A Plagiarism Case Study

By Ted Pedersen
Department of Computer Science
University of Minnesota Duluth
Duluth, MN 55812
tpederse@d.umn.edu

April 19, 2001

The following document presents a case study that identifies various forms of plagiarism that occurred during the course of a project in a graduate-level Computer Science class. It begins with a general discussion of plagiarism and continues with numerous examples drawn from the submitted projects.

Plagiarism

Plagiarism occurs whenever you use someone else's work without giving them proper acknowledgement. The whole of academic life and scholarly activity is based on the presumption that we will give credit to those who provide us with helpful ideas that we can write about, or perhaps even upon which we build ideas of our own. We rarely invent something so new or come up with an idea so revolutionary that we do not owe considerable intellectual debts to others. We must be mindful of these debts and carefully repay them. The remarkable thing is that we can do so simply by giving credit where credit is due.

When someone publishes a scholarly article or a book, or makes software available to others via the Internet, they are rarely motivated by a desire to make money. Instead, they often simply hope that they will make a contribution to the intellectual life of their field of study. However, it should be clearly understood that this is not purely an act of altruism on their part. In exchange, they are hoping for recognition. If a paper, idea, program, or any other intellectual product has influence, the author hopes that their name will live on with it.

Charles Darwin and the Theory of Evolution. Albert Einstein and Relativity. Sigmund Freud and Psychoanalysis. These great names of science are in no danger of being forgotten. However, not all scholars, writers, or artists are so lucky. Many labor in relative obscurity and run the risk of being forgotten, perhaps due to working in a highly specialized field or on very narrowly defined problems. Thus, if you happen to draw upon their work in some way, either directly or indirectly, it is your absolute obligation to see to it that they get the credit they deserve. As you are writing a paper, imagine that it is the sole remaining link to the ideas and sources that you have drawn upon. Will a reader know where the ideas in your paper truly came from? Or will you exterminate the memory of someone whose only offense was to make knowledge available to you? We share a collective responsibility to preserve our intellectual heritage, and those of us that wish to lay claim to having a university education must live up to this at all times.

Case Study

During a graduate-level Computer Science class students were required to work in groups of three or four and produce a written report that summarized several previous approaches to a problem in Natural Language Processing. In addition, the teams were required to implement one such method as a computer program. Both aspects of this project led to clear cases of plagiarism.

Plagiarizing Software

Plagiarism is most often thought of in terms of written work. However, any intellectual product that you draw ideas or information from for a project must be clearly acknowledged, otherwise you are plagiarizing. This includes the images, drawings, and diagrams that you find in a text. It includes video or tape-recorded materials. It also includes computer software. It is increasingly common that source code is published along with articles, or is made available via the Internet. The authors of this code are happy that you would like to us it, but they expect to receive clear acknowledgement!

Two kinds of plagiarism occurred in the class project with regards to software. The first involved a portion of code that was shared among multiple teams and originated in an email message found in the archive of a mailing list on the Internet. No team acknowledged that they were not the original authors of this code, nor did any team give credit to the original source. This is plagiarism. At a minimum, the sender of the email message and the location of the archive should have been clearly indicated in the source code comments of the team. The following is an example of how that could be done:

```
/*
The following is a sentence boundary detection algorithm that was posted
to the corpora mailing list by Tony Rose. The message containing the source code
was downloaded from the following archive:

http://www.comp.lancs.ac.uk/computing/research/ucrel/public/1419.html
*/
```

The inclusion of such a comment relieves you of an intellectual debt, and it also makes your code more useful in that it provides additional information to future users.

The second is perhaps a less obvious situation, but it is still plagiarism. The teams were required to write a program in the Perl programming language. They all decided to use a published C source code implementation of an algorithm as a starting point in their own work. This fact alone requires that an acknowledgement be given to the original authors of the source code, since that material is very clearly aiding and informing the work that the students submitted. It is very likely that the teams would not have been able to implement the algorithm without the assistance of the published C source code, so the failure to acknowledge this fact does not give appropriate credit to what was a very

significant source of information. An appropriate comment in a team's source code might look like this:

```
/*
The following is a Perl implementation of a C algorithm for sentence
alignment that was published in :

Gale, William A. and Kenneth W. Church. 1993. A program for aligning
sentences in bilingual corpora. Computational Linguistics, 19:75-102.

This code is also available for download from Ken Church's web page:
http://www.research.att.com/~kwc/publications.html
*/
```

Plagiarism in Writing:

Copying text that someone else has written without giving proper acknowledgement is the most common form of plagiarism. This often occurs when students attempt to write a paper about a topic they do not fully understand or where they have not done sufficient background research. In such a situation, the student typically sits down with a few materials from different authors and painfully weaves a web of overly close paraphrases and verbatim copying into what they hope will appear to be an original work.

Unfortunately this is quite transparent to someone who knows the background material. Rather than attempting to blindly cut and paste a paper together from multiple sources, a student must first master the subject matter before even starting to write. ***If you find yourself unable to write a few paragraphs about a topic without resorting to copying material from other sources, you are just not ready to write that paper.*** Accept that fact, stop plagiarizing, and do more background work. Come to understand the material well enough so you can sit down and write from an informed and comfortable point of view.

Once you start writing the paper, here is a simple rule of thumb*: **If you copy 3 or more words verbatim from another source, you must enclose that text in quotes and immediately reference the source.*** Including a reference to a paper in your bibliography does not give you license to copy verbatim from that paper! Every phrase, sentence or paragraph that you copy from another source must be enclosed in quotations and you must give immediate attribution to the original source. The reader of your paper must clearly understand that you have taken this material from another source, and they must be given enough information about that source so that they can locate the original material if they wish.

This rule of thumb seems rather daunting to some, so they think that if they rearrange a few words, put in a few extra sentences, then they have freed themselves from the need to properly acknowledge the original source. This is absolutely false. ***Simply rearranging the original text and/or altering a few word choices does not suddenly make you the author of an original work and free you of your intellectual debt.***

When told that a few word substitutions are not sufficient, a student may then attempt to reorder or paraphrase the original material. A typical approach is for the student to read a paragraph from an original source, ponder that for a moment, and then rearrange the phrases or construct a relatively close paraphrase. Then they move on to the next paragraph in the source material, believing that they have put the ideas in their own words. This is absolutely false. ***Simply paraphrasing an original text does not suddenly make you the author of an original work and free you of your intellectual debts.***

What follows are a series of examples of plagiarism taken from submitted class projects. The original source text is listed first, followed by one or more examples from the student teams. I have found examples of verbatim copying, word substitution, phrase rearrangement, and overly close paraphrases to clearly show the wide range of ways that plagiarism can occur. However, this list is not exhaustive, simply an illustration in the hopes that it will clarify what constitutes plagiarism. In my examples, student teams will be referred to with the letters A-F. The original sources will be referred to by the following abbreviations:

FSNLP: Manning, Christopher and Hinrich Schutze. 1999. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA.

Gale and Church: Gale, William A. and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. Computational Linguistics, 19:75-102.

Bisson and Fluhr: Bisson, Frederique and Christian Fluhr. 2000. Sentence alignment in bilingual corpora based on crosslingual querying. Appears in the Proceedings of RIAO (Recherche d'Informations Assistee par Ordinateur), Paris, France.

From page 467 of FSNLP: …but it is also a first step in using multilingual corpora as knowledge sources in other domains, such as for word sense disambiguation, or multilingual information retrieval.

From Team A: Sentence alignment is a first step in using multilingual corpora as knowledge sources in other domains such as word sense disambiguation or cross—language information retrieval.

From Team B: Text Alignment is also the first step in using multilingual corpora as a knowledge source in other domains. These might include word sense disambiguation or multilingual information retrieval.

[*Despite the minor word changes, these are clearly cases where the report author simply copied large portions of the original text and has plagiarized.]*

Here is an example of how attribution could have been handled.

Manning and Schutze (1999) characterize sentence alignment as “a first step in using multilingual corpora as knowledge sources in other domains, such as for word sense disambiguation, or multilingual information retrieval.”

[*Now it is clear that the insight that sentence alignment is useful in other domains is due to Manning and Schutze, and not the submitting team.]*

From page 476 of FSNLP: The method used is to construct a dot-plot. The source and translated text are concatenated and then a square graph is made with this text on both axes. A dot is placed at (x,y) whenever there is a match between positions x and y in the concatenated text.

From Team A: The method constructs a dot-plot. The source and translated texts are concatenated and a square graph is made with this text on the coordinate axes. A dot is placed at (x,y) whenever there is a match between positions x and y in the concatenated text.

From Team C: The method consists of building a dot-plot, i.e, the source and translated text are concatenated and then a square graph is made with this text on both axes. A dot is placed at (x,y) when there is a match.

[Minor word changes do not suddenly bestow ownership of this text to the student teams. They have plagiarized.]

From page 473 of FSNLP: The cost above is then determined in terms of a distance measure between a list of sentences in one language and a list in the other. The distance measure d compares the difference in the sum of the lengths of the sentences in the two lists to the mean and variance of the whole corpus.

From Team A: The cost is then determined in terms of a distance measure between a list of sentences in one language and a list in the other. The distance measure d compares the difference in the sum of the lengths of the sentences in the two lists to the mean and variance of the whole corpus.

[Minor word change example again. Still plagiarism.]

From Gale and Church: The program makes use of the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. A probabilistic score is assigned to each proposed pair of sentences, based on the ratio of lengths of the two sentences (in characters) and the variance of this ratio. This probabilistic score is used in a dynamic programming framework in order to find the maximum likelihood alignment of sentences.

From Team C: Our program uses the fact that longer sentences tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. A probabilistic score is assigned to each pair of proposed sentence pairs, based on the ratio of lengths of the two sentences (in characters) and the variance of this ratio. The probabilistic score is used in a dynamic programming framework in order to find the maximum likelihood alignment of sentences.

From Team B: The basic idea is that shorter sentences tend to be translated into shorter sentences and longer sentences into longer ones. The length of sentences is measured in terms of the number of characters. Sentence pairs are given probabilistic scores, based on the ratio of their lengths and also the variance of this ratio. A dynamic programming framework is then used to find the maximum likelihood alignment.

From Team D: This approach is based on the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. Various possible alignments are considered at every step and the best is chosen. To make the decision a probabilistic score is then assigned to each proposed correspondence of sentences, based on the ratio of lengths of the two sentences (in characters) and the variance of this ratio. This probabilistic score is used in a dynamic programming framework to find the maximum likelihood alignment of sentences.

[*We see here varying degrees of word substitution, phrase reordering, and verbatim copying. However, in no case do the changes rise to the level of creating an original work. When compared to the source it is clear that the team authors owe a great intellectual debt to Gale and Church. Since this debt is not acknowledged this is plagiarism.]*

From page 474, FSNLP: So, in essence, we are trying to align beads so that the length of the sentences from the two languages in each bead are as similar as possible. The method performs well (at least on related languages like English, French, and German).

From Team D: This method in essence tries to align beads so that the length of the sentences from the two languages in each bead are as similar as possible. This method works best on 1:1 alignments, and works well at least on related languages like English, French, and German, …

[*Some rearrangement and variation in word choice, but when compared to the original the plagiarism is apparent. Remember that an overly close paraphrase is just as much plagiarism as is verbatim copying of the text.]*

From page 467 FSNLP: Text alignment is an almost obligatory first step for making use of multilingual text corpora.

From Team A: A significant and almost obligatory first step towards this study is sentence alignment.

[This is plagiarism since the phrase "almost obligatory first step" has been copied without acknowledgement. This short phrase expresses the view of the original authors, not that of the student team, so this fact must be acknowledged. Remember the 3-word rule of them. If you copy 3 or more words verbatim you must acknowledge the source!]

From Gale and Church: Each character in one language gives rise to a random number of characters in the other language. These random variables are independent and identically distributed with a normal distribution. The model is then specified by the mean and standard deviation of the distribution. Let c be the expected number of characters in language 2 per character in language 1, and s^2 be the variance of the number of characters in language 2 per character in language 1. Then the expected number of characters in the sentences translating a group of sentences of length l1 in language 1 is l1c with variance l1s^2.

From Team E: Assume that each character in one language L1 gives rise to a random number of characters in the other language, L2. Also, assume the random variables are distributed with a normal distribution. Suppose in this distribution, c is the expected number of characters in L2 per character in L1, and s^2 is the variance of the number of characters in L2 per character in L1.

From Team D: The model is then specified by the mean and standard deviation of the distribution. Let c be the expected number of characters in language 2 per character in language 1, and s^2 be the variance of the number of characters in language 2 per character in language 1. Then the expected number of characters in the sentences translating a group of sentences of length l1 in language 1 is l1c with variance l1s^2.

[Team D has copied verbatim, while Team E has introduced a few changes in wording and notation. However, both have plagiarized. Even with the small changes made by Team E, the presentation is based entirely on the original source, yet they have not acknowledged this fact.]

From page 471-472 of FSNLP: The method of Gale and Church (1991;1993) depends simply on the length of the source and translation sentences measured in characters. The hypothesis is that longer sentences in one language should correspond to longer sentences in the other language.

From Gale and Church: The program uses the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. A probabilistic score is assigned to

each proposed correspondence of sentences, based on the ratio of the two sentences (in characters) and the variance of the ratio.

From Team E: The method depends on the length of source and translation sentences measured in characters. The hypothesis is that longer sentences in one language should correspond to longer sentences in the other language and shorter sentences tend to be translated into shorter sentences. Based on the ratio of the lengths of the two sentences, a probabilistic score is assigned to each proposed correspondence.

[Team E has blended text from two sources. However, a collage of this sort does not constitute the creation of an original work and the failure to acknowledge the two original sources is plagiarism.]

From Team E: Automatic sentence alignment methods typically face two kinds of difficulties. The problems are that of robustness and accuracy.

From Team A: Automatic sentence alignment methods normally face two major types of difficulties. The first one is robustness…[text removed] …The other problem is that of accuracy.

[Teams A and E appear to think very much alike on this issue. My assumption is that they have plagiarized the same source, although I have not located it as yet.]

From page 473 of FSNLP: Figure 13.3

From Team C: a scanned image of Figure 13.3 in their report, without acknowledgement.

[You must acknowledge the source of any figures or diagrams you use. Failure to do so creates the impression that you created these yourself and is another form of plagiarism.]

From Team F: This algorithm is based on the bivariate population model that is defined as follows: “If for every measurement of a variable X we know a corresponding value of a second variable Y, the resulting set of pairs of variates is called a bivariate population.”

[Team F has the right idea. However, there should be a citation or footnote immediately after this quote that tells explicitly the source of this quote. An earlier citation suggests that where the quote may have come from, but each quote must be separately referenced, even if they all come from the same source]

From Bisson and Fluhr: The statistical model provides the user with a list of documents sorted according to their relevance. The SPIRIT model differs from the vector space

model because it assigns a weight to each database word according to its discriminating power, but does not assign a weight to each word in each document.

From Team G: The statistical model provides the user with a list of documents sorted according to their relevance. The SPIRIT model differs from the vector space model because it assigns a weight to each database word according to its discriminating power, but does not assign a weight to each word in each document.

From Team A: The statistical model provides the user with a list of documents sorted according to their relevance. The SPIRIT model differs from one other well known model, vix., the vector space model, in that it assigns a weight to each database word according to its discriminating power, but does not assign a weight to each word in each document.

[Team G has copied verbatim, while Team A has done a small bit of re-wording. Both are clear cases of plagiarism.]

From Bisson and Fluhr: Crosslingual querying based on bilingual reformulation and sentence alignment are similar problems. In both cases, we must compute the proximity between two texts that are in two different languages. The main difference is that, in information retrieval, the proximity value refers to the semantic intersection between the reference text (query or text selected for dynamic hypertext) and texts stored in the database.

From Team G: Crosslingual querying based on bilingual reformulation and sentence alignment are similar problems. In both cases, we must compute the proximity between two texts that are in two different languages. The proximity between two sentences is computed considering both the intersection and also the portions missing from each sentence.

From Team A: The crosslingual querying based on bilingual reformulation and sentence alignment are similar problems. Both require to compute the proximity between two texts that are in two different languages.

[A bit of rewriting, but still plagiarism.]
