CHAPTER 5

APPLICATIONS OF PARAPHRASE RECOGNITION AND EXTRACTION

Paraphrasing of text is a commonly occurring phenomenon and has wide impact in many areas including Information retrieval and Text Processing. This chapter describes the deployment of Paraphrase Recognition and Extraction systems in various real world applications. Evaluation of open-ended student answers is one such application which would benefit from the usage of efficient Paraphrase Recognition systems. Another potential application of Paraphrase Recognition is the task of Plagiarism detection. Plagiarism in free text has become a common occurrence due to the wide availability of voluminous information resources. Further, sophisticated methods such as paraphrasing and summarization are commonly followed to mask plagiarized text and therefore call for the development of efficient plagiarism detection methods. The SVM based PR systems described in Sections 2.2 and 2.3 have been employed for Student Answer Evaluation and Plagiarism detection.

Online users constantly refer to Frequently Asked Questions (FAQ) databases. For effective access of large-scale FAQ collections, efficient mechanisms are required for matching the user's query posed in a particular language against those available in databases stored in a different language. The UNL based Paraphrase Recognition system described in Section 3.1 has been deployed for cross-language FAQ access. The task of clustering similar news headlines can be viewed as equivalent to extraction of Paraphrases from a large-scale corpus. The PEFHC system described in

Chapter 4 has been used for grouping similar news captions and has been tested on headlines extracted from Google News.

5.1 STUDENT ANSWER EVALUATION

The evaluation of open ended student responses forms an important and challenging part of learning. Automated Student Answer evaluation systems can act as a viable alternative to manual evaluation by saving human effort and also possesses the advantage of ensuring fair assessment. However, the unstructured nature of student responses and the possibility of expressing the same concept in several ways pose challenges to automated evaluation mechanisms. Advances in Natural Language Processing and Information Extraction open up the possibility of developing such mechanisms. The flexibility offered by Intelligent Tutoring systems and the popularity of online courses have further boosted the need for automatic evaluation systems.

Depending on the type of input handled, Computer Based Assessment Systems may be classified as Short Answer Grading systems and Essay Grading systems. This work focuses on the development of a tool for assessing short responses using Paraphrase Recognition techniques. The motivation for the work arises from the fact that student answers for short questions are most often paraphrases of the correct / target answer. This transforms the problem of evaluating short answers to that of recognizing paraphrases.

5.1.1 Answer Evaluation using Paraphrase Recognition

The paraphrased versions of the reference answer are usually considered correct. This makes it feasible to evaluate a student answer by analyzing whether it is a paraphrase of the target answer. Characteristically paraphrase recognition being a two class problem, using a paraphrase recognizer

for evaluation will pronounce the student answer only as right or wrong. Therefore this work concentrates on short answers, specifically answers made up of a single sentence for which it is reasonable to assign a binary score.

The Paraphrase Recognizer detailed in Section 2.2 has been used to develop an Answer Evaluation system. Various lexical, syntactic and semantic features were extracted from the student answer and its corresponding reference. These features were given as input to a Support Vector Machine which classified the input sentences as positive or negative paraphrases. This is equivalent to assigning a full score or zero score respectively. Besides using the best set of 113 features, answer evaluation was also attempted using the reduced set of features (Section 2.3) selected by GA. The block diagram of the evaluation system is shown in Figure 5.1.

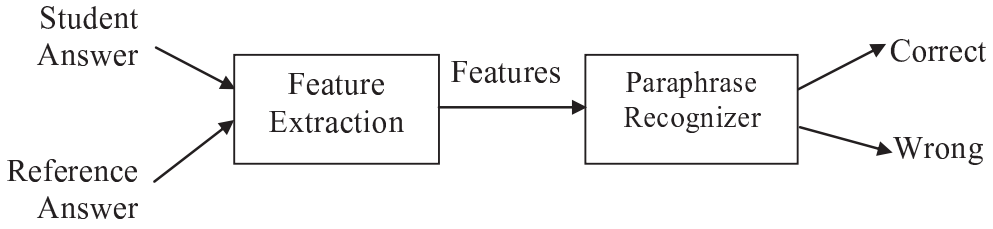

Figure 5.1 Student Answer Evaluation using Paraphrase Recognition

5.1.2 Experiments and Results

In order to assess the performance of the Student Answer Evaluation system, experiments were carried out using two different combinations of features. In the first case, the 113 features which yielded the best performance on the MSRPC (Table 2.3) were used for answer evaluation. In the second case only the 57 features identified by the feature selection technique (Table 2.6) were used. A Support Vector machine with a Radial Basis Function kernel was used for classifying the student answers. To assess the classification accuracy, ten-fold cross validation scheme was utilized.

Evaluation experiments were carried out on two standard data sets, namely the Short Answer Grading Corpus (SAGC) developed by Mohler & Mihalcea (2009) and User Language Paraphrase Corpus (ULPC) developed by McCarthy & McNamara (2008). Some of the earlier Answer Evaluation systems have used these corpora, but the performance was assessed in terms of correlation between human scores and machine generated scores. Because the proposed system produces a binary output, accuracy value has been used to measure the efficiency.

5.1.2.1 Short Answer Grading Corpus

SAGC was developed by Mihalcea and Mohler specifically for short answer grading. The corpus was built by collecting answers from 30 students for 21 questions posed in the Computer Science domain. The answers were graded by two human evaluators on a scale of 0 to 5 and a step size of 0.5. The scores awarded by the individual evaluators were then averaged to a single value. In order to make this corpus suitable for the proposed binary evaluation scheme, one of the possible grades was fixed as the threshold. All answers with scores less than or equal to the threshold were treated as wrong and the remaining ones as correct. To measure the performance of the system, the threshold was varied from 2 to 4 and the accuracy was computed for both feature sets as shown in Table 5.1.

Table 5.1 Performance Evaluation on SAGC

Threshold Value	Accuracy %	
	Using 113 Features	Using 57 Features
2.0	**90.8**	**90.8**
2.5	**86.5**	**86.5**
3.0	**79.0**	77.9
3.5	**78.4**	76.5
4.0	**78.3**	75.7

From Table 5.1 it can be observed that the accuracy of the system increases as the threshold decreases, which indicates that the evaluation system is lenient in nature. When 113 features were used the system performed consistently better. The second feature set which had only 57 features has achieved a comparable performance despite using only half the number of features. It has exhibited the same performance as the first feature set for threshold values of 2 and 2.5. For higher threshold values the difference in accuracy was only about 2 – 3% when compared to using 113 features.

Since a majority of the reported work involves assessment of the answer evaluation system in terms of correlation coefficients, comparison is not feasible. Anyhow, the developers of the corpus have also attempted a binary classification with a threshold of 2.5 using their best system and ten-fold cross validation. An accuracy of 92% was obtained (Mohler & Mihalcea 2009) which is slightly higher than the best accuracy value of 90.8% registered by the proposed system.

5.1.2.2 User Language Paraphrase Corpus

The ULPC consists of 1998 pairs containing target sentences / student responses and was developed as a part of the iSTART system. The corpus was evaluated by human experts along ten dimensions including lexical similarity, syntactic similarity, semantic similarity (McCarthy & McNamara 2008). Of these, the 'paraphrase quality' dimension which is the most comprehensive one has been used in the reported experiments. The scores assigned for this dimension vary from 0 to 6. Similar to the Short Answer Grading corpus, the threshold value used for translating the scores to 0 or 1 was varied between 2 to 4. Preprocessing was carried out to replace garbage answers by a default of 'No Answer' before performing classification. The accuracy values obtained for the two feature sets have been shown in Table 5.2.

Table 5.2 Performance Evaluation on ULPC

Threshold Value	Accuracy %	
	Using 113 Features	**Using 57 Features**
2.0	**77.4**	77.2
2.5	**76.8**	76.6
3.0	69.7	**70.1**
3.5	68.4	**68.5**
4.0	69.8	**69.9**

Akin to the Short Answer Grading corpus, the accuracy is highest for the least threshold value. This shows that the system has a tendency to grade more answers as correct rather than wrong. Unlike the previous case, using 57 features has recorded the best performance at higher thresholds, whereas the system with 113 features has scored well at lower thresholds. With respect to the time aspect, when performing 10-fold cross validation the system with reduced features is faster. On the whole, the system with 57 features has performed better.

The highest accuracy value obtained on the ULPC was 77.4% at a threshold of 2, which is very much higher than the best value of 63.2%, reported by Lintean et al (2010). Even at a threshold of 3 which is prescribed by the developers of the corpus, the proposed system has reached 70.1% as against 63.2% obtained by Lintean et al (2010).

From the experiments it is evident that, a Paraphrase Recognition approach is feasible for Student Answer Evaluation systems. Further the set of features selected by the Genetic Algorithm experiments on the MSRPC have been found to hold good for other corpora such as the ULPC and Short Answer Grading Corpus. This implies that the selected features are ideal candidates for building efficient Paraphrase Recognizers. As expected the

system with reduced features also performs better with respect to the classification time.

The Student Answer Evaluation system is expected to alleviate the burden of human evaluators. It will also make the evaluation process objective by eliminating biases and inter-evaluator variations. Besides complementing the role of human evaluators in traditional examinations, such a tool would be of immense use in Intelligent Tutoring Systems.

5.2 PLAGIARISM DETECTION

The Internet has facilitated instant information access and has also led to the spawning of large amounts of un-structured data, especially text. A major drawback of the easy information access made possible by the Internet is the wide spread prevalence of the phenomenon of copying and reusing information without permission. Plagiarism can be termed as the unauthorized reuse of content or ideas without giving due credit to the original authors. Plagiarism is a major threat to academics and has to be suitably addressed to ensure integrity and authenticity.

Automatic plagiarism detection systems aim to identify plagiarized content present in large repositories. This task is rendered difficult by the use of sophisticated plagiarism techniques such as paraphrasing and summarization which mask the occurrence of plagiarism. In this work, a monolingual plagiarism detection technique has been developed to tackle cases of paraphrased plagiarism. Both Sentence-level and Passage-level approaches have been investigated.

A machine learning based Paraphrase Recognizer which operates by extracting lexical, syntactic and semantic features has been used to detect plagiarism in text passages. The sentence-level Paraphrase Recognition

system described in Section 2.3 has been adapted for determining if two passages have been plagiarized. Two different approaches have been investigated: in the first, the input source and suspicious passages have been split into sentences and the original sentential paraphrase recognition system has been applied. In the second approach, the input passages have been retained as it is and various features extracted from the passages have been used to judge whether the suspicious passage is a plagiarized version of the source.

Since the input for the task of Plagiarism detection consists of passage-level text, the sentence-level Paraphrase Recognition system has been modified to handle passages. The Sentence-level processing algorithm is given in Figure 5.2.

Algorithm: Plagiarism Detection by Sentence-level processing

Input: **Pairs of Original and Candidate Suspicious passages**
Threshold T in the range (0, 1)
Output: **Yes (Plagiarized) / No (Not-Plagiarized)**
Procedure:

- Split the passage pairs into training set (two-thirds) and test set (one-thirds)
- For each pair of passages P_o and P_c:
  - Split the passages into sentences
  - For each sentence S_c in the candidate passage:
    - Determine the word overlap with each sentence of the original passage
    - Identify the sentence from the original passage – S_o having the highest overlap
    - Extract the set of 57 features from the sentence pair (Sc, So)
- Train the SVM Classifier using features extracted from the training set
- Classify the sentences extracted from the test-set passages as 'Plagiarized' or 'Not-Plagiarized'
- For each candidate suspicious passage P_c:
  - Determine the total number of plagiarized sentences - np_c
  - Compute the aggregate score by dividing np_c by number of sentences in P_c
  - If the score exceeds the given threshold T, the candidate passage is declared as a plagiarized version of the original passage P_o

Figure 5.2 Sentence level Plagiarism Detection

The source and suspicious passages were split into sentences. In order to determine the closest matching source sentence for the suspicious passage sentences, the extent of unigram overlap was computed between the sentences in both the passages. For every sentence in the suspicious passage, the source sentence which has the highest word overlap was paired with it. The best set of 57 features identified for paraphrase recognition as described in Section 2.3 were then extracted from the sentence pair. An SVM classifier was used to label the sentence pairs as positive or negative cases of paraphrases. The decisions obtained for individual sentence pairs were then aggregated by computing the percentage of paraphrases among the total number of matched pairs. If the percentage exceeded a given threshold the entire suspicious passage was declared to be plagiarized.

Another scheme termed as 'PassagePlag' which operates directly on passage level text has also been adopted. In this method, paraphrase recognition features were extracted directly from the source and suspicious passage as a whole. Out of the best set of 57 features identified for Paraphrase Recognition, the Tree Edit Distance and Triple Similarity function features have been eliminated for passage level inputs. This is due to the reason that these features are extracted from dependency parse trees which are constructed for sentences and not lengthy passages. In this case the SVM Classifier has been used to directly decide whether the passages are paraphrased and hence plagiarized.

5.2.1 Performance Evaluation

The suitability of the proposed approaches for the task of Plagiarism Detection has been investigated using four different corpora: the Webis CPC corpus which is a subset of the PAN-PC 2010 corpus, a subset of the METER corpus, Wikipedia Rewrite corpus and a subset of the Paraphrasing for Plagiarism (P4P) corpus. The Webis corpus contains a total

of 7859 pairs of source and suspicious passages out of which 4067 have been labeled as positive cases of paraphrasing and the rest as negative. The positive cases vary in length from 28 to 954 words. The corpus has been constructed by crowdsourcing on Amazon's Mechanical Turk. Volunteers were asked to paraphrase passages of text extracted from Project Gutenberg. The generated passages were reviewed to reject as non-paraphrases those cases which were exactly the same or very similar to the original. Of the remaining cases, grammatically correct versions which conveyed the same meaning as the source have only been accepted as paraphrases (Burrows et al 2012).

The METER corpus consists of 1716 text articles extracted from the UK Press Association releases and different newspapers (Clough et al 2002). Each newspaper article is a rewritten version of the corresponding Press Association source(s). The entire corpus collected over a period of twelve months has been grouped into two major domains – courts and show business. In order to reflect the extent of text reuse, each newspaper article has been manually categorized as 'Wholly Derived', 'Partially Derived' or 'Not Derived'. For the purpose of the current study only a subset of 253 articles which have a single source have been chosen similar to the approach followed by Bar et al (2012) and Sanchez-Vega et al (2010). Further the 3-way classification has been converted into two classes: Derived and Not Derived.

The Wikipedia Rewrite Corpus consists of 95 short answers collected from 19 participants for 5 different questions (Clough & Stevenson 2011). The collected answers have originally been labeled as 'Near Copy', 'Light Revision', 'Heavy Revision' and 'Non-plagiarism' depending on their similarity to the reference answer. This 4-way split has been converted into two classes: Plagiarized and Non-Plagiarized.

In order to investigate the performance of the Paraphrase Recognition system in handling various types of paraphrases, experiments were conducted using P4P corpus. The corpus consists of 847 pairs of fragments each containing less than 50 words (Barrón-Cedeño et al 2013). The corpus has been manually annotated at various levels such as: words, phrases, clauses and sentences using the paraphrase typology. Since the original paraphrase recognizer described in Section 2.3 operates at sentence level, types involving only word level or phrase level changes have been eliminated. These include all sub-types falling under morpho-lexicon and miscellaneous categories. Only 10 sub-types falling under Structural and Semantic categories have been considered for evaluation purposes.

The evaluation measures considered here are Accuracy, Precision, Recall and F-measure which are calculated as given in Equations (5.1) – (5.4) where a case of plagiarism refers to a pair of original and candidate passages. True Positive (TP) refers to a plagiarized passage being labeled as plagiarized and True Negative (TN) is a correctly identified case of non-plagiarism. False Positive (FP) refers to non-plagiarized cases labeled as plagiarized while False Negative (FN) is the vice versa.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.2}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5.3}$$

$$\text{F} - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.4}$$

The evaluation of the passage level approach – PassagePlag has been carried out using a 10-fold cross validation approach. On the other hand, since the evaluation of the sentence level approach requires the decisions made on sentence pairs to be aggregated, the corpora was partitioned into three folds of equal number of passages. Similar to the cross-validation technique, three trials were carried out, in each of which two folds were used for training and one fold was used for testing. Paraphrase Recognition Features extracted from paired sentences in the training set were fed to an SVM classifier and used to build a model. This was then used to classify the sentence pairs of the test set. The output decisions produced by the classifier for sentence pairs from the passages were aggregated to arrive at passage-level decisions. The results of the sentence-level as well as passage-level approach on the three corpora have been shown in Table 5.3.

Table 5.3 Performance of Sentence-level and passage-level approaches

Corpus	**Webis Corpus**		**METER Corpus Subset**		**Wikipedia Rewrite Corpus**	
Approach	**Sentence level**	**Passage level**	**Sentence level**	**Passage level**	**Sentence level**	**Passage level**
Accuracy %	79.1	**83.5**	78.7	**81.4**	95.8	**96.8**
Precision %	73.4	**78.7**	82.9	**86.0**	98.2	**100**
Recall %	**93.4**	93.3	**88.4**	**88.4**	**94.8**	94.7
F-measure %	82.2	**85.4**	85.6	**87.2**	96.4	**97.3**

The threshold used to arrive at the passage level decision was varied and the best results were obtained when the threshold was set at 60% for the Webis corpus and 50% for the other two corpora. In the case of the sentence-level approach, the reported values have been arrived at by averaging the results obtained in the three trials. From Table 5.3 it can be observed that the overall performance of the passage-level approach is better than that of the sentence-level approach. Precision and recall exhibit the traditional trade-off with precision being better in the passage-level approach and the sentence-level approach out-scoring in terms of recall, though marginally in the case of Wikipedia Rewrite corpus and Webis corpus. The better performance of the passage-level approach can be attributed to the reason that when passages are paraphrased, a sentence-by-sentence approach may not always be adopted as observed in Burrows et al (2012). Two sentences in the original passage may either be combined or a single sentence maybe split. Hence establishing a one-one correspondence between sentences of the source and suspicious passages becomes difficult. In the current approach, the degree of word overlap between the sentences has been used to pair the sentences for further comparison. Other alternatives such as semantic similarity can be considered, since two sentences which do not have a high word overlap may be paraphrases. When the entire passage is considered as a single entity the candidate sentence pairs need not be identified and therefore better performance is achieved.

In order to benchmark the performance of PassagePlag approach on the three different corpora, the obtained results have been compared with that of Bar et al (2012) and are tabulated in Table 5.4. From the results it can be observed that for the Webis corpus, the proposed passage level paraphrase

recognition approach – PassagePlag has lower performance than Bar et al (2012) which is the best performing system on the Webis corpus.

Table 5.4 Performance Comparison on various plagiarism corpora

Approach	Webis Corpus		METER Corpus		Wikipedia Rewrite Corpus	
	Accuracy %	F-measure %	Accuracy %	F-measure %	Accuracy %	F-measure %
PassagePlag	83.5	85.4	**81.4**	**87.2**	**96.8**	**97.3**
Bar et al (2012)	**85.3**	**86.2**	80.2	85.8	**96.8**	**97.3**

For the METER sub-corpus and Wikipedia Rewrite Corpus also, the performance of PassagePlag approach was compared with that of Bar et al (2012) which is again the best performing approach. The passage-level approach was found to exhibit better or comparable performance. Both of these corpora contain samples belonging to non-plagiarized category as well as varying degrees of plagiarism. The data was folded to two classes: Plagiarized and Non-Plagiarized.

Table 5.5 presents the statistics of true, false positives and negatives for the various corpora. The experimental results indicate that the paraphrase recognition approach used in this work reports a higher number of False Positives and fewer false negatives when compared to Bar et al's system with respect to Webis corpus. The difference is less pronounced with respect to the METER corpus and there is no difference in the case of the Wikipedia Rewrite corpus. The increased number of False Positives reported could be due to the excessive dependence on lexical features. Since the MSRPC has been found to be lenient towards paraphrases with greater word overlap, lexical features were found to be a good choice. But the annotation of the Webis corpus has been carried out to overcome this bias by specifically

labeling duplicates and near duplicates as non-paraphrases (Burrows et al 2012). In the case of the METER corpus even if the candidate passages have a considerable overlap, the presence of additional content in one passage, has led to the passages being labeled as non-plagiarized.

Table 5.5 Performance Statistics for various plagiarism corpora

Approach	Webis Corpus		METER Corpus		Wikipedia Rewrite Corpus	
	PassagePlag	Bar et al	PassagePlag	Bar et al	PassagePlag	Bar et al
True Positives	3795	3654	160	151	54	55
True Negatives	2766	3033	46	52	38	37
False Positives	1026	759	26	20	0	1
False Negatives	272	413	21	30	3	2

In addition to the above experiments, the performance of the paraphrase recognition system was assessed on various categories of fragments available in the P4P corpus. Since all the fragments classified under the Syntactic and Semantic categories are positive cases of paraphrasing, for evaluation purposes samples extracted from the corpus of 1999 negative samples created by Cordeiro et al (2007) in their work on Sentence compression have been used. For each sub-category containing Y positive samples, an equal number of negative samples N was added. The results have been presented in Table 5.6 and have been ranked based on accuracy.

Table 5.6 Performance on a Subset of the P4P corpus

Type	Description	Number of positive pairs	Accuracy %
Coordination	Change in coordinated linguistic units	210	97.8
Punctuation and format	Changes in Punctuation and format	538	97.5
Syntax/discourse structure	Syntax/discourse reorganizations	313	97.2
Sentence modality	Change in sentence modality	35	97.1
Subordination and Nesting	Changes in subordinated or nested units	597	96.4
Direct and Indirect style	Direct and Indirect style variations	36	95.8
Diathesis	Alternations such as Voice change	130	95.8
Negations	Changing the position of negation	33	95.4
Ellipsis	Omission of words or phrases	87	94.3
Semantics based	Different lexicalization of same content	340	91.6

Most of the categories exhibit a good performance with accuracy greater than 95%. The best performing categories are: Coordination, Punctuation and format where the paraphrased versions are very much similar to the original. The two categories with the lowest performance are 'Ellipsis' and 'Semantic based changes'. The reduced performance for the 'Semantic changes' fragments is due to the reason that this is the toughest category to detect as it involves considerable variation from the original. For the Ellipsis category, the omission of words or phrases as in the pair 'the long initial vowel of area' and 'area' results in less overlap and therefore reduced

performance. A study of the inputs which have been wrongly classified as non-paraphrases indicate very low lexical overlap between the original, candidate inputs as well as the presence of phrases. This is demonstrated in the following examples:

- 'it has been endeavored' and 'the attempt has been made' (Category - Semantic changes)
- 'inspire to better attentiveness' and 'excite us to greater diligence' (Category - Ellipsis)
- 'lets stop emptying our heads' and 'don't let us split hairs' (Category -Negations)

The good performance on various corpora as well as different categories of the P4P corpus, indicate the suitability of the current approach for detecting paraphrased plagiarism.

5.3 CROSS LANGUAGE FAQ ACCESS

Queries posed by an user to a FAQ database need not match exactly with the questions available in the database but may be semantically equivalent. The applicability of the UNL matching system described in Section 3.1 has been investigated for accessing a cross-language FAQ database. Since the Russian UNL Converter has been used for UNL enconversion, access to Russian language FAQ has been demonstrated. Though multi-lingual search and translation services are supported in the UNL Explorer interface (Uchida et al 2012), a notable aspect of the proposed system is that it supports sentence level queries and computes the semantic similarity between the UNL forms of the user query and the queries available in the database.

The FAQ of Ozon an online megastore which consists of 104 questions addressing various queries pertaining to online shopping has been considered. The steps in accessing the Russian FAQ using English are shown in Figure 5.3. The user's English language query has been converted to UNL form and compared with the UNL version of the 104 queries in the FAQ. The Russian answer for the matching query was fetched and then translated to English using Google Translation services and presented to the user.

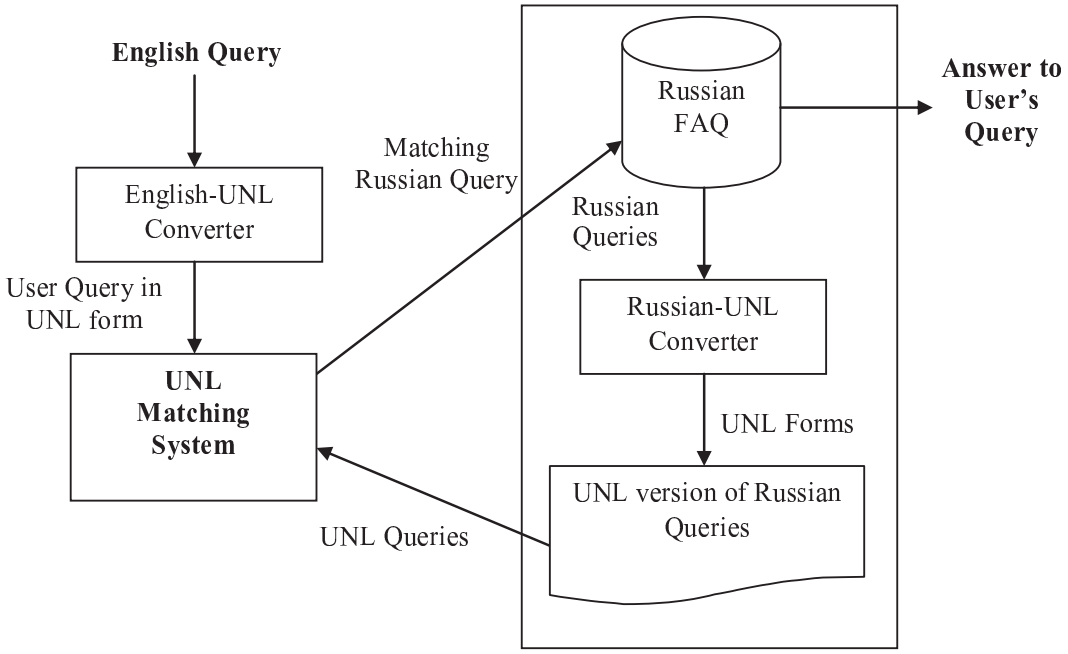

Figure 5.3 Steps in Cross language FAQ Access

Table 5.7 gives a sample set of queries, for which the Russian equivalent was identified by using the UNLPR system. The translation of the identified Russian query listed in the English version of the Ozon FAQ page has also been given. As can be seen from Table 5.7, the advantages of the UNLPR system is that it permits access to cross-language FAQs and that the query need not be an exact translation of the original query as the system checks for paraphrases and not exact matches. This application demonstrates the viability of using PR systems in cross-language FAQ access.

Table 5.7 English Queries and their answers

Query posed	Russian Equivalent	English translation of equivalent Russian query
How to fill an application for the Partner Program?	Как заполнить заявление на партнерской программе?	How to fill out an application to participate in the Partner Program?
Is it possible to combine two orders into one?	Как объединить два или несколько заказов в один?	How to combine two or more orders into one?
Are there any restrictions on the payment of the remuneration?	Какие есть ограничения по выплате Партнерского вознаграждения?	What are the restrictions on the payment of the remuneration of the Partnership?

5.4 CLUSTERING NEWS HEADLINES

The task of grouping short text units based on semantic similarity is equivalent to sentence-level paraphrase extraction and can be applied to various categories of inputs such as Document titles, news headlines and tweets. Grouping of news headlines helps to identify news items conveying the same intent and eliminate redundant ones. Besides the conventional usage of this task in the news domain, it has also been widely used to collect paraphrase corpora as news items reported by multiple press agencies serve as a rich source of paraphrases.

The PEFHC system described in Chapter 4 has been deployed for detecting similar news headlines from a collection of news items. In the first stage of the process, POS tags were assigned to news headlines, verbs were identified and initial clusters were formed. These were then merged based on the similarity between the verbs. Since several news items may describe the same or related events, verbs have been used in this stage. In the next stage,

clusters were divided based on the nouns present in the sentences. This has helped to distinguish events of the same type but involving different entities.

The PEFHC approach for News Clustering has been tested on a set of news captions collected from Google News. Thousand headlines were collected over a period of ten days. The items were spread over areas such as politics, sports, economics, daily events and medicine and belonged to Ninety nine distinct topics. The rigorous variant of the PEFHC system (With WSD, Similarity threshold = 0.2, only top 50% of sentences) was used for Clustering the news headlines. Additionally, to benchmark the performance, the Cosine similarity baseline used by Wubben et al (2009) has also been applied on the same data. This system was found to have the closest performance to the PEFHC approach in Paraphrase Extraction experiments. The results in terms of cluster purity and entropy are shown in Table 5.8.

Table 5.8 Performance Evaluation of News Clustering systems

System	Number of Clusters	Purity %	Entropy %
PEFHC Rigorous variant	162	45.8	0.8
Cosine Similarity–Threshold=0.7	133	34.7	0.9

The PEFHC system has yielded higher purity and the same entropy when compared to the Cosine similarity approach. Since all the sentences from a given reference class are present in the same cluster the purity is better. This shows that the PEFHC approach is better suited for paraphrase extraction from news headlines. A sample cluster formed by the rigorous variant of the PEFHC system and the corresponding reference cluster have been shown in Table 5.9.

Table 5.9 Sample Reference and output clusters

Reference Cluster	PEFHC rigorous variant cluster
1. Learning second language slows brain ageing 2. Speaking two languages keeps brains ageing at bay 3. Learning a second language at any age may slow the brains decline 4. Speaking two languages benefits the ageing brain 5. Being bilingual may help slow brain aging 6. Being bilingual can slow brain ageing 7. Speaking more than one language can keep your brain young 8. Learning a second language in adulthood can slow brain ageing 9. Being bilingual may slow down the brain aging process 10. Speaking second language slows mental aging 11. Brain aging is delayed if you speak more languages 12. Speaking a second language slows age related decline of the brain 13. Secret to slowing brain aging Learn two languages 14. Speaking two languages can slow down ageing of brain 15. Being bilingual may keep brain sharp in old age 16. Learning a new language at any age helps the brain 17. Speaking two languages slows down brain ageing	1. Learning second language slows brain ageing 2. Speaking two languages keeps brains ageing at bay 3. Being bilingual may help slow brain aging 4. Being bilingual can slow brain ageing 5. Speaking more than one language can keep your brain young 6. Being bilingual may slow down the brain aging process 7. Speaking second language slows mental aging 8. Speaking a second language slows age related decline of the brain 9. Speaking two languages can slow down ageing of brain 10. Speaking two languages slows down brain ageing

5.5 SUMMARY

The viability of deploying the designed Paraphrase Recognition and Extraction systems in various real world applications has been investigated. The SVM based PR system which uses lexical, syntactic and semantic features extracted from sentence pairs has performed well in both

the Student Answer Evaluation as well as the sentence-level and Passage-level Plagiarism detection tasks. The PR system has yielded good performance on various short answer and Plagiarism detection corpora. The UNL based PR system has been applied for accessing FAQs available in the Russian language. The task of grouping similar news captions has been modelled as a Paraphrase Extraction problem using the PEFHC system and has been successfully applied on news headlines collected from Google News.