While You Were Sleeping

Project Idea and Motivation:

In a university context, multiple students who attend the same lecture differ in how they take notes and organize them. A student who misses a lecture (say he was sick) will seek to understand the missed course material from a peer's notes. Just as two different people have different perceptions of an event they view simultaneously, two peers in a lecture will have different notes. Therefore, this peer will have different note-taking style and would have recorded different information than what the sick student would have, had he attended the lecture. This could be due to various factors, such as differing extents of prior knowledge, difficulty in understanding a certain concepts more than others (and therefore having to ask a peer sitting in the next seat for clarification), varying perspectives of the same lecture and so on. Therefore, the sick student is likely to miss out on important information that he may not already know. Furthermore, he will not realize this until it is too late.

The idea behind this project is to develop a tool that would compare notes from a set of students. From these notes, the tool derives one topic map that most completely describes the content of that lecture. Multiple set of notes are used for this because, it is very improbable that any one set of notes contains all the information about that day's lecture as discussed above. Reading the topic map gives the sick student a more complete interpretation of that day's lecture, driving the amount of information he misses to a minimum.

New contributions and some innovative thinking in the field of Artificial Intelligence (A.I.) can pave the way to writing a tool that analyzes multiple streams of text (such as multiple sets of notes from the same lecture). The tool then extracts an accurate mapping of the lecture. There are cases where structurally separated text could or could not be describing the same idea with equal probability. Therefore, it may be necessary to implement methods of approximation such as timestamps in the document to make the notes ultimately more structured. Therefore the tool can parse the notes more accurately. However, this should only be used to assist the Natural Language Processing (NLP) engine in differentiating cases that are equally probable. Notes that are partitioned by chronological arguments will not be consistent with notes that are partitioned using spatial arguments. This is because it is common for students to continue writing notes even after the discussion of one idea has been finished because it is uncommon for a professor to leave students with precious lecture time to reflect on what they just said. Students may also revisit an idea which was discussed at a previous time to annotate their notes. Thus, a separation of notes based solely on their timestamps will not accurately match a separation of notes based on what ideas they describe. It is also important to note that each lecture session cannot be considered to be talking about a new topic as a single large topic could span the course of multiple lecture sessions.

Challenges:

1. English:

   English is not a standard language. There are many stylistic variants, all of which represent the same idea. Spelling errors make topic detection further difficult. The converse

is also true: in many occasions, the same word can have multiple meanings. In order to tackle a problem such as this, it becomes important to consider spatial separation of notes to properly map which regions of text correspond to which idea. This problem is more specific to the problem domain. It necessitates context mappings. For example, certain words could have a generalized meaning in colloquial speech and completely different meaning in a subject domain (for example, "clutch" has multiple meanings in vernacular English as well as domain specific meanings in Biology and automotive engineering as demonstrated in http://dictionary.reference.com/browse/clutch). Such multiple meanings necessitate that the tool discern the context of usage of a word, so as to be able to deduce the intended definition.

2. Scope and Tagging:

   The scope of class notes is hardly defined by the duration of each lecture. The course may require that multiple lectures be spent explaining one concept. Therefore, we need a way to tag similar concepts that are identified by different timestamps. This has high relevance to this problem domain and can be rectified with context mappings.

3. Determining the Minimum Sample Size

   How many sets of notes do we have to compare?

   Is this a function of the number of students in the class or is this a constant?

   Can this be done instead, with human editors amending the generated topic maps? Are human editors necessary even if enough notes are compared?

4. Reducing noise:

   Noise can be defined as spots in data which are not as frequent as the rest. However, this is also true of the names of concepts, which are not noise. Therefore, noise reduction is a very tricky feature. This is also specific to this problem domain. One way to tackle the problem is to check for the percentage occurrence of the nodes in question in the total of all notes available for analysis.

What's been done?

Büchler et al. discuss reading into a document and describing as a list of ordered pairs of nodes. This list describes a trail of arguments leading from the beginning to the end of the document (). Liu et al. discuss learning using data mining techniques on web pages which contain information about a specific topic. Their project is aimed at learning about a specific topic by data mining related web pages (). Zaslavsky et al. discuss methods of detecting similarities between different editions of literary works, which has similar objectives to this project (). This publication and others such as the MOSS system help detect similarities between the various input streams, while publications like Büchler et al. discuss how to structure a document. This and the publication by Liu et al. have ideas that may help with this project, but are ultimately dissimilar to this project on the whole. Specifically, the publication by Lui et. al is not very useful because their work has constraints which dictate what the tool can learn (). The publication by Büchler et al. especially will not work for this project because their project is better fitted to record a

discourse or a persuasive argument, which leads the listener through nodes of arguments (). A university lecture on the other hand is a presentation of information, which requires a tool engineered for information storage and organization.

Milestones and Timeline:

Deadline	Milestone
November 15	Produce a topic map from 1 set of notes in English about an event. Tag each node with possible keywords to allow for future expansion. Refine the set of keywords to include only those that are essential for describing a specific node.
December 1	Timestamp the map and alter the tool to consider the timestamp as assistance.
January 15	Derive a topic map from a set of notes written by a student.
February 28	Be able to accommodate a second set of notes and compare the two to deduce a more accurate map. Attempt noise reduction.
March 31	Publication deadline.
If there's more time	Add 5 sets of notes Add multiple sets of notes. Does sample size matter? If so, how is it defined?

Experiment Setup:

Determine the quality of a topic map derived from one set of notes:

Select a group of students from a class as the test group. Select students who are friends, to ensure that they can communicate effectively with each other. Derive a topic map from one student's notes and test how well he understands the topic map derived from his notes before testing how well the others can understand and extract useful information from the map and the notes generated from the map. This experiment will confirm that the tool conveys the original intended meaning of the notes.

Determine the degree of standardization of a topic map:

Repeat the experiment with students who are not friends. This variation will confirm that the tool works universally and can therefore be expanded into broader applications like the *AccessAbility Center* who receive student notes by email all the time.

After implementing the ability to use multiple sets of notes and noise reduction, check with many students of the class if the information in the final notes generated from the map accurately includes all the relevant information and accurately excludes everything that is irrelevant.

Determine the degree of noise reduction:

Repeat the experiment with notes from different lecture sessions. This experiment has applications, especially in first year classes where there are multiple lecture sessions and a student asks a question in one of those sessions, but the question remains unasked in the other. Therefore, the

frequency of this information on the set of notes considered by the tools would likely not be true to the size of the class population that wrote down this information. In such cases, this information should not be excluded as noise as it is still relevant information. Any amount of this information that is not included would be an inadequacy of the tool. The converse is also true: actual noise that shows up with a higher frequency should also be eliminated because they have no relevance to the lecture.