ABSTRACT
Since the advent of the Internet, plagiarism has become a widespread problem in student submissions. Paraphrasing is one of the several types of plagiarism employed by students to mask the original source. In this work, we construct a sub-corpus of paraphrased sentences by extracting all lightly and heavily revised sentences from the Corpus of Plagiarized Short Answers, using modified criteria for sentences. We then apply document similarity measures on this sub-corpus and derive some interesting features of this sub-corpus. Our findings suggest that this sub-corpus is more suited for testing paraphrase detection techniques by providing sentence-level paraphrasing samples instead of the file-level classification provided in the original corpus. Additional sentence samples may also be added to this sub-corpus to achieve variety and scale.

Categories and Subject Descriptors
I.2.7 [Natural Language Processing]: Language Parsing and Understanding

General Terms
Experimentation

Keywords
plagiarism, paraphrasing, similarity measures

1. INTRODUCTION
Plagiarism in colleges and universities is on the rise according to a recent report [1] by the Pew Research Center. Earlier, Maurer et al. [2] have analyzed several surveys carried out at leading universities detailing quantitative estimates of students engaging in plagiarism. Students may attempt several types of plagiarism including, but not limited to, direct copy, modification and paraphrasing. Zahrani et al. [3] have proposed a classification of plagiarism types which includes, among others, paraphrasing, summarization, copy-paste and idea-based plagiarism.

According to the Merriam-Webster Dictionary [4], paraphrasing is defined as “a restatement of a text, passage, or work giving the meaning in another form”. From the stated definition, it is clear that the restatement of a text in different words is paraphrasing, however the extent to which words must be altered in order for a text to be considered a paraphrase is not strictly defined. Therefore, in the context of plagiarism detection, we consider two texts as paraphrased versions of each other if they are semantically equivalent but may not be syntactically identical. By semantic equivalence we imply that the essential information conveyed by the two texts is the same, however they may differ in at least one or more words. An identical approach has previously been adopted in the Microsoft Research Paraphrase Corpus [5] which is a collection of 5801 sentence pairs extracted from news sources and judged as paraphrased or not by human evaluators.

Another collection of plagiarized documents is the Corpus of Plagiarized Short Answers [6]. These short answers are written by students in answer to five computer science tasks; hence it is a valuable resource for investigating paraphrasing patterns in computer science education. There are several other corpora available, e.g., the PAN PC-09 [7] corpus; however these may not be useful for evaluating academic plagiarism due to automatic content generation as well as being not related to education.

In this work, we extract paraphrased sentence pairs from the Corpus of Plagiarized Short Answers using a slightly modified scheme adapted to sentences. We call this collection a ‘Sub-corpus of Paraphrased Sentences’ since this collection is a subset of the original corpus. We then apply document similarity measures to this sub-corpus and derive some interesting features of this sub-corpus. This sub-corpus can be used for testing paraphrasing detection techniques in student work in computer science education. Additional samples can also be added to this sub-corpus to achieve scale and variety.

The rest of the paper is organized as follows: Section 2 describes an analysis of the file-level classification of the original corpus and the methodology of building the sentence-level sub-corpus. Section 3 describes the contents of the sub-corpus, its limitations and ways to extend it. Section 4 describes an application of document similarity measures to this sub-corpus and states some features extracted from this sub-corpus. Section 5 concludes the paper and provides directions for future work.
2. CORPUS ANALYSIS

The Corpus of Plagiarised Short Answers [6] is a collection of short responses of 200–300 words that have been plagiarized by students according to the predefined instructions. It consists of a total of 95 responses to five tasks in total, with each task having 19 responses. Each response is plagiarized according to a level classified as (1) Cut or Near-copy, (2) Light Revision, (3) Heavy Revision, and (4) Non-plagiarized based on the original source which is also provided. Although the size of the corpus is small, it is a very useful resource for researchers since,

- the answers are written by human respondents in contrast to automatically generated corpora,
- being at the student level, the corpus simulates actual writing styles of students (for example, inclusion of typographical errors), and
- the tasks are related to computer science in particular and technology-related areas, in general.

Each of the 95 plagiarized files in the corpus contains a response and is classified according to the four levels mentioned above. However, a file-level classification gives little insight into which parts of the file are exactly copied, revised or are completely non-plagiarized. Hence, we conducted a sentence-level comparison by comparing each sentence of every response with each sentence in the original file of the corresponding task. Each sentence was then labeled as cut/near-copy, lightly revised, heavily revised or non-plagiarized.

2.1 Methodology

Each task in the original corpus was assigned to an evaluator, with the evaluator himself being an instructor in computer science or computer engineering. The evaluators conducted the sentence-level evaluation by first determining sentence boundaries.

<table>
<thead>
<tr>
<th>Task</th>
<th>Original</th>
<th>Light Revision</th>
<th>Heavy Revision</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>It is intended to help reuse existing code with little or no modification.</td>
<td>It was intended to allow existing code to be used again with minimal or no alteration.</td>
<td>The peropos of inheritance in object oriented programming is to minimize the reuse of existing code without modification.</td>
</tr>
<tr>
<td>B</td>
<td>A page that is linked to by many pages with high PageRank receives a high rank itself.</td>
<td>...if a webpage is linked to others that have a high ranking, then it too will receive a high rank.</td>
<td>The algorithm basically works like a popularity contest - if your site is linked to by popular websites, then your site is considered more popular.</td>
</tr>
<tr>
<td>C</td>
<td>The order in which the terms appear in the document is lost in the vector space representation.</td>
<td>and the order that the terms appear in the document is not represented in the vector space model.</td>
<td>the order of words does not matter</td>
</tr>
<tr>
<td>D</td>
<td>It is often used to compute posterior probabilities given observations.</td>
<td>The theorem is often used when we have observations and wish to compute posterior probabilities</td>
<td>It is mainly used to calculate the probability of one event’s outcome given that the previous event has happened</td>
</tr>
<tr>
<td>E</td>
<td>The method takes much less time than naive methods</td>
<td>This is a much quicker method than other more naive method</td>
<td>The advantage being the less time consumption in comparison to other amateur methods</td>
</tr>
</tbody>
</table>

Table 1: Lightly and Heavily Revised Sentence Examples from the Corpus of Plagiarised Short Answers [6]

2.1.1 Sentence Boundaries

The evaluators adopted the following criteria for identifying sentence boundaries in the original corpus for the purpose of sentence-level comparison:

1. A period or dot (‘.’) is considered as a boundary between two sentences, except in case of abbreviations (for example i.e.).

2. In case of numbered lists, each list item is considered as a sentence even though it may be a phrase or part of a larger sentence. In such cases a comma, semicolon or a colon is considered as a sentence boundary. For example the phrase ‘The vector space model has the following limitations:’ was considered as a sentence, although it ends with a colon.

3. Technical texts written using English alphabet but not in English language (such as short computer programs appearing in Task A, or mathematical formulae in Task D) are treated as one sentence in their entirety, although they may span multiple lines.

2.1.2 Sentence Classification

The evaluators compared each sentence in the 19 files with each sentence in the original file and classified these as one of the four types: (a) cut or near-copy, (b) lightly revised, (c) heavily revised or (d) non-plagiarised according to the following criteria:

1. Cut or Near-Copy: A target sentence is considered as a near-copy of a source sentence if the two sentences contain exactly the same words and in the same order.

   Sentences were also considered near-copy if they differed in non-significant parts of speech, for example articles (such as replacing ‘a’ by ‘the’), and in some cases prepositions.

2. Lightly Revised: A target sentence is considered as a lightly revised version of a source sentence if the two...
The total number of all sentence types.

The cut file category contains the highest relative percentage of cut sentences. Moreover, the information conveyed by the cut file category is almost the same. The cut sentences differ in a few words, however the information conveyed by the two sentences is almost the same. For example, in Table 1 (Task C), the word ‘lost’ is replaced by ‘not represented’ in the lightly revised version.

3. Heavily Revised: A target sentence is considered as a heavily revised version of a source sentence if the two sentences differ in a significant number of words and/or their order, however the essential information conveyed by the two sentences is the same. The idea of what constitutes essential information varies from one evaluator to another. Hence we may infer that each file-type may contain more than one type of sentence.

4. Non-plagiarised: A target sentence is considered as a non-plagiarised sentence if it did not match all source sentences in the number and order of words. Furthermore, the information conveyed by the non-plagiarised sentence also did not match that of any sentence in the source file.

Some sentences were considered borderline by the evaluators especially between the lightly revised and the heavily revised categories. In this case a second evaluator reviewed all of these classifications and reassigned sentence categories in order to be consistent. Table 1 lists a sample of lightly and heavily revised sentences from each task. The levels of sentence classification proposed here are similar to the file level classifications of the actual corpus, however they are at the sentence level and give us a finer level of similarity as compared to a file-level classification.

2.2 Results

Task A (File Type) Task B (File Type) Task C (File Type) Task D (File Type) Task E (File Type)

<table>
<thead>
<tr>
<th>Type</th>
<th>Task A</th>
<th>Task B</th>
<th>Task C</th>
<th>Task D</th>
<th>Task E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C L H N</td>
<td>C L H N</td>
<td>C L H N</td>
<td>C L H N</td>
<td>C L H N</td>
</tr>
<tr>
<td>C</td>
<td>28 13 1 0</td>
<td>11 10 2 0</td>
<td>21 22 13 0</td>
<td>33 13 22 0</td>
<td>54 9 3 0</td>
</tr>
<tr>
<td>L</td>
<td>7 17 21 0</td>
<td>0 14 2 1</td>
<td>2 32 30 1</td>
<td>6 36 17 0</td>
<td>0 23 14 2</td>
</tr>
<tr>
<td>H</td>
<td>0 3 10 2</td>
<td>1 0 22 22</td>
<td>0 1 19 1</td>
<td>1 7 6 2</td>
<td>0 3 15 1</td>
</tr>
<tr>
<td>N</td>
<td>4 3 8 101</td>
<td>24 0 6 58</td>
<td>14 7 8 56</td>
<td>22 8 13 43</td>
<td>13 5 7 71</td>
</tr>
</tbody>
</table>

Table 2: Sentence classification according to file types, where C=Cut/Near-Copy, L=Light, H=Heavy and N=Non-Plagiarised

Figure 1 shows the percentage of sentence types in each file type for all tasks. From this figure, it can be observed that on average, for all tasks:

1. The non-plagiarised files contain the highest relative percentage of non-plagiarised sentences. Furthermore, there are very few cut, lightly revised and heavily revised sentences in this category.

2. The heavy revision files contain a mix of heavily revised and lightly revised sentences as a majority of the marked sentences. They also contain some cut and non-plagiarised sentence types.

3. The light revision files contain a significant number of cut or near-copy sentences, suggesting that the participants in the original corpus creation considered a near-copy to be a light revision.

4. The cut file category contains the highest relative percentage of cut sentences. Moreover, this category also contains some percentage of non-plagiarised sentences.

The above observations suggest that the file categories of cut, light revision, heavy revision and non-plagiarised each may not contain a majority of the corresponding sentence types. Rather, we observe that for some categories (light and heavy), there is a mix of both sentence types with significant number of other sentence types. These findings suggest that it would be worthwhile to ignore file-level markings and compare sentence pairs on their own.
3. SUB-CORPUS OF PARAPHRASED SENTENCES

For the purposes of paraphrase detection, we may consider only the revised sentences (i.e., lightly and heavily revised sentences only) for comparison, as these are modified versions of the original sentences. Hence, we suggest to create a ‘Sub-Corpus of Paraphrased Sentences’. This sub-corpus is a subset of the Corpus of Plagiarised Short Answers [6] with each file corresponding to a sentence from the original file and the corresponding revised (i.e., lightly and heavily revised) sentences only.

3.1 Description

The sub-corpus consists of a total of 101 files, with each file corresponding to an original sentence from each task. The number of sentences from the original files for each task is as follows:

- Task A: 14 original sentences,
- Task B: 24 original sentences,
- Task C: 16 original sentences,
- Task D: 19 original sentences,
- Task E: 28 original sentences.

Each file contains the original sentence taken from the original file marked by an ‘O:’, followed by all lightly revised sentences marked as ‘L1, L2, L3…’; and heavily revised sentences marked as ‘H1, H2, H3…’. Figure 2 shows a sample file from the sub-corpus.

3.2 Limitations

Although the sub-corpus contains carefully evaluated sentences, the evaluation is not exactly defined. There may always be some sentences which may be judged as lightly revised by some evaluators and as heavily revised by others. However, for the purpose of paraphrase detection this fine distinction may not matter, as we are interested in the collection of all revised sentences.

Another shortcoming of the corpus is the evaluation of a sentence that is the result of sentence joining. In particular, if in the original corpus, two sentences S1 and S2 are joined together to form a larger revised sentence, it may be difficult for an evaluator to determine whether S1 or S2 forms the central piece of information in the newly formed sentence. In this case an evaluator may consider the newly formed sentence to be a lightly or a heavily revised version of S1, S2 or both. Similarly, in case of sentence splitting, an evaluator may link two or more shorter, split sentences as revised versions of an original larger sentence; however the split sentences may contain partial information from the source sentence only.

3.3 Extension

The sub-corpus is a small and selected version of the original Corpus of Plagiarised Short Answers. Since the size of the original corpus is small [6], therefore the size of the sub-corpus is even smaller.

One way to achieve an extension in the size of the sub-corpus is to compare all sentence pairs in all files across the entire corpus. This may add to the size of the corpus by including sentence pairs that have not been intentionally plagiarised, but are similar versions of each other due to topic similarity. For example, it is possible that two non-plagiarised sentences in two different files may be somewhat similar to each other due to choice of words or the information contained. However, since the non-plagiarised sentences were not plagiarised in the first place, any similarity found between two non-plagiarised sentences would be purely incidental. The resulting collection would then more appropriately be called ‘Sub-corpus of Similar Sentences’.

4. SIMILARITY MEASURES

Clough and Stevenson [6] have given similarity measures for the original corpus based on N-gram overlap and longest common subsequence. Chong et Al. [8] have applied natural language processing techniques to the Corpus of Plagiarised Short Answers with the objective of distinguishing between file categories. For the proposed sub-corpus, we apply N-gram and longest common subsequence based similarity measures to extract some properties of the sentences.

4.1 N-gram Similarity

N-gram based similarity measures [9], [10] can be used to evaluate similarity between two documents. However N-gram based containment measures [11], especially for higher values of N are more suited to document similarity as opposed to similarity measures for sentences. We choose a value of $N = 1$ (unigram) for sentence comparisons due to their shorter length. Furthermore, we normalize all common unigrams by the the length of the original and the revised sentences, respectively.

$$c_1(O, R) = \frac{|S(O) \cap S(R)|}{|S(O)|}$$

$$c_2(R, O) = \frac{|S(O) \cap S(R)|}{|S(R)|}$$

where $O$ = original sentence, $R$ = revised sentence, and $S(O), S(R)$ represent the number of unigrams in original and revised sentences, respectively.

We then evaluate the measure ($c_1, c_2$) as an ordered pair for each sentence in each task. Here $c_1$ represents the containment measure normalized w.r.t. original sentence and $c_2$ represents the containment measure normalized w.r.t. the revised sentence. For example, the sentence H1: This leads to... shown in Figure 2 has ($c_1, c_2$) = (0.846, 0.480) = (84.6%, 48.0%).

Figure 3 shows a plot of ($c_1, c_2$) for each sentence type for Task E. Here, heavily revised sentences are represented by a triangle (△) and the lightly revised sentences by a box (□). It can be seen that most of the heavily revised sen-
sentences lie in the lower left part of the graph, i.e., $(c_1, c_2)$ for heavily revised sentences is less than or equal to (60%, 60%), where the point (60%, 60%) represents a possible dividing point between lightly revised and heavily revised sentences. Likewise, most of the lightly revised sentences have $(c_1, c_2)$ greater than (60%, 60%).

Sentences in the lower right part of the graph in Figure 3 represent sentences for which the value of $c_1$ is higher than 60% while that of $c_2$ is lower than 60%. These sentences may have been paraphrased in a way such that the length of the revised sentence is greater than the length of the original sentence. This could be due to sentence joining or insertion of new information into the sentence. Likewise, sentences in the upper right part of the graph may represent sentence splitting, deletion or summarization.

It may be relevant to mention here that the value of (60%, 60%) simply presents a threshold separating most of the lightly revised sentences from most of the heavily revised sentences for Task E only. Similarity measures for other tasks may produce similar results albeit with different threshold values.

### 4.2 Longest Common Subsequence

Another measure of document similarity is the longest common subsequence similarity (less) measure. We computed the length of the longest common subsequence for each sentence pair in Task E. Similar to N-gram similarity, we normalized this length with the original and revised sentence lengths. The result is plotted in a graph in Fig. 4 which shows $(l_1, l_2)$, where

$$\begin{align*}
l_1 &= \text{length(lcs)/length(original)} \\
l_2 &= \text{length(lcs)/length(revised)}
\end{align*}$$

The results are similar to those Fig 3, however, the heavily revised sentences cluster for a lower value of $(l_1, l_2)$. These results confirm our observations presented earlier.

### 4.3 Observations and Discussion

From the graphs in Figure 3 and 4, we observe that most of the highly revised sentences have a lower containment value with respect to both the original and the revised sentences as compared to the lightly revised sentences. We also note that some sentences may have a higher containment measure with respect to the original sentence as compared to the revised sentence suggesting sentence splitting, summarization or deletion. A similar observation can be made for sentence joining, expansion or insertion. For example, the sentence ‘H1: This leads to...’ represents an example of addition of information since $(c_1, c_2) = (0.84, 0.48)$.

In addition to the above, we note a very basic but important observation: That both lightly and heavily revised sentences have some degree of overlap with respect to the original sentences. One possible reason for this could be that in Computer Science Education in particular (and technical education in general), technical terms cannot be revised or replaced. For example a student cannot replace ‘Dynamic Programming’ with ‘Movable Planning’ or ‘recursive algorithm’ with ‘repetitive method’. Hence, we may expect some degree of string overlap in student work in CS Education which may prompt for the use of appropriate methods in detecting similarity in academic work. Another possible explanation for this overlap could be use of unsophisticated revision methods by students. This issue requires further investigation by analyzing student responses as well as by increasing the size of the sub-corpus. Although these findings have been stated in the context of Task E, they can be generalized by including results from other tasks.

### 5. CONCLUSION AND FUTURE WORK

In this work we constructed a sub-corpus of paraphrased sentences by extracting all lightly and heavily revised sentences from the Corpus of Plagiarised Short Answers. One of our findings was that our sentence-level similarity evaluation offers a finer level of similarity as compared to the file-level categorization of the original corpus. We also applied document similarity-based measures to observe containment patterns of this sub-corpus.

Several directions of future work can be undertaken. A basic application of this sub-corpus could be testing of paraphrase detection techniques, especially in student work. In order to achieve this, the size of the sub-corpus could be increased by adding further sentences from within and outside of the original corpus to include both scale and variety. Moreover, the sub-corpus can also be used to analyze patterns of plagiarism within student work in Computer Science in particular, and technology related areas in general.
6. ACKNOWLEDGMENTS

The authors would like to acknowledge the support provided by the Deanship of Scientific Research at King Fahd University of Petroleum & Minerals (KFUPM) under Research Grant RG-113.

7. REFERENCES


