CHAPTER 1

INTRODUCTION

Natural Language Processing (NLP) is an off-shoot of Artificial Intelligence and focuses on enabling human computer interaction using natural languages. In the present age, the availability of vast amounts of unstructured text calls for efficient Natural Language Processing techniques to overcome the burden of information overload. NLP is alternatively termed as Computational Linguistics and has several applications such as Information Retrieval, Machine Translation, Speech Processing and Sentiment Analysis. Natural Language Understanding (NLU) and Natural Language Generation (NLG) are the two major sub-tasks of NLP. NLU deals with the inference of the semantic intent of text whereas the objective of NLG is to produce human understandable natural language versions of facts stored in computer databases. Though both of these are essential components for carrying out effective natural language dialogue with a computer, several NLP systems lay emphasis on Natural Language Understanding.

1.1 BACKGROUND

Semantic similarity assessment is a key problem in Natural Language Understanding, wherein the goal is to determine whether the input text units are semantically similar. Verification of semantic similarity serves as the foundation for subsequent processing in tasks such as Information Extraction. The major challenges faced in semantic similarity assessment are the rich variability and ambiguity of natural language text. Two text units

which are structurally and lexically very different may convey the same meaning. On the other hand the same text may convey different meanings in different contexts (polysemy). Due to the wide range of transformations possible, detecting whether two sentences are similar in meaning is a challenging task. Given a sentence a semantically equivalent sentence can be formed by the following methods (Mizrahi 2006):

- Reordering of words (Lexical transformations)
- Rearranging the grammatical structure (Syntactic transformations)
- Replacing words with their synonyms or definitions (Semantic transformations)

Paraphrases and Entailment are two common forms of semantic similarity. Two text units are said to paraphrase each other, when exact semantic equivalence can be verified between them as in the case of Example 1.

Example 1: T1: He enjoyed the match
T2: The game pleased him

Though there are word as well as syntax variations, both T1 and T2 are semantically equivalent. Paraphrases are semantically equivalent and can also be considered as bi-directional entailment that is T1$\Rightarrow$ T2 and T2 $\Rightarrow$ T1. In Text Entailment, one of the inputs, usually the shorter one, also termed as hypothesis (H) may be inferred from the longer unit or text (T) as in Example 2 and can be written as T $\Rightarrow$ H.

Example 2: T: Obama congratulates Israel's new President elect Reuven Rivlin

H: Reuven Rivlin has been elected as Israel's President

Paraphrases tend to convey the same meaning but usually differ in terms of words used as well as syntactic structures (Mizrahi 2006). Paraphrases find application in tasks such as Document Summarization, Question Answering, Machine Translation Evaluation and Information Retrieval (Androutsopoulous & Malakasiotis 2010). Paraphrases can occur at word, phrase, sentence and discourse levels. The replacement of a word by its synonym is considered as word level paraphrasing. For example, the words 'buy' and 'acquire' can be considered as word level paraphrases. The phrases 'found a solution to' and 'solved' are examples of phrase level paraphrases. A typical example of sentence level paraphrasing is the following pair of statements "Indian shares fall on monsoon woes", "Monsoon blues hit Indian market". Discourse level paraphrases are longer and involve multiple sentences or paragraphs. The same meaning is conveyed by both the constructs even though the order of the sentences within the constructs may be changed.

Research problems related to paraphrasing are Paraphrase generation, Paraphrase extraction and Paraphrase recognition. Paraphrase generation which is a Natural language generation problem is the process of generating alternative forms of the input text. This finds application in areas such as document summarization and machine translation. Sentence compression is a typical example wherein the same content is conveyed in a condensed form.

Paraphrase Extraction involves the identification or discovery of paraphrases from a large corpus. Paraphrase Extraction serves as a method of acquiring a collection of Paraphrasing patterns which can be subsequently used for Paraphrase Generation and Information extraction. The task of grouping tweets based on similar opinions is an application of Paraphrase extraction. Paraphrase recognition is the task of identifying whether the given

pair of constructs (words, phrases, sentences) forms a paraphrase or not. The task of matching a user query with queries in a Frequently Asked Questions (FAQ) database relies on Paraphrase Recognition.

1.2 MOTIVATION

Paraphrase Recognition (PR) is a pivotal task, as both Extraction and Generation systems require a Recognition component for validation of their performance. Some of the other areas where PR systems find application are:

- Multi-document Summarization –To identify and eliminate common content derived from multiple sources.
- Question Answering –To retrieve the relevant answer to user queries.
- Intelligent Tutoring Systems– To determine whether the student response matches the reference answer in the case of short answers.
- Text Reuse Detection –To detect cases of paraphrased plagiarism in large document collections

Majority of the PR systems focus on phrasal and sentential paraphrases. Existing approaches for Paraphrase Recognition include Vector-Space model oriented methods, logic-based approaches, machine learning techniques, rule-decoding and graph mapping. Of these, machine learning techniques, notably Support Vector Machines (SVM) have achieved considerable success (Androutsopoulous & Malakasiotis 2010) by extracting various features from the input sentences and using these for classification. The limiting factors in developing efficient PR systems is the non-availability

of large scale annotated corpora for evaluation and the requirement of large amounts of knowledge.

Paraphrase Extraction involves the discovery of equivalent text segments from large corpora. An effective Paraphrase Extraction system will benefit NLP applications such as Information Extraction and Document (Blog/Tweet/News-article) Clustering. Previous approaches for Paraphrase Extraction are predominantly based on the Distributional hypothesis and concentrate on phrase-level units. Other approaches include Bootstrapping, which relies on domain-specific lexicons or context rules and Clustering followed by alignment of sentences within each cluster. The large scale of the corpora poses hurdles in Paraphrase Extraction. Therefore, efficient techniques are required to identify possibly similar candidates from large scale corpora and then subject them to further processing to detect exact matches.

Paraphrase Recognition and Extraction are the focal areas of this thesis. Despite the fact that there are numerous PR systems built using a variety of approaches the performance of these recognizers has scope for further improvement. Though machine learning approaches have been successful, they are dependent on the underlying feature combinations and learning methodology. This has motivated the development of alternate PR systems using Neural Networks and intermediate representations such Universal Networking Language and Predicate Argument Structures. Traditional approaches such as k-means clustering which have been previously applied in Paraphrase Extraction either require the number of clusters (k) to be pre-specified or adopt a time consuming approach to determine k. This has inspired the proposal of a Fuzzy Hierarchical clustering scheme for Paraphrase Extraction.

1.3 RESEARCH GOALS AND CONTRIBUTIONS

The objective of this thesis is to design effective mechanisms for sentence-level Paraphrase discovery by employing machine learning approaches. The research goals of this thesis are to:

- Design an efficient Paraphrase Recognizer for sentence level paraphrases by using Machine Learning techniques
- Identify the best set of features for Paraphrase Recognition
- Explore the effectiveness of intermediate representations such as the Universal Networking Language (UNL) and Predicate Argument Structures (PAS) for Paraphrase Recognition
- Design an effective approach for Paraphrase Extraction based on Fuzzy Hierarchical Clustering
- Explore the impact of Paraphrase Recognition techniques on Student Answer Evaluation, Plagiarism detection tasks as well as the viability of using Paraphrase Extraction for news headline clustering.

The major contributions of this research which address the above goals are:

Neural Network based Paraphrase Recognizer: Several machine learning techniques such as Support Vector Machines, Decision Trees and Naïve Bayes technique have been used for constructing paraphrase recognizers. Though Neural Network based learning is very popular in several applications, it has not been fully exploited in the domain of NLP. A Paraphrase Recognizer which has comparable performance with that of SVM recognizers has been designed using Radial Basis Function Neural Network.

Feature Selection for Paraphrase Recognition: Machine Learning based PR systems classify the input text as paraphrases based on lexical, syntactic and semantic features extracted from the input. The performance of PR systems is significantly affected by the choice and combination of features. The best subset of features for Paraphrase Recognition has been identified using a Wrapper method of feature selection by combining Genetic Algorithms with SVM Classifiers.

Universal Networking Language based Paraphrase Recognizer (UNLPR): UNL is an artificial, electronic language proposed by United Nations to function as an intermediate representation and hence supports cross-language computing applications. In this work, Paraphrase Recognition has been achieved by translating both text units to UNL form and then matching these using a machine learning classifier. Various features extracted from the UNL forms of input sentences have been used to classify whether the two text units are semantically similar.

Predicate Argument Structure based Paraphrase Recognizer (PASPR): Predicate Argument Structures capture the semantic roles in a sentence. This enables a deeper comparison of sentences by matching the semantic roles. In this work, a two stage approach has been designed for Paraphrase Recognition by first pairing the Predicate Argument Structures. In the second stage, the sentences were grouped based on the extent of paired and unpaired tuples and features extracted from the sentence pairs in each group were fed to an SVM classifier in order to recognize the paraphrases.

Paraphrase Extraction using Fuzzy Clustering: A novel two-level Fuzzy Clustering technique has been proposed for Paraphrase Extraction. As similar sentences tend to describe the same or similar actions, Fuzzy Agglomerative Clustering based on verbs was performed initially. Divisive Clustering was then applied on the verb-based sentence clusters to identify sub-groups of

sentences which focus on the same nouns. A Support Vector machine based Paraphrase Recognizer was used finally to identify the paraphrases within each cluster.

Applications of Paraphrase Recognition and Extraction: The task of evaluating Student answers is a time consuming and monotonous task which can be simplified by using Computer Based Assessment systems. A short answer evaluation system has been designed based on the fact that Student answers for short questions are most often paraphrases of the correct / target answer. Automatic plagiarism detection systems aim to pinpoint plagiarized content present in large repositories. This task is rendered difficult by the use of sophisticated plagiarism techniques such as paraphrasing and summarization which tend to mask the occurrence of plagiarism. An extrinsic monolingual plagiarism detection technique based on Paraphrase Recognition has been proposed. Information overload is caused due to several news agencies reporting the same events. Effective mechanisms are required for grouping similar news items and filtering redundant ones. Fuzzy Clustering based Paraphrase Extraction has been used for grouping similar news headlines.

1.4 THESIS OUTLINE

The organization of the rest of the thesis is as follows: Section 1.5 presents a survey of related work pertaining to various aspects addressed by this thesis including existing methods for Paraphrase Recognition and the various features employed in Machine Learning approaches. Paraphrase Extraction techniques as well as prior approaches to applications of Paraphrase Recognition and Extraction such as Student Answer Evaluation, Plagiarism Detection and News Headline Clustering have also been described in Section 1.5.

Chapter 2 elaborates on Machine Learning approaches designed for Paraphrase Recognition and their performance evaluation. The first approach uses a Radial Basis Function Neural Network (RBFNN) to classify pairs of sentences as paraphrases based on the features extracted from the input. Its performance has been compared with that of a Support Vector Machine (SVM) based PR system. Since the recognition performance was found to be dependent on the features, a Genetic Algorithm oriented Wrapper based Feature selection strategy was used to identify the ideal feature set and improve the performance of the SVM based PR system.

Chapter 3 describes the design and evaluation of two Paraphrase Recognition systems which rely on intermediate representations. The first approach, UNLPR (UNL based PR) converts the input sentences into UNL representation and then matches these by extracting features from the UNL forms to arrive at a decision. In the second approach, PASPR (PAS based PR) Predicate Argument tuple pairing has been carried out followed by the usage of features extracted from the sentences to detect the paraphrases.

Chapter 4 describes the two-stage Paraphrase Extraction system using Fuzzy Hierarchical Clustering (PEFHC) and its performance evaluation on the Microsoft Research Paraphrase Corpus (MSRPC) and a subset of the Microsoft Research Video Description Corpus. Chapter 5 elaborates on the application of the improved PR system in the tasks of Student Answer Evaluation, Plagiarism Detection and Extrinsic Plagiarism detection. Chapter 5 also proposes the application of the UNL based Paraphrase Recognizer for Cross-language FAQ access. The application of the Paraphrase Extraction system for Clustering News headlines has also been discussed. Chapter 6 presents the conclusions of the thesis along with directions for future work.

1.5 LITERATURE REVIEW

Paraphrase Recognition and Extraction are two challenging areas which have attracted considerable amount of research. This section presents a survey of related work in these two domains, especially pertaining to Machine Learning as well as Interlingua based approaches for Paraphrase Recognition and Clustering techniques for Paraphrase Extraction. The focus of the review is on systems which operate at the sentence-level. A brief overview of previous approaches to applications such as Student Answer Evaluation, Plagiarism Detection and Clustering news headlines has also been presented.

1.5.1 Paraphrase Recognition

Paraphrase Recognition systems are used to determine the semantic equivalence of the input text. Systems used for Recognizing Text Entailment can be employed for Paraphrase Recognition by checking for bidirectional entailment. The popular approaches for Text Entailment and Paraphrase Recognition are classified as follows (Androutsopoulous & Malakasiotis 2010):

Logic-based approaches: The input text is converted into logical representations and then matched using Theorem Provers and resources such as WordNet and FrameNet's frames (Tatu & Moldovan 2005). These approaches are limited by the need for extensive common sense knowledge.

Surface String Similarity Techniques: These compare the input text directly by computing similarity measures such as String Edit Distance and other Machine Translation based metrics (Malakasiotis 2009). This method is successful (Madnani et al 2012) when there is a high degree of lexical overlap between the inputs.

Assessment of Syntactic Similarity: Dependency trees are constructed for the input sentences. These are then compared and tree similarity measures are computed (Wan et al 2006). The accuracy of the underlying dependency parsers impacts the performance of these methods.

Comparison of Symbolic Meaning Representations: In this method, graphs (Haghighi 2005) or frames (Burchardt et al 2007) which represent the semantic relations in the input sentence are constructed. The similarity between the graph representations is then computed. This approach is also dependent on resources such as FrameNet and WordNet.

Machine Learning Approaches: These tend to consider various aspects of the input text by extracting lexical, syntactic and semantic features and employing supervised machine learning strategies (Finch et al 2005, Zhang & Patrick 2005). Because various types of features are used these approaches tend to be more successful than the others.

Decoding Approaches: In this approach, patterns or rules are applied in sequence to transform one input to the other. An example rule is "X likes Y" ⇔ "X is fond of Y". This method is complicated by the necessity for maintaining a rule database and deciding the rule application order (Harmeling 2009).

Of the above mentioned techniques used for Paraphrase Recognition, Machine Learning approaches have found to be more successful (Androutsopoulous & Malakasiotis 2010) than others.

1.5.1.1 Paraphrase Corpora

The Microsoft Research Paraphrase Corpus (MSRPC) is the benchmark corpus used for the evaluation of sentence-level Paraphrase

Recognition systems. Paraphrase corpora are usually constructed by using multiple translations of the source text or by clustering news articles which record the same events. This corpus has been constructed in multiple stages starting from 1,31,27,938 sentence pairs extracted from a collection of Internet news article clusters. An initial data set was formed from the clusters by filtering sentence pairs based on heuristics such as Word edit distance, length ratio and sentence position. The sentence pairs meeting the initial criteria were further filtered based on the length of the sentences, common words and lexical distance bringing the candidate set size to 49,375 pairs. Several features including String Similarity, presence of morphological variants and WordNet lexical mappings were used to classify the input pairs using a Support Vector Machine Classifier. Out of the 20,574 pairs of positive and near miss negative cases, 5801 pairs were randomly chosen for human annotation (Dolan et al 2004, Dolan & Brockett 2005).

The sentence pairs were labeled by two human annotators and the decision of a third judge was used to resolve disagreements. Out of the total collection, 67% of sentence pairs are paraphrases. The corpus has been partitioned into training and test sets. The training set consists of 4076 sentence pairs and test set has 1725 pairs. The number of paraphrases in the training set and test set are 2753 and 1147 respectively.

The Multiple-Translation Chinese Corpus (MTC) which contains multiple English translations of Chinese News articles has also been used either independently (Malakasiotis 2009) or in combination with other sources. Cohn et al (2008) have constructed a corpus by extracting sentences from MSRPC, MTC and translations of Jules Verne's novel. Another corpus has been developed by Cordeiro et al (2007) by complementing the Knight and Marcu Sentence compression corpus which contains 1087 pairs of positive paraphrase cases with an equal number of negative cases. The

Microsoft Research Video Description Corpus has been constructed from multi-lingual descriptions of short YouTube videos by volunteers on Mechanical Turk (Chen & Dolan 2011).

1.5.1.2 Categories of Features

This section briefly outlines the various text features which help to recognize Paraphrases. The features can be classified as Lexical, Syntactic and Semantic. Composite features can be formed by combining multiple categories.

Lexical Features: These characterize the surface similarity or word overlap between the candidate sentences and are best suited when there is high degree of word overlap between the input sentences. The lexical features typically used in paraphrase recognition are given below:

- **Unigram precision and recall** – depend on the number of shared words between the two sentences. Unigram precision and recall are the number of shared words divided by the length of the first sentence and second sentence respectively. Lemmatized unigram precision and recall are calculated after replacing words by their lemmas (Wan et al 2006).

- **Word error rate (WER)** (Su et al 1992) **-** a measure of the number of edit operations required to transform one sentence into another. It is also termed as Levenshtein Edit distance. WER considers the exact order of words while matching the sentences (Finch et al 2005).

- **Position-independent word error rate (PER)** (Tillmann et al 1997) –This measure assesses the number of edit operations

needed to transform one sentence to the other, without taking the word order into account.

- **Bi-Lingual Evaluation Understudy (BLEU)** precision score (Papineni et al 2002) is a measure based on the geometric mean of n-gram matches. After reversing the order of the sentences, the BLEU recall score can be calculated (Finch et al 2005, Wan et al 2006).

- **Longest Common Substring and Subsequence** – measures the length of the longest common sequence of consecutive and non-consecutive words shared by the input sentence pair respectively (Kozareva & Monotoyo 2006b, Zhang & Patrick 2005).

- **Modified N-gram precision -** a variation of the BLEU measure which considers directional n-gram matches between the sentence pair (Zhang & Patrick 2005).

- **N-gram overlap measures** – N-grams are sub-sequences of n-items from a given sequence. N-gram overlap measures the number of shared n-grams between the two sentences (Wan et al 2006).

- **Skip-gram overlap measures** – Skip-grams are non-consecutive sequences of words using a skip distance k. Skip-gram overlap measures are calculated by dividing the number of common skip-grams by the number of word combinations in the sentences (Kozareva & Monotoyo 2006b).

- **Exclusive longest common prefix N-gram overlap** (Cordeiro et al 2007) – This measure computes the number of overlapping

n-grams by disregarding all lower order sub-grams of a maximal n-gram. It is an extension of the simple n-gram overlap measure.

Syntactic Features: These analyze the degree of structural similarity between the pair of sentences. Syntactic features are capable of detecting similarity even when the word order of the input sentences is changed as in the case of Active/Passive transformations. Some of the commonly used Syntactic features are:

- **Dependency tree edit distance -** A dependency tree is a syntactic representation of a given sentence. Dependency tree edit distance assesses the similarity of dependency trees (Wan et al 2006).

- **Dependency relation overlap features -** Dependency relation overlap features measure the extent of overlap of dependency relations which consist of a pair of words with a parent-child relationship within the dependency tree (Wan et al 2006).

- **The morphological variants feature -** identifies the Co-occurrence of morphological variants in sentence pairs. For example, the words “compute” and “computing” are morphological variants (Dolan & Brockett 2005).

Semantic Features: Several Semantic similarity features exist based on the WordNet database. These measures are termed as Knowledge based measures, as they rely on additional resources such as the WordNet dictionary. Semantic features are suitable when the input sentences have less surface level word overlap, but share semantically related words such as synonyms, hypernyms etc. In the WordNet taxonomy, nodes represent

concepts or words and edges represent the relations between the concepts. The Knowledge based measures (Mihalcea et al 2006) include:

- **Leacock and Chodorow measure** (1998) which is calculated in terms of the length of the shortest path between two concepts using node counting and the maximum depth of the taxonomy.
- **Lesk measure** (1986) which is a function of overlap between corresponding dictionary definitions.
- **Wu and Palmer measure** (1994) based on the depth of two given concepts in the WordNet taxonomy and the depth of the Least Common Subsumer (LCS).
- **Resnik measure** (1995) which assesses the information content of the LCS of two concepts. Information content of a concept c, is the probability of encountering it in a large corpus.
- **Lin measure** (1998) which extends Resnik's measure by considering the Information content of two concepts besides the Information content of the LCS.
- **Jiang and Conrath measure** (1997) assessed in terms of the inverse of the Information content of the two concepts and also their LCS.

Various classes of features are typically employed by machine learning techniques for Paraphrase Recognition.

1.5.1.3 Paraphrase Recognition using Machine Learning

Machine Learning is one of the successful paradigms for paraphrase Recognition. Some of the notable work on Machine Learning

based PR systems have been presented here. Finch et al (2005) have employed an SVM Classifier with radial basis function kernel for Paraphrase Recognition. The authors have investigated the suitability of machine translation evaluation measures such as BiLingual Evaluation Understudy (BLEU), National Institute of Standards and Technology (NIST) measure, Word Error Rate (WER) and Parts Of Speech enhanced Position independent word Error Rate (POSPER) for predicting semantic equivalence. Extending the PER feature based on POS information was found to improve the performance and an accuracy of 74.96% has been reported on the MSRPC (Finch et al 2005).

Zhang & Patrick (2005) have used a decision tree classifier to identify paraphrases after transforming the input sentences using canonicalization rules. The rules employed were: replacement of number entities with generic tags, passive-to-active voice change and replacement of specific future tense usages with more generic ones. Lexical features extracted from the transformed sentences were fed to the decision tree classifier. The authors have experimented on the MSRPC and have reported a maximum accuracy of 71.9%. Zhang & Patrick (2005) have concluded that the inclusion of Lexical Semantic features may improve the performance.

Wan et al (2006) in their work on Paraphrase generation have implemented a paraphrase recognition system in order to filter out incorrect results. The major features used include lexical measures such as BLEU, N-gram Overlap and syntactic features such as Dependency tree edit distance, Dependency relation overlap. Though experiments were conducted with several classification techniques, the best performance was exhibited by SVM technique using a polynomial kernel with an accuracy of 75.6% on MSRPC. Dependency tree based features clubbed with bi-gram features were found to exhibit the best performance.

Kozareva & Montoyo (2006a) have analyzed the applicability of various lexical and semantic features for paraphrase recognition using techniques such as Support Vector Machines, k-Nearest Neighbor and Maximum Entropy classifier. The SVM technique was found to perform consistently better than the other techniques. The authors have reported an accuracy of 76.64% on MSRPC by applying voting policy using the three classifiers. The authors have suggested that including syntactic information would be beneficial.

Das & Smith (2009) have used an approach based on alignment of the dependency trees of input sentences augmented with a lexical semantics component. The authors have also attempted a Product of Experts approach by combining the above model with a logistic regression classifier, thereby achieving an accuracy of 76.06% on MSRPC. The authors have advocated the use of lexical overlap features. Heilman & Smith (2010) have utilized tree edit models for Paraphrase Recognition. A greedy search has been employed to detect the shortest sequence of edit operations. The tree edit sequence was then classified by extracting various features and using a logistic regression classifier to yield an accuracy of 73.2% on MSRPC.

Malakasiotis (2009) has used nine different measures including various standard distance metrics and similarity coefficients. These have been computed on shallow abstractions of the input sentence pair generated by using word stems, POS tags and soundex codes of the original words. Additionally scores were calculated based on the shared dependence relations and by treating synonym words as equivalent. Using the Maximum entropy classifier an accuracy of 76.17% has been observed on the MSRPC when all the features were used. The author has recommended the use of BLEU scores and word alignment features for improving the performance further. Wrapper based Feature Selection using strategies such as Forward Hill climbing and

Beam search have been used to determine the best subset of features in terms of F-measure. This has led to the identification of a reduced subset of features but is accompanied by a drop in accuracy.

Madnani et al (2012) have exploited various Machine Translation metrics for Paraphrase Identification. Eight metrics including BLEU and NIST have been fed to a meta-classification scheme composed of three different classifiers namely logistic regression, instance based classifier and SVM using Sequential Minimal Optimization to achieve an accuracy of 77.4% on the MSRPC. The system has also been evaluated on a corpus extracted from Plagiarism analysis, Authorship identification and Near duplicate detection (PAN) 2010 dataset with good results.

1.5.1.4 PR Systems using Intermediate Representations

Universal Networking Language (UNL) has been proposed by United Nations with the objective of developing universally usable computer interfaces. UNL represents the meaning of a sentence in the form of a semantic network with hyper-nodes (UNL Center 2003). As the UNL representation of a sentence is the same irrespective of the language, it can be used as an intermediate language for finding semantic similarity between the sentences. Pakray et al (2011) have developed a UNL based text entailment recognition system. Similarity assessment was done by assigning scores depending on the extent of relation matches. Two relations were said to match exactly if the entire relation with all its arguments match. Wordnet synonyms and expanded relations were used for identifying approximate matches. The system has been evaluated using the RTE-3 (Recognizing Textual Entailment) and RTE-4 datasets with a maximum precision and recall of 60%. Pakray (2011) has extended the UNL matching system for the Answer Validation task by grouping similar relations and considering word synsets as well as Named Entity matches.

Singh et al (2012) have assessed sentence similarity by performing UNL matching. A three stage scoring system has been used where attribute and word matching scores contribute to universal word scores which in turn are used for calculating relation scores. Precision and recall scores were computed by dividing the aggregate relation score by the number of relations in the UNL forms of each of the input sentences from which the F1-score was calculated. The system has achieved a correlation of only 0.1936 on the MSRPC due to difficulties in UNL conversion. Dan & Bhattacharyya (2013) have used lexical and syntactic features in addition to semantic features extracted from UNL graphs for measuring similarity. A linear regression model was built from training data and then used for prediction on the test data set with good results only for short sentences.

Several other approaches exist for Paraphrase Recognition. Rus et al (2008) have employed a text entailment recognition approach for Paraphrase detection. Entailment was detected by mapping sentences to graph structures and then determining whether one sentence was subsumed within another by applying graph isomorphism algorithms. The entailment scores for Text A with respect to Text B and vice-versa were averaged to determine whether A and B form paraphrases. An accuracy of 70.61% has been observed for the MSRPC. Cordeiro et al (2007) have framed a metric termed as ‘Sumo’ for detection of paraphrases which has been found to work well for symmetric and asymmetric paraphrases. The Sumo metric was computed based on the one-gram exclusive word overlap. The performance of the metric has been evaluated on the extended Knight & Marcu Corpus (KMC), extended version of the MSRPC and a combination of the two corpora. An accuracy of 98.4% has been achieved for the extended KMC.

Fernando & Stevenson (2008) have utilized a matrix similarity method for paraphrase detection. The semantic similarity values between all

pairs of words have been computed using knowledge based measures and an accuracy of 74.1% has been reported. The authors have suggested the incorporation of syntactic features to improve performance. Socher et al (2011) have used Recursive Auto Encoder which is a recursive neural network to construct feature vectors for both words and phrases from the parse trees of the input sentences. A similarity matrix has then been computed between the vectors of both sentences. Since the number of vectors depends on the sentence size, the computed similarity matrices vary in size. Dynamic pooling has been carried out to bring the similarity matrices to a fixed size. In the last stage a soft-max classifier was used to make a decision based on the pooled similarity matrix. An accuracy of 76.8% has been obtained on the MSRPC.

Burchardt et al (2007) have carried out frame semantic analysis to represent the predicates and arguments in the sentence as frames and roles. This process helps to overcome word-level variations of a semantic concept. Graph matching was then carried out by extracting various features from these semantic representations. Amoia & Gardent (2005) have extended the Xerox Incremental Parser with information from VerbNet and WordNet so as to produce the same semantic representation for paraphrases. Matching of a modified version of Conceptual Graphs which consists of concepts and relations between them has been used by Boonthum et al (2003) for Paraphrase Recognition.

Wang & Zhang (2009) have exploited the technique of PAS matching by constructing Predicate Argument graphs for recognizing text entailment. The graphs were further decomposed into Predicate trees and Argument trees. The tree structures in the two sentences of the input pair were matched by treating them as semantic dependency triples. The approach has been tested on the Recognizing textual Entailment (RTE)-3 and RTE-4

corpora and has resulted in precision values of 70.3% and 79.7% respectively. Hickl et al (2006) have designed a system termed as 'GroundHog' for the RTE task which also works by aligning Predicate Argument structures in addition to using lexical, syntactic and co-reference information.

Rios et al (2011) have employed the TINE metric designed for automatic evaluation of machine translation besides other lexical metrics, Named entities and chunking for the RTE task. The TINE metric combines lexical matching and semantic role matching. The verbs were first aligned and then the similarity between their arguments was computed by using a cosine similarity approach. The performance of Rios et al's system was found to approach the average performance of other systems on the RTE1, 2 and 3 data sets.

Qiu et al (2006) have utilized a supervised framework focused on matching predicate argument tuples for detecting dissimilarities between sentences and detecting paraphrases. Initially the most similar predicate argument tuples were paired and the unpaired tuples were then examined by an SVM based dissimilarity classifier to judge the significance of extra information. The system labeled the input sentences as paraphrases, if there were very less or no unpaired tuples. The system has exhibited an accuracy of 72% on the MSRPC. Yadav et al (2012) have proposed an extension of Qiu et al's approach by distinguishing between paired, unpaired, loosely paired tuples and determining the significance of unpaired tuples for sentence similarity establishment.

The performance of various Paraphrase Recognition techniques described above with respect to the MSRPC has been summarized in Table 1.1. The best performance has been reported by Madnani et al (2012).

Table 1.1 Performance Evaluation of various approaches on MSRPC

Technique	Accuracy %	F-measure %
Finch et al (2005)	75.0	82.7
Zhang & Patrick (2005)	71.9	80.7
Wan et al (2006)	75.6	83.0
Kozareva & Montoyo (2006a)	76.6	79.6
Das & Smith (2009)	76.1	82.9
Heilman & Smith (2010)	73.2	81.3
Malakasiotis (2009)	76.2	82.9
Madnani et al (2012)	77.4	84.1
Rus et al (2008)	70.6	80.5
Fernando & Stevenson (2008)	74.1	82.4
Socher et al (2011)	76.8	83.6
Qiu et al (2006)	72.0	81.6

From a study of various Paraphrase Recognition techniques the following conclusions have been made: Methods using Machine Learning Classifiers, especially Support Vector Machines consistently achieve a good performance and Neural Network based techniques are underexplored. Combining various categories of features yields better results than using any single class of features and Feature Selection approach can be used to identify non-redundant set of features. PR approaches which use Intermediate representations especially UNL are applicable across languages. PAS matching which is widely used in the RTE task can be explored for Paraphrase Recognition.

1.5.2 Paraphrase Extraction

The task of Paraphrase Extraction or acquisition aims at extracting paraphrases from a given corpus. Previous Approaches to Paraphrase Extraction can be classified based on various schemes:

- Technique used: Common approaches employ Distributional Hypothesis, Bootstrapping methods and Alignment based procedures.
- Unit of extracted text: Sentential and Sub-Sentential. In the later case, the techniques focus on extracting equivalent phrases or in some cases even words (Androutsopoulos & Malakasiotis 2010).
- Nature of the corpus: The corpora may be Monolingual Comparable which is obtained from several sources and Monolingual parallel, where the source is the same and variants are obtained through different translations and Bilingual corpora (Wang & Callison-Burch 2011). The corpus can again be: text obtained from multiple translations, news / event descriptions, Speech or Video descriptions (Max et al 2012).

1.5.2.1 Approaches for Paraphrase Extraction

Several Paraphrase Extraction approaches exploit the Harris's Distributional hypothesis, which states that words in similar context tend to have the same meaning (Harris 1954). Using this approach, words or phrases which share the same context are declared as paraphrases. Bootstrapping methods rely on seed patterns for acquiring paraphrases. A set of positive and negative seed patterns have to be supplied manually or extracted automatically in this method. Alignment based procedures work by aligning

sentences from comparable or parallel corpora and then applying the Distributional hypothesis.

Distributional Hypothesis based methods

Lin & Pantel (2001) have employed the distributional hypothesis approach and used the dependency parse tree paths as contexts to extract phrase level paraphrases. This approach has been found to extract several incorrect patterns and therefore requires further filtering (Madnani & Dorr 2010). Shinyama et al (2002) have picked up anchors such as names, numbers and dates from text and have then identified the phrases which share the same anchors. The authors have observed that focusing only on named entities limits the number of extracted patterns. In the extended version of this work (Shinyama & Sekine 2003) co-reference resolution and structural restrictions on dependency trees were used. The system was evaluated by collecting paraphrases from pairs of Japanese news articles and achieved a precision of 62%.

Bhagat & Ravichandran (2008) have extracted paraphrases, by constructing a feature vector for each phrase based on its context. Equivalent phrases were then identified by computing the cosine similarity between the feature vectors of phrases. This approach was found to perform better than Lin & Pantel (2001) and Szpektor et al (2004) with an accuracy of 70.79% on a test set of randomly selected paraphrases from the output. But the system was found to extract a large number of redundant patterns. Metzler & Hovy (2011) have deployed the Distributional hypothesis in a Hadoop-based framework to operate on large-scale corpora. This approach requires the initial set of phrases to be given as input though it has registered a high coverage of 86% when evaluated on paraphrases extracted for verb phrase chunks from news articles.

Bootstrapping methods

Barzilay & McKeown (2001) have matched similar sentences from multiple translations and chosen the common words from these sentences as positive seeds and other words as negative seeds. From these seeds, positive/negative context rules have been framed using which sentential paraphrases were picked. The system was evaluated on 500 paraphrase pairs selected randomly from 9483 pairs and 86.5% of the reported pairs were declared as paraphrases when annotated by two human judges. This system has been tested only on monolingual parallel corpora and is constrained by the need to identify negative seeds also. Szpektor et al (2004) have adopted the bootstrapping approach by using terms from a domain-specific lexicon and coupling these with frequently co-occurring noun phrases to form seed slots and then extracting the templates. The authors have reported that the yield of the approach depends on the number of anchor sets considered. The system was tested by generating templates for verbs from Reuters corpus and yielded an average precision of 44.15%. Increasing the anchor sets was found to improve the yield but also resulted in an increase in computational time. Keshktar & Inkpen (2010) have adopted the method used by Barzilay & McKeown (2001) to extract Paraphrases of emotion terms by using emotion words as seeds and then analyzing the context of these seeds to acquire equivalent terms. The method was tested using newspaper headlines annotated with respect to six emotions and has obtained an average precision of 84% and recall of 89%. However this method is also limited by the need to specify the seeds.

Alignment techniques

Barzilay & Lee (2003) have applied hierarchical complete-link clustering to cluster the sentences describing the same type of event. Sentence level paraphrases were discovered by applying Multiple Sequence Alignment

on pairs of sentences from each cluster, constructing word lattices and then matching them. This method has been found to perform better than Lin & Pantel's system (2001) with an improvement of 38% with respect to correctness, when evaluated on 50 pairs randomly selected from the templates generated by both systems. However the construction and processing of lattices is computationally expensive (Madnani & Dorr 2010). A similar sequence alignment technique has been employed by Regneri & Wang (2012) to first extract sentence-level paraphrases from which phrase-level paraphrases were further extracted. This approach performs well with an accuracy of 85% and F-measure of 72% but relies on discourse information and hence is not suited for processing sentence level text units.

Wubben et al (2009) have compared Clustering against pair-wise matching for extracting paraphrases from news corpora. A k-means Clustering approach was used to subdivide already existing clusters of headlines. Sentence-level Paraphrases were then extracted by matching all possible sentence pairs within each cluster. In the alternate approach, cosine similarity was computed between all sentence pairs within the pre-existing clusters. The pair-wise matching approach was found to have better performance in the experiments conducted by Wubben et al with a precision of 76% as against the k-means clustering results of 66% on a dataset containing headlines extracted from Google News.

From the study of related work, it has been observed that paraphrase Extraction approaches based on the Distributional Hypothesis though suitable for extraction of phrase-level units are not well suited for sentence-level units. When moving up from the phrase level to the Sentence level, detecting similar contexts becomes challenging. Also in cases such as Tweets, there may be no immediate context available. In the case of Bootstrapping approaches, identifying the seeds or rules is tedious, especially

when the corpus is dynamic and grows incrementally. Therefore, Alignment approach is better suited for the extraction of sentence level paraphrases.

1.5.2.2 Fuzzy Clustering

For sentence-level paraphrase extraction, applying the Alignment approach directly on a large corpus is infeasible. Hence it is better to first apply clustering and then match the sentences within the cluster. Clustering divides data into related groups so as to maximize the similarity within each cluster and minimize the similarity between clusters. Two major categories of Clustering are Partitioning methods and Hierarchical techniques. In partitioning methods, iterative relocation is used to assign objects to a pre-specified number of clusters. Hierarchical Clustering may be Agglomerative, wherein similar clusters are merged at every step or Divisive which involves the splitting of clusters (Manning et al 2008). Hybrid techniques have also evolved by combining the individual methods (Berkhin 2006).

Traditional Clustering algorithms create a hard partitioning of data in which each object is assigned to only one cluster. Fuzzy Clustering is an alternate, wherein a soft partition is constructed with each object belonging to multiple clusters with different degrees of membership. In line with the original categorization of Clustering, two popular variants of Fuzzy Clustering are Fuzzy C-means and Fuzzy hierarchical approach. The Fuzzy C-means approach is a fuzzy variant of the original k-means partitioning approach and aims to construct a Clustering which minimizes the intra-cluster variance by using the objective function of Equation (1.1):

$$J(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n}(\mu_{ik})^{m} \parallel x_k - v_i \parallel^2 \qquad (1.1)$$

where J(U, V) represents the fuzzy partitioning (U) of n data objects into c clusters, with V being the vector formed from the centroids of the clusters (Torra 2005). μ_{ik} represents the membership of the k^{th} object x_k in the i^{th} cluster with centroid v_i and m is the fuzziness parameter. This technique is limited by the need for specifying the number of clusters-C. Though techniques for automatically detecting C exist, they become infeasible in large corpora.

The hierarchical clustering methods operate on the inter-cluster similarity matrix which is typically reflexive and symmetric. One of the methods of creating a hierarchical clustering is to convert the similarity matrix between clusters into an equivalence relation by computing its transitive closure and then determining the alpha-cut. But this involves additional computation and also results in a hard partition (Gagula-Palalic 2008). Some of the previous work on Sentence Clustering has utilized the hierarchical approach. Seno & Nunes (2008) have used an incremental approach for clustering sentences from multiple documents belonging to Portugese language. The first cluster was created with the first sentence, and each subsequent sentence was either added to an existing cluster based on cosine similarity and word overlap measures or assigned to a new cluster. Representative terms from the cluster were used to establish the centroid for each cluster. The major advantage of this work is its incremental nature.

Geiss (2011) has employed Latent Semantic Analysis (LSA) for Sentence Clustering by considering word usage patterns instead of the typical word overlap measures. Singular Value Decomposition was applied on a Term-by-Sentence matrix followed by Hierarchical Agglomerative Clustering. The author has reported the superior performance of LSA approach when compared to the traditional Vector Space Model technique. Though hierarchical in nature the approaches followed by Geiss (2011) as

well as Seno & Nunes (2008) assign a sentence to a single cluster only. This may not be suitable when a sentence can belong to multiple clusters.

Bank & Schwenker (2010) have proposed a fuzzified version of the agglomerative algorithm, where after merging the two most similar clusters C_i and C_j to create C_{ij}, only the similarity value S_{ij} was set to zero and the remaining values corresponding to the clusters C_i and C_j were retained. This permits clusters C_i and C_j to take part in subsequent mergers. In order to prevent further merger of C_i or C_j with C_{ij} the similarity between C_i, C_j and C_{ij} was set to zero. When C_i is a part of several clusters-C_{ij}, its membership in each of the parent clusters was computed by normalizing the similarity between C_{ij} and current parent p by the sum of similarities of C_{ij} with all parents. This approach supports the assignment of an object to multiple clusters and also ensures finite clustering.

Rodrigues & Gama (2007) have proposed an Online Divisive Agglomerative Clustering which performs incremental clustering of time series data. A semi-fuzzy approach has been used for assigning objects to clusters by computing the distance between the object and the candidate clusters during division. Based on the distance computation, an object may be assigned to multiple new clusters with memberships as per Equation (1.2).

$$\mu_{in} = \mu_{ip} / n_i \tag{1.2}$$

where μ_{in} and μ_{ip} are the membership values of object i in the new cluster n and original parent cluster p respectively and n_i is the number of new clusters to which object i is assigned. The approach has been found to work well for time-series data and has the advantage of re-computing the similarity matrix only for the cluster which is being split.

Fuzzy approaches have been previously applied to both document clustering as well as Sentence Clustering. Rodrigues & Sacks (2004) have proposed Hierarchical Hyper-spherical Fuzzy C-Means (H^2-FCM) approach for clustering documents. Hyper-spherical FCM (H-FCM) uses a cosine similarity coefficient for computing distances between clusters instead of the Euclidean measure. In order to overcome the problem of fixing the number of clusters –C, the Hierarchical variant initially forms a set of clusters using H-FCM and then merges similar clusters. Though the approach was found to be scalable it still requires an estimate for 'C' to be specified. Skabar & Abdalgader (2013) have proposed a Fuzzy Clustering algorithm for relational data where sentence similarity values are used rather than the vector space representation of the sentences. The approach combines the concept of Gaussian Mixture Models with a Graph representation to determine the membership of objects in clusters in terms of the Page-Rank score. This approach also requires the specification of the number of clusters. The system was tested on a dataset of quotations and achieved better performance than k-means and Spectral Clustering with v-measure of 65.2% and F-measure of 63.6%.

Bordogna et al (2006) have proposed an extension of the Fuzzy C-means algorithm for performing incremental and hierarchical clustering. The number of clusters –C, at the lowest level of the hierarchy has been fixed automatically based on the degree of overlap between news items. A fuzzy hierarchy of news clusters has been constructed by combining clusters at lower levels. The approach has the advantage of being incremental and has reported a high precision of 98% and recall of 50% when tested on documents extracted from the Open Directory Project dataset. The shortcoming of the approach is its usage of vector space models and cosine similarity thereby neglecting the aspect of semantic similarity.

A study of the literature shows that fuzzy clustering is more suitable for natural language input than crisp approaches. Likewise the hierarchical Clustering approach, which automatically establishes the number of clusters using thresholds on the similarity metric and also supports incremental data is better than partitioning approaches. Hence the fuzzified agglomerative approach proposed by Bank & Schwenker (2010) and the Divisive approach of Rodrigues & Gama (2007) have been adapted during Fuzzy Clustering for Paraphrase Extraction.

1.5.3 Applications

Paraphrase Recognition has applications in several NLP tasks such as Multi-document Summarization, Machine Translation Evaluation, Answer Evaluation, Plagiarism detection, Question Answering and Information Retrieval. PR systems are also used in Paraphrase Extraction to verify that the identified candidates are Paraphrases. Paraphrase Extraction systems are typically used in Information Extraction systems to determine similar content and in Paraphrase Generation systems to produce equivalent versions of the input. This section outlines previous approaches to some of the above mentioned tasks such as: Student Answer Evaluation, Plagiarism detection and Clustering of news headlines.

1.5.3.1 Student Answer Evaluation Approaches

Evaluation of Student answers is a time consuming and monotonous task which can be simplified by using Computer Based Assessment systems. Though computerized evaluation of objective answers is a commonly adopted practice, grading of short and long answers in examinations is usually performed manually.

Latent Semantic Analysis (LSA), assessing word-to-word Similarity and Information Extraction are the pre-dominant methods used for Student Answer Evaluation. Sukkarieh et al (2003) have successfully employed an Information Extraction approach for evaluating University of Cambridge Local Examinations Syndicate (UCLES) answers. The marking was based on Information Extraction, wherein a set of rules or patterns were associated with a question which defined the various ways of answering it. The student answer was graded based on the presence of these patterns. This required considerable effort in terms of writing patterns and employed only primitive pattern matching based on bag-of-words approach.

LSA technique which has been used in several Intelligent Tutoring Systems follows a Bag-of-Words approach and ignores the word order information. Kanejiya et al (2003) have extended the conventional LSA method by considering the word order as well as syntactic neighbourhood. The approach termed as SELSA (Syntactically Enhanced LSA) was found to have comparable performance to LSA with a maximum observed correlation of 0.47 as against a correlation of 0.51 for LSA technique using a short answer corpus developed at University of Memphis.

Perez et al (2005) have proposed a technique termed as Evaluating Responses with BLEU (ERB) for assessing student answers. The n-gram co-occurrence between the student answer and reference answer was computed using the BLEU metric. This method was found to outperform the LSA method with a mean correlation coefficient of 0.47 in contrast to LSA's score of 0.43 on various datasets constructed from Spanish exams. Rus et al (2009) have applied word to word relatedness measures such as LESK, HSO and VECTOR to evaluate student answers. Assigning word weights combined with VECTOR measure yielded the best performance. The system was tested

on User Language Paraphrase Corpus (ULPC) and has recorded a correlation of 0.606 against that of LSA's 0.555.

Mohler & Mihalcea (2009) have used traditional knowledge based and corpus based measures for short answer grading. Among the knowledge based measures, Jiang and Conrath measure resulted in the highest correlation of 0.45 on a computer science assignment dataset. Out of the corpus based methods, Explicit Semantic Analysis (ESA) based on Wikipedia performed better with a correlation of 0.47 in comparison to LSA approach's correlation of 0.41. An extended version of this work has been reported by Mohler et al (2011) wherein graph alignment features have been combined with lexical semantic similarity measures. Employing graph alignment features was found to improve the correlation to 0.52 for the same computer science assignment dataset by using Bag of Words as well as Alignment features to construct SVM models.

Nielsen et al (2009) have viewed the problem of grading student answers as one of detecting text entailment relationship between the student answer and the reference answer. Answers were divided into facets and the student answer was then analyzed for correspondence to the target with respect to each of these facets. This approach has recorded an accuracy of 63.8% on the RTE-3 test dataset. The same authors have used a machine learning approach by extracting lexical and syntactic features and obtained an accuracy of 67.1%. The c-rater system developed by Sukkarieh & Blackmore (2009) has adopted a rule based concept-matching approach between the target answer and the normalized student answer. The system has achieved an average agreement with human ratings ranging between 69% and 98% on a corpus constructed from short answers of school students.

Though c-rater and the work by Nielsen et al (2009) have used Text Entailment detection, they operate by splitting the answer into concepts

and then perform matching. The iSTART (Interactive Strategy Training for Active Reading and Thinking) system has employed the Paraphrase recognition strategy to assess whether the student's understanding of a concept matches the target (Boonthum 2004). Similar to c-rater and Nielsen's work, iSTART has also carried out a fine-grain analysis by matching triplets within the sentence.

The literature study indicates that majority of student answer evaluation systems use the LSA approach and its variants. There are very few systems which exploit the fact that in student answer evaluation, correct answers can be viewed as paraphrased versions of the reference answers. This has motivated the development of a short answer evaluation system using Paraphrase Recognition.

1.5.3.2 Plagiarism Detection methods

Plagiarism can be termed as the unauthorized reuse of copyrighted content without giving due credit to the original authors. Plagiarism of text can be broadly classified into two categories: Literal and Intelligent (Alzahrani et al 2011). Literal Plagiarism usually includes copy-paste operations and is usually easy to detect. Intelligent Plagiarism on the other hand is much more difficult to identify and may involve translation, summarization and paraphrasing. One of the most difficult to detect and relatively less addressed forms is Paraphrased Plagiarism in which the original content may be completely reworded and altered beyond recognition (Barrón-Cedeño et al 2013).

The objective of Plagiarism detection systems is to ensure the originality of text content. Such systems are categorized as Intrinsic and Extrinsic detectors. Intrinsic detectors attempt to identify plagiarism by analyzing the writing style variations within a single document. Extrinsic

detectors compare a suspicious document against a set of source documents and identify the plagiarized portions, if any, by first choosing a set of candidate documents and then assessing similarity.

Extrinsic Plagiarism Detectors usually characterize un-structured documents using various categories of textual features such as lexical, syntactic and semantic (Alzahrani et al 2011). The most popular lexical features are character and word n-grams, while Parts-of-Speech (POS) information is used extensively to extract syntactic features. Semantic features depend on thesaurus like WordNet to typify word relationships. In order to retrieve the candidate source documents for matching against the suspicious document, traditional Information Retrieval techniques based on Cosine Similarity, Vector Space Model and Fuzzy Retrieval may be used. Once the candidate documents are identified they can be compared exhaustively using techniques based on String matching, Vector Similarity computation, Syntax, Semantic, Fuzzy and Structural feature based methods. Of these, Semantic and Fuzzy methods are more effective in detecting complex types of plagiarism including paraphrasing and re-structuring besides the simpler copy-paste forms.

Clough et al (2002) have carried out some of the earliest experiments in this domain and have also constructed the METER (MEasuring TExt Reuse) corpus. More recently the Plagiarism analysis, Authorship identification and Near duplicate detection - Plagiarism Corpus (PAN-PC) competitions have generated considerable interest in this domain and have led to the development of several successful systems which work on large scale document collections. Some of the approaches used include winnowing, hash function computation, finger-printing and exact matching at various levels such as character-n-grams, word-n-grams, sentences

(Potthast et al 2009, 2010). Various commercial and free plagiarism checkers such as Turnitin, Cross-Checker and Copyscape are available online.

Despite the large number of Plagiarism detection alternatives the identification of paraphrased plagiarism has not been fully addressed (Barrón-Cedeño et al 2013). As the amount of lexical variation between the text units increases, plagiarism detection becomes tougher. In an effort to focus on paraphrased plagiarism, subsequent PAN competitions have introduced multiple cases of simulated plagiarism which were created by workers on Amazon's Mechanical Turk by rewriting original text content. In the PAN 2011 corpus, 71% of the plagiarism cases are paraphrased ones (Potthast et al 2011).

Burrows et al (2012) have adopted the crowdsourcing approach to create paraphrased versions of text passages for constructing the Webis Crowdsourced Paraphrase Corpus (CPC). The corpus was originally developed as a part of the PAN 2010 competition to test the efficiency of plagiarism detection systems. The authors have also assessed the performance of various paraphrase similarity metrics for automatically filtering the generated paraphrases. The metrics include normalized edit distance, n-gram comparison based measures such as simple word n-gram overlap, BLEU metric and Longest Common Prefix n-gram overlap, besides the Sumo metric and various asymmetrical paraphrase detection functions proposed by Cordeiro et al (2007). Burrows et al (2012) have concluded that using a combination of these metrics with a machine learning classifier yields the best results.

Bar et al (2012) have combined three categories of features based on the content, structure and style for measuring text reuse. Content based features were generated by comparing the text content. These include string similarity measures, greedy string tiling, word and character n-gram features,

Wordnet based semantic similarity measures besides Latent Semantic Analysis and Explicit Semantic Analysis. Structural similarity was assessed in terms of word pair order, distance as well as stop-word and Parts Of Speech n-grams. Stylistic similarity was determined using sentence, token length properties, function word frequencies and vocabulary richness measures such as sequential Type-Token ratio. The approach was tested on three different corpora namely: Webis CPC, Wikipedia Rewrite Corpus and subset of METER corpus. Combining the three categories was found to yield best results in two out of three cases.

From a study of related work, it is observed that paraphrased plagiarism though common has not been addressed satisfactorily. Hence there is a need for efficient plagiarism detection approaches which can handle paraphrased plagiarism.

1.5.3.3 News Headline Clustering Methods

In the domain of news reporting, multiple news agencies report the same information. This leads to information overload and requires the grouping of similar content and elimination of redundancy. A similar situation occurs with respect to social media such as Twitter, where multiple users tweet on the same events. These tweets can be analyzed to detect trending events and similar sentiments. In both cases, Paraphrase Extraction approach can be used to group semantically similar content. The Microsoft Research Paraphrase Corpus itself has been constructed by clustering and filtering news headlines.

Barzilay & Lee (2003) and Bordogna et al (2006) have employed Hierarchical Clustering and a modified version of Fuzzy C-means approach respectively for clustering news items (Section 1.5.2.2). Naughton et al (2006) have used Hierarchical Agglomerative Clustering and computed cosine

similarity using a bag of words approach and have achieved a precision of 50%. The authors have reported that considering the sentence order within the news report and using a TF-IDF weighting measure has improved the performance further.

Wubben et al (2009) have clustered and aligned news headlines to extract paraphrases. News headlines collected from Google News have been aligned using two different approaches: k-means clustering and pair-wise cosine similarity computation. The k-means approach requires the number of clusters to be identified by using efficient cluster stopping criteria. The pair-wise matching approach has been found to perform better than the clustering approach provided the context of the news headline is also used. Boraa et al (2012) have applied clustering based on frequent terms for grouping news headlines. The key terms in each document were identified and if the term is more frequent in un-clustered documents, a new cluster is formed. Experiments conducted on Reuters news headline datasets as well as scientific abstracts datasets have indicated good performance of the frequent term clustering algorithm on small datasets. In the case of larger datasets, the k-means algorithm was found to perform better.

From a study of related work, it is observed that though Paraphrase Acquisition approaches have been previously applied for grouping similar news articles they are limited by the underlying clustering technique which is usually partitioning based (Wubben et al 2009) or hard in nature (Barzilay & Lee 2003, Naughton et al 2006). Even in techniques using a Fuzzy scheme such as Bordogna et al (2006), the semantic similarity aspect has not been taken into account. In order to effectively address the incremental, voluminous and ambiguous nature of news items, Paraphrase Extraction techniques using Fuzzy Hierarchical Clustering may prove to be effective.

1.5.4 Summary

A detailed survey of work related to Paraphrase Recognition methods, Extraction techniques and relevant applications in Student Answer Evaluation, Plagiarism Detection and Clustering News headlines has been carried out. With respect to Paraphrase Recognition, machine learning approaches have been found to register good performance. The fact that supervised classification techniques such as Support Vector Machines, Naïve Bayesian technique have been used extensively in PR systems while Neural Network schemes have been less explored for the purpose motivates the work on using a Radial Basis Function Neural Network Classifier for Paraphrase Recognition. Also since the performance of machine learning classifiers is affected by the features used for learning, a Feature Selection approach has been proposed for identifying the best set of features. Traditional PR systems focus on monolingual similarity. In order to handle multi-lingual inputs a system using UNL representation and combined with machine learning has been designed. Semantic role-based comparison using PA structures has also been explored for sentence matching at a deeper level.

The survey of Paraphrase Extraction approaches indicates that Clustering followed by Alignment is better suited for sentence level paraphrase extraction. Existing methods either employ partitioning based clustering or rely only on word overlap measures for semantic similarity computation. This has inspired the proposal of a Fuzzy Hierarchical Clustering approach which works by computing word semantic similarity.

The tasks of Student Answer Evaluation as well as Plagiarism detection can be viewed as applications of Paraphrase Recognition. But very few systems have adopted this approach. Existing systems for Student Answer Evaluation focus on LSA while Plagiarism detection systems rely on word overlap which is suitable only for Copy-Paste plagiarism. This has

motivated the work on solving the above problems using a Paraphrase Recognizer. Clustering of news articles can be viewed as an application of Paraphrase Extraction and previous approaches to the problem rely either on crisp clustering or focus only on word overlap. This has encouraged the application of the Fuzzy Clustering based Paraphrase Extraction system for Clustering news articles.