation disambiguation respectively. A rule base database having the UNL attributes is generated on the basis of Punjabi
morphology and UNL relations are resolved on the Noun and verb morphology rules in a specific format for the conversion of Punjabi sentences to UNL expression. Converting French sentence into UNL expression (Gala 2004) used a rule based approach and a syntactic parser to maintaining the sentence structure.

The UNL framework provided by UNDL foundation is a language independent framework which uses a set of language specific analysis rules for morphological analysis, generating a syntax tree and providing appropriate rule for extracting UNL relations. Many languages such as Arabic (Alansary et al 2013), French (Blanc 2005), Bangla (Ali et al 2011; Mridha et al 2010) are included in the UNL Framework. This UNL framework was initially developed by UNDL foundation in which the rules were framed using regular expressions. To perform the UNL Enconversion process from any natural language, the rules are modified by adding the language specific features to the rules to obtain the language independent semantic representation. Ali et al (2011) designed a rule based approach in which lexical and syntactic features based rules are used to enconvert a natural language sentence into a UNL representation. Ali et al (2012) extended the approach of UNLization using a Predictive Preserving Parsing (PPP) technique which again uses a set of language specific rules and grammars for Bangla to UNL enconversion. The PPP technique analyzes at each step of conversion by keeping the main predicate of the sentence under consideration.

A statistical approach on the UNL enconversion uses syntactic features and dependency parse tree to identify the phrases. The conditional probability estimated for each relation is given as a feature vector which consists of features such as phrase type, head word, voice and dependency path similar to the features used in semantic role extraction (Nguyen & Ishizuka 2006).
The UNL enconversion techniques discussed above use a rule based approach that depends on the structure of a sentence. Although many languages use rule based approaches for UNL enconversion, the rules are applied to a specific domain and are fine-tuned to adapt to another domain. Highly-skilled linguists are required, in order to create, enhance, and maintain these rules. Most of the rules are designed based on the syntactic structure information of the natural language sentences. While designing the syntactic rules for agglutinative languages, it is difficult to frame generic rules for a relation of same concept pair that can occur with different forms of the sentences.

In the same vein, the UNL framework is based on the set of rules, specific for each language and hence need to be provided with hand-coded language specific rules to convert a natural language sentence into UNL representation. The use of statistical approach requires probability values for extracting semantic relations. However, the statistical supervised approach does not deal with the context explicitly and requires a large amount of data for probability scoring. Supervised approaches require tagged data to obtain the correct label. Therefore, it is difficult to attain supervised approaches for resource constrained languages. On the other hand, unsupervised approaches do not require any tagged data to start with. Unfortunately, to the best of our knowledge, UNL enconversion has not been attempted using unsupervised approaches. However, semantic relations, other than UNL, have been experimented with many unsupervised approaches which we will discuss in the following section.

2.3.2 Unsupervised Approaches to Semantic Relation Extraction

The development of unsupervised learning methods for the semantic relation extraction task has become an important and popular area of research. The main advantage of these methods is that they can learn a model
using only unlabeled data, as unlabeled text in digital form is in abundance, as unlabeled text in digital form is in abundance, while labeled datasets are usually expensive to construct.

Hasegawa et al (2004) proposed an unsupervised approach for discovering relations between named entities from a newspaper domain. This approach employed a clustering technique to cluster named entity pairs according to the similarity of context words intervening between them. The relation discovery process was based on the assumption that pairs of co-occurring named entities in similar context can be grouped together in a cluster.

Sekine (2006) and Shinyama & Sekine (2006) presented two unsupervised approaches to Information Extraction (IE) known as ‘On-demand IE’ and ‘Preemptive IE’ respectively. The on-demand IE system (Sekine 2006) extracts salient relations from the text based on a user query. The system used a newspaper corpus and retrieved relevant documents, based on a user query and then applies POS tagger, a dependency analyzer and an extended NE tagger to extract patterns from the relevant documents. These extracted patterns are then arranged into a set of similar patterns by applying paraphrase recognition. Shinyama & Sekine (2006) (pre-emptive IE) apply NER, co-reference resolution and parsing to a newspaper corpus in order to extract relations between NEs. The unsupervised approach used unrestricted relation discovery in order to discover all possible relations from texts. The extracted relations are grouped into pattern tables in NE pairs expressing the same relation. This approach uses clustering in order to cluster the semantically similar relations.

Etzioni et al (2008) introduced an open Relation Extraction (RE) system known as TEXTRUNNER - an unsupervised approach to RE by using the Web as a corpus. TEXTRUNNER consists of three key modules: self-
supervised learner, single-pass extractor and redundancy-based assessor. Self-supervised learner module produces a classifier by using a small sample corpus without any hand-tagged data. The single-pass extractor module makes a single pass over the whole corpus to extract tuples of all possible relations from corpora. A redundancy-based assessor module assigns a probability score to each trustworthy tuple based on a probabilistic model of redundancy in the text (Downey et al 2005).

Giovannetti (2010) proposed a hybrid unsupervised approach for semantic relation extraction from Italian and English texts. This approach introduced the open and closed lexico-syntactic patterns to find the hyponymy, meronymy and co-hyponymy semantic relations.

All these approaches extract the semantic relations with the use of lexico-syntactic features. External knowledge sources are required for tagging various features such as POS tag, entities, etc. The analyses reveal that unsupervised approaches require more runs to obtain the correct context of the semantic relations. However, the unsupervised approaches lead to more noisy instances. When attempting them for morphologically rich and partially free word-order languages, it is difficult to learn the context where the semantic relations can occur. Thus, to overcome these difficulties, semi-supervised pattern based approaches are investigated. The following section describes the semi-supervised learning approaches to semantic relation extraction (Auger & Caroline 2008) which require a minimal amount of training data and learns the new context through matching and scoring.

2.3.3 Bootstrapping Approaches to Semantic Relation Extraction

There are several bootstrapping algorithms that have been developed primarily for the purpose of extracting tuples of terms between which a specific relationship exists. This section briefly discusses some of the
successful relation extraction bootstrapping algorithms: DIPRE (Brin 1998), SnowBall (Agichetin & Gravano 2000), Espresso (Pantel & Pennacchiotti, 2006). Each of these algorithms bootstrap for a single relationship type.

2.3.3.1 Dual iterative pattern relation expansion

The Dual Iterative Pattern Relation Expansion (DIPRE) bootstrapping system was developed by Brin (1998) to extract tuples of authors and book titles from raw HTML pages. All relation tuples that match with the identified patterns are extracted without any selection scoring. Each extraction pattern generated by DIPRE is represented as a five tuple:

<order, url-prefix, prefix, middle, suffix>

The order flag determines whether the author appears before or after the title in the matching text fragment. The url-prefix corresponds to the URL of the document, the pattern is identified in. The prefix corresponds to the 10 characters preceding the author (or title). The middle is the text between the author and title pair, and the suffix consists of 10 characters after the title (or author). This five tuple will match an author-title pair if there is a document matching the url-prefix and contains text matching the regular expression:

prefix<author | title> middle <title | author> suffix

where both author and title strings must also match regular expressions.

As each extraction pattern is associated with a url-prefix, each pattern can only be applied to certain web pages.

DIPRE bootstrapping begins by identifying all sentences where both elements of a seed tuple occur. In Brin’s (1998) experiments, five author...
and title pairs were used. Using these sentences, a set of patterns was generated. For all seed examples, candidate patterns which share the same order and middle context are clustered together and combined into one candidate pattern by identifying the longest common url-prefix, prefix, and suffix they share. If the resulting combined length is below a threshold, or the pattern matches only one seed tuple, sub clusters are formed to generate more specific patterns.

To evaluate DIPRE, Brin (1998) extracted author-title pairs from a repository of 24 million web pages. The initial five seeds, identified 199 occurrences and generated three patterns. These patterns then extracted 4047 unique author-title pairs. In the second iteration, Brin (1998) restricted the document sets to a smaller sample (2 million documents), and in the third to those containing the term books (156,000 documents). Over the three iterations, a total of 15,257 unique books was extracted. Manual filtering of the extracted pairs was also performed to remove incorrect author names.

2.3.3.2 Snowball

Based on DIPRE, Agichtein & Gravano (2000) developed the SnowBall system to bootstrap tuples of organizations and their locations from newspaper articles. Snowball extends the DIPRE system by incorporating pattern and tuple scoring functions. Before bootstrapping can begin, a named entity tagger is applied to the document set to tag the instances of organizations, locations and persons.

Pattern in Snowball, are presented as a 5-tuple:

<prefix, tag1, middle, tag2, suffix>

which contains one organization NE and one location NE are either tag1 or tag2, and vectors associating weights with terms appearing in the prefix, suffix and middle contexts.
This pattern representation is used for generating patterns and extracting new tuples. As in DIPRE, during the bootstrapping process, patterns are identified using a clustering method over the set of seed occurrences. An occurrence is identified, if the location and organization in a seed tuple are identified in the same sentence with the correct NE tags. Therefore, the performance of Snowball is dependent on the performance of the NE tagger. The centroids of each of the occurrence clusters are then used as the extracting patterns. No pattern ranking or selection is performed at this stage.

To identify new tuples, Snowball scans all sentences containing a location and organization tag to identify text segments that are similar to at least one centroid pattern. If the similarity between the sentence and a centroid pattern is above a minimum similarity threshold, the location-organization pair is considered a candidate tuple. Each candidate tuple is then scored using the confidence metric (Equation (2.1)), where \( P \) is the set of patterns that generated the tuple and \( C_i \) is the context associated with an occurrence of the tuple that matched \( P_i \) with a degree of similarity, \( \text{Match}(C_i, P_i) \)

\[
\text{confidence (tuple)} = 1 - \prod_{i=0}^{\left| P \right|} \left( 1 - \text{confidence}(P_i).\text{Match}(C_i, P_i) \right) \quad (2.1)
\]

This metric incorporates the confidence values of the extraction patterns that identified the tuple \( (P_i) \), which is based on the \( RlogF \) metric in Riloff (1996a) (Equation (2.2)). The sets \( pos \) and \( neg \) correspond to the correct and incorrect seed matches identified by the pattern, respectively.

\[
\text{confidence(pattern)} = \frac{|pos|}{|pos| + |neg|}.\log_2(|pos|) \quad (2.2)
\]
Using these metrics, Snowball discards all candidate tuples with low confidence. The remaining tuples are added to the list of extracted relations, and the bootstrapping process repeats using the entire list of relations as seed tuples.

Agichtein & Gravano (2000) compared Snowball with the DIPRE on a large collection of newspaper articles. Both the systems were started with five organization-location tuples. The confidence thresholds in Snowball for selecting candidate tuples after each iteration resulted in slightly higher precision in comparison to DIPRE after the first iteration. In Snowball, the confidence threshold can be varied in each iteration depending on the precision or recall requirements. The performance of Snowball was evaluated on 100 randomly selected tuples, and the majority of the errors was related to the performance of the named entity tagger. Therefore, bootstrapping systems which process raw text and do not depend on the accuracy of NLP tools are favorable, especially in domains where the tools are significantly less accurate than those for Newswire.

2.3.3.3 StatSnowball

Zhu et al (2009) developed the Statistical Snowball (StatSnowball) system to bootstrap tuples for traditional relation extraction like Snowball to extract pre-specified relations and open information extraction (Open IE) to identify the general types of relations. StatSnowball extends the Snowball system by defining general patterns, and the use of $l_1$-norm penalized maximum likelihood estimation for the selection of features and patterns.

The patterns in StatSnowball, are represented as a three tuple:

$$ (e_i, e_j, \text{key}) \ i \neq j $$

where $e_i$ and $e_j$ are two entities and key is a set of keywords that indicate the relationship.
StatSnowball takes these patterns as input as seeds to learn the extraction patterns and the relation tuples. The $l_1$-norm regularized Maximum Likelihood Estimation (MLE) is applied to identify the confidence of the patterns and features. The new extraction patterns are generated by the use of Markov logic network, which is the key component of StatSnowball.

Zhu et al (2009) compared the StatSnowball with the Snowball for traditional relation extraction and the O-CRFs for Open IE on a large web data corpus of news articles. The StatSnowball system has several advantages over the Snowball system. In StatSnowball, the probabilistic foundation provides a principled approach to evaluating and selecting patterns. In Snowball, however, the confidence measures lack an elegant interpretation and they are difficult to be applied to general patterns. StatSnowball automatically learns the weights of the generated patterns, which are represented as formulae in MLN, while Snowball applies some heuristic rules to assign the weights.

The quality of the extraction patterns was investigated using two competitive criterias – specificity and coverage. Specificity dealt with the identification of high quality tuples, that the pattern can be. Coverage dealt with the identification of a statistically non-trivial number of good relation tuples, that the pattern can be. In Snowball, the generated patterns are mainly based on keyword matching. These strict patterns can have very high precision, but the coverage (i.e., recall) is very low. Therefore, StatSnowball performs better than Snowball and O-CRF in bootstrapping the tuples of Open IE and traditional relation extraction.

2.3.3.4 MultiSnowball

Liu & Yu (2010) developed the MultiSnowball bootstrapping system to bootstrap the tuples of multi-type relation extraction. This approach
starts with only one pattern and iteratively finds new similar patterns for each of the relations. The researchers knew the inherent weaknesses in the snowball system and hence developed a MultiSnow ball system which deals with multiple relations and restricts these relations to be of a particular type such as professional relations. However, though they could extract multiple type relations, they could only do strict keyword matching to extract relation tuples and the extracted relations were only between entity pairs.

The pattern for profession is defined as:

\[ < \text{Org}; \text{RelKeyword}; \text{People} >, \]

where \( \text{Org} \) means an organization, \( \text{RelKeyword} \) is the relation keyword.

The pattern is defined as 7-ary tuple:

\[ < \text{left}; \text{tag}_1; \text{mid}_{1,2}; \text{tag}_2; \text{mid}_{2,3}; \text{tag}_3; \text{right} > \]

where \( \text{tag}_i \) are 3 required slots, marked according to their relative position. \( \text{left} \), \( \text{mid}_{1,2} \), \( \text{mid}_{2,3} \), \( \text{right} \), are four normal slot sequence, which can be empty.

The pattern confidence is used to measure the confidence of the relation type candidates.

\[
\text{Conf}(K) = 1 - \prod_{i=0}^{P} \frac{1 - \text{Conf}(P_i)}{|\text{SptTE}_i|} \quad (2.3)
\]

where \( P = \{P_i\} \) stands for the patterns that have inferred keyword \( K \), \( \text{SptTE}_i = \{\text{SptTE}_b\} \) is the relation tuples that have been extracted by \( P_i \), \( |\cdot| \) is the count. A pre-defined threshold \( \tau \) is used to select the relation keyword get a higher precision than \( \tau \).
2.3.3.5 Espresso

Espresso is a weakly-supervised bootstrapping algorithm proposed by Pantel and Pennacchiotti (2006) to learn the lexical patterns and relation tuples from newswire and chemical articles. It mainly emphasizes generality and weak supervision. It is a only general purpose system, that identifies a wide variety of binary semantic relations from classical (is-a and part-of relations) to specific and domain oriented relations (chemical reactants, and position succession relations). The Espresso algorithm initially starts with a set of seed patterns to extract the seed instances from the corpus. The Pointwise Mutual Information (PMI) measure has been introduced to estimate the strength of the association between all patterns and all instances. Pantel & Pennacchiotti (2006) define the reliability of a pattern \( p \), \( r_\pi(p) \), as its average strength of association across each input instance in \( i \) and \( I' \) weighted by the reliability of each instance \( i \):

\[
 r_\pi(p) = \frac{\sum_{i\in I'} \left( \frac{pmi(i, p)}{\max_{pmi}} \cdot r_i(i) \right)}{|I'|}
\]

(2.4)

where \( r_i(i) \) is the reliability of instance \( i \) (defined in Equation (2.4)) and \( \max_{pmi} \) is the maximum pointwise mutual information between all patterns and all instances. \( r_\pi(p) \) ranges from [0,1]. The reliability of the manually supplied seed instances are \( r_i(i) = 1 \). The pointwise mutual information between instance \( i = \{x, y\} \) and pattern \( p \) is estimated using the following formula:

\[
 pmi(i, p) = \log \frac{|x, p, y|}{|x, *, y| \cdot |*, p, *|}
\]

(2.5)
where \(|x, p, y|\) is the frequency of pattern \(p\) instantiated with terms \(x\) and \(y\) and where the asterisk (*) represents a wildcard. The set of highest scoring patterns \(P'\), is then selected for extracting the instances.

\[
\pi(i) = \frac{\sum_{p \in P} \left(\frac{p_{mi}(i, p)}{\max_{p_{mi}}} \ast r_{\pi}(p)\right)}{|P'|}
\]  

(2.6)

Espresso finally selects the highest scoring \(m\) instances, \(I'\), and retains them as input for the subsequent iteration.

The Espresso system was tested with TREC and CHEM datasets. The quality of the extracted instances is evaluated manually by assigning different scores range from 0, 1, and \(1/2\) based on the relevancy of matching.

The state-of-the-art bootstrapping systems of semantic relation extraction discussed above dealt with entity-entity relations which are domain specific. These relation types are identified using lexico-syntactic patterns which are defined using the lexical and syntactic information associated with the words and the context. Moreover, entities are identified using a Named Entity Recognizer. In order to extract semantic relations that can exist between any types of constituents in morphologically rich language text these systems do not have enough knowledge because of the defined patterns.

Furthermore, the essential tasks to improve semantic relation extraction task are word sense disambiguation and multiword expressions. These tasks are important to identify the semantic relations that can occur in a correct context and to identify multiple words that acts as a single entity. The following sections describe the works done on word sense disambiguation and multiword expression identification.
2.4 WORD SENSE DISAMBIGUATION

Yarowsky (1995) applied Decision Lists to WSD. In this work, each condition corresponds to a feature, the values are the word senses, and the weights are calculated with a log-likelihood measure indicating the probability of the sense given the feature value. The input of the Yarowsky's algorithm is a set of labeled examples, also called seeds, and a set of unlabeled examples, typically about the 90% of the total. The algorithm consists of an iterative process in which a decision list learner is built with the corpus of seeds and applied to the unlabeled data. In the next iteration the algorithm gets the rules of the seeds plus those of the best confidence acquired from the unlabeled set and a new learner is built. The process is repeated until reaching some training parameters. One of the drawbacks of bootstrapping is the difficulty of guarantee theoretically the learning process.

Another bootstrapping approach (Mihalcea, 2004) uses co-training and self-training for disambiguating various senses of ambiguous words with a set of labeled and unlabeled data. Tatar (2006) applied the Yarowsky algorithm to disambiguating senses iteratively and uses a Naïve Bayes Classifier to classify the context based features for disambiguation. Navigli (2009) proposed an approach for verb sense disambiguation using semantic roles and WordNet along with syntactic features. The use of bootstrapping for Word Sense Disambiguation first needs a set of samples tagged with appropriate features and labeled with the correct sense. The small set of labeled examples can be generated manually (Hearst 1991) or from the automatic selection with the aid of accurate heuristics (Yarowsky 1995). The approaches discussed above dealt with disambiguating either nouns or verbs and the senses of the ambiguous words are disambiguated using the external
knowledge sources such as WordNet. The features used for disambiguation process are mostly based on the syntactic structure information.

2.5 MULTIWORD EXPRESSION IDENTIFICATION

In this section, we discuss the works dealing with MWE identification. A bootstrapping approach recommended by Heid (1999) extract German bigram MWEs where the regular expressions are applied over the single term candidates. And performs the relevance filtering using POS tag and lemma information. The context based expansion was utilized by Heid (1999) for extracting bigrams from possible single word term candidates using regular expressions. The single word term candidates are themselves extracted by regular expressions that are defined using specific morphemes.

Another bootstrapping approach proposed by Rasooli et al (2011) use probabilistic PMI measure to tackle the data sparseness during the identification of compound verbs and light verb constructions where the clustering of MWEs was performed using K-means clustering. The corpus considered for Persian MWEs is annotated with POS tags and some morpho-syntactic features. Instead, in our approach, we annotate the corpus using POS tags and morpho-semantic features to tackle all types of MWEs.

Riloff & Shepherd (1997) introduced a bootstrapping algorithm for extracting semantic lexicons from the parsed data. Their algorithm exploits the observation that members of a semantic category often co-occur with each other in common syntactic constructions: conjunctions, appositives and nominanl compounds. This algorithm takes as input a five seed words that are members of the semantic category of interest ad a parsed corpus, and iteratively identifies additional terms that are hypothesized to be members of the same category. All of the matching patterns are used to identify new candidate terms, and only top scoring terms are extracted.
Riloff & Jones (1999) described a multi-level bootstrapping by incorporating both candidate term and candidate pattern scoring functions. In their approach, two separate bootstrapping loops are introduced. The inner loop corresponds to the mutual bootstrapping algorithm where all the extracting patterns are scored and all of the terms matching the top scoring pattern are extracted. The outer loop also referred as meta-bootstrapping, which identifies the top five terms using a scoring function to generate new seeds.

The above discussed algorithms initially start with seed words and iteratively expand them to extract the n-gram multiwords. However, in this thesis, we describe a multi-stage bootstrapping approach which uses word based features to define patterns and iteratively expand various word based features to obtain the n-gram MWEs. We use various association and context scoring schemes to validate the MWEs and to expand them iteratively.

From the analysis, the above discussed approaches dealt with only specific MWE types. While identifying MWEs from a morphologically rich language text, several challenges and issues have to be considered. Unlike English, agglutinative languages do not have capitalization information and thus it is difficult to identify the complex named entities which may be MWEs.

Section 2.6 describes the works attempted on the complex structure of sentences such as phrases and clauses identification with rich semantic features.

2.6 NESTED GRAPH IDENTIFICATION

Blanc (2000) described the French UNL Deconverter in which the issues in representing the semantic characteristics of predicative concepts are
discussed. Dikonov (2008) discussed the representation of UNL graphs by segmenting complex graphs into simple graphs by applying rules based on propositions. Coreferential links are also considered in segmenting the UNL graphs. (Jain & Damani 2008) described the identification of scopes by relative positions in a phrase structure tree. The author classified the relations into cumulative and others.

Similar to the graph based semantic representation discussed above Chein et al (1997) presented a general framework for nested graphs provided with morphism, which is a mapping between two graphs that induces reflexive and transitive relations for defining nested conceptual graphs in preorder. The simple conceptual graphs were generalized by reasoning of objects based on projection operation. Krishna et al (2011) described a survey on frequent sub-graph discovery which compares the popular graph mining algorithms. The author also discussed the essential factors of various graph algorithms for discovering the subgraphs.

Most of the approaches on nested subgraph identification are rule based. As discussed earlier, rules have limited knowledge which is difficult to design for a generic domain. These representations have also been used to identify the referents of an entity or a pronoun. The following section describes the works done on anaphora and coreference resolution.

2.7 ANAPHORA RESOLUTION

In this section, we discuss various approaches attempted for anaphora resolution. Li & Shi (2008) proposed a Conditional Random Field (CRF) statistical model to resolve Chinese personal pronouns which creates a feature vector based on number gender agreement features, distance and head noun features. However, the non-referring candidates are filtered out using a set of rules. Murthi et al (2007) attempted a machine learning to resolve
person pronouns in Tamil which determines the salience weight of the antecedents using the grammatical features associated with the head of the noun phrase. Bin et al (2010) proposed a twin candidate based learning model to resolve event pronouns in English in which the verb predicates are considered as antecedents. The position, lexical and syntactic features are explored with the structural information using the composite kernel method to identify the antecedent of anaphors.

The existing approaches on anaphora resolution dealt with only single pronoun type and followed a syntactic structure information in the form of syntactic tree and dependency parse tree. While dealing with morphologically rich language text, the syntactic structure and grammatical information alone fail to identify the correct referent of a pronoun. This may lead to incorrect mapping of antecedent or an ambiguity between more than one referent. None of the approaches are attempted to resolve all types of pronouns. One such similar problem is the coreference resolution in which the referents of an entity can be anywhere in the document and it need not be a single referent. The following section describes the works on coreference resolution using bootstrapping approaches.

2.8 COREFEERENCE RESOLUTION

Bergsma & Lin (2006) presented a bootstrapping approach to identify coreference entities based on syntactic paths and word associations. The dependency paths and node sequence are used to identify the co-referent path between entities in the parse tree. Moreover, gender/number information and semantic compatibility are determined using a probabilistic behavior. Another bootstrapping procedure proposed by Kobdani et al (2011) for co-reference resolution uses the word association information to identify mentions and labeled using a self-trained approach.
Muller et al (2002) performed coreference resolution using a co-training algorithm, which puts the features into disjoint subsets when learning from labelled and unlabeled data. The mention-pair model proposed by Soon et al (2001) classifies links (pairs of two mentions) as co-referent or dis-referent, followed by a clustering stage in which link decisions are made by distinguishing the entities.

Only limited works have been attempted on coreference resolution using bootstrapping. The bootstrapping approach follows a number/gender and syntactic path information to identify the coreferent of an entity. When processing the morphologically rich language text, it is difficult to identify all the coreferents of an entity with only agreement and syntactic information. Referents can be complex noun phrases and does not always satisfy the agreement information. Therefore, in order to overcome these difficulties a rich semantic feature is required.

We next discuss the summarization task which includes the single-document, multi-document and query focused multi-document summarization using various techniques.

2.9 SUMMARIZATION

In this section, we discuss the approaches that are carried out for both extractive and abstractive summarization methods. Hendrickx et al (2008) extended an approach for graph based summarization, where coreference relations are utilized for sentence compression. The relevancy and redundancy are expressed in the links connected with the content units represented as nodes. The sentence compression module outputs full sentences, which are simply added to the summary (Hendrickx et al 2008). Erkan & Radev (2004) presented an extractive summarization procedure,
where the salient measures are determined using the graph-based centrality measures of sentences.

Canhasi & Kononenko (2011) presented a frame graph model for extractive summarization where the relation similarity is measured for sentences and are ranked using page rank algorithm iteratively. The existing graph based approaches use node information such as word and its frequency of occurrence, and link information such as in-degree, out-degree connecting each node in a graph. They use semantic role frames where the relation similarity is measured for selecting and ranking the sentences.

Chali & Joty (2008) attempted an unsupervised approach for the query based summary, where the EM algorithm is used for clustering the documents using various lexical, lexico-semantic and cosine similarity features, and ranking the sentences using the K-means algorithm to generate summaries.

Rosner & Camilleri (2008) generated a query based multi-document summary by extracting the frequent items from a clustering of a document set based on the weighted tfidf. The irrelevant documents for a query are filtered out and the important sentences are extracted from the selected documents for the summary. Radev et al (2004) proposed a centroid based multi-document summarizer called MEAD, which generates summaries using cluster centroids produced by the topic detection and tracking system.

Another approach to summarization uses a directed document semantic graph, where entities/concepts are represented as nodes and relations between entities/concepts, such as “is-a” and “related_to” are represented as edges. While entities/concepts are extracted from the parsed sentences, the relations among them are identified, using a set of heuristics (Mohamed & Rajasekaran 2006). However, while the document centric graphs are merged
to form a global graph, the redundant subgraphs are not handled, and this results in the summary containing repetitive information.

From the detailed study of the existing works of the NLP tasks, it is challenging to handle all these tasks for a morphologically rich, agglutinative and a resource constrained languages. The next section describes the objectives behind our research works.

2.10 OBJECTIVES

- To exploit the morphological features and semantic information of word based and context based features to perform higher level of NLP tasks of morphologically rich languages

- To overcome the resource scarcity problems (required for supervised approaches) and to handle the flexible structure of morphologically rich languages, a semi-supervised approach is designed to handle the NLP tasks

- To design a pattern based bootstrapping approach with the use of morphological and semantic information which requires generic pattern representation, scoring schemes for selection and filtering of patterns and generation of new patterns for each NLP task.

- To handle complex NLP tasks, the basic bootstrapping approach is modified to multi-stage bootstrapping.

2.11 OVERVIEW

In this thesis, we focus on the minimally supervised pattern based bootstrapping approaches to handle various NLP tasks. To achieve these
goals, we first use the original single stage bootstrapping algorithm for semantic relation extraction and word sense disambiguation. Our contributions rely on the morpho-semantic features, pattern representation, matching and scoring, and the generation of new patterns. We also use this bootstrapping approach to learn various graph operations that can be carried out in the important components identification which is necessary for the abstractive summarization tasks. Next the single-stage bootstrapping is modified to two-stages to handle the tasks such as anaphora and coreference resolution, and complex structure identification from semantic graphs. Further, the basic bootstrapping is modified to three-stages to handle the multiword expression identification.

This thesis attempts to tackle NLP tasks by exploiting special characteristics of morphologically rich languages. In this thesis, we use Tamil as an example to show how computational approaches for such morphologically rich languages need to be different. Our initial work used the special characteristics to build rule-based systems. However as it is the case with most rule-based systems, only limited types of natural language sentences could be tackled. As a result of our experience in building the rule-based systems, we were able to identify linguistic features that could be effectively used for the NLP processing of morphologically rich languages.

In order to overcome the limitations of rule-based approaches, we next attempted to explore machine learning approaches to the problem. One of the common machine learning approaches used for languages such as English is supervised learning. Supervised approaches require large labor intensive annotated and labelled corpus which is not available for resource scarce languages such as Tamil. Unsupervised approaches on the other hand, take a long time to converge to a solution. We first attempted an unsupervised approach to one of the NLP tasks. From our experience with unsupervised
approach, we found that partially free word order characteristic of morphologically rich language did not lend itself to fast convergence to a solution. In this context, we concluded that semi-supervised approaches that require limited number of trained samples could be attempted.

![Diagram](image)

**Figure 2.1 Bootstrapping Diagram for NLP Tasks**

We adapt the basic semi-supervised bootstrapping approaches for NLP tasks. NLP is basically based on patterns of word formation, phrase formation, sentence formation and document formation. The patterns can specify the internal structure of the instance we hope to discover or specify the context in which that instance can occur. In short, NLP is about discovering these patterns for analysis, representation and interpretation. Therefore, in this thesis, we attempt to design a pattern based bootstrapping approach. Hearst (1992) used a set of example patterns where exact matching and partial matching was performed to extract and label the instances from the
test data iteratively. Exact matching was performed by matching all tuples of
the example patterns with the input instances. Instances that are not labelled
during exact matching are then given for partial matching. Scoring helps in
the selection of patterns for partial matching. During partial matching, new
patterns are generated and the unlabeled instances are labelled in an iterative
manner. This iterative process continues until no input instances are available
in the test corpus or no new patterns are generated. We use the appropriate
features to design three bootstrapping approaches for NLP tasks as shown in
Figure 2.1. In this work, we have used Hearst’s basic single-stage
bootstrapping methodology for semantic relation extraction and word sense
disambiguation tasks with modifications to the pattern representation and
scoring. We have also designed a new two-stage bootstrapping approach for
nested semantic graph identification, anaphora resolution and coreference
resolution where again the pattern representation and scoring have been newly
designed. In addition, we have also designed a three-stage bootstrapping
approach to identify variable length multiword expressions.

We have attempted to carry out natural language processing tasks
for a morphologically rich, resource scarce language using a single semi-
supervised pattern based bootstrapping approach, we have also tempted to use
the same approach for an important application of Natural Language
Processing – Summarization.