A Story Matching System for the Victorian Laptop

by

Petra Chong

Submitted to the Department of Electrical Engineering and Computer Science
on May 18, 2000, in partial fulfillment of the
requirements for the degrees of
Master of Engineering in Electrical Engineering and Computer
Science
and
Bachelor of Science in Electrical Engineering and Computer Science

Abstract

In this thesis, I designed and implemented a topic-based story matching system for
the Victorian Laptop. A corpus of historical stories is divided into categories and
single-word indicators of those categories derived from the corpus. An index of topic
keywords and an index of proper nouns (also taken from the corpus) is built. The
system scans the user input text for topic keywords and proper nouns and decides
based on the relative proportions of the two, whether a topic-keyword-based match
or a proper noun-based match is to be returned. Some artificially constructed stories
were input to the system as well as some actual travelogue, to test the effectiveness
of this story matcher compared to the previous-generation one in the same project. I
also discuss the challenges present in the process of designing a computer system to
perform a highly subjective task like story matching.

Thesis Supervisor: Justine Cassell
Title: AT&T Career Development Associate Professor of Media Arts and Sciences
Acknowledgments

Thanks are due to the following people whose assistance I could not have done without:

- Justine, for believing and encouraging.
- Jennifer Smith and Julie Burch, UROP supervisors, who got me started.
- Mike Ananny, for many excellent discussions.
- Erik Mueller, for help and guidance.
- everybody in GNL!
- and of course my parents without whom MIT would not have been possible.
Contents

1 Introduction 8
  1.1 Scenario ................................................................. 8
  1.2 Overview of the Victorian Laptop ................................... 9
  1.3 Statement of Problem ................................................ 10

2 Background 12
  2.1 Traditional Information Retrieval and Narrative ............... 12
    2.1.1 The Current Story Matching System ............................. 14
  2.2 Full Understanding is a Difficult Problem: Is it Necessary? ........ 15

3 Related Research 17
  3.1 Systems of Computational Knowledge Representation ............ 17
    3.1.1 Theory of Scripts ................................................... 17
    3.1.2 Expectation Packages .............................................. 18
    3.1.3 The WordNet Lexical Database .................................. 19
  3.2 Systems for Text Classification and Story Matching ............ 19
    3.2.1 The AutoSlog Project .............................................. 19
    3.2.2 The WISE Project ................................................ 21

4 Design of the Victorian Laptop Story Matching System 23
  4.1 The Corpus ............................................................. 23
  4.2 Underlying Assumptions .............................................. 25
  4.3 Design Alternatives .................................................. 25
6.3 Discussion and Suggestions for Further Work ............... 72
6.4 Evaluation From the Point of View of a Human Audience ........ 75
List of Figures

3-1  Expectation Package representing the situation “Having a tea party”  19

5-1  Data structure representing index of keywords to topics and stories  46
Chapter 1

Introduction

1.1 Scenario

I am visiting Boston for the first time. As the plane touches down at Logan, I take out my Victorian Laptop, insert a new sheet of paper, and begin to write my first impressions of the city.

We have just arrived in Boston! And about time too; we were delayed for two hours because of bad weather, mostly strong wind – I hear these New England winters are something to behold. Fortunately we made it in safe and sound. The plane came in over the water, and I got to see Boston Harbour in the sunset.

I pause to think; what I have written on the sheet of paper appears on the screen of the Laptop, and the following passage appears on the other side:

*Amelia Burnett, Journal of a Visit to the United States, 1846-1848* Source: Massachusetts Historical Society

Saturday 31st, Oct., at 5. A.M., a sailor from the mast’s head, called out 'Land' delightful sound, and it was amusing to watch the general movement which followed the intelligence. I sat on deck till dark, when I could perceive the distance light on Thatcher’s Island off Cape Ann. We went to our berths early, but at half-past 10 I was woke up, as I had requested, by the mate, to see Boston Light which had just appeared in sight. I went on deck, where Clara and Olivia joined me, and there remained, until we
cast anchor in the stream close to the wharf, at 1-1/2 Sunday morning. The sky though hazy was sufficiently lit up by the moonlight to enable me to perceive the general appearance of the harbour. Boston Light is a very large one and shone like a bright star: marking our course for us; the coast is very low, and dotted with islands, which continued the whole way to Boston. They are well cultivated, and look prettily in the day-time, but of course I could only see their general form. Soon we got amongst the shipping, which ranged outside the wharves, reminded me of the effect of the Thames at London. Lights could be seen, here and there, amidst the darkness of buildings, but it was too dark for me to judge of the appearance of the city. After talking of all we should do the next day, we again retired to our berths, but it was some time before I could sleep; I kept trying to realize that I was actually in Boston Harbour; and about to see places and persons that I had so often heard of, and longed to visit.

This is interesting; there were no airplanes in the 19th century, so of course everyone would have sailed in to Boston. This makes me look at my circumstances differently. I continue writing:

Considering that it used to take days to get from Europe to Boston, a two-hour flight delay seems rather paltry. I don’t see the Boston Lighthouse that she mentions – is it still around somewhere?

The next time I pause, the Laptop obliges with another segment of travelogue, relevant to my writing, that opens a window to the world of the nineteenth-century traveller in New England. Through the course of the interaction, I learn that travelling in the nineteenth century was far different than now; people did it for different reasons (I am here on business), and they did different things. My visit to Boston is colored by this awareness of the past.

1.2 Overview of the Victorian Laptop

In the latter half of the 19th century, it was common for upper-class travellers to carry portable wooden writing desks with them. These desks were used to write travel diaries, journals, and letters to be sent home. The Victorian Laptop extends
this idea. The user of the Victorian Laptop writes a story, and is presented with an authentic historical story that is relevant to what she has written.

The Victorian Laptop consists of a pen-input laptop computer housed in an antique Victorian writing desk. The user writes a story on paper that is placed on a digitizing tablet connected to the laptop. Writing is done with a digital pen, which transmits the strokes of handwriting to the laptop. The computer performs handwriting recognition to obtain the text of the user’s story. The story text is then analyzed (a description of what this analysis entails will follow in a later section) for matching with the computer’s database of stories.

The computer contains a database of historical travel journals. This database is searched to find a story that is relevant to the user’s story. This story is then returned to the user; the user’s story is displayed on one half of the screen, and the returned story is displayed on the other half for comparison.

It is hoped that presenting the user with a historical story about similar subject matter will stimulate her to think about her experience in a different context, and encourage her to write more. Every effort has been made to keep the interaction as similar to the experience of writing in a diary as possible; the technology is intended to be unobtrusive.

1.3 Statement of Problem

This thesis will discuss only the specific problem of story matching. The story matcher will have to be designed taking into account the project paradigm as a whole, but issues of external interface design of the Victorian Laptop device are not discussed as part of that.

The main underlying issue is how to decide if two stories match or not, in order to return a database story that is relevant to the user’s story. The use of the term “match” implies that the two stories contain similar information. If so, what sort of information is this? How similar does the information in the two stories have to be in order for them to “match”? What is a way of measuring and quantifying this
information? These issues are discussed.

Also discussed is how to use domain-specific knowledge to improve matching. The Victorian Laptop corpus is entirely historical and consists only of travel journals. The Victorian Laptop device is intended to be a computationally-enhanced travel journal. Therefore, there are some assumptions and generalizations that can possibly be made about the expected input. We will examine how these issues affect the design of the system, and the quality of the matching.

The Victorian Laptop story matching system should not require “strong artificial intelligence”; the system is not required to completely “understand” the story in the same way that a human does, knowing the exact context of the story, all the emotional nuances, and all the real-world knowledge extant in and implied by the story. The only criterion that it has to satisfy is that the user must find the match meaningful. Previous research has shown that as long as a match is related in some way to the user story, the user will find it meaningful. Therefore, deep understanding is probably not necessary to produce a match that the user will find meaningful.

The main engineering issue that this project presents is the design of a system for text retrieval of narrative. In order to do this, the characteristics of narrative that differentiate it from other sorts of text are also studied. Most text retrieval systems handle news text, which tends to have a standard structure and style – these constraints are not usually present in narrative text.
Chapter 2

Background

2.1 Traditional Information Retrieval and Narrative

"Traditional" information retrieval systems as described by Pao [15] usually depend on keyword-based indexing. In these systems, important words in each text are collected into an index which records which texts contain what words. A user looking for a document containing certain information has to specify exactly what keyword to search for. She then receives a list of all the documents that contain these words.

This form of indexing and retrieval is mainly used with scientific and technical documents, for example, in library information retrieval systems. There has not been much work with similar retrieval of narrative texts. Scientific fields usually have a widely accepted, and very specific terminology. Keyword-based information retrieval depends heavily on this. In a scientific document, ambiguity is frowned upon, as the intended purpose of a scientific document is to convey to the reader exactly how an experiment was performed in a clear way, so as to enable the reader to duplicate the results of the experiment. Content-bearing words in a scientific document therefore usually have specific meanings that are well-defined, and have become well-defined in the years that it took for the field to become well established in the scientific community. Therefore, for a field that is not yet fully defined, the exact terminology
may not yet be known. Consistent with this, Pao finds that “poor retrieval can be expected for emerging and interdisciplinary subjects”.

This form of information retrieval requires certain assumptions to be true which narrative does not satisfy. For example, Pao states that an important attribute of text to be indexed is the presence of a list of bibliographic references, and that this practice of citation provides a convenient way of quantifying subject relationships. Any scientific paper is required by convention to cite other documents from which content was drawn; this serves a legal purpose (preventing plagiarism) as well as allowing interested readers to find out more about the subject. Titles and abstracts, also present in such documents (also by convention), usually contain domain-specific keywords.

Keyword indexing does not handle synonymy and ambiguity well. Most words are used in more than one way, and what a word means depends on the other words present in the same document. For example, the word “tree” as used in a document which contains the words “fruit”, “conifer”, and “forest” probably means a large plant with a woody stem. The word “tree” as used in a document containing “chess”, “intelligence”, “computation” and “search” probably means a data structure used to hold information about the state of a board game. Pao states that free-text indexing fails specifically on this point [15][p. 105], as it does not take into account context.

In this thesis, context is taken to mean “the company a word keeps”. In the example above, the word “tree” in one context may be found to co-occur with the words “fruit”, “conifer” and “forest”. “Context” here refers to the “sense” of the word, the real-world concept which that word is supposed to embody or describe. Ideally, given a word in a story, it should be possible to make predictions about what other words may occur in the same story.

Narrative texts do not, as a rule, possess any of the above characteristics that make keyword indexing optimal. The author of a narrative is not bound by any conventions that state specifically what words may be used. Neither is she bound to explicitly cite references to other narrative literature; often such quoting is left unstated for the reader to discover and make the connection. Ambiguity and synonymy are also
often present in narrative, as the use of simile and metaphor are well-known tools of the creative writer. This makes the narrative interesting to read, as different readers will interpret it differently. However this is also a cause of narrative being difficult to analyze using word-based techniques.

2.1.1 The Current Story Matching System

The previous generation Victorian Laptop story understanding system uses plain keyword matching. Keywords are stored in an index which keeps pointers to exactly which stories contain which keywords. User input to the Laptop is searched for the presence of any of these keywords. The story that contains the most keywords present in the user input text is returned as the best match. In case of ties, the best match is picked randomly from the possible candidates. If the user input text does not contain any of these keywords, the system is unable to return a match. The strength of a particular story as a match is the number of keywords in that story that are found in the user input story (duplicates allowed). The strength of a story as a match is normalized by the length of that story, because it was intended that shorter entries should return shorter matches.

The keywords were derived partly by hand and partly automatically. A list of all the words in the corpus was made using the WordSmith tools [22]. Stopwords (like “an”, “the”, “if”, “then” – words that do not contain meaning by themselves, only in relation to others) were removed from this master list, then the list was subjected to minimal vetting by hand. Names of people who were only known to the authors were removed (mostly names of relatives and friends), for example. No other words were removed; no attempt was made to cull words based on frequency. Even if a word appeared only once in the corpus, if it was not obviously a name of a person who was not a historical figure (“Thoreau”, for example, was retained as a keyword), it was kept as a keyword regardless of its significance in the corpus.

This system sometimes gives reasonable results. For example, the phrase “Boston

---

1There is support for weighting of keywords, but this was not implemented and all the keywords currently have the same weight.
Common” is a keyword. If the user writes about “Boston Common”, she will probably get a story in which the Boston Common is mentioned somewhere. This may or may not be interesting. If the Boston Common is the location for where the story happened, then the user might find this to be interesting. However, a story that contains the phrase “Boston Common” may well mention the Common just incidentally at the beginning, and go on to describe how the author went hiking in Maine. Keyword indexing would fail to pick this up. Therefore, the keyword indexing system is hampered by lack of context.

The plain keyword indexing system demonstrates a lack of focus. Suppose we have a story that contains five instances of the word “boat”. It would be easy to infer from this, that the concept represented by “boat” is a concept that is present in the story to a greater degree than other concepts. What about a story that contains one instance of each of the following words - “boat”, “ship”, “ocean”, “berth”, “cabin”? A human reader would know that “boat” is also a concept present in the story to a greater degree. The plain keyword indexing system would not be able to identify this, because it does not recognize that a single semantic concept can be evoked by more than one word. The plain keyword indexing system is fragmented, and does not make use of the idea that words carry context with them, and that certain words co-occur in normal usage.

It is possible that providing some context to the system, in the sense of providing it with knowledge of which words are semantically related, may improve matching. This thesis describes an attempt to provide the system with a limited idea of context, and evaluate whether matching improves or not.

2.2 Full Understanding is a Difficult Problem: Is it Necessary?

McCarthy [9] and Mueller [13] give reasons for why the state of the art in natural language systems is inadequate for deep understanding by computers. The most im-
mediate problem is the management of complexity. Any word can be used in more than one way; if one sentence contains ten words, we have a complicated combinatorial problem on our hands. Most of the possible interpretations are not sensible; it is difficult to weed out the implausible interpretations (Domeshek, Jones and Ram [18][pp. 93-94] quoted by Mueller [13]).

There is a lot of real-world knowledge that is inherent in a narrative. This real-world knowledge would be required for a system that needed to make inferences, answer arbitrary questions, summarise and perform other high-level tasks. The Victorian Laptop system does not need to do these things. It needs to produce a related story that the user will find meaningful; it does not have to suggest why the story is relevant to the user’s. Bers [1] found in the SAGE project that when a user is presented with a story that is relevant in only some aspects to her story, the user tends to attribute relevance to them according to her personal experiences. Therefore, the Victorian Laptop does not have to parse all of the information present in a story. It does not have to understand all the implications of the story. It has to extract enough information to enable the match it returns to be felt by the user to have some relevance.
Chapter 3

Related Research

In this chapter several related methods of knowledge representation and text retrieval are described. Most of them have as basic principles of their design the fact that knowledge and language can be divided into categories according to real-world concepts. All of these systems make use of the fact that language is used in predictable ways to describe situations and objects in the world.

3.1 Systems of Computational Knowledge Representation

3.1.1 Theory of Scripts

Schank [21][pp. 6-7] proposes that human understanding of stories is based on \textit{scripts}. A script is a generalization of a common situation and a set of expectations about what happens in these common situations. For example, a script could be used to encode our expectations of the events that take place when one goes to a restaurant for dinner; the restaurant script may contain the following events: sitting down at the table, perusing the menu, ordering, being served and paying the bill. Scripts allow people to predict what is going to happen next in a common situation. This is a useful concept for a system that understands travel narrative to have, as many of the things that travellers do are well-defined and are done by more than one traveller.
Any event may appear in more than one script; it therefore has a different context in each. Its context is the events surrounding it. Therefore a certain event may have different significance depending on what other events it is next to. For example, “ate dinner” has different significance if it follows “ordered a meal” versus “went home for the day”.

3.1.2 Expectation Packages

Gordon, a student of Schank, proposes a design [2] for a browsing interface to a digital library that makes use of an underlying knowledge representation of linked knowledge structures. Gordon’s knowledge structures are called Expectation Packages. Expectation Packages are clusters of words related to a particular situation or script (see Figure 3-1 for an example). The five fields of an Expectation Package are Events, Places, People, Things and Miscellaneous. The values of these fields are nouns or verbs that represent things or actions that are found in the situation described by the Expectation Package. Therefore, an Expectation Package represents a context for the terms contained within it. In Gordon’s interface, Expectation Packages provide a way of browsing a digital library that makes use of the knowledge inherent in the content of that library. While this thesis is not concerned with browsing databases, the idea that information inherent in a domain forms a well-connected set is useful and interesting, because in this system of knowledge representation, every piece of knowledge represented has a context. Its context is the other pieces of knowledge that it is connected to in the network. The definition of such a network is subjective (Gordon states this explicitly). Gordon built his Expectation Packages by hand, because he built his knowledge structures from a pre-existing database of images. There is no intellectual reason why Expectation Packages should have to be hand-built; the issue at present is a practical one, that there is no good pre-existing database that effectively captures concepts and the interconnections between them the way that Gordon’s Expectation Packages require.
<table>
<thead>
<tr>
<th>Events</th>
<th>Conversation, Tea parties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Places</td>
<td>Living rooms</td>
</tr>
<tr>
<td>People</td>
<td>Guests, housewives, socialites</td>
</tr>
<tr>
<td>Things</td>
<td>Coffee tables, sofas, tea, teapots</td>
</tr>
<tr>
<td>Misc.</td>
<td>Rumor</td>
</tr>
</tbody>
</table>

Figure 3-1: Expectation Package representing the situation “Having a tea party”

### 3.1.3 The WordNet Lexical Database

Miller et al. [11] built a large database for English that grouped words according to topic. WordNet is described as an “on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory”. Nouns, verbs, adjectives and adverbs are grouped into sets that each represent a single underlying lexical concept. For example, “eat”, “consume” and “ingest” are all classified as words that describe the action of eating. At present, WordNet only matches words within the same part of speech. That is, WordNet does not recognize that “boat” (the noun) and “sail” (the verb) are descriptive of the same concept. Also, its classifications are often inconsistent. For example, “hotel” and “apartment” are not recognized as both being descriptions of different types of lodging. “Hotel” is seen as an instance of “building” but apartment is seen as a type of “lodging”. Disambiguation is a major problem with WordNet; one word may have many senses, and without knowing beforehand which sense is being used, it is a significant computational problem to decide which sense is the salient one.

### 3.2 Systems for Text Classification and Story Matching

#### 3.2.1 The AutoSlog Project

Riloff [19] has built an information extraction application called AutoSlog that extracts predefined types of information from previously annotated text.\(^1\) This is a

\(^1\)The annotation process requires a human user to mark sentences which s/he considers to be relevant to the meaning of the text. AutoSlog then parses these marked sentences to obtain the
sort of “text skimming” which is different from in-depth understanding. It picks out the parts of a text that are relevant to a constrained domain. AutoSlog makes use of CIRCUS [5], a conceptual sentence analyzer that extracts information, building a dictionary of “concept nodes” that recognizes patterns and expressions specific to the domain. AutoSlog was used to extract information from a database of news stories for the 4th Message Understanding Conference. It was incorporated into a system to classify newspaper articles into two categories; whether an article was about a specific terrorist incident or whether it was not. AutoSlog was able to extract linguistic information that would be difficult for a human reader to identify, for example, that the phrase “was assassinated” was often indicative of a description of a specific incident, and that the word “assassins” was not. AutoSlog was also able to deduce that the concept of “guerrilla” is a perpetrator of murder. AutoSlog is thus able to connect the idea of “guerrilla” with that of “perpetrator” and “murder”. AutoSlog uses some heuristics for pattern extraction that may be potentially useful. The AutoSlog heuristics may yield patterns that could be used to extract information from narratives.

An important observation about AutoSlog is that it picks out information that is relevant to a specific domain. The text classification system that made use of AutoSlog-extracted information was handling a binary problem; the text was either in the domain or not. It operated on the assumption that topics had “indicators” (words or phrases that indicate the text being in that domain, in their system called Relevancy Signatures). Conceptually this is an assumption that is easy to accept.

The interesting thing about the Relevancy Signatures in this system is that they are sometimes found to be very specific in terms of verb tense and form. For example, it was found that the signature “< X > was assassinated” was indicative of the topic of terrorism, but that the signature “< X > assassinated < Y >” was not. This is a subtle distinction that a human reader may not have picked up. However, their choice of topic was very well-defined and constrained; their definition of “assassination” referred to specific accounts of individual acts of terrorism, which did not include information.
generic descriptions of non-specific occurrences of terrorism.

The AutoSlog/MUC corpus consisted of news stories, which have a fairly well-defined style of writing. The Victorian Laptop corpus consists of narrative text, which may behave differently (though this has not been tested).

3.2.2 The WISE Project

The WISE Project [23] is a Web-based version of the SAGE Project [1] which allowed children to create characters that would function as interactive storytellers. When a child created a character, she designed the interaction that the character would provide to a user, as well as the stories that the character would tell in response to a user story. To facilitate matching, the user was asked to define keywords for their story, as well as use WordNet to define which sense their keyword takes.

The explicit specification by the user of the keywords and their senses simplifies the story matching problem immensely, as it removes the problem of word sense disambiguation. Knowledge of the exact keywords and their senses enabled the WISE system to obtain synonyms, hypernyms and homonyms for keywords, thus extending the number of words that the system could recognize. Shah, in [23], discusses the difficulties of using WordNet without dynamic word sense disambiguation; it turned out to be impossible to use WordNet on words obtained from the user input story, instead using it on words in the characters’ pre-written stories, which were already disambiguated.

Like the Victorian Laptop, this story matching system was set in the context of a larger goal. In the case of WISE this goal was to provide an understanding listener for children who would provide them with a relevant story. The system was tested on adult users (college students).

Evaluation of WISE found that verb extraction was helpful in increasing the relevance (as perceived by users) of matches. Shah suggests that this is because users consider “plot elements” when judging relevance of matches. Though he does not define the term “plot elements” further, this phrase intuitively suggests that these correspond to “happenings” or “events” (an idea that will be discussed in a later
WISE is an interesting system to examine in terms of information retrieval, because it handles narratives. Most other information retrieval systems work with informative text, where the style of writing is fixed and is not a major component of interest.
Chapter 4

Design of the Victorian Laptop
Story Matching System

4.1 The Corpus

The Victorian Laptop corpus consists of about 300 segments (ranging in length from one paragraph to several pages) of excerpts from historical travel journals. The journals range in time period from the 1950s to the 1930s. They are almost all by amateur writers, all in the first person, and all are narrations of actual events (non-fiction). Most were obtained from the Massachusetts Historical Society.

Most of the stories by amateur writers are disorganized, switching subject midway in a paragraph or a sentence. They tend not to exhibit a structured form; sometimes they are written in a fairly “stream-of-consciousness” fashion. Because they were originally intended for private consumption (letters home to friends and family, or private journals), they make extensive reference to a cast of characters who are well known to the authors without elucidation of who those characters are.

Of important note is the fact that the entire corpus is historical. There are two major differences that this can cause. The first is that the style of the language used in the corpus is different from contemporary style. The second is that the content of the corpus, the things described, are different from those usually described in modern travelogue. Consider the following passages. The first is written by Amelia Burnett
You may fancy me looking out on either side; being Sunday morning, the streets were deserted, and the general appearance of the town was far from cheerful, but after passing the long Cambridge bridge, the aspect of things improved. The wooden houses looked cheerful and are prettily built with piazzas, and bright green blinds to the windows.

To tell you the truth, travelling scares the crap out of me. That may sound like a strange admission from someone running a website called "Big Adventures", but, well, there you have it. The thought of arriving alone in some smelly dirty city on the far side of the planet where no one speaks anything even resembling a Latin-based language, with no hotel reservation, no idea where the good hotels are, not even an idea of whether the good hotels will have a toilet, and oh by the way, we don't take credit cards... well, I'm sure you get the picture.

Comparing both these excerpts, it can be seen that both the style of writing and the content described are very different. There are several reasons why the content of the corpus is different from contemporary accounts of travelling. Firstly, there are many things that exist today that were not existent then. The best example of this is forms of transportation. Because all of the stories in the corpus come from 1930 or before, none of the stories in the corpus will ever mention aeroplanes. Secondly, the sorts of activities that people did while they were travelling were different. There are many accounts in the corpus of attending lectures and visiting factories while on holiday. Possible reasons for this were that the length of a trip was usually longer (since it took longer to get there), and that people travelled not only to see new sights but to learn new things, while today's travellers tend to travel out of escapism.

There are four sorts of stories in the corpus. These may be roughly characterized as narratives, descriptions, anecdotes, and reflections. Narratives tend to describe the actions of a protagonist, usually but not always the author. The actions take

---

1 The names for the categories are my own but certainly others have made similar propositions.
place over a prolonged time period. Descriptions tend to be static; when something (usually a physical object of some sort) is described it is described as how it is at a single particular time. Anecdotes are related to narrations; in addition to describing a series of actions, there is usually some higher-level information conveyed, like humor or a moral. Reflections are introspective and are largely descriptions of the author’s internal state.

From analysis of the corpus, it was found that for stories which are narratives, the verbs contain a lot of the content. For stories which are descriptive, the nouns tend to be more important in indicating content. Anecdotes and reflections contain much high-level information, and it is hard to pick out this information at only the word level.

4.2 Underlying Assumptions

There is no restriction on the input beyond it being an account of travelling. Besides that, we allow users to write about more than one topic in their stories, and use free text with no restrictions on the sort of words they can use. This is bearing in mind that the overall effect of the device is to stimulate expression in writing.

To simplify the task, only information contained in the corpus will be used to build the knowledge base for this matcher; the objective of this project is not to build the ultimate knowledge base.

4.3 Design Alternatives

4.3.1 An Event-Based Matcher

One of the possibilities that was considered was an event-based matcher. This was based on the philosophical theory of events, which uses the term “event” to denote a

\[ I \text{ use } high-level \text{ here to mean information that is based not on the direct meanings of words used in the story, but information that has to do with implied meanings and inferences. For example, metaphor, simile and writing style.} \]
particular mapping from language to things in the world. Language is used to convey information about real-world happenings; different sorts of happenings are described by different language structures.

The idea of Events, Processes and States described by Parsons [16] is an attempt to differentiate the different sorts of occurrences. In Parson’s theory an Event has a “culmination” and a “holding”; States have no concept of duration, only existence or nonexistence (a State is either there or it is not), and a Process is related to an Event in that it has a culmination and a holding but also a duration (Processes take some time to happen).

Another concept of an Event was given by McNeill [10][Chapter 5, pp. 76-92] quoting von Wright’s calculus of events. von Wright’s definition of an Event is a change of state over time. It is represented as two successive states separated by an interval of time.

These concepts of events largely relate descriptions of actions and situations to time. If a certain amount of time is observed and segmented, each segment should contain a number of happenings which are events. Theoretically, it should be possible to construct a matcher that matches stories based on the similarities of events contains in them. The idea of Event and State may provide a way to extract information from stories and use this information as the basis to decide how similar two stories are.

Parsons [16], McNeill [10][Chapter 5, pp. 76-92] and Rothstein [20] all mention that events are related to verb phrases. Verb phrases can be divided into two categories; phrases containing instances of the verb “to be”, and phrases containing other verbs. This categorization is made based on the fact that the verb “to be” is a copular verb and has little meaning when not used in conjunction with another word, as compared to other non-copular verbs like “run”, “eat”, “exist”. This categorization is my own but has been described by many others. Roughly, phrases containing instances of the verb “to be” usually correspond to descriptions (States) and phrases containing instances of other verbs correspond to Events.

The possibility considered was to build a system that automatically extracts verb phrases (using some specification of English grammar) from all the stories in the
corpus, builds an index of these phrases, and matches by extracting phrases from the user story and returning the corpus story that contains the most similar phrases.

### 4.3.2 A Topic-Based Matcher

It was noticed that the stories in the corpus tended to fall into broad categories. Furthermore, there tended to be certain words that occurred repeatedly in these categories, that could possibly be useful in distinguishing those categories. This is not a new idea; this is the basis for most information retrieval based on keywords, but most information retrieval systems do not deal with narrative text.

To build a topic-based matcher, indicators for the various topics need to be found from the corpus – either by an automatic process, or through a human intermediary. After the indicators are obtained, indices describing the mapping of indicators to topics as well as the mapping of topics to stories would be built. The matcher would count the number of individual words in the user story that are indicators of some topic and return a story from the topic with the most indicators found.

### 4.4 Comparing the Two Possibilities

The two approaches are philosophically different. The topic-based matcher adds some real-world knowledge to the system, as it groups words according to topics which presumably are reflective of separate concepts “in the world”. However, that real-world knowledge has to come from somewhere; current computational knowledge bases are not good enough, so a human intermediary has to be used. This means that the corpus will have to be annotated by hand. Compared to this, the event-based system is automatic provided a good specification for English grammar can be found. This also makes the event-based system more extensible; if new stories are added to the corpus, the automatic extraction system can be re-run to generate the new indices. However, with the topic-based system, the new items will have to be annotated by hand, topics assigned and topic indicators found, unless the matching system is good enough to be unambiguous (which is highly unlikely).
Conceptually, the topic-based matcher is simpler to implement. The event-based matcher needs extensive specification of what an event is (at best, this is a vague philosophical notion and hard to quantify). It is not clear what the actual mapping is between language structures and real-world concepts in the case of events. Any real-world occurrence could be divided up into smaller events, and it is difficult to tell what the maximum granularity should be. For example, in a description of a meal at a restaurant, should the event described by “had a bite of steak” be accorded the same importance as the event described by “ordered the wine”? In any description of a story, there will be certain events that are less important. It is not easy to decide which events should take precedence.

The grouping of words and phrases according to topics provides limited capability for handling noise. Sometimes a word may be used in a metaphorical sense, as in “the field was carpeted with flowers”. Suppose “carpeted” is originally indicative of *house* and “flowers” and “field” are indicative of *countryside* – this sentence would trigger the detection of the topics *house* and *countryside*, but because more identifiers of *countryside* than *house* were triggered, would still preferentially trigger *countryside*, which is the desired effect. Of course, this fails if too many words are used in unusual ways and if they all happen to be, by chance, triggers of the same topic.

For a topic-based system, the list of indicators will never be exhaustive. Furthermore, we would be deriving the list of topic indicators from the corpus, which is not large (about 300 stories).

There is more to stories than just events and categories. One thing that needs to be considered when examining travel narratives is the role of proper nouns. Proper nouns are usually found as place-names. In travel narrative, travellers may tend to write about the places that they have visited. Thus, it may not be a good idea to simply match on topics or events without taking proper nouns into consideration.

Metaphor is a characteristic of text that would not be handled well by either system. It is impossible to encode all of the high-level relationships between metaphor and the subject being described. However, to this date there have not been any information retrieval systems that are capable of handling metaphor and simile securely.
in free text, so this lack would not be isolated to this system alone.

4.4.1 What About WordNet?

The WordNet lexical database [11] records the associations between words and concepts. This seems like it would be a useful thing to use, as it may help solve the problem of the knowledge base being too small as it is derived solely from the corpus.

An example of a good use of WordNet would be testing if “boat”, “sail”, “berth”, “deck” and “ocean” were all related conceptually. While WordNet is not exhaustive, it should be expected to give a better picture of “the world” as understood through the English language, than the corpus would. The WordNet database encodes that words may have more than one sense. For example, “deck” could be used to describe a part of a ship, or a part of a house. In order to properly utilize WordNet we need to be able to tell, from the text, what sense of the word is being used.

Li, Szpakowicz and Matwin [7] propose a WordNet-based algorithm for word sense disambiguation. They describe several heuristic rules used to determine the intended meaning of a noun object in a given verb context. Their algorithm produces correct noun meaning more than 72% of the time.

Unfortunately there is a practical barrier to the effective use of WordNet in the Victorian Laptop story matching system. The issue is with the WordNet Applications Programming Interface (API) provided by the creators of WordNet. WordNet provides functions that allow a user to query the WordNet database to determine the synonyms, meronyms and holonyms of a word. This query function takes two arguments, a word and an indicator of what part of speech it is (noun, verb, adjective or adverb). Unfortunately, if this function is called with a word and an incorrect part of speech 3, the entire program crashes fatally. This is documented in the WordNet manual [12]. Because the system is running under the Windows 95 operating system, we cannot use command-line redirection of the WordNet process. Therefore, until this problem is fixed in a future version of WordNet, or until the Victorian Laptop

---

3 For example, “posh” as a noun instead of an adjective. This is an actual example verified by experimentation.
system is ported to the UNIX operating system, it is not possible to make use of WordNet to disambiguate.

Even if WordNet is used, the computational complexity of the calculation is significant. Below is shown some sample WordNet output and the discussion following explores the issues of what has to be calculated to test whether “boat”, “sail”, “berth”, “deck” and “ocean” are all related conceptually. For each word, WordNet is queried with the following parameters: find synonyms of word as a noun, find meronyms of as a noun, find synonyms of word as a verb.
 wn boat -synsn
Synonyms/Hypernyms (Ordered by Frequency) of noun boat

2 senses of boat

Sense 1
boat
=> vessel, watercraft

 wn boat -hmern
Meronyms of noun boat

1 of 2 senses of boat

Sense 1
boat
   HAS PART: boat whistle
   => vessel, watercraft
      HAS PART: anchor, ground tackle
       HAS PART: fluke, flue
       HAS PART: shank, stem
      HAS PART: bilge
      HAS PART: bilge keel
      HAS PART: bow, fore, prow, stem
      HAS PART: hull
       HAS PART: keel
       HAS PART: keelson
       HAS PART: rib
       HAS PART: rider plate
      HAS PART: rudder
       HAS PART: rudder blade
       HAS PART: rudderpost, rudderstock
       HAS PART: tiller
      HAS PART: wale, strake
      HAS PART: washboard, splashboard
     => craft
       => vehicle
         HAS PART: splashboard, dashboard

 wn boat -synsv
Synonyms/Hypernyms (Ordered by Frequency) of verb boat

1 sense of boat

Sense 1
boat
=> ride

Doing the above for the example given (full transcripts omitted for brevity), we would find that “sail”, “berth” and “deck” are parts of “boat”. The connection between “boat” and “ocean” could be made by noting that “ocean” is a body of water, and “boat” is a water-craft.

For each of the results returned by these queries (noun synonyms, verb synonyms, noun meronyms) it is necessary to check if any of the other words being considered match, as well as their noun synonyms, verb synonyms and meronyms. Combination-
ally this becomes too much to handle if one considers that a single story may have
tens of different words, all of which have to checked against each other, as well as
each other’s synonyms (both noun and verb) and meronyms.

WordNet presents a lot of information, and it is necessary to sort through all this
information to find the relevant items.

4.5 A Design Decision: Topic-Based Matching

Based on the above reasoning it was decided to build a topic-based matcher without
using WordNet. The motivation for this is much like the motivation for WordNet in
that we are grouping words by concept, albeit very crudely and unspecifically. The
topic indicators would, for simplicity’s sake, be obtained from the corpus.
Chapter 5

Implementation

5.1 The Corpus

5.1.1 Culling and Segmentation

The original corpus consisted of historical first-person narratives and historical guidebooks. It was decided that the guidebook entries should be removed from the corpus, as they were not narratives and for the most part provided statistical data about New England that was not immediately relevant to the purpose of the device (which is to encourage creative expression in the user, not merely to provide information). This is an example of the sort of passage that was removed from the corpus:

Many bridges link Charlestown, South Boston, and the main with the peninsula. These structures are among the peculiarities of the place, in their fashion, their number, and their length. The first one which was built was that over Charles River to Charlestown, 1,503 feet long. The Old Cambridge Bridge, across Charles River to Cambridge Road, is 2,758 feet in length, with a causeway of 3,432 feet. The South Boston Bridge, which leads from the Neck to South Boston, is 1,550 feet long. The Canal Bridge between Boston and Lechmere Point, is 2,796 feet, and from its centre another bridge extends 1,820 feet, to Prison Point, Charlestown. Boston Free Bridge to South Boston is 500 feet; and Warren Bridge to Charlestown is 1,390 feet. Besides these bridges, a causeway of a mile and a half extends from the foot of Beacon street to Sewell’s Point, in Brookline. This causeway is built across the bay upon a substantial dam.
After the guidebook stories were removed from the corpus, the remaining stories (mostly travel journals) were segmented further if it was judged necessary. The original corpus was segmented using the Segmenter software tool [4]. Automatic segmentation was not always successful. Some of the segments that Segmenter returned were too short consisting of a sentence or two, fairly far removed from their context. Other segments were excessively long (several pages). Some hand segmentation was therefore necessary. When hand-segmenting, a long text was segmented if there was a major change in topic between paragraphs; in this case, the change in topic marked the segmentation boundary. In the case of entries which were descriptions of a trip that lasted several days, the boundaries between descriptions of separate days’ events were the segmentation boundaries. Long descriptions of a single day were usually kept as a single segment.

Even after the corpus was re-segmented, there are still some very long segments. This is because the reader could not find a suitable place to break the segment. This is, of course, a highly subjective decision. A different reader may well have found different ways of doing so. A better way of segmenting the stories might have been to get more than one reader to read the corpus and decide on segments taking into account the decisions of all of them (to obtain a sort of “weighted average”).

5.1.2 Annotation

A human reader (the author) read the entire corpus several times. Each story segment was assigned a short description that was fairly general (for example, “parade” or “description of scenery” or “food”, not “happiness at returning home”). After the entire corpus was annotated, the descriptions were collated into topics. The set of topics found was:

arrival, art, autumn, ball (in the dancing sense), books, buildings, christ-
mas, church, cold, college, commencement (graduation), comparison, concert, countryside, culture, departure, disaster, family, fashion, food, fourth-of-july, friends, home, house, hot, journey, landmarks, language, lecture, money, music, new-year, night, party, politics, rainy, shopping, sightseeing, spring, storm, summer, sunny, technology, thanksgiving, theatre, town, transportation, war, water, winter, working

The stories were then grouped by topic into separate text files, before extraction of topic indicators commenced.

This process was done keeping in mind that care is needed in the choice of categories. We need to choose categories that are universal and fairly objective. For example, *sea* is a good choice of category, and *happiness* may not be. The reason is that the *sea* is a name for an object in the real world; what one person experiences as the sea is the same as what another would experience as the sea. This is not the case for the concept of *happiness*; it is not a concrete thing, and its causes are far from standard. Therefore the system’s categories should be general, objective and universal – things which everybody should be expected to know about. This also increases the chance that the user will write about one of them – the user cannot write about something that she does not know, so our categories should endeavour to be as common to human experience as possible. In retrospect, some of the categories are not suitable because they are too general (for example, *comparison* and *sightseeing*).

Not all stories were able to be assigned topics; in particular, the set of entries by Henry James were very difficult to categorize. These stories were highly metaphorical and consisted of the author’s thoughts as he walked through Boston. An example is discussed later on in this section.

It was not always just the words in a story that led to the assignation of that story to a particular topic. In some cases the indicators of a topic in a particular story were at a very high level. Consider the following story:

---

[3] However, it may be possible to detect expressions of happiness by an author. For the moment, detecting the emotional affect of the text will not be considered.
Saturday 31st, Oct., at 5. A.M., a sailor from the mast’s head, called out ‘Land’ delightful sound, and it was amusing to watch the general movement which followed the intelligence. I sat on deck till dark, when I could perceive the distance light on Thatcher’s Island off Cape Ann. We went to our berths early, but at half-past 10 I was woke up, as I had requested, by the mate, to see Boston Light which had just appeared in sight. I went on deck, where Clara and Olivia joined me, and there remained, until we cast anchor in the stream close to the wharf, at 1-1/2 Sunday morning. The sky though hazy was sufficiently lit up by the moonlight to enable me to perceive the general appearance of the harbour. Boston Light is a very large one and shone like a bright star: marking our course for us; the coast is very low, and dotted with islands, which continued the whole way to Boston. They are well cultivated, and look prettily in the day-time, but of course I could only see their general form. Soon we got amongst the shipping, which ranged outside the wharves, reminded me of the effect of the Thames at London. Lights could be seen, here and there, amidst the darkness of buildings, but it was too dark for me to judge of the appearance of the city. After talking of all we should do the next day, we again retired to our berths, but it was some time before I could sleep; I kept trying to realize that I was actually in Boston Harbour; and about to see places and persons that I had so often heard of, and longed to visit.

The above quote is definitely about arrival. However, there are no words in it that, on their own, mean “arrival”. In this story, the elements that indicate the concept of “arrival” are much harder to define. “... a sailor from the mast’s head, called out ‘Land’ delightful sound...” is one such element. Reasoning about why this is evocative of arrival requires some associations at a very high level. We know that “mast’s head” is part of a ship, and the fact that the sailor called out “Land” implies that the sighting of land is an unusual occurrence – in other words, that they have not seen land for some time. This, coupled with the fact that they are on a ship, suggests that they are arriving. Also the author’s last sentence “I kept trying to realize that I was actually in Boston Harbour ... to see places and persons that I had ... longed to visit” suggests anticipation and the fulfillment of such.

Hence in the above story, the concept of “arrival” cannot be picked up from individual words. Even processing at the sentence level would require a lot of real-world knowledge, of a kind that is very difficult to represent computationally – in the above example, knowledge of the fact that ship voyages are long. The implications
this has for this system are that matches may be one-way. This means that the above story may be returned as a good match for one in which the user has used a designated "arrival" keyword, but it itself may not be detected as being of the topic "arrival" if it is used as input. This is why the human intermediary is needed to assign topics to stories; because the mere presence of keywords is not enough. Of course, this means that we will sometimes fail to detect stories that are about a certain topic.

Another example of this occurred in many of the stories that were assigned the topic "family". There are several stories in which the author describes her family by name; it is only by the repetition of these names throughout the corpus that we know that these names come to indicate the topic of family. Thus, such associations that occur across stories are not detectable through keywords.

Certain topics, while very concrete and universally understandable, present the problem that it is virtually impossible to document all of the words that could trigger that topic. Examples of these are food and lecture. There are an infinite number of different types of food. There are an infinite number of subjects that a lecture could be about. Because the corpus is restricted in terms of the time periods described and location of travelling described, we can select as keywords words that are indicative of these restrictions upon the domain. This simplifies our task but does not solve the problem completely. For example, "chowder" is a word that we would expect to be related to food in New England. It is reasonable to expect that a contemporary visitor to New England would encounter chowder at some point. However, if we wanted to compare the types of food available now in New England (say, the user writes about the fabulous sushi dinner they had last night) with the types of food available then, this system would fail to pick it up this association.

The obvious thing to do here is to pick keywords that are more generally indicative of categories. For example, "silverware", "waiter" and "napkin" are items that are likely to be found at any restaurant, and are thus indicative of food. This is not so for nouns that are descriptive of the type of food; those are extremely specific and virtually uncountable. However, adding this sort of knowledge to the system requires the designer of the knowledge base to codify his knowledge of the concept of food and
make a guess about what is most likely to occur. These words did not occur in the corpus in the stories that were categorized as being about food, probably because the author tended to describe only the unusual things about her situation (the new types of food) and not the things that were common (waiters and silverware and napkins).

One of the largest failings of the topic-based keyword system is its failure to pick up on high-level content. There were several stories in the corpus that compared life in the place being visited to life at home. For example:

*The American girls have much more ease and grace of manner than the English generally have, this may partly be accounted for by their being brought forward earlier; they receive their own visits, do their own shopping etc., from the age twelve or even earlier. They remind me somewhat of the Portuguese girls, though much prettier; but their figures are slight, their hair abundant and beautifully dressed, and their complexions generally pale. They show much taste in dress, and wear pretty ornaments in ribbon and lace, mixed with their hair, which set them off exceedingly. You never notice that vulgarity of appearance so common amongst the lower class in England, often at theatres and concerts, do I look around amongst an audience composed in a great measure of shopkeepers and mechanics’ families, and only detect their class, from a more drawling pronunciation, and a few of their remarks: - none of those shocking vulgarisms so common in England.*

The above passage compares the appearance of young women in America and England. This could be categorized under the topic *fashion* as that seems to be a major point of comparison. However, the act of comparison that takes place in the story is interesting in itself, because this sort of comparison of the way of life between the new place and home is something that travellers do a lot of. Again, this example requires analysis at the sentence level to pick this up – the pattern "< X > has more < Y > than < Z >" is a possible candidate for indication of such comparison. Although there was a topic *comparison* in the list of topics of the corpus, *comparison* is very difficult to detect at the word level.

Many of the stories by Henry James are of this form; they are very introspective and *imply* a lot of things, but do not state them outright. For example:
It was on two or three such loitering occasions, wondering and invoking pauses that had, a little vaguely and helplessly perhaps, that changed crest of Beacon Hill for their field—it was at certain of these moments of charged, yet rather chilled, contemplation that I felt my small cluster of early associations shrivel to a scarce discernable point. I recall a Sunday afternoon in particular when I hung about on a now vaster platform of the State House for a near view of the military monuments erected there, the statues of Generals Hooker and Devens, and for the charm at once and the pang of feeling the whole backward vista, with all its features, fall from that eminence into gray perspective. The top of Beacon Hill quite rakes, with a but slightly shifting range, the old more definite Boston; for there seemed no item, nor any number, of that remarkable sum that it would not anciently have helped one to distinguish or divine. There all these things essentially were at the moment I speak of, but only again as something ghostly and dim, something overlaid and smothered by the mere modern thickness. I lingered half-an-hour, much of the new disposition of the elements here involved being duly impressive, and the old uplifted front of the State House, surely, in its spare and austere, its ruled and pencilled kind, a thing of beauty, more delightful and harmonious even than I had remembered it; one of the inestimable values again, in the eye of the town, for taste and temperance, as the perfectly felicitous "Park Street" Church hard by, was another.

From this passage one can pick up that the author is comparing Old Boston with new developments. This was written by Henry James in 1907. The feeling of nostalgia is important in characterizing this passage, but this is something that would not be picked up by the topic-based keyword matching system. The indicators of nostalgia in this passage are at an incredibly high level—"something overlaid and smothered by the mere modern thickness", "much of the new disposition... being duly impressive, and the old uplifted front of the State House".

Good creative writing often contains words used in a less than obvious way. For example, the use of the phrase "I lingered half-an-hour..." creates a much different mood than would have been made by the use of "I stayed for half an hour..." instead. The use of this, coupled with words such as "ghostly", "dim" and "gray perspective" paints a picture of the authors seeing the shadows of old Boston overlaid upon the new.

From these examples we may postulate that the topic-based keyword-matching system will perform reasonably well with stories that describe concrete things con-
cretely. Stories which contain much implied knowledge will be hard to classify because they do not contain concrete keywords.

5.2 Extracting Topic Indicators

5.2.1 Method

As the corpus was small (about 300 stories), it was decided to extract topic indicators manually for two reasons: firstly, 300 stories is a manageable size for manual extraction (which is simple to do but tedious), and secondly statistical extraction only works for corpora of much larger size (in the thousands of stories). The latter is because the accuracy of statistical extraction depends on the assumption that the corpus is a good sample of all possible stories. The larger the corpus is, the better a sample it is of the universal set of all possible stories. If the corpus is not large, the distribution of the words in it will not be close to the distribution of the words in actual usage, and we will get “key” words that appear many times in the corpus that are there many times merely because of chance.

Although statistical extraction was not used as a means of extraction itself, a statistical list of words in each topic was produced by the WordSmith Tools [22]. A consistency analysis was also done (described in the following section); this compared all words in each topic against each other topic. Keyword identification proceeded in two phases; first, from the consistency analysis and second, from individual topic wordlists. This was because the consistency analysis identified larger trends, and the topic wordlists were more specific and provided detail that the consistency analysis often missed.

Consistency Analysis

The WordSmith Tools consistency analysis tabulated how many different topics a word appeared in. Words that appeared in many topics were not considered good topic indicators – if the word was found in a user text, there would be too many
possible topics that it could indicate. For example, the word “made” was found many times in many topics. It was not picked as an indicator for any topic, because there were too many contexts it could be used in. We want keywords that distinguish the different categories.

If a word appears in only two or three topics, it could possibly be an indicator of that topic. This was not always the case but was a good place to start. For example, the word “steamer” appeared many times under the topics of “transportation” and “water”. This word was indeed picked as an indicator for those topics. However, this was not always the case. The word “Harriet” appeared many times in the categories of “friends” and “music”, but was not selected as an indicator of these categories. This is because the author of many of the stories about “friends” and “music” had a good friend named Harriet, with whom she visited other friends, and went for many concerts. “Harriet” was therefore indicative of the writer’s personal circumstances, not of the topic she was writing about.

**Topic Wordlists**

Using WordSmith, a list of words (excluding stopwords) was created for each topic. These wordlists were vetted by a human reader to determine the actual keywords. The words in the lists were ordered according to frequency. However, because the corpus is so small, the number of stories in a topic was usually small (less than 10), so frequency was not a good indication of whether a word was a keyword for that topic or not. For example, in the set of stories embodying the topic “commencement”, the word “students” appears only once. Despite the low frequency, it was still taken as a keyword for this topic. In this case, the topic “commencement” had very few stories in it (fewer than 10), so frequency analysis is probably not valid here. The most frequently occurring word in this category is “class”, and that occurred only five times, so the reason for the low frequency is that there are too few stories. In categories for which there were many stories, the words which occurred with low frequency (only one or two occurrences in that topic) did indeed seem to be irrelevant. This suggests that if the corpus were larger, statistical analysis would work better.
There is another reason that frequency of a particular word in a topic wordlist may not be indicative of its “keyness” for that topic. This is because stories are allowed to have more than one topic. When this happens, the words for all those topics appear in the wordlists for all those topics. This could be overcome by better segmenting, but not always – as mentioned before, the stories are mostly amateur-written, and there are some which are not organized according to topic. For example, there are stories which switch back and forth between topics at the sentence level.

In short, the consistency analysis and keyword lists were good places to start but far from definitive. The lists still had to be extensively vetted by a human reader. The human in the loop is, however, far from fallible; the choice of keywords is subjective, and might be completely different if another person is doing it. As mentioned above, the decision to discard a word in a topic wordlist as a keyword list is made if it seems that the word is not related enough to the topic. In the case of topics with few stories, it is possible that there are some words for which it is not immediately obvious that there is a direct connection with the topic, but statistically it could be possible, and the only reason it does not show up in the topic wordlist is the small size of the corpus.

The method that was eventually used to find keywords was rather labor-intensive. It took (estimated) about 50 hours to read the corpus several times, assign categories, make word lists and phrase lists and read those to decide on the final keyword indicators. One possible way to decrease dependence on a human reader would have been to use some already existing knowledge base to derive the keywords from the categories. This is a different approach; it is non-corpus-based in the sense that only the categories are derived from the corpus, but the indicators of those categories are not.
5.2.2 Finding Good Topic Indicators

Single Words

It was found that word inflection was significant in indicating topic. For example, “church” was found to be indicative of the topic church, and that “churches” was not indicative of church but rather the topic buildings. This is because in the corpus, there are many occurrences of phrases like “going to church”, “went to church”, etc. in stories that were tagged as being in the topic church, but phrases like “many beautiful churches” in stories describing buildings or landmarks in the places being visited. Another example was the word “volumes”. This was found in the topic books, but the single word “volume” was not found to be significant of any topic.

Due to this reasoning, at first, keywords were not stemmed. In practice, however, preliminary testing of the matcher showed that many keywords were not being detected because not all of the inflected forms of a keyword were present in the keyword list. Rather than adding all the inflected forms of a keyword to the keyword list, it was decided that it was simpler to just stem all the keyword (and all input to the matcher), realising that this might add some noise to the matcher (because then “serve” and “serving” would be detected as triggering “service”).

The number of keywords for each topic varied. For example, the topic friends had much fewer keywords than the topic food. This is because most of the stories about friends described friends by name (and names are not accepted as keywords for reasons described previously), and many stories about food tended to describe the sorts of food eaten, spectacular meals, etc. This does not mean that the topic food is easier to detect than the topic friends; it may even be more difficult, because the possible list of foods is not exhaustive and never will be. More general keywords should be used.

Phrases

The choice of single words as indicators was not arbitrary. Bearing in mind the results from the AutoSlog project [19], phrases were considered as possible candidates for
indicators. To examine whether there were topic indicators that were larger language structures than single words, a program was written that would extract phrases from parsed text, and run on all the stories from each topic. Noun and verb phrases (subject to some limitations) were extracted using a Finite State Machine built on top of the Libero template software [3] and written in the Java programming language. The resulting lists of phrases were read by a human reader who decided whether they were considered indicative of the topic or not.

Phrases were considered as topic indicators only if they did not contain any single words which by themselves would be good topic indicators. For example, the phrase "went to afternoon service" was considered a possible indicator for the topic church because each individual word in it is not usually associated with church, but the collection of those words together means something more than their individual meaning. The phrase "went to church" would not, under this criterion, be considered a good indicator of church because it contains the keywords 'church' that is sufficient by itself to indicate this topic.

The results of this were not sufficient to justify implementing this as part of the matching system. There were very few, if any, such phrase signatures for topics. Most of the topics had no phrase indicators. To include this as part of the matcher would require adding a parsing module to the system to parse all user input, as well as the phrase extractor. Furthermore, the "dictionary" of topic indicators would have to include parse information, which takes up much more space than raw text information.

This would be a viable option if there were strong indicators of topics which were phrases. In this case, there were none that could be detected from the corpus. This is probably due to the fact that the corpus is small, and so the number of stories in individual topics is also small. It could be concluded that the individual topics do not contain enough text to provide enough samples of phrases that could be strong indicators. This does not mean that only single words can be good indicators of topics. It depends on the topic; some topics may have metaphors that are often used to describe the topic. As a hypothetical example consider the topic of death; a common
euphemism for death is “kicked the bucket”, which would be a good phrase indicator of this topic. A large enough collection of texts might contain several occurrences of this phrase, which could then be indicative of this phrase being “key” to this topic.

5.2.3 Topics, Indicators and Context

The grouping of words under topics works under the assumption that all indicators of a topic will be the contexts for each other. This means that if the story is seen as a sample description of the real world, then in real life we would expect a higher probability of seeing the words in the story together. For example – if we have a story with “Thanksgiving”, “parade”, “police”, “horseback”, we might expect that the real-life concept of Thanksgiving may be related to horseback and that if we experienced one, we would experience the other at the same time.

5.3 Matching

5.3.1 Matching Only On Topics

The matcher was implemented in stages, with iterative design. Features were added, the matcher was tested, and changes were made if necessary. The basic idea for the matcher is that it contains indices which encode the mappings from words to stories. The first generation of the matcher contained only one index, which encoded the mapping from keywords to stories and topics. A conceptual diagram of this index is in Figure 5-1.

Each keyword maps to a list of pairs. Each pair maps a particular story to a particular topic. One story may have more than one topic, and one keyword may be indicative of more than one topic. In this implementation, the user input text is scanned for topic keywords. When any keyword is found, the list of stories that it maps to is added to a list of possible
match candidates. When all possible keywords have been found, the final list of match candidates is sorted by topic. The topic that is the most often triggered (the topic that is mapped to by the largest number of keywords in the user stories) is selected as the best match topic. The stories in the final list of match candidates that correspond to this topic are then searched for the keywords in the user story; the story that shares the largest number of keywords in the user story under this topic is then returned. The matcher contains a record of which stories have been returned, and if the best match story has been returned before in this session, another story is found.

### 5.3.2 Taking Proper Nouns Into Account

The problem with this matcher is that it does not take into account proper nouns, which are fairly important in the domain of travel, as they encode place-names. It is fairly expected that the user will choose to write about places that they visit. A better matcher would recognize that proper nouns were present and take those into account while matching. Proper nouns and topic keywords are orthogonal features;
both may be present, and the type of match returned may be different depending on which feature is used.

Of course, the best match taking both features into account would be a story that is about the same topic, and contains the same proper nouns as in the user story. If this is not possible, which feature should take precedence? It is difficult to quantify which feature should be more content-bearing. A good heuristic might be to consider that it takes more topic keywords to indicate a topic than it takes proper nouns to indicate a description of a particular location. This was used in the implementation; if the strength of a topic match was twice or greater than the strength of the proper noun match, then the topic-based match was chosen.

This was used in the next iteration of the matcher. An index of proper nouns was added. This index was built from the word list of the previous generation matcher. This time, both proper nouns and topic keywords were detected in the user input story. A story (in the corpus) was considered to be “detected” if an indicator of it (proper noun or topic keyword) was found in the user input story. The same story in the corpus could thus be detected multiple times – either if multiple words in it were detected in the user story, or if the same word it in was detected multiple times in the user story.

The counts of detected stories are kept separately for proper nouns and topic keywords. The most commonly detected stories are tallied for each case (proper nouns and topic keywords) and compared at the end. The “strength” of a possible match is defined as the number of times that story was “detected” in the user story. Therefore there are two possible strongest matches: the story which is the most triggered according to the number of proper nouns in the user story, and the story which is most triggered by the topic keywords in the user story. These two are compared; if the strength of the topic keyword match is more than twice than the strength of the proper noun match, the story returned is the topic keyword match.

If there was a story which matched on both proper nouns and topic keywords, this story was returned.
Chapter 6

Evaluation

The evaluation of this system should be done from two angles. The first is microscopic—an analysis of how the matcher performs given different sorts of stories, bearing in mind the constraints of the corpus, the user input and the knowledge base. The second is from a higher-level point of view—testing of the system by a human audience to see whether the matcher performs its function *within the Victorian Laptop project*, of evoking creativity in writing. Due to time constraints only the first was performed, but the second will be discussed.

It is difficult to evaluate the performance of something as subjective as story matching. With this in mind it was decided that the best course of action would be to present the matches given by both the new matcher as well as the previous generation matcher. The input story is presented, then the new matcher output and the old matcher output, and finally a discussion (as objective as possible) of how the two compare to each other. In the evaluations of responses to individual stories, not just the performance of the matcher is discussed but also the appropriateness of the keywords found, the quality of the segmentation, etc.

The system was tested with some artificially constructed stories and some actual travelogue. “Artificially constructed” means that the stories were written to fulfill specific requirements as to topic or number of keywords. In these cases, being “on topic” meant that the story addressed some single specific abstract concept (not a specific location).
6.1 Constructed Stories

In these cases, the constructed stories were constructed such that they contained few proper nouns in proportion to other words, as the point of this is to test how well the topic-based matching works. Where mentioned, “keywords” indicates keywords found from the corpus through the process of assigning topic indicators discussed in the previous chapter.

In the following transcripts, the following should be noted: the text in teletype font denotes the non-story output of the matcher that describes its internal working. Also, the topic keywords have been stemmed using the Porter Stemmer \[1\] [17][pp. 130–137].

6.1.1 Strongly on topic, containing keywords

The following story was written as an example of a story containing many keywords indicative of the topic *arrival*.

I arrived in Boston last October for the first time. I was glad to have finally reached my destination after so many hours of travelling. As the plane landed I thought about how I was finally going to see the people I had always wanted to visit. Sure enough, my host family was waiting at the airport to welcome me.

New Matcher Output

*Topic keywords found:*
*[arriv, reach, travel, land, peopl, visit, famili, wait, welcom]*

*Proper nouns found:*
*[boston]*

*Best possible topics are:*

---

\[1\] The Porter Stemmer stems words by attempting to remove common suffixes. It uses heuristic rules to perform stemming, and thus the resulting word stems may not correspond to actual word stems in normal English usage. It also sometimes stems words unnecessarily. For example, the word “bay” is stemmed to “bai”. However, it is consistent, even if it is sometimes incorrect. For example, the Porter Stemmer stems “arrival” and “arrived” to “arriv”.

49
Monday, the 1st, November, the anniversary of my arrival in this country. I left my kind friends and came down to New York with Mrs. Sandford and Mr. Birsted. An amusing little adventure occurred on board which I must relate to you. On entering the cabin, I noticed rather a handsome looking man; between forty and fifty years of age who seemed to be a foreigner. He looked at me very earnestly, but I left a minute after, to sit outside with Mrs. Sandford. In a few minutes Harry came and told me that there was a Colonel King on board, who had been several years in South America in Bolivar’s services, and afterwards in Buenos Ayres, and he wished to know whether I could speak Spanish. I said perhaps Portuguese would do, and Harry brought back the Colonel in a few minutes, and introduced him to us all. He sat down by Mrs. Sandford, and an animated conversation soon began between him and the Padre about the Mexican War, the affairs of Buenos Ayres etc., and he told us many curious adventures and particulars of his own history, by which I found out that he was a daring adventurous person, naturally shrewd and quick, though entirely self educated, and one who had seen much of the world. By and by he asked Mrs. S. to let him sit by me, as he wished to compare Spanish and Portuguese. He accordingly began, but in a short time commenced compliments, and ended by asking me to marry him. He told me that such sudden likings had often occurred before, and that he had taken such a fancy to me, that he was certain no one else would suit him. Promised that I should travel where I liked, and was sure that he would make me a very good husband. I told him, I dare say he might, but that I, on my part had no fancy for such sudden affairs, and that I did not think, I should like him well enough to marry him. Well he continued some time in the same strain until we reached New York; tried to persuade me to write to him, and at last told me that he should find out where I was staying in Boston, and hoped he might come to see me. I shall take care and not let him know where I am, for he is one of those determined persons, that might prove very annoying. Mrs. Sandford caught a few words of our conversation and was curious to know what he had been saying. She enquired of her husband about him, who says all that is known of him is that he is a man of great courage and enterprise, and that he returned from South America two or three years back with a considerable fortune, a great deal of which he has invested in land, but of his adventures there, no one knows anything. We arrived in New York late, took tea, and then harry escorted me as far as the Stalkers. I found them all in the parlour expecting me; Mr. and Mrs.
S. received me very cordially, and introduced me to a Mr. Newberry, a gentleman lately arrived, and now partner with Mr. S. in business. He is a gentlemanly man, been a great deal abroad, and converses well. Emma looked as pleasant and amiable as she always does, and seems to have enjoyed her trip to the south very much. She offered to assist me in getting anything I required before leaving for Skeneateles, and I feel very grateful to her for her unvarying kindness to me. The next day I had to go to the Sandford’s to get my trunks, which I had left there. Dr. Broteiro came in, and talking about the sights of New York, Mrs. S. enquired if he had seen the Ethiopian Serenaders, and on his replying in the negative, invited both him and me to accompany herself and family there that evening. We both accepted very willingly, and I got a hack, and went up to the Stalker’s with my goods and chattels, and arranged them all in my room, and then set off again without my dinner, as the Stalker’s do not dine till nearly six, and the Sandford’s live two miles further down. I found Dr. Broteiro there, Lucy and her cousin just come from Sing Sing, and two other gentlemen. We made a cheerful party, and set off, nine in number to the theatre. These Ethiopians are really white men, dressed as blacks, and their music gave us great pleasure. They play on the accordian, banjo, cymbals, and bones, and make exquisite music. Their voices are fine, and the harmonies are very beautiful. We went home to a nice little supper, and Dr. Broteiro escorted me home in the cars, and took leave of me as he was going south. Since then I have been busy shopping, and visiting. Seen my old friends in New York, made acquaintances with the Porters, cousins of the Burnetts, and got every thing ready to go to Skeneateles, when the sad news of Mr. Burnett’s illness put a stop to it. I have attended a concert of Herz and Sivori and enjoyed it highly. But I must say farewell for the present dear friends, as my eyes are tired with writing so much. I shall send you some more journals by the next opportunity.

Old Matcher Output

Keywords found: [arriv, boston, glad, visit]
Match strength = 0.0069204154

*Amelia Burnett, Journal of a Visit to the United States, 1846-1848* Source: Massachusetts Historical Society

It was about seven in the evening when we arrived, and we were both tired and hungry, so were glad to sit down to tea, and have our plates filled with raspberries and currants.

The new matcher detected quite a few topic keywords indicative of the topic *arrival*. The story it returned, however, was very long and not only about arrival.
There were several elements that were common to the user story and the returned story, such as the ideas of family and travelling.

The old matcher output is very short and the sense of “arrived” used in this is more returning home than arriving in a new place (small-scale change vs. large-scale change).

6.1.2 Strongly on topic, containing few keywords

The following story was written as an example of a story about a topic food but not containing many keywords defined as indicators of this topic in the corpus.

Last night we went for dinner at a Japanese restaurant in Porter Square. We had some very good sushi and the price was excellent. It was comparable to what I had in Japan when I was there last. We ordered a large selection and were not disappointed.

New Matcher Output

Topic keywords found:
[night, dinner, restaur, good, larg]
Proper nouns found:
[squar]
Best possible topics are:
countryside
food
By topic match, story 5 was detected 5
By PN match, story 6 was detected 2
Best match is by topic

Eva Blackwelder, Tour of Massachusetts, 1938-1939 Source: Massachusetts Historical Society
[sightseeing in Cambridge, Quincy, Natick, Wellesley, other activities in Boston]
12/12/38 Wylie Allen took us to lunch at Durgin Park no 30 N. Market below Faneuil Hall over the Market. A large place where rich and not so rich come to lunch breakfast or dinner. Dinners cooked behind a long counter and wood tables with little white table cloths. Had wonderful
chicken lobster. I browsed in the market afterwards and found everything very find but not cheap. For 150 years Durgin Park has been a restaurant.

Old Matcher Output

Keywords found: [dinner, larg]
Match strength = 0.0060790274


Oct. 17. Dined with Joseph Barrell, Esq., by appointment. A large company of gentlemen present, among whom were Judge Lowell, Judge Iradell, Judge Sullivan. An elegant dinner & delicious high flavored wines.

Both systems returned matches that were about dining. However, the new matcher returned a story that was more specific to the user story – in this case, a description of dining in a restaurant. Although the user story did not contain many topic keywords, the ones that it did contain seem more strongly indicative of the topic. For example, “dinner” and “restaurant” are general indicators of the act of eating; the types of food that one may consume for dinner or in a restaurant are varied, but if “dinner” or “restaurant” is mentioned then it is definite that the act of eating is occurring.

This shows that general words work better as topic keywords.

Also of note is the fact that countryside was detected (incorrectly) as a possible topic for the story. This was because the keywords “night”, “good” and “large” were found through the WordSmith consistency analysis to be indicative of countryside. This just indicates that in the corpus these words are associated with countryside, but in other usage, they may not be. They are very general words applicable to many situations, and the decision to include them as keywords indicative of this topic was made because they occurred so many times within stories of that topic.

6.1.3 A story with no topic

The following story from the corpus was selected as an example of a story with no topic. It was taken from “The Traveler’s Life”, an online travel diary by Steve
I stayed by myself on the steps for a half hour or so, enjoying the sunshine. I wasn’t alone per se – the steps of the public library are a popular meeting place – but I was starting to feel isolated. Several attractive women had found similar perches nearby, but, alas, I couldn’t work up the nerve to approach any of them. I could feel that slippery spiral of loneliness starting to build. Deep inside, I knew I wouldn’t be the one to approach a complete stranger. What I was waiting for was someone to come up and talk to me. Unfortunately that didn’t happen, so I retired back to the hotel.

In the evening we amused ourselves looking at the Naushon book, which records the names of all the visitors to the Governor, receives their literary contributions, and records the many gay parties, and merry deeds enacted here. Mr. Swain continues his old fondness for charades, and has composed several, which he has set us to guess; and in this puzzling amusement, a walk on the piazza by moonlight, and some songs from Maryanne and Miss Drinker, our first evening was passed, and M. and I retired to bed in high spirits, looking forward to much enjoyment during our visit. At a quarter past six the next morning, the sound of the gong roused us from our slumbers, and by seven the whole family was at the breakfast table. The meals here are very pleasant; everything of the best and plenty of it, and perfect neatness and order, without pretence to style. After breakfast the ladies of the house disappear, and the visitors amuse themselves as they like. Mr. S. generally sits on the piazza, reading, smoking, or talking. There are plenty of books in the library, and I amused myself with Willis’s works, lately presented by the author to the Governor after his visit on the island. Then we played my favorite game, billiards,
Mr. Swain and Maryanne against Bruce Upton and myself. At eleven the horses were ready and Mr. S., Maryanne and I went off riding. The day was a bright one; the sky without a cloud, just like our sky in Portugal; our horses were small and active, and well acquainted with the forest paths. It was something new to me to be riding through such thick woods, composed principally of oak and beech, covered with grey moss: every now and then we would come upon some open ground where we could enjoy a fine canter, and after one of these my horse carried me towards a tree, and before I was able to turn away, a large branch across the path put a stop to my further progress, my horse passed under it, but I should have been violently thrown, had I not seized the branch with my hands, and bending back upon the saddle, allowed the horse to pass from under me, when I jumped to the ground. Mr. Swain called out "Very well done", and with his assistance I soon mounted again. Dinner at one, and in the afternoon we again took a ride on horseback, and I enjoyed it even more than the morning. The shadows produced beautiful effects amongst the open ground, enjoying new beauties at every change. There is a pretty glen near the house, where I went twice to read; rocks and trees interspersed reminded me of dear Cintra, and perched up in an old tree, Maryanne and Annie Upton enjoyed the quiet beauty, and listened to my reading of Willis. There was one very beautiful walk through the woods, and over three bridges which connect some little islands with the larger one. The sea runs up into the land, and in an opening of the woods above is arranged a pretty arbour, where we sat listening to the ripple, and watching the sparkling of the sea, half concealed from us by the trees. Here we would talk of Robert, so beloved and so deeply regretted. I visited his grave, alone in the woods, with Miss Swain; his Mother often goes there to weep. Near the house in another direction, is a pretty little lake called Mary's lake, where we went rowing one afternoon. You may judge from what I have told you how pleasantly our time was passed, the last afternoon we girls drove down to the south beach where there are bathing houses, and took a delightful bath in the sea, how we enjoyed it! I took a great liking to Eliza Swain and Miss Drinker, and had some long conversations with them; Mr. Swain made himself very agreeable, and kept us well amused. On Saturday morning the Uptons and ourselves left Naushon with regret and returned to New Bedford. We found Aunt Amelia and Sis well, and all much as usual. In the evening got a note from Lucy Sandford expressing her willingness to accompany me to Niagara under her brother's escort. I was perfectly happy, as I had long been making every inquiry as to the possibility of visiting Niagara, and as yet heard of no party going.
Old matcher output

Keywords found: [staye, enjoi, public, public librari, librari, found, build, hotel]
Match strength = 0.004950495

Amelia Burnett, Journal of a Visit to the United States, 1846-1848 Source: Massachusetts Historical Society

Friday afternoon. Eben left us. We were all sorry as he had become a general favorite, and seemed to have enjoyed himself greatly. In the evening, the Dr. going into his study, found a note for each of us left by Eben, with some little remembrance. Mine was a mother-of-pearl paper knife.

Neither matcher seemed to return a strongly relevant match. Again, the new matcher returned a very long excerpt that contained many keywords just by virtue of its length. In this case, it would have been better to return a story about the Boston Public Library, but because there were so many topic keywords found, the topic-based matcher took precedence. The topic-based matcher actually picked a topic for this story that was not too far off the mark (friends) but that was largely serendipitous.

The old matcher returned a short excerpt containing two keywords, “enjoy” and “found”. Neither of these was used in the same way that they were in the user story. “Enjoy” and “found” are both words that are very general and can be used in many situations. Therefore, they may need additional clarification in order to be good keywords.

6.2 Tests On Actual Travelogue

Both matchers were tested with excerpts from an online travel journal by Steve Leroux [6]. The excerpts selected were chosen such that they were fairly short segments about a single topic, or as close to that as possible.
6.2.1 Excerpt 1

We (a co-worker and I) flew in late Thursday. A quick, aggressive cab ride from Logan Airport to the Copley Square Hotel didn't leave me with much in the way of initial impressions. Freeways and traffic, the scourges of every American city. At least the Bostonians are trying to hide theirs with the "Big Dig" – a ten-billion-dollar project to bury all the expressways under the city. The Copley Square hotel is small and pleasant – nice, but not quite charming enough to be called quaint. The rooms are functional but the hotel is definitely lacking some of the amenities of its larger brethren. $160 wasn't exactly a bargain – but then again, it wasn't too bad. But then again (again), the company was picking up the tab for this one. After checking in we met downstairs to find a late dinner. The doorman directed us towards Skipjack's, a few blocks down the street. The weather was warm – a typically humid East Coast evening. Great for seeing the late-night sites, but not so good for sleeping.

New matcher output

Topic keywords found:
[cab, ride, hotel, leav, citi, citi, hotel, pleasant, charm, room, hotel, compani, met, dinner, street, weather, coast, even, good]
Proper nouns found:
thursdai, coplei squar, squar, hotel, american, citi, bostonian, citi, coplei squar, squar, hotel, room, hotel, street, east]
Best possible topics are:
friends
By topic match, story 150 was detected 10
By PN match, story 0 was detected 26
Best match is by PN

Eva Blackwelder, Tour of Massachusetts, 1938-1939 Source: Massachusetts Historical Society

Arrived in Boston 11/20/38 and went directly to see the Charlesgate East apartment. There were no vacant ones, must write and tell Miss Mildred Vincent....We spent three hours vainly looking for an apartment, all were rented by the year. We even went to Cambridge across the Charles River where we saw a poor place in a fine building, where the only bed was a sofa. We finally went to the hotel Westminster ($3.50) in Copley Square across from the Copley Plaza Hotel, a gorgeous hotel taking up a whole square block. (rates $7-8-12) We expected a F.H.A. a rate at the Westminster. We had room without bath. It is a very shabby hotel, traveling salesman
type. No closets but a wooden wardrobe. The faucets in the basin were old and leaked and everything was worn. The public bath was a mess, the floor and tiled walls spotted. My heart sank, what would we get for less? That night we went over to the Copley Plaza to get a Boston map (free) and I asked the clerk if he knew of a furnished apt. house or a real estate agent. He was about to put down a real estate agent when his face brightened and he gave me the name of a friend Mr. Sawyer, Fensgate Hotel, Kenmore 4460 (the Standard Commonwealth Corporation) I called next morning and he said he had a furnished apt., which was vacant for two weeks. Dec. 8th it was rented. I went to see it, usually it rents for $110 a month he would let us have it for $21 a week everything included. 11/21/38 If we could only have it all the time we are here! It just suits us. It seems luxury to us for there is a drawing-room, a kitchen and a lovely new bath room all spiffy and fresh and clean. It is in the center of Boston 12 Commonwealth Ave. Facing the Hamilton Statue in the parkway of the Ave. there are two broad lawns of grass lovely high trees and a pathway down the center of the green. The street is faced with tall narrow houses, three to four windows in width, brown stone or brick, partition walls and stand four to seven stories high the roof about on a level, with masses of chimneys shooting into the air. Some of the houses have entrances flat with the side walk, some go down a flight or are approached by high stone steps. From our windows [706] Apt we can see the Public Garden with its long wide lake the equestrian statue of George Washington in greenish bronze standing boldly in front and way beyond the garden we can see Boston common with its rising green up to the State House with its golden cupola and on the north of the Common on the hill, house upon house stone and brick, rise one after the other a huge mass of masonry like an old Italian town. On the south of the garden and beyond the Boston Common shops shops shops, and beautiful ones, as fine as in New York. The people look prosperous and those one meets on the streets in this part of town are very smart looking both men and women. Many of the older women wear flat heels and a serious walking expression and many have a high-bred look. ..... [descriptions of Waltham, Harvard Square]

The old matcher output was identical to the new matcher output. This is probably because most of the proper nouns found by the new matcher were the same as the keywords found by the old matcher. At any rate, the match seems to be quite a good one – Copley Square and hotels figure widely in both the user story and the returned story.

However, the topic matcher is not doing a good job. This user story should have been classified under the topic arrival, not friends. The strongest reason for this
mis-classification is that the indicators of arrival in this story are not found in the corpus at all, so would not have been picked up during the analysis of the corpus. Some indicators of arrival here are “flew in”, “initial impressions” and “checking in”. They are all phrases, and one of them ("flew in") is a phrase that would never be found in the historical corpus.

This demonstrates one failing of the technique of deriving the topic indicators from the corpus. Because the corpus is historical, modern indicators of arrival would never be present in it. Also, as mentioned before, the small size of the corpus limits the number of possible words in it.

6.2.2 Excerpt 2

The entrance to Skipjack’s is framed in blue neon that is nearly painful to look at. Inside, however, is a nice upscale seafood restaurant. We shared a grilled sashimi tuna appetizer (I’m not sure if something can be grilled and still be considered sashimi...) and then hit the main course. I went for a nice light seafood salad. My coworker went for the lobster. I guess I have a thing or two to learn about dining on the company’s credit card. Anyway, the food was decent but not spectacular. A true feast of Bostonian proportions would come later.

New matcher output

Topic keywords found:
[blue, look, restaur, lobster, dine, compani, food]
Proper nouns found:
[pain, bostonian]
Best possible topics are:
food
By topic match, story 39 was detected 5
By PN match, story 281 was detected 2
Best match is by topic
*Amelia Burnett, Journal of a Visit to the United States, 1846-1848* Source: Massachusetts Historical Society

*Wednesday. We had a delightful day. At 2 P.M. got into an omnibus with a number more of our Cambridge friends, and went into Boston to Agas-
siz's first lecture. The title of them is "The Plan of Creation as developed in the Animal Kingdom". In this he gave us the general division; Animal Life, and the characteristics of each, illustrated by diagrams. My pleasure was greatly impaired by his strong foreign accent, and his occasional hesitation in finding words to express himself. It deprived him of eloquence, and I longed to hear him change into French. Instead of returning to Cambridge, Uncle proposed to take us to the theatre. I was delighted, and so were Harriet and Eliza Nye. The others went home. Uncle took me to a tea and coffee saloon, where I had an excellent tea. He then left me at a confectioner's whilst he went to a restaurant to dine. I read the papers till Harriet and Eliza came in, and Uncle joined us to take us to the theatre called the Howard Athenæum. The house is very pretty, and just opened, having been built entirely in four months. We had Henry 4th of Shakespeare, and a most capital Falstaff of the name of Hackett. He kept us laughing with his excellent personation of the character. The Prince of Wales was good, and Hotspur excellent; and my enjoyment of the whole play very great.

Old matcher output

Keywords found: [pain, look, light, dine, bostonian]
Match strength = 0.003726708

Amelia Burnett, Journal of a Visit to the United States, 1846-1848 Source: Massachusetts Historical Society

Saturday. Packed my things, bade farewell to my very dear friends, the Websters whose kindness has been unvarying, and went in to Boston with Clara. After calling on the Cunningham's, and doing some shopping, I dined at Aunt Prescott's, and at half-past three, Dr. Webster came to accompany me to the cars. I made the journey without any adventure, amusing myself looking at the country, and reading the history of a "Feather"; arrived at New Bedford about seven, put myself and baggage into a carriage, and took Aunt Amelia by surprise, she not having received my letter announcing my coming. I found her looking very well, the house prettier than ever, and little Sissy terribly spoilt.

The results returned by both matchers mention dining. The new matcher output is slightly more relevant than the old matcher output, as it mentions dining in a restaurant. The old matcher picked up on the keywords "look" and "dine", neither of which is very important in the returned match, which is mostly about visiting friends.

Neither match is terribly relevant. While the new matcher output does mention
dining, this is not the main focus of the story – the main focus of the story seems to be a description of Agassiz’s lecture, and the author’s opinions about his eloquence. This admittedly is due to the way stories were categorized; the rather simplistic heuristic was used of assigning a story to a topic if it contained any information about that topic, even if it was not the most primary information in the story.

6.2.3 Excerpt 3

I had grand plans for the next morning. I would wake up very early, before 5, and snap photos of the slowly awakening city before getting back to business at 8. Early morning is a special time for me – probably because, more often than not, I sleep right through it. I set the alarm for 5, and, true to form, slept right through it. Well, actually, I turned the alarm off around 4:30 after sleeping fitfully in the muggy room. Apparently I should have left the air conditioner on through the night but, being a longtime resident of the northwest coast, I had no experience with such strange devices. I finally managed to drag myself out of bed at 7 to take a short stroll. A block east of the hotel is the Trinity Church, a gorgeous old building. It’s a famous Boston landmark and is, officially, one of the 10 greatest public buildings in the US. It somehow manages to combine delicacy with oppression: light spires sit on heavy foundations; sombre stained glass lurks amid earthy stones.

New matcher output

Topic keywords found:
[morn, earli, citi, earli, morn, room, left, air, night, coast, bed, hotel, church, build, public, build]
Proper nouns found:
[citi, well, room, east, hotel, church, old, boston, public]
Best possible topics are:
countryside
town
By topic match, story 165 was detected 12
By PN match, story 122 was detected 11
Best match is by PN

Amelia Burnett, Journal of a Visit to the United States, 1846-1848 Source: Massachusetts Historical Society

61
When I sent off my journal, I mentioned that I was then waiting for the Cunninghams’ escort to come on to Boston, and was getting somewhat impatient of the delay. At last, the 4th of May your birthday dear Harriet, was fixed as the day, and by five that morning I was up and dressed waiting for the carriage. I bade adieu to the Stalkers, and got into the vehicle about a quarter after five, and gave a last look at New York streets, quite deserted in the early morning, as I passed through them on my way to the wharf. I felt a little timid on arriving there, at finding myself alone; but everything connected with travelling is conducted with so much method here, that I found no difficulty in getting myself and luggage on board the steamboat Traveller, bound to New Haven, and was soon joined by Mr. and Mrs. C. The view of New York from the Harbour is very fine, and from the city to the opening into the Sound, there is much to attract the eye, in the pretty country seats and gardens on both sides of the East River. We passed Hell Gate, which you may remember is spoken of by Washington Irving, but which disappoints one, as the whirlpool occasioned by the meeting of opposite currents is but slight to what one would imagine from the fearful name. It was very cold on deck, and I was glad when breakfast was announced; a long table was spread with a variety of good things, which seemed to be well appreciated. Neither Mrs. Cunningham nor I had brought any work or books, I had anticipated much enjoyment from the views, but was disappointed, as the shores, both on the Connecticut and Long Island side, were low and flat, and the wind blew sharp and cold. However, I got the stewardess to lend me some old magazines, and with their help and a great deal of walking the deck, the time passed till we reached New Haven, very prettily situated in a bay formed by two projecting Headlands. There is Yale College, of which Dr. Dwight, author of the System of Theology, was once President. We had little time to look at the view, as the cars started almost immediately, and left New Haven far behind in a few minutes. I was much pleased at coming this route, as it gave me an opportunity of seeing the Connecticut valley, famed for its fertility and beauty. But the season is still too early for seeing the country to advantage; nothing looked green but the grass, and the woods of pine and fir; the fields were not ploughed, nor the cattle turned out. Still I enjoyed the scene, and could form some idea of the fertility of the valley in summer. Hartford, one of the capitals of Connecticut, looked prettily built on the shores of the river; the railroad is carried close to the river on the opposite side from Hartford for many miles, then is carried across the river and continues to follow its course to Springfield, a very pretty country town, much resorted to by new married couples, as Tunbridge Wells is in England. There we had twenty minutes allowed us to refreshments, and found a room well supplied with eatables of all sorts upstairs. We took oysters and milk, and were soon seated in the cars again. All the way to Boston, through Worcester, Framingham, and Watertown, the views continued to be very pretty, and many large pieces of water, surrounded
by trees, gave life and variety, to the scene. We reached Boston at 7 in
the evening, much tired, from being seated so long in the cars. I went to
Mrs. Chas. Cunningham's to sleep, and finding the family all gone to
the Opera, went to bed immediately, and slept soundly after the fatigues
of the day. The next morning I met them at the breakfast table and had
to answer many questions about New York, and listen to their enthusi-
amism about the Italian Opera, which I find is now almost the only topic
discussed by the Bostonians. After breakfast, I went to see Aunt Prescott,
who seemed pleased to see me; Cousin William and his wife also came in
to welcome me. From thence I walked out to Cambridge and found them
all well. The Websters manage their domestic affairs admirably, though
with only one servant everything goes on the same as before, and this they
accomplish by getting up at five, and attending to household affairs the
first two or three hours of the day. How great was my delight at finding a
large parcel directed to me from Fayal; and that I had no less than sixteen
letters to read from the island and my dear friends in Portugal. How can
I think you all as I ought for writing me so fully! I did nothing but read
all day and felt myself transported home again. Since then I have been
very busy working, reading, talking to the dear Websters, and visiting my
Cambridge friends. The country is looking very pretty, though the weather
is still cold. The leaves are opening, the fruit trees covered with blossoms,
and the meadows of a beautiful green; and the view from my window as
I write, is beautiful, how I wish you could see it! I must leave off for the
present, feeling myself quite unable to write now, but you shall hear every
detail henceforward.

6.2.4 Old matcher output

Keywords found: [citi, busi, morn, form, well, room, air, resid, east,
hotel, triniti church, church, old, boston, public, build, light]
Match strength = 0.010729615

Eva Blackwelder, Tour of Massachusetts, 1938-1939 Source: Massachusetts
Historical Society

11/24/38 Thanksgiving day, our first without a turkey, it seemed strange.
Took a drive about Boston around the Common pass the pold Park Street
Church along Tremont Street where the shops are, the movie houses into
old Boston with its narrow streets, old houses and quaintness although
sombre, reminded me of the gay French quarter of New Orleans. Every
once in awhile one seemed to be steering right into a house at the end
of a street but the street divided there and ran around the house on both
sides. we saw the old markets which were closed for the Day. The modern
buildings here and there topped the old building but did not seem out of
place for they looked like great palaces surrounded by an old town. Many of the streets are one way and I saw one street only wide enough for one car, like the streets of Pompeii in Italy.

The input story itself cannot be categorized under any of the topics that are in the topic dictionary of the new matcher. It seems to be mostly describing the concept of morning and how it fits into the author’s daily routine. Morning would have been a possible topic for inclusion into the topic dictionary, but there were no stories in the corpus that were about this topic, so it was not included.

Again, the new matcher returned a very long excerpt. Examination of the matcher output shows that the new matcher considered the best topic matches for the user story to be countryside and town – quite a contradiction. The most likely reason for this is, again, that the excerpt is so long it contains words that are indicators of both topics, and some of those are out of context.

The old matcher output is actually more relevant than the new matcher output – the returned story is at least set in Boston. The keywords found are actually very similar to the ones found by the topic-based matcher – in this case, it seems that the topic-based matcher got confused because of the presence of some words that, in other contexts, could indicate the topic countryside.

6.2.5 Excerpt 4

I eventually found myself walking through the Boston Public Garden. It is billed as the U.S.’s first public garden, but really it’s just a bunch of lawns and flowers packed with tourists. If you’re into American history there are some statues of dead people. The swan boats (huge pedal boats that slowly navigate the small central lagoon) looked vaguely interesting – if only for a chance to chat with one of the cute boat captains. In the end, though, they were just too kitschy for me. Ultimately, I left the Public Garden feeling rather unimpressed. Vancouver’s Stanley Park has infinitely more character (and a slightly smaller tourist density).
New matcher output

Topic keywords found:
[walk, public, garden, public, garden, flower, pack, histori, statu, peopl, boat, boat, look, interest, boat, captain, left, public, garden, feel]
Proper nouns found:
[boston, public, public garden, garden, public, public garden, garden, american, captain, public, public garden, garden, park]
Best possible topics are:
friends
By topic match, story 134 was detected 15
By PN match, story 6 was detected 14
Best match is by PN

Eva Blackwelder, Tour of Massachusetts, 1938-1939 Source: Massachusetts Historical Society

Saw the wonderful new Post office between Milk St. and Water St. there is a column in the middle of the Square in front of the Post Office topped by an eagle, it was erected to the memory of Angel who formed the society for the prevention of cruelty to animals. I was told that in Summer a man was always on hand at the monument with a bucket of water to water any passing horse....

12/13/38 At every turn in Boston there is something interesting to note. Under the shadow of church doors, there are men selling flowers, daintily budded roses and gardenias and at the corner of the Public Library on Copley Square an Italian has his hot roasted chestnuts steaming over a charcoal fire. Thousands of pigeons feed in front of the new Statehouse facing Boston Common and squirrels make their home in the big trees in the Public Garden ..... [sightseeing in Medford, Lexington, Concord]

11/15/38 In Lexington on the Common is an old monument erected by the state in 1799, the oldest memorial of the Revolution. It bears the names of the Lexington heroes and below "To the memory of their Fellow Citizens who fell on this field, the first Victims of the Sword of British Tyranny and Oppression 19 April 1775 The Die was Cast!!! The Blood of these Martyrs In the cause of God and their country, Was the Cement of the Union of these States, then colonies and gave the spring to the spirit, firmness and resolution of their Fellow Citizens. They rose as one man to revenge their brethren's Blood and at the point of the sword to assert & Defend their native Rights they nobely dar'd to be free!! the contest was long, bloody & affecting. Righteous Heaven approved the solemn appeal, Victory crowned their arms and the Peace Liberty and Independence of the U.S.A. was their Glorious Reward. 1799 There is also a Minute Man (Kitson's) in Lexington. We passed Louisa Alcotts old house and the house
Old matcher output

Keywords found: [found, walk, boston, public, public garden, garden, public, public garden, garden, flower, american, boat, boat, small, look, boat, public, public garden, garden, park, charact]  
Match strength = 0.012285012

*Amelia Burnett, Journal of a Visit to the United States, 1846-1848 Source: Massachusetts Historical Society*

Saturday morning, I returned in the boat, with a little girl, given to my charge, and at the wharf found a carriage waiting for us. After a short chat with Aunt Prescott, I went to the Cunningham’s, and found Clara and Maryanne just returned from Milton, where they had spent the previous day.

There is no doubt that the new matcher did better here. The only keyword in the old matcher output is “found”. The new matcher output describes the Boston Common, the Public Library and the Public Garden, which are either mentioned in the user input story or close by. The old matcher does not pick up that some words are more important than others; clearly “found” should not be as important in the story as “Boston Public Garden”. In fact, when examining the user input story, “found” is not important at all.

This may provide evidence to show that if the proportion of proper nouns to keywords is above a certain value, the proper nouns should take precedence. However, estimating this value is difficult and cannot be determined theoretically.

### 6.2.6 Excerpt 5

One of the great things about Boston, being an old city, is the great architecture. The Trinity Church is just one example of many beautiful old churches that are scattered throughout the Back Bay. By the way, the Back Bay is the neighborhood containing Newbury Street, the Trinity Church, and my hotel (among other things). It’s a fairly wealthy old district and is built on a huge landfill that used to be a great big bay.
They built a dam on the bay, to control the water levels in the Charles River, but when that didn't work they figured they'd just keep adding dirt. Forty years later, the Back Bay was born.

New matcher output

Topic keywords found:
[citi, church, beauti, church, bai, bai, street, church, hotel, built, bai, built, bai, water, river, work, year, bai]
Proper nouns found:
[boston, old, citi, church, old, church, bai, bai, newburi, street, church, hotel, old, bai, bai, charl river, year, bai]
Best possible topics are:
countryside
town
By topic match, story 165 was detected 61
By PN match, story 2 was detected 27
Best match is by topic

After breakfast, an open carriage was ordered, and we drove out to see the lions. First we passed the grand Battery, and on to the esplanade where the view is beautiful. It reminded me very much of the views from some of the heights in Lisbon. From where we stood, at the edge of the cliff, we looked down on the roofs of the houses in the narrow streets beneath; beyond was the river, narrow immediately in front of Quebec, a town looking very like Cacilhas at Point Levi, from whence it widened into a bay, much resembling that of Palmella. The ships too, were anchored in the stream, another point or resemblance. The day was bright and clear, but very cold; but we were all in high spirits, and I was all eyes and ears. We met the Highland regiment on its way to parade. They were not in their national costume, which is rather cold for a Canadian winter; but wore the scarlet coat with tartan trousers and a scarf tied across the shoulders. They looked well with their bonnets and waving black plumes, as they marched to the sound of some of their national airs. We drove along a fine road, with many handsome houses and gardens on each side, to the plains of Abraham, where we stopped to view the spot where Wolfe fell, and then took up a position where we could have a good view of the troops. It was an exciting scene to me; watching the manoeuvres of these well trained soldiers, on the very plains of which I read in history, and never thought I should see, whilst beyond was the hilly country round Quebec and below one of the largest rivers in the world. On our return from the plains, we visited the citadel, said to be the strongest on this continent, and from what I could judge, it merits its reputation; then visited two Catholic churches, (no novelties to me), and returned to lunch. In the afternoon we drove out
to see the Falls of Montmorenci, seven miles below Quebec. It is built on a headland, called Point Diamond, formed by the junction of the river St. Charles to the St. Lawrence. The sides of the cliff are all built upon, and on a bright day, the roofs (covered with tin) sparkle in the sunlight. We crossed the St. Charles on a long bridge, passed a very long French village, called Beaufort, and then came to the Montmorenci, and crossed it above the Fall. We were somewhat disappointed in it; for though the height is considerable, 180 feet, the amount of water is small, and we were too late in the season to see the scenery to advantage. The Fall is close to the St. Lawrence, which is here divided by a long island, called the Ile d'Orleans. The next day it rained, so we walked out but little, and at five in the afternoon re-embarked for Montreal. We arrived there early, took breakfast at a French Hotel, and amused ourselves walking about town till twelve, when we went on board the Ferry boat for La Prairie. Here we took cars for St. John's, very antiquated affairs, which took us through a dreary looking country at a very slow rate. At St. John's we embarked in the steamboat Burlington on Lake Champlain; but unfortunately passed the finest scenery at night. However, we enjoyed one beautiful view next morning before reaching Whitehall; the lake narrows very much and the shores projecting, make it form an angle; just here is the most perfect subject for a picture I have ever seen. In the foreground, a finely wooded point, with a real Indian Wigwam amongst the trees; behind, on the opposite shore, a large saw mill establishment, backed by high hills, on the other side a range of mountains richly wooded, and the morning mist still in the hollows. I can give you a very imperfect idea of the scene, but we all exclaimed at its beauty. At Whitehall we were hurried on board the canal boat, which canal unites Lake Champlain with the Hudson at Troy. This day was a tedious one to us all, but we were glad of the opportunity of becoming acquainted with canal boat travelling. A cabin extends almost the whole length of the boat, the upper end being divided by a curtain for the use of the ladies, and a kind of a screen near the door hiding the water's department from public gaze. Every one on board eats and sleeps in this cabin, and the scenes at night in a very crowded boat, are sometimes very ludicrous. My wonder was where the cooking could be carried on, for no appearance of it was to be seen, and the fare was really very good. We amused ourselves reading and watching our fellow passengers in the cabin, and sitting on deck where we had to lower our heads every five minutes for the bridges. We were much interested in watching the boat ascend and descend the locks. I do not wonder Brindley was so proud of his canals, for they are a beautiful invention. It was ten at night before we reached Mechanicsville, where we took cars for Troy. Being drawn by horses, it was nearly midnight before we reached the Hotel. After a good warming by the fire, we were shown to our rooms, and slept soundly till five the next morning, when the waiter's knock at the door unwillingly roused me from my slumbers. We went down to the boat, and I had the satisfaction of seeing the Hudson
by daylight. It is a glorious river, and very varied in its beauty. The
tints of the woods here, were still bright, though in Canada, they had
quite disappeared. We passed the Catskill mountains, twelve miles back
from the river, many towns and villages, surrounded with well cultivated
country, and after passing Newburgh where the river widens considerably,
entered the narrow defile, leading to the Highlands. High cliffs overhang
the water's edge, the boat passes the Crow's nest, a high and precipitous
hill so called from a hollow at its summit, and then turning an angle
comes upon West Point. The view a short distance below this place is the
most perfect on the river. The Hudson here forms a lake, surrounded by
bold mountains, and West Point jutting out into the stream, completely
conceals the outlet.

Old matcher output

Keywords found: [boston, old, triniti church, church, beauti, old,
church, bai, neighborhood, newburi, street, triniti church, church,
hotel, old, built, built, water, charl river, river, year, bai]
Match strength = 0.020179372

Rev. Francis Higginson, Massachusetts Bay Colony, 1629 Source: Everett
Emerson, Letters from New England, 1976

And as for fresh water the country is full of dainty springs and some great
rivers and some lesser brooks, and at Masathulets Bay [at Charlestown]
they digged wells and found water at three foot deep in most places, and
near Salem they have as fine, clear water as we can desire, and we may
dig wells and find water where we list.

In this case, the old matcher seems to return a better match. Once again the
length of the new matcher output is a problem. The old matcher returns a story that
is primarily about water, which is mostly detected by the occurrences of the words
"bay" and "water" and "river".

The topic-based matcher should have classified this as being under the topic water.
A possible reason for it failing to pick this up is that some of the keywords present
in the user story, like "bay" are indicative of countryside as well as water. The user
input story does not contain any keywords that can distinguish topics water and
countryside, and in fact water is not selected as one of the most common topics in the
user story. However, if we were to have as indicators only words that distinguished
topics, it would be difficult to detect anything at all.

6.2.7 Excerpt 6

I was up late the next morning and checked out of the hotel. Around
11:30 I called Jen but she wasn’t home. I tried again a little after noon,
but still no answer. I gave up on her and went in search of lunch. My
intended destination was Chinatown. I was hoping to find some dim sum
and peruse the tea shops. On my way, though, I passed Legal Seafood – a
Boston landmark that my coworker had mentioned on our first night in the
city. A bit on the expensive side, but I figure I deserved to treat myself to
the bounty this fine city had to offer. The menu was amazing. Everything
looked good. With the help of my waitress I settled on Boston bluefish
pate for an appetizer. While I waited for the next course, the waitress
came over and talked for a while. I told her I had a free afternoon and
she recommended I take the T (what the locals call the subway system)
and visit Harvard. I told her about my experience with beautiful women
on the subway and she said, "That’s nothing! You should try the orange
line.” I made a mental note to fit the orange line into my travel itinerary
and then dove into the main course: steamed lobster (complete with bib),
crab legs, mussels and clams, with corn on the cob and cole slaw. The
lobster disappeared in short order, but I had trouble finding room for the
rest after my big appetizer. I seem to recall that there was also some clam
chowder in there somewhere, too.

New matcher output

Topic keywords found:
[morn, hotel, home, gave, lunch, tea, shop, pass, night, citi, citi,
look, good, wait, talk, visit, harvard, beauti, travel, steam,
lobster, corn, lobster, room]
Proper nouns found:
[hotel, peru, boston, citi, citi, boston, harvard, room]
Best possible topics are:
countryside
By topic match, story 158 was detected 20
By PN match, story 246 was detected 12
Best match is by PN

Sept. 5. A large carriage and four horses took our travelling party out in the country. We set out immediately, and passed over the new Charles River Bridge through the Town of Medford, over the weirs and down through Menotomy to the house of Mr. Cragie in Cambridge, where we were cordially received and hospitably entertained by the modest and opulent proprietor. We walked up to his summer house, a pretty piece of ornamental architecture situated on an eminence, once a reservoir of ice, built by Nathaniel Tracy, Esq., when he was proprietor of and resided at this superb and delightful seat. Our Maryland gentlemen were perfectly enchanted with it. I think one may safely assert that, after Beacon Hill in Boston, this spot presents the most beautiful, extensive, and variegated landscape in the world. From Mr. Cragie's we went to the colleges, to take a view of the library, which is said to be the best in America, both with respect to the number of books and the selection. The books are arranged with great propriety and elegance. There are several paintings and engravings at the end of the room that attract one's attention, particularly a portrait of the Cardinal Bentvolio, executed by Smybert, from the original by Raphael. It is certainly an excellent painting, and does much honor to the copyist. We took a cursory view of the cabinet of curiosities; the collection is small and boasts but few rarities. We next visited the Corporation Room, which is adorned with several handsome paintings and engravings. In this room stands the elegant Orrery made by Mr. Joseph Pope, a watch-maker of Boston, which for mechanism and excellence of workmanship is said to equal any thing of the kind in the world. After looking into the Commons or Eating Room, and into the chapel, which occupy the whole first floor of the building, we embarked in our carriages, and returned to Boston, via Roxbury. The gentlemen from Maryland expressed themselves much pleased with the excursion.

Old matcher output

Keywords found: [morn, hotel, call, gave, tea, pass, boston, side, fine, citi, look, boston, afternoon, call, visit, harvard, beauti, corn, room]
Match strength = 0.015267177

Amelia Burnett, Journal of a Visit to the United States, 1846-1848 Source: Massachusetts Historical Society

Sunday. Went twice to church; in the morning with the Prescott's to hear Dr. Frothingham a Unitarian preacher, who has great mannerism
and wants the power of rivetting the attention. In the afternoon to Dr. Gannet's, a very fine preacher. Unitarianism is quite the fashionable creed in Boston; the only city in the world where it is so, in the south it is looked upon with as much horror as it is in England.

Neither matcher did well in this case. The topic-based matcher failed to classify this as being mostly about the topic food. Again this is because there were several indicators of this topic present in the user input story that were not listed as keywords – examples of these are “waitress”, “appetizer”, “pate”, “mussels”, “cole slaw”. Food is an example of a topic for which it will never be possible to list all the varieties and descriptions of food.

The old matcher, on the other hand, returned a short story which contained only “morning” and “Boston” as keywords. Neither of which was very relevant to the story returned.

6.3 Discussion and Suggestions for Further Work

The topic-based matcher sometimes performs better than the old keyword matcher, but not consistently. One problem that kept showing up is due to story length. The stories in the corpus need to be re-segmented; many of them are so long that they contain many topic keywords, some of which are used out of context. Thus, more than one topic is detected, and the chance that this is actually what the story is about is low.

Certain keywords are also indicative of more than one topic, and if there are no keywords to distinguish the topics, the topic-based matcher may detect the wrong topic. A possible way of fixing this is to change the new matcher algorithm to disregard words that are indicative of more than one topic in favor of words that distinguish the topics. Alternatively, keywords that distinguish topics could be given higher weights than keywords that are common to more than one topic.

Another issue is the fact that there may be insufficient keywords to indicate topics. As mentioned before, all the topic keywords are taken from the corpus, which is not very large. More stories would present a clearer picture of the topic division and
their indicators. A larger corpus might also show up several new categories. The problem of overly general keywords may be solved using collocations. Certain words tend to occur close to each other, and the combination of both words together may be indicative of a certain concept even if each separate word is not. A collocation is a combination of words that tend to occur close together. In certain cases, collocations can be used to disambiguate word senses. For example, "baggage" and "luggage" have similar meanings as nouns, taken on their own. However, "baggage" can be preceded by "emotional"; the resulting phrase means something entirely different from the usual sense of "baggage".\footnote{This example was taken from [8].} Suppose we had found that "baggage" was indicative of the topic transportation; if we encountered the collocation "emotional baggage", we would know not to detect the topic transportation based on this.

The statistical analysis using the WordSmith tools is not infallible. As mentioned previously, statistical analysis with the WordSmith tools was used as a starting point for selecting keywords, but the fact remains that no keywords were picked which were not in the statistical analysis. Also, not all the words in the corpus were read through – any word which only appeared once (there were several hundreds of these) was not considered a possible topic keyword. It is possible that given the small size of this corpus, that assumption was wrong.

It seems that most of these issues could be elucidated by enlarging the corpus. More stories means more information with which to build the knowledge base. No corpus will ever be exhaustive, but once the corpus is large enough, statistical heuristics may be used that may not be accurate if used in this case. For example, consider the following situation. If the WordSmith consistency analysis returns results that show that word \( W \) is present many times in topic \( T \), we may guess that word \( W \) is a good indicator for topic \( T \). That is, we would guess that if word \( W \) appears in a story, topic \( T \) is likely to be a topic of that story. However, this guess may be wrong. This depends on the accuracy of the "prediction" made by the WordSmith consistency analysis; was word \( W \) present many times in topic \( T \) because it really is used to describe topic \( T \) in regular usage, or was it present many times just because
of chance. As the corpus size grows, we may expect it to present a more accurate picture of “normal” English usage.

As mentioned previously, the matcher either matches by topic or proper noun, depending on the relative numbers of keywords and proper nouns present. Currently the matcher is set to match by topic if the strength of the topic match is twice that of the proper noun match (where strength is measured by how many occurrences of that indicator are in the user story). This ratio was arbitrary; the heuristic was used that it should take more keywords to specify a topic than it takes proper nouns to specify a description about a particular place. The correct value for the ratio should properly have been determined experimentally.

The issue of picking up high-level information from stories, such as relationships between people, things and places, still has not been solved. It is unlikely that this is going to be solved by a category-based keyword matcher, unless the categories chosen are very fine-grained. Relationships between characters and things in a story are best captured not by individual words but by the relationships between words. This suggests that a phrase- or sentence-level analysis may be necessary to pick up this sort of information.

Ultimately, it seems that the design of a computational system to work with material that is subjective and not easily quantifiable is very sensitive to the information that is provided to it. In the topic-based matcher we see problems arising from the fact that the categories are not hard-edged. When we get topic indicators that are indicative of more than one topic (categories therefore overlap), the matcher quite often makes the wrong decision. This is something that is not dependent on the efficiency of the computational algorithm or the speed of the processor – it is dependent upon the human view of the information contained in and embodied by the topics and their indicators.

At this point, the most useful thing to add to the system would be more extensive real-world knowledge, most especially the relationships between concepts and how this is embodied in language. Word sense disambiguation with WordNet would be a good candidate, provided the WordNet API is fixed, or a new version released.
There is an alternative to WordNet; ThoughtTreasure, a database that attempts to address the problem of representing common sense in a computationally readable form. Its database contains concepts, English and French words and phrases, assertions and scripts that try to “make the human world more understandable to computers” [14]. Some items in ThoughtTreasure are:

A hotel room has a bed, night table, minibar, ...
One hangs up at the end of a phone call.
Rough synonyms for food are foodstuffs, groceries, chow, grub, ...
The sky is blue.

A ThoughtTreasure Java API was very recently released (December 1999), and this is a very promising future direction. In concept at least, ThoughtTreasure contains much information that seems useful to a story understanding system.

6.4 Evaluation From the Point of View of a Human Audience

The above comparison of the old and new matchers is highly subjective. It is done from one person’s point of view, which may be warped because of too much knowledge of the problem. This system is intended to function within the larger context of an interactive travel journal; ultimately the measure of its success will be whether it inspires users to write more, and to write more creatively.

We do not have a clear idea of what human users find “relevant”. What one person finds to be relevant another might not, depending on differing personal circumstances. However, if a large enough community of users is surveyed, trends emerge which, if followed, will increase the chance of the matcher giving “relevant” matches to some majority of users.

There are two aspects of this matcher that should be evaluated by human users. The first is how good the matches are. This may be done by surveying human users, by having them rank matches made by three sources: the new matcher, the old matcher
and a human-made match (the “best” story selected by a human reader according to that reader’s definition of “best”). A way to collect a lot of information anonymously would be to have this survey online, where anonymous users could rank matches and have the results saved and statistically analyzed. Topics and keywords might also be obtained this way – although it is impossible to “prove” that this is more accurate than any other way of obtaining topics, it is at least connected with the human experience of stories, which, in this project, is the most important characteristic to keep in mind.

The second aspect that needs to be evaluated is how well the device stimulates creative writing with the new matching system. This is even more difficult to evaluate than the quality of matching. To do this, the length of user’s stories with and without the matcher’s prompting should be measured, as well as analyses of how much of the information (in terms of story) provided by the device was used in the user’s stories. The ideal scenario is for the user to have a meaningful interaction with the device, and leave feeling that they have seen their travel experience in a new light.
Bibliography


