Preliminary Investigation Toward an Automatic Notetaking System

Andrew Fister
Department of Linguistics
University of Illinois
at Urbana-Champaign
afister2@uiuc.edu

Roxana Girju
Department of Linguistics
University of Illinois
at Urbana-Champaign
girju@uiuc.edu

1 Introduction

Notetaking as a phenomenon is a remarkable extension of the human brain's memory structure and ability. One definition states that "Notes can be defined as short condensations of a source material that are generated by writing them down while simultaneously listening, studying, or observing"(Piolat et. al., 2005). In the data structure created when taking notes, linguistic content, which usually conveys most of the meaning, intermingles with non-linguistic content(arrows, circles, bullets, etc.) that generally works to connect ideas together. There are many research problems that make notetaking very intriguing indeed. How does the source material of the notes(lecture, meeting, book, introspection) correspond to the notetaking content? Which semantic relations do non-linguistic elements of notes capture, and why aren't these elements encoded as linguistic elements? What is the best digital representation of notetaking? What is the best machine learning algorithm to use in order for a computer program to learn to produce notes? And lastly, how would this algorithm differ from similar types of algorithms like those used in automatic summarization(Mani and Maybury, 1999), indexation, and highlighting?

The end goal of this project is to create a predictive learning model that can automatically generate a coherent and useful set of notes on a given linguistic input, such as a lecture or a book. In today's internet age, information overload is an increasing issue that affects us all. The modern citizen cannot possibly sift through all the information that is out there, and would benefit from a software application that can distill personalized notes based on documents that contain useful information. This application would be a boon to educators that desire to improve the educational experience by supplementing educational materials with automatically generated notes for reviewing material. Linguistics researchers would benefit from a thorough analysis of the linguistic structure of notetaking, about which there is an extreme paucity of work done to date. Natural Language Processing would benefit from the contribution of a primarily linguistic model for automatic notetaking.

2 Previous Work

(1) Shortening of words through the use of abbreviations and symbols.

(2) Omission of finite forms of the copula.

(3) Omission of articles, both definite and indefinite.

(4) Omission of (unstressed) pronouns, especially personal pronouns.

(5) Omission of (unstressed) finite forms of the auxiliary verb do in questions.

(6) Omission of multi-word groups and entire phrases from sentences, even of all phrases but one.

(7) Nominalization of verb (-phrase)s–often combined with genitivization of NP-subjects or with addition of as–resulting in the conversion of sentences into NPs.

(8) Combinations consisting of two reduced sentences linked together in topic + comment form.

(9) Replacement of relative clauses by participal constructions.

(10) Conversion of active sentences into passive ones.

Listing 1: Summary of grammatical reductions observed in notetaking in Janda(1985)

There is little work by linguists on notetaking. To my knowledge, there is only one thorough linguistic analysis of notetaking, done in 1985 by Richard Janda, whose own definition of notetaking states that "the purpose in taking notes is normally to have a potentially permanent record of at least the salient points of a lecture". Janda approaches an analysis of notetaking by considering notetaking to be a register of human language. That is, notetaking is a marked mode of expressing human language that differs from unmarked language in specific and systematic ways. He refers to the notetaking register as

note talk(NT) and compares note talk to baby talk(BT; how adults talk to a baby) and foreign talk(FT; how native speakers of a language speak to non-native speakers of a language). The main difference between these two speech registers and the note talk register is that note talk has no "no expressive, upgrading, or even clarifying processes," and Janda proposes to call note talk a "reduced register". Janda(1985) bases this analysis on a collection of notes from a variety of classes and students and delineates ten particular kinds of grammatical reductions that occur systematically in the notetaking register. These grammatical reductions are delineated in Listing 1. Janda concludes "the only grammatical reduction found in NT is that affecting elements that are fairly easily recoverable, syntactically-as well as mostly very empty, semantically." An interesting reductive factor not specifically noted by Janda is that it appears that often, linguistic content is reduced syntactically into non-linguistic content, while retaining the essential semantic expression. This will be a focus of the present research, along with the verifying Janda's conclusions.

Another interesting analysis of notetaking comes from Roger Shuy(1998). This work describes an analysis of note-taking from the perspective of showing what about a set of notes is *not* like taking notes. That is, as Janda(1985) notes, "Just as not all speech addressed to infants is BT, so not all language used in writing notes is NT". The analysis concerns a court case in which a deciding factor to the integrity of the plaintiff's case was whether a set of five pages of notes was taken shortly after the conversation ended, purely from the memory of the note taker, or if they were taken based on a recording of the conversation. The plaintiff claimed under oath that the notes were taken shortly after the conversation, while the defendant claimed that they must have been taken with the aid of a tape recorder. Shuy(1998) determines that if the notes exhibit consistently the reductions roughly consistent with those in Listing 1, then the notes are an example of note talk generated spontaneously at the time of the conversation. However, if the notes were taken from a recording, there would be at least some instances where the normal systematic reductions of note talk would not be present. And indeed, what Shuy(1998) found was that there was very many instances of non-reduced linguistic forms in the notes, and he concluded that the notes were taken with the aid of a tape recorder, and that the plaintiff was lying. This analysis is an interesting and convincing application of the kind that Janda pursues.

It is important to note here that most of the previous work done in notetaking has been in the field of psychology. There are several areas of interest in psychology related to research in notetaking: analysis of mental processes during notetaking(Piolat et. al., 2005), performance of different notetaking strategies(Davis and Hult, 1997; Kiewra et. al., 1995; Nye et. al., 1984; Kobayashi, 2006), and measuring notetaking performance among different student populations(Dunkel et. al., 1989). These studies are very informative going forward in determining how to approach setting up notetaking experiments, data collection, and possible ways to model notetaking computationally.

3 Data Processing

3.1 Raw Data Collection

In order to establish an empirical basis for the eventual training of a machine learning model for notetaking, a hand-built corpus is being constructed with the hope of creating a computer-readable notetaking corpus, which will be a resource of value to researchers in and of itself. The current data collection involves capturing the lecture and notetaking interaction that takes place within a classroom, specifically an undergraduate anthropology course on Sephardic Jews. The properties that make this particular class interesting and applicable in informing a machine learning system are: (1) the lectures are narratives on history and culture, which tend to deal in concrete topics as opposed to abstract topics; (2) the lectures are presented almost completely orally from personal lecture notes written on a legal pad, and few notes are written by the lecturer on the blackboard; (3) the lectures are coherent discussions of a particular topic informed by an assigned reading for that particular lecture. These conditions promote an optimal environment for particularly creative notetaking with mostly linguistic content, connected together logically by the narrative form of the lectures and more generally by the coherent topic structure of the entire course.

Each lecture is recorded using a digital voice recorder, placed near the front of the classroom. After each lecture, the notes taken by each student are requested, and those who wish can submit them. Images of the notes are then taken using a digital scanner and the lectures' audio data are also transferred to a computer.

A simple transcription of each lecture's audio is then recorded in a text file. This transcription includes disfluencies, which are generally conflated into a few general types, in order to make parsing easier. Short pauses are indicated by commas and thought completions indicated by newlines. This transcription approach expedites the labor-intensive process of transcription and keeps the effort required to parse the large amount of text to a minimum. A time-aligned transcription using a proper transcription program(like the open source Transcriber program) should be considered for processing future audio recordings.

```
<text style="align:margin-top">Start in 1840's</text>
<target id="17">
<text style="align:margin-top;size:small">means Jew by faith of Judaism</text>
</target>
<mark type="arrow" id="16">
<mark type="box"><text>Israelite</text></mark>
<text>Universelle</text></mark></mark><line/>
<indent><target id="16">
<text>huge influence</text></target><line/>
<target id="16"><text>private charity organization</text></target>
<target id="18">
<text style="size:small">spread all over Africa, Ottoman Empire, etc.</text>
</target><line/>
<mark type="arrow" id="18">
<target id="16">
<text>established by wealthy in France</text>
</target></mark><line/>
```

Listing 2: Excerpt from XML representation of notetaking.

The notes from each lecture are transcribed into an XML format. The purpose of this XML representation is to encode the non-linguistic elements of the notes into a format that makes the semantics of these elements easily decodable by a computational algorithm. The goal at this point is not to annotate or parse the structure of the linguistic content of the notes. It is instead to represent all the non-linguistic content and isolate the linguistic content into segments that are completely uninterrupted by any markings or formatting conventions. The intention in constructing the annotation procedure in this way is to enable the interpretation of the semantic interaction between linguistic and non-linguistic elements.

3.2 Linguistic Relations of Non-linguistic Objects

For example, in one of the simplest cases of this type of interaction, which occurs in most kinds of writing, we can consider what happens when there are two lines of text separated by a line break. What is the linguistic relationship between the two lines of text in this case? Fairly intuitively, we know that there should be a serial conjoining of the two lines of text in most cases, especially where the first line of text is not a complete phrase or sentence. Other non-linguistic elements facilitate different and more complex relations that may encompass the morphology, syntax, semantics, and pragmatics of the linguistic elements they pertain to.

We can also examine a more complex case for our edification. In note-taking, one of the most common non-linguistic structures is the arrow. An arrow has the potential to connect any two linguistic elements on a single page and acts as a vector between the two elements. For this reason, any type of one-way semantic(in the non-linguistic sense) relationship is a valid interpretation of the arrow. These pre-linguistic one-way semantic relations include(but not exhaustively): attachment, indicating some kind of adjoining of one element to the other; entailment, indicating that one element is true in every case when the other element is true; and sequence, which indicates that under some linear scale, one element precedes the other. The interesting task in this case is to find out in which contexts these pre-linguistic relations are represented by a particular arrow found in a page of notes, and then to determine the specifically linguistic relations(syntactic adjoining, metnoymy, chronological ordering) are used by arrows in which contexts. Also, it would be interesting to find out if there is any variation in how different note takers encode these relations.

3.3 XML Representation of Notes

The representation of notetaking given here is meant to be tailored to future computational processing, and certain considerations were taken to ensure this. First, the annotation is done in XML, which makes the representation very readable by a computer. Second, in the course of annotation, if there are any perceived spelling or grammatical errors in a particular linguistic element, the original text written by the note taker is retained, but an XML attribute proposes a replacement for the problematic text. This will make it possible for either the original text or a more easily parsable version of the text to be analyzed as the research demands.

The top element in this representation is the <note> element, which contains the notes taken by one note-taker during a single lecture, and contains attributes that represent metadata identifying aspects of the individual's notetaking session. These attributes include 'subject', the initials of the note-taker, and 'date', the date the notes were taken, as well as 'doodle', which indicates whether or not the note taker made any doodles. All the direct children of a <note> element are <page> elements, which contain all the notes on a single page. All <page> elements are required to have an 'id' attribute that indicates the page number. Optionally, if the page of notes is divided into columns, <column> elements may be used to divide the page into its constituent columns, with a required 'id' attribute identifying which column is being identified, moving left to right across the page as the value of the 'id' attribute increases.

The approach taken in this representation is to encode coherent text segments as units under a <text> tag. But what does it mean to have a coherent text segment? It means that no linguistic element of any sort may intervene between the Whether it be an arrow with its source or target evidently toward a substring of a text segment, a cirlce around a single word in a sentence, or a line break.

The <text> element has several optional attributes. As mentioned before, there is an 'edit' attribute that proposes spelling and grammatical corrections. There is a 'style' attribute that describes any properties of the text that are relatively abnormal compared with the rest of the text on the page. These values include font face, alignment, and font size. Any non-ASCII characters are represented with <symbol> tags. Each symbol tag must contain a 'name' attribute whose value is a string uniquely identifying the character represented in the notes.

With a few exceptions, non-linguistic elements are represented with a <mark> element, surrounding the span of elements targeted by the non-linguistic element, with a 'type' attribute specifying the type of non-linguistic element. What if a non-linguistic element connects one or more other elements? In that case, the <mark> element representing the non-linguistic element is assigned a number with an 'id' attribute. Then, one or more <target> elements are placed around the element(s) that the non-linguistic element "points to", with an 'id' attribute matching its corresponding <mark> element. The <mark> element may also contain a 'style' attribute with a weight value that determines the evident emphasis given to the element relative to other elements on the page.

The exceptions to the <mark> element are the <bullet>, <indent>, and <line> elements. The <bullet> element is a convenient substitute for the <mark> element when denoting bullet points that usually precede elements in a list. This element has an optional 'style' attribute with a weight value that indicates the evident emphasis placed on the bullet relative to other bullets in the list. The <bullet> element always has empty content. The <indent> element indicates that the contained content is indented one unit of space relative to the immediately surrounding content. The <line> element indicates a line break, and its content is always empty.

The distinctions that motivate the decision by the annotator to use the elements summarized above require the fine-grained interpretive ability that only a human can currently provide. By using an unambiguous, solely descriptive annotation on the notes initially, the annotation remains agnostic regarding the semantic interpretation of the notes, but becomes easily analyzable by a computer program for semantic and other linguistic content.

4 Topic Analysis

With the data as described above in hand, a number of interesting analyses can be done. The analysis underway at the moment involves identifying global and local topics that occur in the lectures as well as the notes, and then to map the topic structure of the lectures to the notes. This kind of topic analysis will shed light on the methods of representation that people use to connect together in their notes the ideas from the lecture.

4.1 Lecture Topic Annotations

In order to take a principled approach toward topic analysis, we should first consider the raw input that the note takers processed, which is the lecture. Given the lecture transcription discussed in the previous section, it is possible to annotate a lecture's topics. Using an XML representation, it is possible to specify topics at differing granularities. The lecture annotation's top level element is the <lecture> element, which contains a 'date' attribute indicating the date of the lecture. All other elements in this annotation are <topic> elements, which contain 'name' elements denoting the topic name. <topic> elements can be placed in a hierarchical configuration to portray subtopics all the way down to a word level if so desired. It is important to note that topic names are not unique in this annotation scheme, as topics often repeat multiple times throughout the course of a lecture.

The decision regarding which spans of text constitute the discussion of a single topic is rendered mostly through the intuition of the annotator. However, a principle is applied in this matter: (1) the name of each topic must exist as a word or phrase within the span of text of the corresponding <topic> element; (2) the language content(semantic and lexical) of the span of text of the <topic> element must relate closely to the domain indicated by its 'name' attribute.

4.2 Note Topic Annotations

The goal in annotating topics in the notes that correspond to the lectures annotated as described above is to find an alignment between the two. There is a need that the set of topic names used in the note annotation be a subset of the set of topic names in the lecture annotation. And again there is also a need to facilitate ease of parsing by a computational algorithm.

The annotation of notes for topics is fairly simple and consists in adding an optional 'topic' attribute to those <text> elements that pertain to the particular spans of text encapsulated by the <topic> elements in the lecture annotation. The annotation of notes for topics is more systematic than the annotation of the lectures is, and follows these principles: (1) the semantic content of the target <text> element(in context with surrounding text elements, if necessary) must correspond to the semantic content of the text encapsulated by the corresponding <topic> element in the lecture topic annotation; (2) the topic name must be linguistically recoverable in whole or in part(abbreviations, acronyms, etc.) from the set of topic names defined in the corresponding lecture topic annotation.

It should be noted that under this annotation scheme, there will be many <text> elements that meet criterion (1) but not criterion (2). In this case, it should be noted that the same process that can be used to automatically parse

the semantic content of the connection between linguistic elements through non-linguistic elements will likely easily extend to the ability to associate the topics attached to <text> elements to other semantically related <text> elements.

4.3 The Topic Matrix

One method of analysis of the annotation method described above is to generate a summary of the topic coverage evident in the lecture and of the notes of each note taker. By listing each topic covered by the lecture and marking off each topic covered by each note taker, it's possible to get a sense for variation in topic coverage and also topic overlap between note takers. In the first lecture analyzed, it is evident that notetaking occurs in waves, with important topics occurring in clumps, most likely because a very important topic radiates importance to its surrounding topics. More specifically, in this lecture that was approximately one hour of relevant speech, with 102 topics identified overall, there were 8 topics covered by all five analyzed note takers, and 10 topics covered by four of the five note takers.

5 Future Research

An important point that the analyses of this project will focus on is that notetaking indeed differs significantly from the process of the summarization of a piece of text(for instance, creating an abstract from an entire academic paper). Since notetaking and summarization are similar kinds of tasks, it is important to delineate, empirically, exactly the differences between them and demonstrate the advantages of notetaking and summarization. One such difference is that notes are generally intended only for private use, while summaries are generally written to be read by someone other than the writer(unless the summary is written as a note). In order to carry out this kind of comparison, one approach would be to compare manual notetaking data from a series of lectures to the results of running various automatic summarization algorithms. While these are two different kinds of things, it would be useful in that the intent of the summarization algorithms would cause results to be produced that reflected manual summarization methods. To balance the possible inadequacies of automatic summarization compared with manual notetaking, manually generated summaries could be compared as well. Compared with manual notetaking, manually generated summaries could be compared as well. Once it is known how tasks like automatic summarization, highlighting, indexation and others differ from notetaking, an automatic notetaking system can be constructed using novel and established machine learning algorithms.

6 Acknowledgements

Many thanks goes to Professor Mahir Şaul of the Department of Anthropology for facilitating and encouraging the research during his class sessions, the students in his class for facilitating the recording of the class sessions, and especially to those students who volunteered to lend their notes every week.

References

Annie Piolat, Thierry Olive, and Ronald T. Kellogg. 2005. *Cognitive effort during note taking.* Applied Cognitive Psychology, 19(3):291-312.

Inderjeet Mani and Mark T. Maybury. 1999. *Advances in Automatic Text Summarization.* The MIT Press. Cambridge.

Keiichi Kobayashi. 2006. *Combined Effects of Note-Taking/-Reviewing on Learning and the Enhancement through Interventions: A meta-analytic review.* Educational Psychology, 26(3):459-477.

Kenneth A. Kiewra, Stephen L. Benton, Sung-Il Kim, Nancy Risch and Maribeth Christensen. 1995. *Effects of Note-Taking Format and Study Technique on Recall and Relational Performance.* Contemporary Educational Psychology. 20(2):172-18.

Martha Davis and Richard E. Hult. 1997. *Effects of writing summaries as a generative learning activity during note taking.* Teaching of Psychology, 24(1):47.

Patricia Dunkel, Shitala Mishra, and David Berliner. 1989. *Effects of Note Taking, Memory, and Language Proficiency on Lecture Learning for Native and Nonnative Speakers of English.* TESOL Quarterly. 23(3):543-549.

Pauline A. Nye, Terence J. Crooks, Melanie Powley, and Gail Tripp. 1984. *Student Note-Taking Related To University Examination Performance.* Higher Education, 13:85-97.

Richard D. Janda. 1985. *Note-taking English as a simplified register.* Discourse Processes, 8(4):437-454.

Roger W. Shuy. 1998. *What we do with English when we take notes: evidence from a civil lawsuit.* Studia Anglica Posnaniensia: international review of English Studies, Adam Mickiewicz University Press.