Automated Creation of Wikipedia Articles

by

Christina Sauper

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2009

Author . . . .
Department of Electrical Engineering and Computer Science
December 18, 2008

Certified by . . . .
Regina Barzilay
Associate Professor
Thesis Supervisor

Accepted by . . . .
Terry P. Orlando
Chairman, Department Committee on Graduate Students

Automated Creation of Wikipedia Articles

by

Christina Sauper

Submitted to the Department of Electrical Engineering and Computer Science
on December 18, 2008, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

This thesis describes an automatic approach for producing Wikipedia articles. The wealth of information present on the Internet is currently untapped for many topics of secondary concern. Creating articles requires a great deal of time spent collecting information and editing. This thesis presents a solution. The proposed algorithm creates a new article by querying the Internet, selecting relevant excerpts from the search results, and synthesizing the best excerpts into a coherent document. This work builds on previous work in document summarization, web question answering, and Integer Linear Programming.

At the core of our approach is a method for using existing human-authored Wikipedia articles to learn a content selection mechanism. Articles in the same category often present similar types of information; we can leverage this to create content templates for new articles. Once a template has been created, we use classification and clustering techniques to select a single best excerpt for each section. Finally, we use Integer Linear Programming techniques to eliminate any redundancy over the complete article.

We evaluate our system for both individual sections and complete articles, using both human and automatic evaluation methods. The results indicate that articles created by our system are close to human-authored Wikipedia entries in quality of content selection. We show that both human and automatic evaluation metrics are in agreement; therefore, automatic methods are a reasonable evaluation tool for this task. We also empirically demonstrate that explicit modeling of content structure is essential for improving the quality of an automatically-produced article.

Thesis Supervisor: Regina Barzilay
Title: Associate Professor

Acknowledgments

I am grateful to my advisor, Regina Barzilay, for her constant support and advice. The energy and enthusiasm she contributed to this project were key to the success of this journey.

I would like to thank Alok Baikadi and Frank Stratton for reading countless drafts, providing technical suggestions, and attempting to keep my stress level down. Likewise, I would like to acknowledge all of my friends and colleagues at CSAIL for their support with both this thesis and life in general, including Branavan, Erdong Chen, Harr Chen, Zoran Dzunic, Jacob Eisenstein, Yoong Keok Lee, Tahira Naseem, Benjamin Snyder, and David Sontag.

Finally, I would like to dedicate this thesis to my family: to my father, Metro, who gave me a love of science; to my mother, Deborah, who gave me a love of language; and to my sister, Elizabeth, who has always believed in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As more people have gained access to the Internet, Wikipedia has become a significant resource for information on a wide variety of topics. This electronic encyclopedia contains over 2.3 million articles and is viewed by an estimated 10.65 million people per day. The collaborative approach for Wikipedia development allows for continuous improvement and expansion of its content.

This model of collective editing is particularly beneficial for mainstream articles. The article for *Barack Obama*, a figure who is active in the news, has been edited approximately 750 times in the past month alone. Other famous figures have also received significant attention; the article for *Britney Spears* has been edited over 16,800 times in total.

On the other hand, articles about new or specialized topics are often either nonexistent or present as simple “stub” articles, containing a brief description of the topic and often only a few sentences long. The problem is more severe in other languages, where the total number of articles is only a fraction of the English data. For example, Chinese Wikipedia contains 171,000 articles, while Arabic Wikipedia contains only 57,000.

1.1 Automatic Generation

To remedy this problem, we propose an automated method for producing new Wikipedia articles, selecting information from across the Internet to synthesize coherent, logical articles. This type of generation will allow articles to be quickly created as soon as a topic becomes of interest. Instead of a stub article consisting of a few sentences, it will be possible to create a complete, well-rounded article containing information on multiple facets of the topic.

At the core of our approach is a method for using existing human-authored Wikipedia articles to learn a content selection mechanism. For each Wikipedia category, our model identifies topics that should be included in an article, specifies their order, and provides a search query for their extraction. For instance, it learns that articles in the *Diseases* category often contain sections *diagnosis*, *causes*, *symptoms*, and *treatment*, commonly presented in that order. We induce these patterns automatically by analyzing the structure of existing Wikipedia articles. This information is stored in a template and used each time a new article is created for the given category.

A key challenge in selecting material for each section is filtering search results to match the style and content of existing Wikipedia articles. For each section, the model identifies query terms that yield related web pages and then attempts to select salient, topic-specific paragraphs from the retrieved pages. This extraction is formulated as a supervised learning task: given the search engine output for a topic of interest – e.g., *Bacillary Angiomatosis & Symptoms* – the model is trained to identify text fragments that are most similar to text in corresponding *Symptoms* sections of existing Wikipedia articles in the *Diseases* category. Finally, we put the sections together using an integer linear programming algorithm for redundancy elimination. Some sample articles created by this system are presented in Appendix A.

1.2 Evaluation

We evaluate our approach by generating articles in two Wikipedia categories: *American Film Actors* and *Diseases*. The generated articles contain between 12 and 100 sentences; this is consistent with short Wikipedia articles. The results of both automatic and human evaluation are encouraging: articles generated by our system are close to human-authored Wikipedia articles in quality of content selection. We also empirically demonstrate the benefits of structure-aware content selection over structure-agnostic approaches.

Currently, due to the extractive nature of our system, Wikipedia's copyright policy prevents us from contributing our automatically generated articles. However, our articles can be useful starting points for human editors; with additional sentence-level paraphrasing, these articles could be directly submitted to Wikipedia.

1.3 Key Contributions

This system is inspired by ideas from several areas in natural language processing. Text generation tasks include concept-to-text content planning and text-to-text extractive summarization. As in many content planning approaches, the content selection component of our system is driven by a template. For example, when generating an article about *Diseases*, we want to include information about causes, diagnosis, symptoms, and treatments. Like extractive summarization, we attempt to fill each of these areas with some fragments of human-authored text taken from a large array of documents, the Internet. Our techniques for search and query generation draw inspiration from web-based question answering systems, which attempt to answer specific questions using Internet information. Our scope is much broader than this, however, as we are trying to gather a variety of information as put forth in our template.

This thesis presents three major contributions to these areas. First, the dynamic creation of templates for each category is far more efficient than in previous work where templates were created manually. Second, content for each article is selected from the Internet, rather than a closed set of data. Finally, we present an application of integer linear programming (ILP) for redundancy elimination.

1.3.1 Dynamic template creation

In most content-planning applications, templates are crafted manually. This process is detrimental in two ways: creating templates for each domain (in this case, each category) is extremely tedious, and human analysis and expectations may not actually provide the most useful content summary. Instead, we automatically analyze the existing human-authored Wikipedia articles to determine their structure. By clustering sections based on title keywords, our system identifies common content areas and creates a final template consisting of section titles and the queries to retrieve information for those sections. Templates created in this manner are general enough that they are applicable for an entire category.

1.3.2 Internet-based content selection

The Internet offers an unlimited wealth of documents on every subject imaginable. In many traditional generation applications, however, this resource remains untapped as the system design requires a closed set of information, either a database in the case of concept-to-text generation, or a number of pre-selected documents in the case of text summarization. Other applications, such as web-based question answering, do use the Internet as a resource; however, in this case, they search for a very narrow piece of information rather than a well-formatted summary of a topic.

Using the Internet to find a general summary requires completely new methods for obtaining and filtering information during the content selection stage. Our approach combines methods from information retrieval and supervised machine learning to achieve robust performance in the presence of noise.

1.3.3 Redundancy Elimination

Because we are combining information from several sources, it is likely that there will be redundancy between excerpts in the final document. This can be frustrating or even confusing to a reader. When attempting to eliminate this redundancy, the main problem arises in determining which of two equivalent sentences to remove. To deal with this, we create an ILP formulation which assigns some value to each sentence, then attempts to maximize the total value of a document, given constraints on redundancy and paragraph size. ILP has been used previously in summarization tasks, but its application to global redundancy removal is new.

1.4 Thesis Overview

The remainder of this thesis is organized as follows: In Chapter 2, we discuss related work in the areas of Text Generation, Web-Based Question Answering, and Global Inference using ILP. In Chapter 3, we provide a detailed explanation of our model framework, including the three main components: template creation, excerpt selection, and article assembly. In Chapter 4, we discuss several experiments at both section and article levels to examine the quality of generated articles. Finally, in Chapter 5, we recapitulate the main ideas and contributions of this thesis and discuss a few directions for future research.

Chapter 2

Related Work

In this thesis, we draw on previous work in three main areas. First, because we are generating individual articles based on a large amount of available data, we must look at text generation, both concept-to-text, for our content planning component, and text-to-text, for the summarization component. Next, our data is drawn from the Internet, which is the subject of a large body of information retrieval and web question answering work. Finally, to optimize our final articles, we use Integer Linear Programming, which recently has been employed in several NLP tasks.

2.1 Text Generation

2.1.1 Concept-to-Text Generation

Content selection is a fundamental task in concept-to-text generation [28]. It determines what type of information should be conveyed in the output of a natural language generation system. While traditional methods for content selection are rule-based [20, 28, 30], recent work has focused on corpus-based approaches for this task [3, 12].

Duboue and McKeown [12] create biographical summaries based on fact sheets and descriptive texts. The fact sheets contain brief snippets of information in certain categories, while the descriptive text contains a corresponding written biography. The goal, then, is to find the intersection of both fact sheet and biography to produce a short, coherent summary.

In their approach, Duboue and McKeown use the fact sheet to create knowledge frames, which are graphs representing properties of the article subject. Values within the path are clustered to find pieces of text which contain a change of word choice correlated with a change in frame data. This indicates a significant piece of data within the text, which should be retained in the summary. Overall, content selection in this approach is highly dependent on reliable data present in the fact sheet and descriptive text; its main limitation is that without either of these components (as might be the case for an up-and-coming actor), the approach will not produce an optimal result.

Barzilay and Lapata [3] present a complex statistical model for content selection. They evaluate their work on a mapping between a database of American football statistics and associated sports news reports. To select which statistics are important to present in the final document, they use a series of constraints to identify related or dependent statistics. They then use a graph-theoretic minimum cut approach to find the globally optimal solution. Unlike our approach, this method generates content based on available information. However, this approach is not feasible in our case, where we do not have all information available prior to content selection.

2.1.2 Text-to-Text Generation

Single Document Summarization

There are numerous approaches to single document summarization. Some use specific methods of representing data [2], while others focus on classifying based on lexical features [14]. There is also an important distinction between sentence fusion [4, 18], which creates new sentences from several existing ones, and sentence extraction [16], which extracts whole sentences from the original document as a unit. Our approach focuses on lexical classification and sentence extraction.

Kupiec et. al. [14] create a generic document summarizer by extracting sentences using a statistical classification algorithm. They are given a variety of scientific and technical papers in order to create abstracts similar to those created by professional abstractors. They assume that each sentence in a professionally-written abstract is inspired by a certain sentence within the original document, then attempt to extract these inspirational sentences to form a complete abstract.

To evaluate which sentences are most relevant, a Bayesian classifier is trained with five features (see Table 2.1). Many of these features, including length, thematic words, and proper names, are very similar to those used in our system for paragraph classification. These are used to select a user-defined percentage of input sentences for an abstract.

Table 2.1: Sentence features used in Kupiec et. al. [14]

Feature	Description
Sentence Length	True if the sentence is longer than some threshold.
Fixed Phrases	True if sentence contains any of a series of 26 ‘indicator phrases’ or follows a section header with specific keywords (e.g., “results” or “summary”).
Paragraph Feature	Paragraph-initial, paragraph-medial, or paragraph-final.
Thematic Word Feature	True if a sentence contains many thematic words (content words appearing most frequently in the document as a whole).
Uppercase Word Feature	True if a sentence contains several instances of proper names.

Teufel and Moens [31] design a system to create abstracts for scientific conference articles in computational linguistics. In many approaches to news article summarization, specific newsworthy events are picked out and used as the basis for a summary. In scientific papers, however, there are no such similar events, as the focus is on new ideas. Instead, Teufel and Moens attempt to restore the structure of discourse by analyzing rhetorical context. For example, based on common dimensions of a scientific paper, such as *problem structure*, *intellectual attribution*, and *scientific argumentation*, sentences are annotated as belonging to one of several categories (see Table 2.2).

These scientific categories are the most important content areas within a scien-

Table 2.2: Annotation scheme for rhetorical status in Teufel and Moens [31]

Label	Description
Aim	Specific research goal of current paper
Textual	Statements about section structure
Own	Neutral description of own work
Background	Generally accepted background
Contrast	Comparison with or weaknesses of other work
Basis	Agreement or continuation of other work
Other	Neutral description of other work

tific document. Summaries (or abstracts) can then be created by selecting a certain number of sentences from each category. The type of summary depends on the proportion of each type; for example, a summary for experts in the field might contain more relating to prior work, while a summary for a non-expert might contain more background information. This approach of choosing different content areas is similar to our own work, although because our work draws information from many sources, it is better to search for material specifically for each section, rather than to perform one general search and hope to retrieve information on each.

Multi-Document Summarization

Summarizing multiple documents is an extension of single-document summarization, and therefore contains many of the same distinctions. Because our system deals with text from many Internet sources, this is a prime example of multi-document summarization. One key problem in this area is determining related information between documents. Repeated text can be indicative of important information that needs to be summarized.

McKeown and Radev [19] use information from several news articles to create a final summary document. They use MUC templates from the terrorist domain as inputs, where each template contains fields extracted from news articles, such as *perpetrator* and *victim*. The system combines information from several templates in text form by using certain operators to process the data, such as contradiction, agreement, refinement, and change of perspective. If a relevant combination operator is found, pieces of information from both templates appear in the final summary, along with cue phrases which connect them. Chronological information is also used when combining information; articles published later are assumed to know more facts than those published earlier. This approach provides an interesting method of combining data and relying on chronological information to assess quality of information. Templates provide an easy method of determining which information should be directly compared and which operators should be used to combine the text. In a case where there are no templates, however, this approach may break down. It may be possible to fill in templates and then use this method for combination; however, success would depend on accuracy in template completion.

Radev et. al. [27] create a centroid-based summarization technique. Input is a group of clusters created by CIDR, which uses a modified TF*IDF formulation to cluster news articles on the same event. The authors define a "centroid" as the portion of words in a cluster with count*IDF $> X$. Those documents containing a high percentage of centroid words are considered more indicative of the topic of the cluster. Sentences are ranked by centroid content, and the top sentences are extracted in chronological order from the publish time of the original documents. Additional filters are used to eliminate sentences which are subsumed in information content by other sentences in the corpus. The use of centroids allows for coverage of sentences with words unique to the topic of the articles to summarize. Because this is a shallow lexical feature, however, there may be some false positives in the writing; this can be compensated for by having more articles. Also, the quality of the result is dependent on having fairly distinct topics between clusters.

Barzilay and McKeown [4] also work on summaries of news articles. As a preliminary step in their work, they identify sentences containing common information across multiple documents. The number of similar sentences found across all documents is generally proportional to the importance of the sentence in the final summary. Once several similar sentences are found, they are aligned using dependency trees, and the closest matching subtrees are combined. This process is information fusion, where a new sentence is generated by grammatically and coherently combining important portions of the input sentence, while keeping the length of the new sentence as short as possible. This approach will work well for news articles and other single content area domains; however, it may not work well to ensure that all aspects of a general biography or disease article are covered.

2.2 Web-based Question Answering

Our techniques for Internet search and answer synthesis are related to research on web-based question answering [1, 5, 6, 21, 22]. Specifically, issues related to query reformulation, answer selection, and answer synthesis are common to both Q&A and our task.

The focus of our work, however, is different: while the goal of traditional question-answering is to provide a short answer to a given question, we aim to generate a comprehensive, well-rounded overview article. This difference motivates substantial changes in technique. One new challenge is to identify both the specific sections to be included and also the best way to combine them into a coherent article.

Radev et. al. [26] design a web-based question answering system which processes data in several steps. First, the user query is converted to a more web-appropriate form (e.g., *What is the largest city in Northern Afghanistan? → (largest OR biggest) city "Northern Afghanistan"*) using a series of transformation operators. These operators include insertion, deletion, replacing, bracketing, ignoring, and swapping; several of these operators use related words from WordNet. Next, questions are analyzed to determine the type of information expected. Then, documents are retrieved from the Internet and searched for the textual units that contain answers. Answers are extracted, and relevant sentences are split into phrases. Finally, phrases are ranked to increase the position of the correct answer. The query and extraction are similar to that in our work; however, the goal of our work requires more elaborate and longer responses to form a coherent article. In addition, our work does not specify questions or pieces of information to retrieve about a topic; we must determine the relevant information.

Brill et. al. [7] and Dumais et. al. [13] also use question reformulation as a preliminary step; however, in this case, the reformulation creates several expected answer phrases (e.g., “Who created the character of Scrooge?” → “created +the character +of Scrooge”, etc.). These queries are performed in a search engine, and resulting front page summaries are mined for 1, 2, and 3-grams. These are filtered based on expected result type, similar answers are merged, and the top ranked n-gram is returned. It is interesting to note that there is no deep semantic understanding required here; because of this, it is possible to retrieve correct answers even in some cases where traditional understanding-based approaches fail to process the data correctly.

2.3 Integer Linear Programming

Integer Linear Programming (ILP) is a method of optimizing an objective function subject to one or more constraint functions, where each function is a linear equation. ILP has been successfully employed in a variety of language generation tasks. In paraphrasing or summarizing tasks, this method may be used to select certain sentences or words to appear in a final document, while maintaining global coherence or length constraints. In classification tasks, a classifier is used to propose labelings for items; however, in certain cases, these labelings may conflict. ILP is used to find a globally optimal labeling for all items [25, 29].

Dras [11] creates a document paraphrasing model (“reluctant paraphrasing”) using ILP. The goal of this model is to take a coherent document and compress it to the least extent possible to ensure that certain length, style, and readability requirements are met, while keeping the meaning of the text intact. This algorithm uses three levels of work: sentence-level rules for rewriting text, global constraints on length and readability, and an objective function to minimize total effect of paraphrases. This paraphrasing model is similar to the one proposed in our system; however, Dras assumes that the initial document is fully coherent, where in our system there may be redundancy.

Clarke and Lapata [9] use a similar technique for compression at a sentence level.

Their work presents three compression models formulated with ILP. The first uses a trigram language model, the second a significance model based on degree of embedding within the sentence structure, and the third a fully supervised discriminative model. Each compression model contains constraints to choose sequences of words from the original sentence.

For example, the basic language model is as follows:

Let $\mathbf{x} = x_1, x_2, \ldots, x_n$ denote a source sentence to compress. For each word in the sentence, define a binary decision variable, with the value 0 when the word is dropped, and the value 1 when the word is included.

$$\delta_i = \begin{cases} 1 \text{ if } x_i \text{ is in the compression} \\ 0 \text{ otherwise} \end{cases} \qquad \forall i \in [1 \ldots n]$$

Next, define three binary decision variables, for each combination of words in the sentence. These reflect groups of words in the final document as a trigram language model.

$$\alpha_i = \begin{cases} 1 & \text{if } x_i \text{ starts the compression} \\ 0 & \text{otherwise} \end{cases} \qquad \forall i \in [1 \ldots n]$$

$$\beta_{ij} i = \begin{cases} 1 & \text{if sequence } x_i, x_j \text{ ends the compression} \\ 0 & \text{otherwise} \end{cases} \qquad \begin{array}{l} \forall i \in [0 \ldots n-1] \\ \forall j \in [i+1 \ldots n] \end{array}$$

$$\gamma_{ijk} = \begin{cases} 1 & \text{if sequence } x_i, x_j, x_k \text{ is in the compression} \\ 0 & \text{otherwise} \end{cases} \qquad \begin{array}{l} \forall i \in [0 \ldots n-2] \\ \forall j \in [i+1 \ldots n-1] \\ \forall k \in [j+1 \ldots n] \end{array}$$

Then, create an objective function to maximize. This is the sum of all trigrams occurring in the source sentence:

$$\begin{aligned}\max z = &\sum_{i=1}^{n} \alpha_i \cdot P(x_i|\text{start}) \\ &+ \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} \gamma_{ijk} \cdot P(x_k|x_i, x_j) \\ &+ \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} \beta_{ij} \cdot P(\text{end}|x_i, x_j)\end{aligned}$$

Constraints must be created to ensure that there is exactly one word to begin the sentence, exactly one word must end a sentence, and any words within a sentence are preceded and followed by one word each. Also, a length constraint is enforced to maintain a decent sentence length.

To avoid ungrammatical sentences, it is also necessary to formulate additional sentence-wide constraints; otherwise, verbs might be included without arguments or modifiers might be included without their heads. The ILP formulation is solved for each sentence to be compressed within a document; this yields an 'optimal' sentence according to the particular model used.

Sentence by sentence compression may work well with a small document; however, in a large document, it may be tedious to compress each individual sentence. Additionally, there are factors across sentences which could be used to create a more coherent document or further compress sentences; for example, knowing which information in a sentence is redundant with other sentences in the document could lead to a more effective compression.

2.4 Summary

While this thesis is based on previous work in text generation, web question answering, and integer linear programming, each of these areas lacks some features that are necessary for our model.

For text generation, both content planning and summarization has not been effectively done using the Internet as the primary data source. The open nature of the Internet results in a number of problems, including the inability to evaluate all possible combinations of information for our document and even to know all of the information that exists for each topic.

In web question answering, the Internet is used as a data source, but only very specific pieces of information are requested. This allows for more specific query transformation, and therefore, more specifically relevant results. To retrieve more general information about a topic, the topic must be broken up into several smaller subtopics, which we attempt to address in our model. In addition, our goal is a large body of information rather than a single short answer.

Finally, in Integer Linear Programming, work has been done on sentence and document compression, but not from the standpoint of summarizing multiple documents. This poses additional challenges of removing redundancy between sections.

We will attempt to address each of these areas in our model, which is presented in the next chapter.

Chapter 3

Algorithm

In this chapter, we will introduce our algorithm for generating articles. This algorithm contains three main components: Template creation, excerpt selection, and article assembly (see Figure 3-1).

To begin the article creation process, we first create a template which details the necessary sections. Existing articles of the same category serve as examples of proper article structure; specifically, many articles within the same category share the same sections. Therefore, we can build a template for a new article without having prior knowledge other than the general category.

Once a template is created, we must fill in each section with data from the Internet. This process is performed for each section independently. First, the search query provided in the template is used to gather candidate excerpts, which are scored using a classification algorithm. Then, the excerpts are clustered to find the most relevant excerpt for the section.

The selected excerpts are then combined into a coherent article. Because we draw information from several sources, it is possible that redundant information is contained across sections. The final step, therefore, is to eliminate this redundancy using Integer Linear Programming (ILP). This technique allows removal of all but the single most valuable sentence to convey each piece of information.

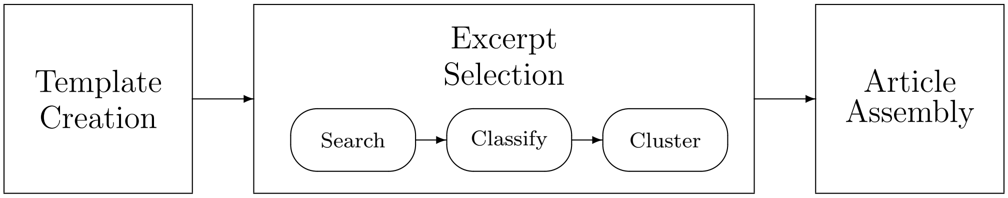


Figure 3-1: Overview of algorithm components.

3.1 Template Creation

The first task in creating a new article is to determine what information needs to be present. A typical Wikipedia article is composed of one short excerpt containing introductory information about the topic followed by sections for each facet of the topic. These sections may be common to many articles, such as *early life* for biographies or *treatment* for diseases, or they may be specific to a more narrow category (e.g., Filmography, which is common to actors), or an individual (e.g., Vice Presidency 1993-2001). In our new article, we want to address a range of topics represented by these sections, without providing irrelevant information.

To create a template for the new article, we use all articles in the given category as examples of good article structure, and we assume that some elements of the structure of any given article will recur in others. This is obviously not true in the case of the very specific sections (e.g., Vice Presidency 1993-2001), so we attempt to find the most common sections between all articles in the category. In addition to section titles, we also generate queries which can be used to find related information for each section. These will be used each time we want to fill in data for an article.

3.1.1 Section Selection

To determine the most common sections, we look at section titles across articles in our given category. Those which are shared in many articles are likely to be the most relevant to the category, while the less frequent or unique sections are not. Although many titles are standardized across articles, some related sections may have different titles. For example, there are many “Biography” and “Filmography” sections, but information relating to a person’s early life may be found in “Early Life”, “Childhood

Years", or even "Childhood and Family Life".

To deal with this, we extract all section and subsection titles from these related articles and stem them using the Porter stemming algorithm [24] to allow comparison between different forms of the words. We cluster them using the direct clustering algorithm provided in CLUTO, a publicly available clustering toolkit. [1] Similarity between elements is defined by cosine similarity between word vectors, where the length of each vector is equal to the total number of unique words, and each element within the vector is the count of that word in the title. To name each cluster, we simply take the most frequent individual title from the cluster.

After clustering, we ignore clusters with internal cosine similarity less than a given threshold to limit the tolerance for variation in titles. Clusters with low internal similarity contain very different titles; therefore, it's not likely to find similar information within the sections. Conversely, clusters with high internal similarity are easy to identify as being similar. For example, one cluster contains "Responsible Drug Use" and "Psychological Drug Tolerance", which both contain *drug*, but because there are also sections named "Drug Therapy", there are also several titles in the same cluster relating to *therapy*, such as "Positive airway pressure therapy" or "Additional Therapy". Because most of the titles relating to drugs have little to do with those relating to therapy, the internal cluster similarity is low. Choosing this category for a template would lead to poor results due to the mixed training data. Some example clusters are shown in Figure 3-2. For this experiment, we use a similarity threshold of 0.5.

Finally, we select the most populated clusters to become our template. Because these sections appear in many articles in the given category, we assume that they are representative of the information likely to appear in these articles as a whole. Also, because we do this on a per-category basis, we are likely to find those sections which are category-specific, such as Filmography.

These templates need to be created only once per section; any future articles created in the same section will use the same template. Templates generated for

[1] CLUTO is available at http://glaros.dtc.umn.edu/gkhome/views/cluto.

Causes
Genetic Causes
Genetics
Causes and Mechanisms
Mechanisms
Causes of Injury
Causative agent
Cause of Ephonia
Environmental causes

(a) Samples from Causes

Treatment
Emergency Treatment
Treatment Research
Research
Treatment and Prevention
Medical Treatment
Benefits of methadone treatment
Methadone
Other Treatment Options

(b) Samples from Treatments

Diagnosis
Differential Diagnosis
Symptoms and Diagnosis

(c) Samples from Diagnosis

Symptoms
Signs and Symptoms
Signs

(d) Samples from Symptoms

Figure 3-2: Sample title clusters from *Diseases* category.

American Film Actors
Biography
Early Life
Career
Personal Life

(a) Section Template for *American Film Actors*

Diseases
Symptoms
Diagnosis
Causes
Treatment

(b) Section Template for *Diseases*

Figure 3-3: Sample section templates for *American Film Actors* and *Diseases*.

American Film Actors and *Diseases* are shown in Figure 3-3.

3.1.2 Query Selection

For each article to be created, we will need to find text for each section of the template. To do this, we will need queries for each section. Because we will not have information about the new article except its category, we need a general query that can be used for all articles in that category. Search engines do not handle noisy queries well, so irrelevant information should not be included. For example, in Figure 3-4, we compare two potential queries for a *personal life* section of an actress. Because Emma Watson is unmarried and has no children, the second query produces several erroneous results, while the first query yields more reasonable text for the section. With this in mind,

Emma Watson Personal life - Vox
Emma Charlotte Duerre Watson[1] (born 15 April 1990) is an English actress who rose to fame playing the role of Hermione Granger in the Harry Potter film ...
Emma Watson - Wikipedia, the free encyclopedia
3 Personal life; 4 Filmography; 5 Awards; 6 References; 7 External links ... Emma Watson was born in Paris, France, daughter of Jacqueline Luesby and Chris ...
Emma Watson, Pictures, Bio, Video, Wallpaper, Harry Potter News
Emma Watson on the Web. ... 2007 cover of Tatler magazine with a full fashion photo spread and a feature interview detailing her career and personal life. ...
Emma Watson photos: Watson Personal Life
Dec 13, 2007 ... Watson Personal Life. Emma Watson was born in Paris, France to Jacqueline Luesby and Chris Watson, English lawyers who lived in France at ...

(a) Search Results for " Emma Watson personal life " – Query generated from title text

Parenting techniques mirror marriage quality— Sheknows.com
Just because a married couple has a good relationship when a child is born is no guarantee the marriage will stay that way as their child grows older, ...
Divorce reduces chance of new, successful relationship— Sheknows.com
In particular, a previous marriage or children from a previous relationship, reduce the chances of a new relationship. Moreover, the prospects are slimmer ...
Entertainment : The Ins And Outs Of Emma Watson's Relationships ...
He especially cares about the needs of children, mothers, and families, and he wants a love partner who values marriage, home, and family as much as he does ...
The Emma Watson Forums
The Emma Watson Forums _ Advice _ Marriage At A Young Age .... I guess it's good for them because they get to see how their relationship progresses. ...

(b) Search Results for " Emma Watson marriage child relationship" – Query generated from body text

Figure 3-4: Search results using both query generation methods for actress Emma Watson (Hermione in the *Harry Potter* series). Because Emma is young and unmarried, results generated from body text in 3-4(b) are less relevant than those generated from title text in 3-4(a).

we evaluated two methods of extracting search queries from data; one using section titles, and the other using section body text.

Section title queries

Our first method of query generation is based on section titles. While creating the template (as described in Section 3.1), we assign each cluster a label, the most popular exact title in the cluster, which becomes the title of that section. Popular section titles have the benefit of being very generic; in order to be applicable to many articles, a title cannot refer to any unique properties relating to a given topic. This avoids any specific references to subjects that may be somewhat common but not ubiquitous, such as *marriage* or *children* in a *personal life* section.

Body text queries

The body text of a section has the potential to generate a more specific query. While section titles contain a few generic words to choose from, body texts contain many specific keywords relating to the topic. For example, most *early life* sections contain information about *family* or *school*, even though these are not mentioned explicitly in the title.

To select a few useful query words from the large number of total words in the section, we must use a metric to determine how useful any particular word is; in this case, we use information entropy.

The information density function, I, is a measure of the information provided by a word in a document. We split our training data into two portions: sections which match the current type, and sections which do not. For example, when trying to determine a query for a *personal life* section in the *Actors* category, all sections in every article within the *Actors* category is marked either positive, if it is a *personal life* section, or negative, if it is not. I is given by the following formula, where a is a fraction of the positive examples and b is a fraction of the negative examples:

$$I(a,b) = -a\log(a) - b\log(b)$$

Words which appear in every section regardless of type will have a low I, while words specific to a section type will have high I.

Given a word w, we partition the data into four sets:

p_0 – positive examples not containing w

n_0 – negative examples not containing w

p_1 – positive examples containing w

n_1 – negative examples containing w

In addition, we define the following counts:

p – number of positive examples ($p_0 + p_1$)

n – number of negative examples ($n_0 + n_1$)

Now, we can compute the probabilities of a word appearing (*pos*) and not appearing (*neg*) in a section of the desired type, weighted by the amount of information that conveys.

$$neg = \frac{p_0 + n_0}{p + n} \cdot I\left(\frac{p_0}{p_0 + n_0}, \frac{n_0}{p_0 + n_0}\right)$$

$$pos = \frac{p_1 + n_1}{p + n} \cdot I\left(\frac{p_1}{p_1 + n_1}, \frac{n_1}{p_1 + n_1}\right)$$

This can be used to compute the total entropy, D:

$$D = neg + pos$$

Entropy increases when a word is descriptive either in presence or absence in the given section. For example, the word *death* appears very rarely in *biography* sections; however, this absence makes it a very good predictor of whether or not a given excerpt should be a *biography* section. Therefore, the entropy of *death* is high. Because we only want words which are descriptive by presence, we look at only *pos* from this formula.

Some examples of words and their *neg*, *pos*, and D values for the Early Life section of *American Film Actors* are in Table 3.1. Notice that the keywords we expect to see in the section have high *pos* values, while the keywords we expect are more common in other sections have high *neg* values.

To select the optimal words for a body text query, we first find the entropy of each word that appears in the section. We select the top k words from *pos*, then use them to build all possible queries of 1-3 words. For example, in *personal life*, some queries are as follows:

- marriage children relationship
- romantic wife rumor

Keyword	*pos*	*neg*
born	0.257	0.139
year	0.186	0.377
attend	0.143	0.361
school	0.132	0.402
life	0.048	0.514
home	0.030	0.532
death	0.013	0.559
plane	0.002	0.561
socialite	0.001	0.561
airport	0.001	0.561

Table 3.1: Positive and negative entropy values for select keywords from the *early life* section of *American Film Actors*.

- son drama

Selecting the best query

Given the section title query and all body text queries for the given section as candidates, we need to determine which query best represents the entire section. Using each query, we retrieve text excerpts from the Internet for that section of several existing articles in Wikipedia. Then, we select the best resulting excerpt for each article by comparing these excerpts with the corresponding section from the original Wikipedia article using cosine similarity. Where s is the query score for a single article, there are n candidate excerpts, and sim_i is the similarity between candidate excerpt i and the existing Wikipedia section,

$$s = \max\left(sim_i\right) \qquad \forall i \in [1 \ldots n]$$

The query score is equal to the sum of the individual scores. That is, where q is the total query score and there are m articles,

$$q = \sum_{j=1}^{m} s$$

Once we have computed a score for each query, we select the query with the highest total score. The query scoring process is illustrated in Figure 3-5.

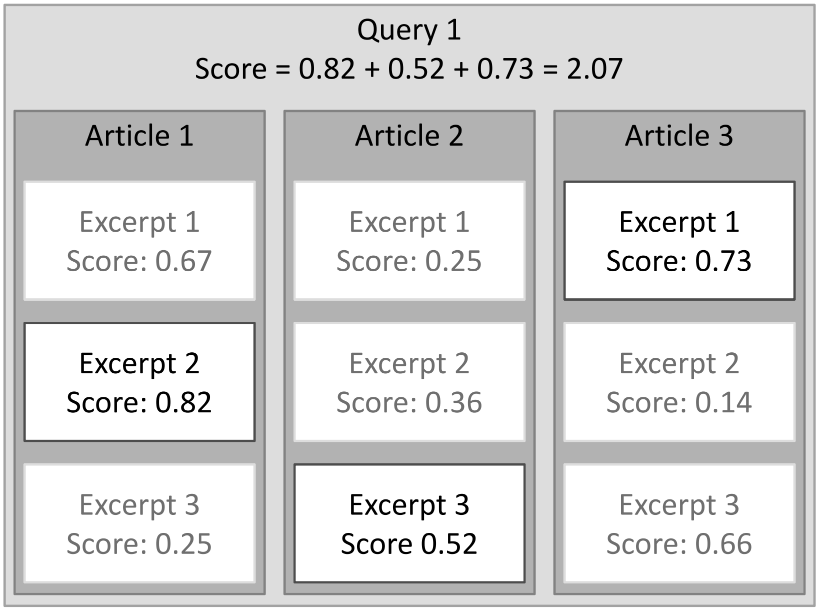


Figure 3-5: Query selection: Scoring of a single query. For each section, a search is performed using the query. The query score is the sum of the highest cosine similarities in each article.

	Avg.	Articles w/ sim.		
Query	Similarity	≥ 0.35	≥ 0.50	≥ 0.80
"personal life"	0.412	55.7%	33.8%	6.6%
"marriage children relationship"	0.285	35.4%	11.6%	0.7%

Table 3.2: Comparison of title text ("personal life") and body text ("marriage children relationship") queries.

When running this experiment on the *personal life* section of the Actors category in Wikipedia, the best candidate queries were *personal life* (section title) and *marriage children relationship* (body text). With this data, the section title query outperforms the body text queries, as shown in Table 3.2.

As mentioned before, this is caused by creating a noisy query from body text; the words in the body text query may be relevant to most articles, but for the rest, the extra query words become noise, lowering the quality of the search results.

The completed template for each category is comprised of the relevant sections and the optimal query for each section.

3.2 Excerpt Selection

Once the template has been created, we must fill each section with data from the Internet. This process (shown in Figure 3-1) is repeated once for every section in the document.

The first step is to retrieve information from the Internet for the section. Using the query from the template, we perform the search and select candidate excerpts.

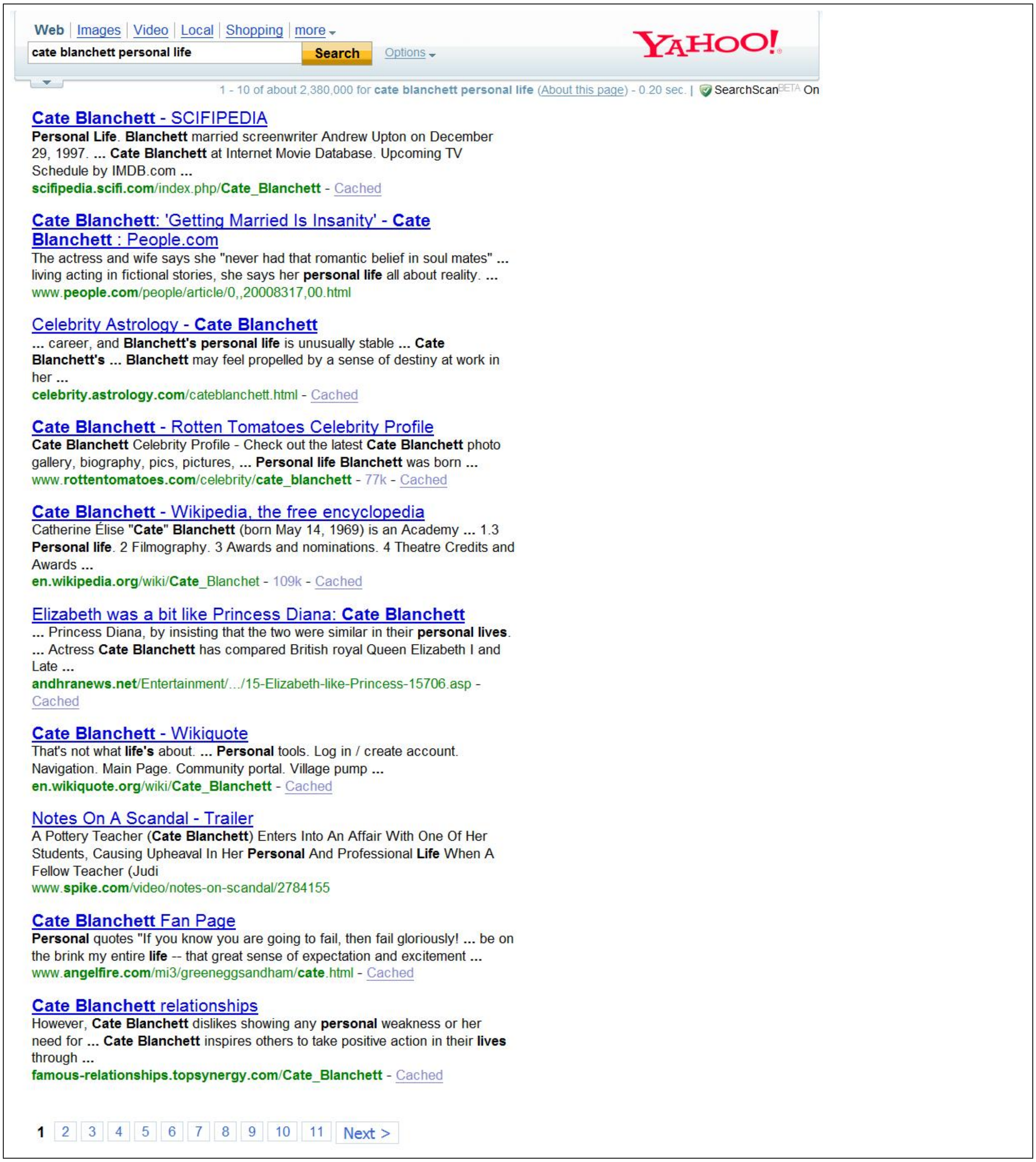
Next, we must evaluate the quality of each candidate excerpt. To do this, we use a classification algorithm which determines whether a section is representative of features we expect from the section. Each section is given a score from 0 (not representative) to 1 (very representative).

While these scores are often accurate, they are based on lexical features, so there may be false positives. This is analogous to well-disguised spam being allowed through an email spam filter. The third step of excerpt selection, therefore, is to cluster candidate excerpts to expose the false positives and select one excerpt which best represents the section.

3.2.1 Search

Using the query defined in our template, we perform a search using Yahoo! and select the first ten results (the first page of search results). An example of the search results can be found in Figure 3-6. Because we are training and testing on Wikipedia data, we filter Wikipedia itself out of the search results. From each page, we extract chunks of text, some of which may appear in the final document.

Because the results are in HTML format, we attempt to distinguish paragraph boundaries using certain indicators, such as the use of the `<p>` (paragraph) HTML tag, several `<br>` (break) tags in a row, or blank lines in the file. We accept chunks as excerpts if they have 3-15 sentences and 15-400 words in total, and we ignore any excerpt which has an excessive amount (greater than 10%) of punctuation. An example of excerpt retrieval can be found in Figure 3-7. We retrieve an average of 143 excerpts per search.


Web | Images | Video | Local | Shopping | more
Search
Options
YAHOO!
SearchScan BETA On
1 2 3 4 5 6 7 8 9 10 11 Next >

cate blanchett personal life

1 - 10 of about 2,380,000 for **cate blanchett personal life** (About this page) - 0.20 sec.

Cate Blanchett - SCIFIPEDIA
Personal Life. **Blanchett** married screenwriter Andrew Upton on December 29, 1997. ... **Cate Blanchett** at Internet Movie Database. Upcoming TV Schedule by IMDB.com ...
scifipedia.scifi.com/index.php/**Cate_Blanchett** - Cached

Cate Blanchett: 'Getting Married Is Insanity' - **Cate Blanchett** : People.com
The actress and wife says she "never had that romantic belief in soul mates" ... living acting in fictional stories, she says her **personal life** all about reality. ...
www.**people.com**/people/article/0,,20008317,00.html

Celebrity Astrology - **Cate Blanchett**
... career, and **Blanchett's personal life** is unusually stable ... **Cate Blanchett's** ... **Blanchett** may feel propelled by a sense of destiny at work in her ...
celebrity.astrology.com/cateblanchett.html - Cached

Cate Blanchett - Rotten Tomatoes Celebrity Profile
Cate Blanchett Celebrity Profile - Check out the latest **Cate Blanchett** photo gallery, biography, pics, pictures, ... **Personal life Blanchett** was born ...
www.**rottentomatoes.com**/celebrity/**cate_blanchett** - 77k - Cached

Cate Blanchett - Wikipedia, the free encyclopedia
Catherine Élise **"Cate" Blanchett** (born May 14, 1969) is an Academy ... 1.3 **Personal life**. 2 Filmography. 3 Awards and nominations. 4 Theatre Credits and Awards ...
en.wikipedia.org/wiki/**Cate**_Blanchet - 109k - Cached

Elizabeth was a bit like Princess Diana: **Cate Blanchett**
... Princess Diana, by insisting that the two were similar in their **personal lives**. ... Actress **Cate Blanchett** has compared British royal Queen Elizabeth I and Late ...
andhranews.net/Entertainment/.../15-Elizabeth-like-Princess-15706.asp - Cached

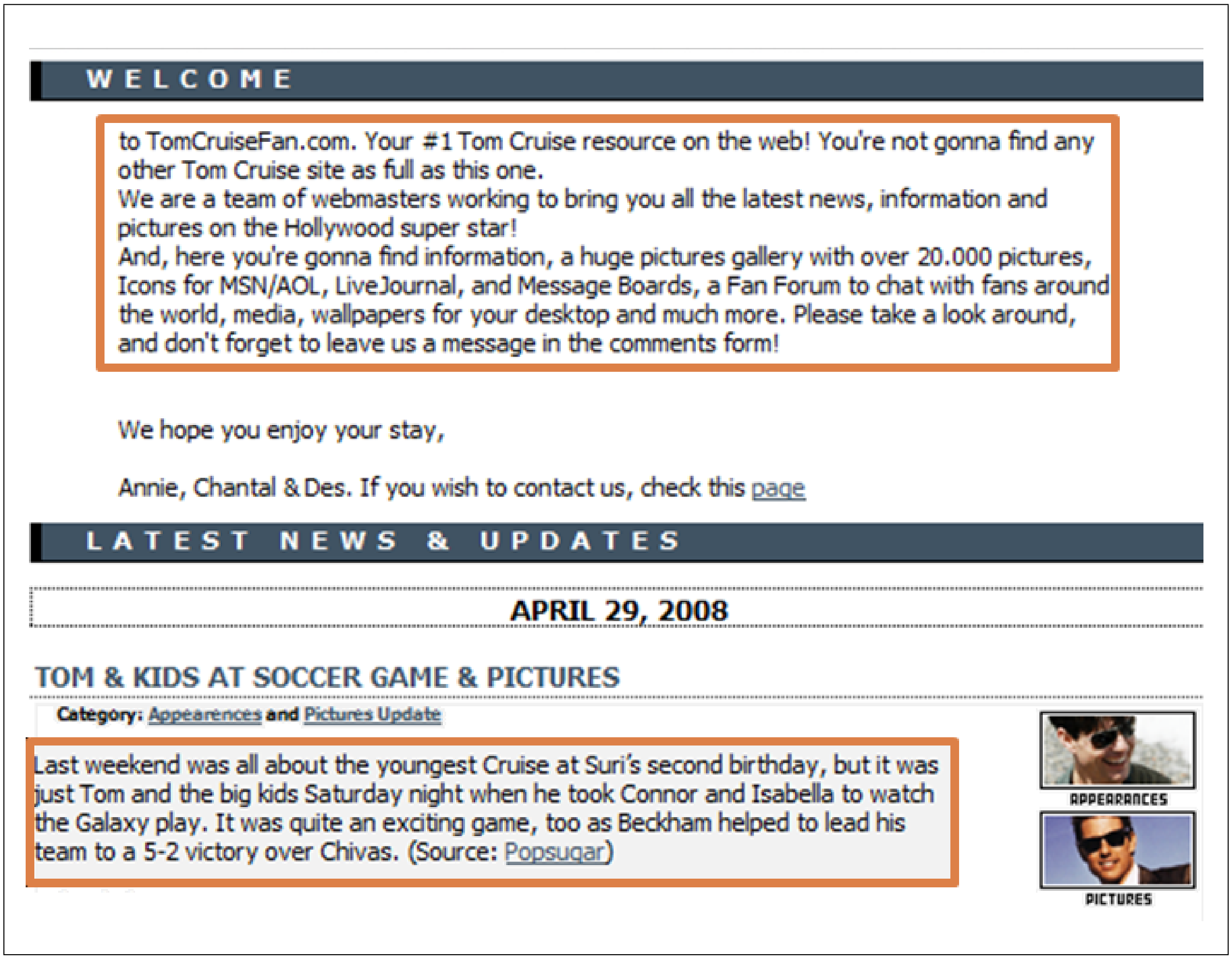
Cate Blanchett - Wikiquote
That's not what **life's** about. ... **Personal** tools. Log in / create account. Navigation. Main Page. Community portal. Village pump ...
en.wikiquote.org/wiki/**Cate_Blanchett** - Cached

Notes On A Scandal - Trailer
A Pottery Teacher (**Cate Blanchett**) Enters Into An Affair With One Of Her Students, Causing Upheaval In Her **Personal** And Professional **Life** When A Fellow Teacher (Judi
www.**spike.com**/video/notes-on-scandal/2784155

Cate Blanchett Fan Page
Personal quotes "If you know you are going to fail, then fail gloriously! ... be on the brink my entire **life** -- that great sense of expectation and excitement ...
www.**angelfire.com**/mi3/greeneggsandham/**cate**.html - Cached

Cate Blanchett relationships
However, **Cate Blanchett** dislikes showing any **personal** weakness or her need for ... **Cate Blanchett** inspires others to take positive action in their **lives** through ...
famous-relationships.topsynergy.com/**Cate_Blanchett** - Cached

Figure 3-6: Sample results from the search "cate blanchett personal life". Excerpts are taken from each of these results except the Wikipedia article.



WELCOME

to TomCruiseFan.com. Your #1 Tom Cruise resource on the web! You're not gonna find any other Tom Cruise site as full as this one.
We are a team of webmasters working to bring you all the latest news, information and pictures on the Hollywood super star!
And, here you're gonna find information, a huge pictures gallery with over 20.000 pictures, Icons for MSN/AOL, LiveJournal, and Message Boards, a Fan Forum to chat with fans around the world, media, wallpapers for your desktop and much more. Please take a look around, and don't forget to leave us a message in the comments form!

We hope you enjoy your stay,

Annie, Chantal & Des. If you wish to contact us, check this page

LATEST NEWS & UPDATES

APRIL 29, 2008

TOM & KIDS AT SOCCER GAME & PICTURES

Category: Appearences and Pictures Update

Last weekend was all about the youngest Cruise at Suri's second birthday, but it was just Tom and the big kids Saturday night when he took Connor and Isabella to watch the Galaxy play. It was quite an exciting game, too as Beckham helped to lead his team to a 5-2 victory over Chivas. (Source: Popsugar)

Figure 3-7: Excerpts selected from http://www.TomCruiseFan.com (shown in orange boxes). Note that not all selected paragraphs are relevant to the final topic; excerpts are simply groups of sentences.

3.2.2 Classification

Once we have a selection of excerpts to choose from, the next task is to determine which are reasonable for the given section. Note that our excerpts can be any blocks of text between boundaries, so many of the excerpts will be completely irrelevant to the topic. For example, in many cases there is a paragraph about the site itself as part of a standard header or footer.

To express how representative each excerpt is of the desired section, we use a publicly available maximum entropy classifier[2] to score the likelihood of the excerpt appearing in the final document. Our features are as follows:

- Unigrams
- Bigrams
- Count of numbers
- Count of proper nouns
- Count of personal pronouns
- Count of exclamation points
- Count of question marks
- Count of topic mentions (any word from article title)
- Length in words
- First appearance of words in text (percent of total length)

To train our classifier, we rely on two sets of data. The first is from existing Wikipedia articles from our training corpus (see Section 4.1). We use all sections from these articles as training examples. Sections of the same type that we are classifying are marked as positive examples, and the rest are negative examples. This data set allows us to make distinctions between sections within a single article.

The second data set is created by performing a web search on each training article, in the same way as described in Section 3.2.1. Resulting excerpts are compared to the actual Wikipedia section using cosine similarity. Those with low similarity are added to training data as negative examples. This allows us to better process irrelevant data such as standard website headers.

[2] MaxEnt classifier available at http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

The classifier is trained on both sets of data with equal weight put on each example from both sets.

3.2.3 Clustering

Given the scores from our classifier, we observe that some excerpts are rated more highly than they deserve. These excerpts may use promising keywords or repeat a name several times without providing any useful information. Therefore, we use a clustering algorithm to expose these false positives.

Excerpts from the Internet generally fall into three classes. First, there are the genuinely and obviously useful excerpts, which contain the topic name and facts relevant to the section at hand. In general, these excerpts are rated highly by the classification algorithm, with exact score dependent on the quality of information. Second, there are excerpts which are completely irrelevant to the topic at hand, such as general website information or discussion of other topics. These will not be rated highly by the classification algorithm, especially when the topic name is not mentioned at all.

The third class is more difficult to identify. Some excerpts contain only links to information about a person without giving any details (e.g., “Find everything you want to know about *topic* by clicking here!”). These pages use topic names liberally in order to convince search algorithms that their content is worthwhile. Our classifier may also score these excerpts highly, since multiple topic name mentions often means the excerpt is relevant to the topic. The problem is essentially the same as the one faced by email spam filters. Spammers attempt to make their messages look legitimate, so they are able to pass the filter.

Selecting the best excerpt for the final document is not necessarily straightforward, as the highest-rated excerpt may fall into this third category. To prevent these excerpts from becoming part of the final document, we notice that the most correct results for a particular section should be similar to each other. For example, if we are searching for the symptoms of a specific disease, we are likely to retrieve multiple excerpts which report the correct data, but relatively few excerpts which contain wrong or misleading data. We attempt to exploit this observation by clustering our resulting excerpts based on word frequency.

The number of candidate excerpts varies greatly between sections and articles. While the average number of excerpts is 143, the total may be as high as several hundred or as low as 50, depending on the amount and types of information available on the Internet. To ensure high internal cluster similarity, we would like fairly small clusters. We attempt to normalize this by setting the number of clusters equal to the number of excerpts divided by the desired number of results per cluster. In our evaluation, we use approximately ten results per cluster.

As in title clustering, we employ the direct clustering algorithm provided in the publicly available CLUTO clustering toolkit. [1] Words are stemmed using the Porter stemming algorithm [24] and common stop words are filtered out. For a similarity metric, we use cosine similarity between word vectors.

3.2.4 Final Selection

The scores of the paragraphs (p_i) within a cluster (c) are averaged to compute a cluster score. The highest scoring paragraph (E) within the highest scoring cluster (C) is selected as the final excerpt.

$$\begin{aligned} C &= \operatorname*{argmax}_{c\in[1\ldots k]} \left(\frac{1}{n} \sum_{i\in c} p_i \right) \\ E &= \operatorname*{argmax}_{i\in C} p_i \end{aligned}$$

3.3 Article Assembly

The final step in our algorithm is to combine the excerpts chosen for each section of the template. Because we select excerpts for each section individually, there may be some overlap between them. For example, the *early life* section of a child star's article may contain similar information to her *career* section.

Our goal is to eliminate redundancy while preserving the article structure, by filtering content at the sentence level. When considering which of the mutually redundant sentences to eliminate, we need to take into account the likelihood of a sentence to be selected for the final document, as measured by the classification score of the corresponding excerpt. This is important because sentences with high semantic overlap may exhibit stylistic differences. One may be more appropriate for inclusion in the article than another. In addition, we need to make sure that the size of the resulting article is consistent with the length of human-authored Wikipedia articles.

To select sentences that satisfy all these constraints, we employ integer linear programming (ILP). This framework is commonly used in generation and summarization applications where the selection process is driven by multiple constraints [8, 17].

Let e_{s_i} be the probability that sentence s_i is extracted. We consider the probability of a sentence to be equal to the probability of the corresponding excerpt. We represent sentences included in the output using a set of indicator variables x_{s_i} that are set to 1 if s_i is extracted, and 0 otherwise. The score of a document is the sum of the scores of all the sentences:

$$\sum_{s_i} x_{s_i} e_{s_i} \tag{3.1}$$

Our inference task is solved by maximizing the overall score of sentences in a given document:

$$\text{argmax} \sum_{s_i} x_{s_i} e_{s_i} \tag{3.2}$$

subject to:

$$x_{s_i} \in \{0, 1\} \; \forall \; s_i \tag{3.3}$$

We augment this basic formulation with two types of constraints. The first type of constraint ensures that mutually redundant sentences are not included in the system output, while the second type expresses constraints on section length, preventing over-trimming.

Redundancy Constraints

We place constraints to ensure only one of a pair of similar sentences is kept. The degree of overlap between sentences is measured using cosine similarity denoted by *sim* in the equation:

$$(x_{s_i} + x_{s_j}) sim_{s_i,s_j} \leq 1.0 \tag{3.4}$$

If sentences s_i and s_j have cosine similarity $sim_{s_i,s_j} > 0.5$, retaining both sentences in the final article would cause $x_{s_i} + x_{s_j} = 2$ and $2sim_{s_i,s_j} > 1.0$, which is not permitted. If only one of the sentences appears, $(x_{s_i} + x_{s_j}) = 1$, and the constraint is satisfied. On the other hand, if the cosine similarity between s_i and s_j is $sim_{s_i,s_j} \leq 0.5$, then $2sim_{s_i,s_j} \leq 1.0$, and it is possible to retain both sentences.

Length Constraints

Our length constraint controls the compression rate of the redundancy elimination process. This constraint also regulates the distribution of deletions across different sections. Without such regulation, the article may end up with sections that are too short and incoherent. For instance, when two excerpts share some information, the lower scoring excerpt may have almost all of its sentences removed.

For each excerpt e, we require that at least some fraction (t) of the sentences in the excerpt (denoted by $N(e)$) must be kept.

$$\forall e, \sum_{s_i \in e} x_{s_i} \geq N(e)t \tag{3.5}$$

Notice that the fraction t is not known in advance. However, we estimate this parameter from our development data. In our experiments this number is set to $\frac{1}{3}$.

Together, these constraints enforce that there will be low redundancy, excerpts will remain a decent size, and only the better of two redundant sentences will be kept. The objective function (Equation 3.2) retains as many sentences as possible without violating these constraints.

Solving the ILP

In general, solving an integer linear program is NP-hard [10]. Fortunately, there exist several strategies for solving ILPs. In our study, we employed *lp_solve* [3], an efficient Mixed Integer Programming solver which implements the Branch-and-Bound algorithm. We generate and solve an ILP for every article we wish to generate. For documents of average size (approximately 558 words), this process takes under 1 second on a 3 GHz machine.

3.4 Summary

Overall, our algorithm is the combination of three separate modules: Template Creation, Excerpt Selection, and Article Assembly. In the template creation step, we generate a template for a general article category based on the structure of pre-existing articles in that category. To do this, we cluster the existing titles and select the most popular clusters with internal similarity above a threshold. Additionally, we select a query based on either body or title text for each section of the template.

The second step, Excerpt Selection, is composed of three sub-modules. First, the Search sub-module uses the corresponding template query to retrieve a set of excerpts from the web. Second, the Classification sub-module scores each excerpt based on a set of lexical features. Each score represents how representative an excerpt is of the 'typical' section. Third, the Clustering sub-module clusters excerpts based on cosine similarity, to expose false positives and ensure that the main concepts are repeated across several excerpts, implying that several sources agree on the facts. From these clusters, one excerpt is selected based on its cluster and classification score.

The final step, Article Assembly, puts all of the section excerpts together into one document. The primary concerns during this step are removal of redundant information and ensuring that the best duplicate sentence is chosen for retention. To accomplish this, we use an ILP formulation with constraints to preserve minimum paragraph size.

[3] `http://lpsolve.sourceforge.net/5.5/`

Once this algorithm is complete, we have transformed an article title and category into a complete article. In the next section, we will discuss our evaluation methods, including metrics for both individual sections and complete articles, and experimental results.

Chapter 4

Experimental Setup

We evaluate our system by generating new articles for several topics already existing in two different Wikipedia sections, the *Diseases* category and the *American Film Actors* subcategory of *Living People.* This provides a good basis for comparison between the generated article and the existing article. To gain a comprehensive view of the strengths of this system, we evaluate the accuracy of both individual sections and entire articles. In this section, we describe our data, baseline, and evaluation metrics.

4.1 Data

For training and testing our algorithm, we use a corpus of articles from Wikipedia. Specifically, we use the *Diseases* category and the *American Film Actors* subcategory of the *Living People* category. We choose *Living People* because the data in these biographies tends to be well-structured into common subsections. Furthermore, new articles will often appear in *Living People*, as most deceased people have already had their lives well documented in Wikipedia and other encyclopedias. Likewise, in the *Diseases* category, there are several common article subsections.

In *Living People*, the *American Film Actors* category contains the most articles; therefore, it provides substantial data for training and testing. This category contains 2150 articles, including a total of 8648 sections, for an average of 4 sections per article.

Section	Training	Evaluation	Total
American Film Actors			
Career	578	145	723
Personal Life	462	116	578
Early Life	486	121	607
Biography	161	40	201
Diseases			
Treatment	106	27	133
Causes	98	24	122
Symptoms	177	44	221
Diagnosis	43	11	54

Table 4.1: Training and testing corpus sizes for each section of *American Film Actors* and *Diseases*.

Within *American Film Actors*, we focus on *Career* (found in 723 articles), *Personal Life* (found in 578 articles), *Early Life* (found in 607 articles), and *Biography* (found in 201 articles), as these are the most common sections across the category, and therefore can form an article template.

In *Diseases*, there are 523 articles, with a total of 2283 sections, also an average of 4 sections per article. Within this category, we focus on *Causes* (found in 122 articles), *Diagnosis* (found in 54 articles), *Symptoms* (found in 221 articles), and *Treatment* (found in 133 articles), the most common sections.

For each section and the articles as a whole, we select data from the Internet to serve as supplementary negative examples. For each article or section, we perform a web search using the query specified in the template. Any resulting excerpt with cosine similarity less than a designated threshold is used as a negative example. In this experiment, we use a threshold of 0.2. These supplementary results generally cover web-specific information such as copyright notices, general navigation instructions, website introductions, and so on.

For our per-section evaluation, we train our model on 80% of each section and test on the remaining 20%. For the complete article evaluation, we compile a list of all articles containing at least one of the above sections. We train on 80% of this list, and we test on the remaining 20% of articles. See Table 4.1 for a complete list of training and testing data.

4.2 Baselines

We compare our system's performance against several baselines, to illustrate different features of our approach. The first baseline verifies the template selection portion of our system, while the second tests the final clustering and ILP formulation. In addition, we create an Oracle, which selects the optimal (closest to existing Wikipedia) result from the web search. This provides an upper bound for the performance of the classification and clustering portion of our algorithm.

4.2.1 Search

Our first baseline, *Search*, relies solely on search engine ranking for content selection. Using an article name as a query (e.g., *Bacillary Angiomatosis*), this method selects the web page ranked first by the search engine. From this page, we select the first K paragraphs, where K is the average number of sections in articles of the given category. This yields a document of comparable size to the output of our system. Despite its simplicity, this baseline is not naive; extracting material from a single document guarantees that the output is coherent and does not have any redundancies. Moreover, a document which is highly ranked for a generic search query may readily contain a comprehensive overview of the topic.

4.2.2 No Template Generation

Our second baseline, *No Template Generation*, is a variation of our method that does not use a template; therefore, there are no structural constraints on content selection. To generate a new article, this method searches the web once, using the article name as a query (e.g., *Emma Watson*). We use the complete Wikipedia article text in the given category as positive examples, and we draw negative examples from dissimilar web results as for our full system. After performing a web search, we cluster the results using CLUTO's direct clustering algorithm and select one paragraph each from K clusters with the highest average classification score. This baseline follows previous work on biography generation [5].

4.2.3 Max Score

Our third baseline uses a simplified version of our template completion algorithm. This simplification uses the same excerpt retrieval and classification steps as the complete algorithm, but simply selects the excerpt with the highest classification score as the final excerpt, instead of performing clustering and ILP.

This approach should perform identically to the complete algorithm in most cases; however for those excerpts which incorrectly appear to be extremely relevant, it should perform worse. For example, some Internet pages contain links to forums or fan sites for a variety of actors. These sites attempt to insert relevant keywords to attract search engines; however, they contain no relevant data. They may score highly in classification, due to the relevance of the keywords, but they will not be similar to other high-scoring excerpts. Therefore, this baseline measures the importance of the clustering step.

4.2.4 Oracle

In addition to these baselines, we compare against an Oracle system. For each section present in both the template and the human-authored Wikipedia article, the Oracle selects the excerpt with the highest cosine similarity to the existing Wikipedia article. This excerpt is the optimal automatic selection from the results available, and therefore represents an upper bound on the excerpt selection task. Not all Wikipedia articles in the given category contain all of the sections in the template; therefore this system may produce a shorter article than our algorithm. The Oracle system generates an average of 2 sections per article.

4.3 Evaluation Metrics

To test whether our final article is reasonable, we use both human and automatic evaluation. Our automatic evaluation assesses the content of a large set of system output by comparing it with the corresponding Wikipedia article. Human evaluation of a smaller selection of articles allows us to determine the quality of the prose selected and appropriateness for each section of content.

4.4 Rouge Scores

Rouge (Recall-Oriented Understudy for Gisting Evaluation)[15] includes several measures to automatically evaluate the quality of a summary by comparing it to one or more reference summaries. Rouge has been used extensively at the Document Understanding Conference (DUC), specifically DUC-2001, DUC-2002, and DUC-2003. In these cases, Rouge correlates well with human evaluation.

For this experiment, we choose to use Rouge-N, with $N = 1$. Rouge-N is defined as n-gram coverage between a test summary and one or more reference summaries. The formula for Rouge-N is as follows:

$$\text{Rouge-N} = \frac{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

In this case, we have only one reference summary, the existing text in Wikipedia. We use the publicly available Rouge toolkit to compute unigram overlap recall, precision, and F-score for each test example in comparison to the example from Wikipedia. These scores are averaged per section in the case of section analysis, or per category in the case of complete article analysis.

4.5 Section Selection

Because the basis of our approach is to choose sections independently, a high quality article can be formed only when the individual sections are chosen well. Therefore, part of our evaluation is on each section separately.

Using the training and testing data specified earlier, we generate articles with our complete system and compare the results to the *Max Score* and *Oracle* baselines.

4.6 Complete Article Generation

To fully evaluate our system, it is important to determine the quality of the article as a whole in addition to each section individually. This allows us to prove the quality of the template generation step. If the generated article does not cover a good percentage of the information in the article, our template may be unreasonable.

We use both human and automatic evaluation. It is difficult to judge the quality of a generated article automatically because of the vast amounts of information available on any given topic. Information within the generated article may be of excellent quality without matching the existing article completely. However, in order to reasonably evaluate our results, we need evaluation on many test data points. Therefore, we use automatic evaluation to judge overall content overlap of a large number of articles, and we use human evaluation to examine the quality (through content and readability) of a smaller selection of articles.

4.6.1 Human evaluation

For human evaluation, we asked eight evaluators to rate articles on a scale of 1 (worst) to 5 (best) for both content and readability. Content was defined as information breadth and accuracy, while readability covered tone, redundancy, and transitions between excerpts. This evaluation method is applied to a set of 16 articles in the *American Film Actors* category. These metrics have been used in previous work [23].

4.6.2 Automatic evaluation

For this experiment, we generate complete articles and compare performance with the *No Template*, *Max Score*, and *Oracle* baselines. For each article, we use Rogue to evaluate the generated article text against the original Wikipedia articles.

Chapter 5

Results

5.1 Section Selection

The results of section selection are shown in Table 5.1. Unfortunately, many of these results are not statistically significant. This suggests one of a few things. First, there may simply be not enough data. Proving statistical significance with as few test cases as we have here is difficult, especially when there is high variance between articles. This is the case here, where quality of information available on the Internet is highly topic-dependent. Additional tests could combine several types of actors or several types of diseases to achieve more test points. Second, the clustering sub-module may need to be improved by changing the number of clusters, clustering algorithm, or final cluster selection. Finally, the clustering sub-module may be unnecessary, if the classifier score is sufficient to distinguish a better excerpt.

Comparison to the Oracle shows that there is room for improvement in final paragraph selection. This is likely the effect of the classification sub-module. Designing new features to more accurately filter out false positives and more reliably recognize reasonable excerpts may be required for the score to improve.

In addition, it is possible that even though the Wikipedia article does not contain a specific piece of information, it may also be valid information for an article. Likewise, not all information about a subject is present in the Wikipedia article. Because of this, Rouge may not be an optimal evaluation metric; this is the reasoning behind

Section	Recall	Precision	F-measure
American Film Actors			
Biography∗			
Max-Score	0.38	0.46	**0.36**
Full Model	0.31	0.53	0.34
Oracle	*0.46*	*0.62*	*0.48*
Careers			
Max-Score	0.39	0.37	0.32
Full Model	0.34	0.39	0.32
Oracle	*0.45*	*0.55*	*0.44*
Early Life			
Max-Score	0.51	0.36	0.38
Full Model	0.50	0.41	**0.41**
Oracle	*0.68*	*0.68*	*0.64*
Personal Life			
Max-Score	0.34	0.26	0.25
Full Model	0.31	0.28	0.25
Oracle	*0.51*	*0.49*	*0.44*
Combined			
Max-Score	0.41	0.34	0.32
Full Model	0.37	0.38	0.32
Oracle	*0.53*	*0.57*	*0.50*
Diseases			
Causes∗			
Max-Score	0.25	0.31	0.22
Full Model	0.25	0.34	**0.23**
Oracle	*0.37*	*0.34*	*0.31*
Treatment			
Max-Score	0.29	0.22	0.21
Full Model	0.30	0.21	0.21
Oracle	*0.43*	*0.21*	*0.24*
Diagnosis			
Max-Score	0.33	0.29	0.30
Full Model	0.33	0.29	0.30
Oracle	*0.50*	*0.33*	*0.34*
Symptoms			
Max-Score	0.29	0.20	0.18
Full Model	0.26	0.20	**0.19**
Oracle	*0.45*	*0.21*	*0.25*
Combined			
Max-Score	0.27	0.22	**0.20**
Full Model	0.25	0.22	0.19
Oracle	*0.42*	*0.26*	*0.27*

Table 5.1: Section selection Rouge scores broken down by section and combined average (weighted by number of test points) per category. The methods against which our model has significantly different results are indicated with $*$ for $p \leq 0.055$. The Wilcoxon Signed Rank Test is used for statistical significance.

providing human evaluation of complete articles.

5.2 Complete Article Generation

5.2.1 Automatic Evaluation

Automatic evaluation results (shown in Table 5.2) indicate that our model outperforms the baselines. Interestingly, the outputs produced by the baseline systems exhibit problems of different types. Outputs produced by the *No Template* baseline often focus on a narrow aspect of the topic at the expense of breadth. For instance, some articles in the *American Film Actors* category consist solely of information which falls into the *Filmography* section while skipping other topics typically reported in human-authored articles. This trend in the baseline outputs support our hypothesis that explicit modeling of content structure in the article generation process is essential for improving its coverage.

The performance of the *Search* baseline changes dramatically from article to article. This is expected since this method does not filter in any way the output of a search engine: if the top ranked article provides a good description of the topic, then the selection mechanism succeeds. Otherwise, the method yields poor results. These results are consistent with previous findings in Question-Answering literature that demonstrate the need of additional analysis of search results.

Performance of the *Oracle* is lower than expected. This is a result of the definition of the *Oracle* system – if a section from the template is not present in the human-authored article, no excerpt is included for that section. Since the comparison is done against the full article, the *Oracle* is penalized for sections that are not covered.

5.2.2 Manual Evaluation

The results of human evaluation are shown in Table 5.3. The ranking between different systems is consistent with the result of the automatic evaluation – our system performs better than *No Template* and *Search* baselines, but is outperformed by *Oracle*.

	Recall	Precision	Rouge F-measure
American Film Actors			
Search	0.13	0.45	0.17 $*$
No Template	0.24	0.44	0.27 $*$
Oracle	0.33	0.70	0.40 $*$
Full Model	0.45	0.51	**0.42**
Diseases			
Search	0.31	0.38	0.25 $*$
No Template	0.42	0.32	0.29 $*$
Oracle	0.32	0.38	0.28
Full Model	0.41	0.38	**0.32**

Table 5.2: Results of complete article automatic evaluation. Rouge scores are obtained by comparing automatic outputs to actual Wikipedia articles in two categories — *American Film Actors* (215 articles) and *Diseases* (46 articles). The methods against which our model has significantly different results are indicated with $*$ for $p \leq 0.02$. The Wilcoxon Signed Rank Test is used for statistical significance.

	Content	Readability
Search	1.66	2.46
No Template	2.86	3.03
Oracle	4.25	3.96
Full Model	**3.40**	**3.53**

Table 5.3: Results of complete article human evaluation. Content and readability are rated between between 1 (worst) and 5 (best). The evaluation was performed on 16 articles from the *American Film Actors* category.

The low readability scores of *Search* are surprising. We expected low content scores since this method extracts text from a single article in contrast to other, more diverse methods; however, we expected high readability scores since this method takes a contiguous block of text. Manual inspection of the results reveals that multiple extracted fragments contain headers and navigational information. This may indicate a need for a better excerpt extraction procedure.

5.3 Summary

The best comparison when attempting to generate a Wikipedia-style article is Wikipedia itself. Given the test subset of our Wikipedia corpus, both *American Film Actors* and *Diseases*, we generate new articles for each section and each topic. Then, we compare our generated article to the existing article to determine the topic coverage, using ROUGE, a metric for word overlap.

We evaluate our results in three experiments: automatic section evaluation, automatic complete article evaluation, and human complete article evaluation. Because the end goal is a complete article, the complete article evaluation should be a better measure of the system as a whole. However, a good article is comprised of good sections, so we attempt to evaluate sections individually as well.

Section selection evaluation indicates that work may be needed to perfect the clustering sub-module, as there is no significant difference in performance with and without the sub-module. On the other hand, the evaluation itself may leave something to be desired, as amount of data may have been insufficient for a valid comparison.

With complete article evaluation, on the other hand, there is a significant difference between our model and the no-template baseline for both human and automatic evaluation. This baseline is designed to evaluate the method of using a template for each category; these results do support a template-based approach.

Chapter 6

Conclusions and Future Work

This thesis describes an algorithm for generating Wikipedia-style articles. It uses a three-step process consisting of Template Creation, Excerpt Selection, and Article Assembly. First, a template is created using existing articles as examples of good article structure. Then, each section is completed using information from the Internet. Finally, the pieces are combined to form a single, cohesive article. This algorithm shows promise in selecting a breadth of information from across the Internet to create new, relevant articles.

Manual and automatic evaluation show that system-produced articles provide informative overviews of a topic, consistent in style with Wikipedia articles. Further analysis reveals that content selection driven by structural patterns observed in human-authored documents yields significantly better performance than a structure-agnostic approach. In addition, the automatic evaluation shows strong correlation with manual evaluation in most cases. This result encourages utilization of fast, inexpensive evaluation methods while developing new algorithms for this task.

This work opens several directions for future research. First, *Diseases* and *American Film Actors* exhibit fairly consistent article structures, which is successfully captured by a simple template creation process. With categories that exhibit structural variability, however, more sophisticated statistical approaches may be required to produce better templates. Second, the current Wikipedia repository contains links to the original sources and the paraphrased versions produced by human contributors. This is an excellent resource for learning how to paraphrase the source articles and fuse information on a more refined level. If this line of research is successful, Wikipedia can be directly populated with system outputs. Finally, instead of creating articles from scratch, it would be possible to edit certain sections of existing articles which are missing or contain outdated information. An updating procedure could seek out new information and update articles as it becomes available.

Appendix A

Sample Generated Articles

The following are articles generated by the full algorithm in the *Diseases* category. Each of these is only present as a stub in Wikipedia.

A.1 3-M syndrome

Diagnosis

DNA banking is the storage of DNA (typically extracted from white blood cells) for possible future use. Because it is likely that testing methodology and our understanding of genes, mutations, and diseases will improve in the future, consideration should be given to banking DNA of affected individuals. DNA banking is particularly relevant in situations in which molecular genetic testing is available on a research basis only. No laboratories offering molecular genetic testing for prenatal diagnosis of 3-M syndrome are listed in the GeneTests Laboratory Directory. However, prenatal testing may be available for families in which the disease-causing mutations have been identified in an affected family member in a research or clinical laboratory. [1]

Causes

Three M syndrome is thought to be inherited as an autosomal recessive genetic trait. Human traits, including the classic genetic diseases, are the product of the interaction of two genes, one received from the father and one from the mother. In recessive disorders, the condition does not occur unless an individual inherits the same defective gene for the same trait from each parent. If an individual receives one normal gene and one gene for the disease, the person will be a carrier for the disease, but usually will not show symptoms. (Again, in the case of Three M syndrome, some carriers may exhibit some mild symptoms associated with the disorder.) The risk of transmitting the disease to the children of a couple, both of whom are carriers for a recessive disorder, is 25 percent. Fifty percent of their children risk being carriers of the disease, but usually will not show symptoms of the disorder. Twenty-five percent of their children may receive both normal genes, one from each parent, and will be genetically normal (for that particular trait). The risk is the same for each pregnancy. Individuals who carry a single copy of the defective gene for Three M syndrome (heterozygotes)

[1] "3-M Syndrome – GeneReviews – NCBI Bookshelf", `http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=gene\&partid=1481`. Retrieved on 2008-06-28.

may exhibit some mild physical findings associated with the disorder (e.g., subtle craniofacial abnormalities and/or unusually slender bones). [2]

Symptoms

Three M syndrome is an extremely rare inherited disorder that appears to affect males and females in equal numbers. Approximately 25 cases have been reported in the medical literature since the disorder was first described in 1972. The name "Three M syndrome" refers to the last initials of three researchers (J.D. Miller, V.A. McKusick, P. Malvaux) who were among the first to identify the disorder and report their findings in the medical literature. Many of the symptoms and physical features associated with the disorder are apparent at birth (congenital). In some cases, individuals who carry a single copy of the disease gene (heterozygotes) may exhibit mild symptoms associated with Three M syndrome. [2]

Treatment

In most cases, Three M syndrome is diagnosed shortly after birth, based upon a thorough clinical evaluation, identification of characteristic physical findings (e.g., low birth weight, short stature, characteristic craniofacial and skeletal malformations, etc.), and/or a variety of specialized tests, such as advanced imaging techniques. Specialized x-ray studies may detect, confirm, and/or characterize certain craniofacial malformations (e.g., dolicocephaly, maxillary hypoplasia) as well as other skeletal abnormalities often associated with the disorder such as distinctive malformations of the vertebrae, the long bones, the ribs, and/or the shoulder blades. Identification of the gene that causes Three M syndrome may eventually lead to molecular genetic testing to confirm a suspected diagnosis. The treatment of Three M syndrome is directed toward the specific symptoms that are apparent in each individual. Treatment may require the coordinated efforts of a team of specialists. Pediatricians, physicians who specialize in treating skeletal disorders (orthopedists), dental specialists, and/or

[2] "CIGNA - Three M Syndrome", `http://www.cigna.ca/healthinfo/nord150.html`. Retrieved on 2008-06-28.

other health care professionals may need to systematically and comprehensively plan an affected child's treatment. In some cases, orthopedic techniques, surgery, and/or other supportive techniques may be used to help treat certain skeletal abnormalities associated with Three M syndrome. Surgery and/or supportive measures may also be used to help treat or correct certain craniofacial, digital, and/or other abnormalities associated with the disorder. In addition, in affected individuals with dental abnormalities, braces, oral surgery, and/or other corrective techniques may be used to help treat or correct such malformations. Genetic counseling will be of benefit for affected individuals and their families. Family members of affected individuals should also receive regular clinical evaluations to detect any symptoms and physical characteristics that may be potentially associated with Three M syndrome or heterozygosity for the disorder. Other treatment for Three M syndrome is symptomatic and supportive. [2]

A.2 Ablepharon macrostomia syndrome

Diagnosis

Ablepharon-Macrostomia Syndrome may be diagnosed at birth based upon a thorough clinical evaluation, identification of characteristic physical findings, and/or specialized imaging techniques. For example, in some cases, computerized tomography (CT) scanning may be helpful in demonstrating absence of the zygomatic arch, improper union of portions of the upper and lower jawbones (maxillary and mandibular prominences), etc. During CT scanning, a computer and x-rays are used to create a film showing cross-sectional images of tissue structure. Thorough examination and specialized testing may be conducted by eye specialists (ophthalmologists) to appropriately characterize eyelid malformations (ablepharon or microblepharon), detect any additional or associated eye abnormalities, and ensure appropriate preventive steps and/or prompt treatment. The treatment of Ablepharon-Macrostomia Syndrome is directed toward the specific symptoms that are apparent in each individual. Treatment may require the coordinated efforts of a team of specialists who work together to systematically and comprehensively plan an affected child's treatment. [3]

Causes

The term syndrome derives from the Greek and means literally "run together," as the features do. The term syndrome is most often used when the reason that the features occur together (pathophysiology) has not yet been discovered. A familiar syndrome name often continues to be used even after an underlying cause has been found. Many syndromes are named after the physicians credited with first reporting the association; these are "eponymous" syndromes. [4]

[3] "CIGNA - Ablepharon Macrostomia Syndrome", `http://www.cigna.com/healthinfo/nord1093.html`. Retrieved on 2008-06-28.

[4] "Syndrome", `http://www.economicexpert.com/a/Syndrome.htm`. Retrieved on 2008-06-28.

Symptoms

In infants with Ablepharon-Macrostomia Syndrome, characteristic craniofacial features may include absence or severe underdevelopment of the upper and lower eyelids (ablepharon or microblepharon) as well as absence of eyelashes and eyebrows; an unusually wide, “fish-like” mouth (macrostomia); and/or incompletely developed (rudimentary), low-set ears (pinnae). Abnormalities of the eyes may occur due to, or in association with, ablepharon or microblepharon. Individuals with AMS may also have additional characteristic features including abnormally sparse, thin hair; thin, wrinkled skin with excess (redundant) folds; webbed fingers with limited extension; and/or malformations of the external genitals. In some cases, additional features associated with AMS may include absent or abnormally small (hypoplastic) nipples and/or abdominal wall abnormalities. Although the exact cause of Ablepharon-Macrostomia Syndrome is not fully understood, some cases suggest that the disorder may be inherited as an autosomal recessive genetic trait. [3]

Treatment

Ablepharon-Macrostomia Syndrome may be diagnosed at birth based upon a thorough clinical evaluation, identification of characteristic physical findings, and/or specialized imaging techniques. Such specialists may include pediatricians; ophthalmologists; specialists who diagnose and treat disorders of the skin (dermatologists), the male and female urinary tracts and the male genital tract (urologists), and the gastrointestinal tract (gastroenterologists); plastic and/or reconstructive surgeons; physical and occupational therapists; and/or other health care professionals. Specific therapies for the treatment of AMS are symptomatic and supportive. In some cases, plastic and reconstructive surgery may possibly be performed to correct certain malformations such as abnormalities of the eyelids, mouth, and/or ears. In some cases, surgery may also be performed to correct other eye abnormalities, malformations of the fingers, certain skin abnormalities, malformations of external genitalia, and/or ventral hernias. Other treatment is symptomatic and supportive. Genetic counseling will be of benefit for affected individuals and their families. [3]

A.3 Bacillary angiomatosis

Diagnosis

The presence of neutrophils adjacent to the blood vessels is noteworthy and may be an important clue to this diagnosis. Granular material resembling fibrin may be beside the neutrophils. This is the bacterium, observed best with either Warthin-Strry silver or Grocott-silver methenamine stain. A similar histologic pattern may be evident in affected oral mucosa, lymph nodes, liver, spleen, bone marrow, larynx, gastrointestinal tract, peritoneum, diaphragm, and bronchial mucosa. B henselae and B quintana, the etiologic agents of BA, may stain positively with a specific antiserum against the cat scratch bacillus; however, BA is avascular proliferation, not a formation of stellate abscesses without granuloma formation as is cat scratch disease. In addition, patients with cat scratch disease do not respond to antibiotics, as do patients with BA. The organisms causing BA resemble the agent of verruga peruana and Oroya fever (bartonellosis), Bartonella bacilliformis, in producing a histologically similar vascular proliferation, in having a gram-negative wall structure observed by electron microscopy, and in tending to grow in clumps visible by light microscopy. Bartonellosis is transmitted by an insect vector (a Peruvian sandfly) present only in a mountainous region of Peru near the city of Oroya and is first evident within erythrocytes, producing its febrile manifestation (Oroya fever). [5]

Causes

The related bacteria B. henselae was first identified several years ago as the cause of cat-scratch fever. It also can lead to bacillary angiomatosis in AIDS patients. Bacillary angiomatosis caused by this bacteria is transmitted to AIDS patients from cat fleas. [6]

[5] Robert A. Schwartz. "Bacillary Angiomatosis", http://www.emedicine.com/DERM/topic44.htm. Retrieved on 2008-06-28.

[6] "Bacillary Angiomatosis", http://www.healthatoz.com/healthatoz/Atoz/common/standard/transform.jsp?requestURI=/healthatoz/Atoz/ency/bacillary_angiomatosis.jsp. Retrieved on 2008-06-28.

Symptoms

Most patients are infected with HIV and have CD4 cell counts of less than 200L. The duration of symptoms before diagnosis is usually several months.

Symptoms resulting from skin, subcutaneous, mucosal, and osseous lesions include the following: Raised red or purple lesions in the skin that bleed when traumatized; similar lesions in the oral mucosa, tongue, oropharynx, nose, penis, or anus; and bone pain, frequently in the forearms or legs.

Symptoms resulting from visceral involvement may include the following: Asymptomatic; fever, chills, malaise, night sweats, anorexia, and weight loss; abdominal pain, nausea, vomiting (peliosis hepatis); jaundice secondndary to biliary obstruction as a result of external compression of periportal lymph nodes; intra-abdominal mass and gastrointestinal bleeding; abdominal cramps, tenesmus, and bloody diarrhea (colonic bacillary angiomatosis); psychiatric symptoms, such as exacerbation of depression or new-onset psychosis; personality changes, including anxiety and irritability, headache, trigeminal neuralgia, seizures, or back pain (central nervous system bacillary angiomatosis); and difficulty in breathing secondndary to laryngeal obstruction.

Underlying disease conditions may include the following: Commonly, a history of HIV infection, organ transplantation, leukemia, or chemotherapy; bacillary angiomatosis developing prior to HIV seroconversion in some patients; or apparent immunocompetence in some patients.

Bacillary angiomatosis was reported in a patient who was HIV-seronegative but had idiopathic thrombocytopenic purpura, had undergone splenectomy, and had been administered long-term systemic prednisone. Another recent report described an immunocompetent child with infected facial wound, in the vicinity of which bacillary angiomatosis lesions had developed Similar lesions also appeared at the donor site of the skin graft, which was grafted on the facial wound. Multiple leg ulcers caused by bacillary angiomatosis without a history of direct contact with cats in an adult immunocompetent man has also been reported. [7]

[7] KoKo Aung. "Bacillary Angiomatosis", `http://www.emedicine.com/DERM/topic196.htm`.

Treatment

BA often responds to therapy with oral erythromycin, although other oral antibiotics and antituberculosis medications, including tetracycline, trimethoprim-sulfamethoxazole, and rifampin, may also be effective. While BA is treatable and curable, it may be life threatening if untreated. [8]

Retrieved on 2008-06-28.

[8] Robert A. Schwartz. "Bacillary Angiomatosis", `http://www.emedicine.com/derm/byname/Bacillary-Angiomatosis.htm`. Retrieved on 2008-06-28.

A.4 Brittle Nails

Diagnosis

The dermatologist needs to make the diagnosis of brittle nails based on the patient's history and certain non-pathognomonic clinical features. The presenting complaints of patients with brittle nails are often their inability to grow long nails and a description of their nails as soft, dry, weak, or easily breakable. More objective clinical features seen in brittle nails are onychoschizia (transverse splitting), onychorrhexis (longitudinal splitting), and nail plate surface degranulation. Brittle nails have been divided into several types including an isolated split at the free edge, lamellar splitting of the free edge, transverse splitting of the lateral edge, and multiple crenellated splitting that resembles the battlements of a castle. [9]

Causes

Crushing the base of the nail or the nail bed may cause a permanent deformity. Chronic picking or rubbing of the skin behind the nail can cause a washboard nail. Long-term use exposure to moisture or nail polish can cause nails to peel and become brittle. Fungus or yeast cause changes in the color, texture, and shape of the nails. Bacterial infection may cause a change in nail color or painful areas of infection under the nail or in the surrounding skin. Severe infections may cause nail loss. Viral warts may cause a change in the shape of the nail or ingrown skin under the nail. Certain infections (especially of the heart valve) may cause splinter hemorrhages (red streaks in the nail bed). Liver disease can damage nails. Thyroid diseases including hyperthyroidism or hypothyroidism may cause brittle nails or splitting of the nail bed from the nail plate (onycholysis). Severe illness or surgery may cause horizontal depressions in the nails (Beau's lines). [10]

[9] Hendrik Uyttendaele, Adam Geyer, Richard K. Scher. "Brittle nails: pathogenesis and treatment", `http://findarticles.com/p/articles/mi_m0PDG/is_1_2/ai_110152641`. Retrieved on 2008-06-28.

[10] Kevin Berman. "Nail Abnormalities", `http://www.nlm.nih.gov/MEDLINEPLUS/ency/article/003247.htm`. Retrieved on 2008-06-28.

Symptoms

An easy way to treat brittle nails it to apply lotion to the nail. There are certain nail solutions and lotions which will aid in preventing the nails from becoming brittle. Another treatment option for brittle nails is a vitamin known as Biotin. Be sure to read the label on the vitamin prior to taking it to ensure that doing so would not be detrimental to one's health. For example, pregnant women and women who are nursing should avoid Biotin. [11]

Treatment

The treatment of brittle nails is often difficult. If no precipitating or contributing factors can be elucidated and if the brittle nails have been present for many years, available treatments are often ineffective. However, the initial approach to the treatment of brittle nails should focus on the removal of any exogenous factors that may cause or exacerbate nail fragility. Patients should be instructed not to wash hands frequently and to avoid contact with water or other dehydrating chemicals. Rehydration of the nail plate, cuticle and surrounding nail fold can be obtained by soaking the nails in lukewarm water followed by application of an effective moisturizer. Alpha-hydroxy acid containing moisturizer and preparations that contain hydrophilic substances such as phospholipids have been successfully used for this. Occasionally, the once a week use of nail enamel is encouraged to slow water evaporation from the nail plate. It is also recommended that the patients keep their nails short, and clip them after soaking them in lukewarm water. [9]

[11] "Brittle nails: cause, treatments, how to prevent, symptoms", `http://www.mamashealth.com/nails/britnails.asp`. Retrieved on 2008-06-28.

A.5 Bronchopulmonary dysplasia

Diagnosis

Chest X-rays may be helpful in making the diagnosis. In babies with respiratory distress syndrome (RDS), the X-rays may show lungs that look like ground glass. In babies with BPD, the X-rays may show lungs that appear spongy. [12]

Causes

Development of BPD is not limited to RDS survivors. Any newborn infant who has serious respiratory problems in its first few days after birth is at risk of developing BPD. Although BPD is most common in premature babies, it can occur in full-term infants who need mechanical ventilation and oxygen under pressure for problems such as neonatal pulmonary hypertension. [13]

Symptoms

Bronchopulmonary dysplasia occurs most often in premature newborns who had severe lung disease at birth, such as respiratory distress syndrome, particularly in those who needed treatment with a ventilator for more than a few weeks after birth. The delicate tissues of the lungs can become injured when the air sacs are over-stretched by the ventilation or by high oxygen levels. As a result, the lungs become inflamed and additional fluid accumulates within the lungs. Full-term newborns who have lung disease (such as pneumonia) occasionally develop bronchopulmonary dysplasia. [14]

Treatment

No treatment is specific for bronchopulmonary dysplasia (BPD); rather, treatment is supportive. Doctors provide treatment to help the baby breathe better. They

[12] "Bronchopulmonary Dysplasia", `http://kidshealth.org/parent/medical/lungs/bpd.html`. Retrieved on 2008-06-28.

[13] "Bronchopulmonary Dysplasia (BPD)", `http://www.legalpointer.com/displaymonograph.php?MID=159`. Retrieved on 2008-06-28.

[14] "Bronchopulmonary Dysplasia", `http://respiratory-lung.health-cares.net/bronchopulmonary-dysplasia.php`. Retrieved on 2008-06-28.

make sure the baby has enough oxygen; is properly fed, kept warm, and treated for infections; and is given the right amount of fluids and nourishment. This gives the baby's lungs time to mature. Babies with BPD are usually treated in a special intensive care unit (ICU) for newborns (neonatal ICU). Some babies may be treated in the neonatal ICU even before doctors know that the babies have BPD, because they may have been born premature, needed treatment for respiratory distress syndrome (RDS), or had other problems. Doctors can tell if a baby has BPD only after the baby is several weeks old. [15]

[15] "Bronchopulmonary Dysplasia (BPD)", http://www.daviddarling.info/encyclopedia/B/bronchopulmonary_dysplasia.html. Retrieved on 2008-06-28.

A.6 Chorea acanthocytosis

Diagnosis

If there are other affected siblings this also suggests the diagnosis. Not all persons show the presence of Acanthocytosis. However, the majority of persons with Chorea-acanthocytosis will show an increased presence of muscle creatine phosphokinase. The VPS13A gene is the only gene known to be associated with the condition. As molecular genetic testing for the gene is only available on a research basis, availability of the test needs to be kept under review. [16]

Causes

Neuroacanthocytosis is a very rare disorder usually inherited as an autosomal recessive genetic trait. However, the medical literature documents one family that may have inherited the disorder as an autosomal dominant genetic trait. Human traits, including the classic genetic diseases, are the product of the interaction of two genes, one received from the father and one from the mother. In dominant disorders, a single copy of the disease gene (received from either the mother or father) will be expressed "dominating" the other normal gene and resulting in the appearance of the disease. The risk of transmitting the disorder from affected parent to offspring is 50 percent for each pregnancy regardless of the sex of the resulting child In recessive disorders, the condition does not appear unless a person inherits the same defective gene for the same trait from each parent. If an individual receives one normal gene and one gene for the disease, the person will be a carrier for the disease, but usually will not show symptoms. The risk of transmitting the disease to the children of a couple, both of whom are carriers for a recessive disorder, is 25 percent. Fifty percent of their children risk being carriers of the disease, but generally will not show symptoms of the disorder. Twenty-five percent of their children may receive both normal genes, one from each parent, and will be genetically normal (for that particular trait). The

[16] "Neuroacanthocytosis disorders", `http://www.cafamily.org.uk/Direct/n238.html`. Retrieved on 2008-06-28.

risk is the same for each pregnancy. [17]

Symptoms

Mean age of onset in choreoacanthocytosis (ChAc) is about age 35 years, although ChAc can develop as early as the first decade or as late as the seventh decade. It runs a chronic progressive course and may lead to major disability within a few years. Some affected individuals are bedridden or wheelchair dependent by the third decade. Life expectancy is reduced and several instances of sudden, unexplained death, or during epileptic seizures have been reported. Age at death ranges from age 28 to 61 years. [18]

Treatment

Phenytoin has been considered the drug of choice and generally is used in dosages similar to those used in epilepsy; however, it also has been shown to be effective in smaller doses. Satisfactory response has been obtained with other anticonvulsants, particularly carbamazepine; acetazolamide also may be used In one case, haloperidol worsened paroxysmal kinesiogenic dyskinesia. [19]

[17] "Neuroacanthocytosis", `http://www.cigna.com/healthinfo/nord975.html`. Retrieved on 2008-06-28.

[18] "Choreoacanthocytosis", `http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=gene\&partid=1387`. Retrieved on 2008-06-28.

[19] Ismail Mohamed. "Chorea in Children", `http://www.emedicine.com/neuro/TOPIC644.HTM`. Retrieved on 2008-06-28.

A.7 Heterotopic ossification

Diagnosis

X-Ray: During the early stage, an x-ray will not be helpful because there is no calcium in the matrix. (In an acute episode which is not treated, it will be 3- 4 weeks after onset before the x-ray is positive.) Laboratory Tests: Also, not very helpful. Alkaline phosphatase will be elevated at some time, but in patients who have had fractures or spine fusion recently, this is not diagnostic The values will often be quite high but unless weekly tests are done this peak value may not be detected Initially the value may be only slightly elevated Bone Scan: The only definitive diagnostic test in the early acute stage is a bone scan. When the initial symptoms are an acute inflammatory process with swelling and increased temperature, the differential diagnosis is thrombophlebitis. It may be necessary to do a bone scan and avenogram to differentiate which is present, and it is even possible that both could be present simultaneously. Clinical Exam: The swelling tends to be more proximal with little or no foot/ankle edema; whereas, in thrombophlebitis the swelling is more uniform throughout the leg. [20]

Causes

Study on heterotopic ossification has suggested that it may be linked to injuries to the spinal cord, along with neurological conditions. It appears that mixed signals in the body stimulate normally dormant osteoprogenitor cells, causing them to start growing bone. When these cells are in the soft tissues of the body, it results in heterotopic ossification. The condition often appears in the form of periarticular ossification, especially around the site of hip injuries. [21]

[20] "Heterotopic Ossification", http://www.spinalcord.ar.gov/Publications/FactSheets/sheets1-5/fact1.html. Retrieved on 2008-06-28.

[21] "What is Heterotopic Ossification?", http://www.wisegeek.com/what-is-heterotopic-ossification.htm. Retrieved on 2008-06-28.

Symptoms

In addition, the bone scan will show heterotopic ossification seven to ten days earlier than an x-ray. The three-phase bone scan is perhaps the earliest method of detecting heterotopic bone formation. However, in some cases, an abnormality may be detected in the early phase which does not necessarily mean it will go on to form heterotopic bone. Another finding, often misinterpreted as early heterotopic bone formation, is an increased (early) uptake around the knees or the ankles in an early spinal cord injured patient. It is not clear exactly what this means because these patients do not develop heterotopic bone formation. It has been hypothesized that this may be related to the autonomic nervous system and its control over circulation. [22]

Treatment

The effect of the Didronel is to prevent calcium from being deposited in the bony matrix that has already been formed Therefore, it is essential to make the diagnosis as soon as possible (preferably before any calcium shows up on x-ray) and start the Didronel immediately. Didronel will do nothing to remove calcium that has already been deposited. It is a preventative drug, and has no effect on existing ossification. It also has no effect on the underlying process which produces the bony matrix. There are no known side effects that would prohibit usage. Many physicians recommend prophylactic use of Didronel in all acute spinal cord injuries, but because of the cost this may not be practical. Some patients complain of nausea the first week, but this is rarely severe enough to stop treatment and usually subsides in a few days. There is no uniform agreement on how long the Didronel should be continued In most cases, there will be a brief flare-up of the heterotopic ossification following discontinuing the Didronel and some increase in the amount of calcium deposited There are no completely reliable tests to indicate that the heterotopic ossification is inactive and treatment can be safely stopped However, if the treatment was continued long enough this calcium deposition will be of minimal clinical significance. The patient needs to

[22] "Heterotopic Ossification - SCI InfoSheet #12", http://www.spinalcord.uab.edu/show.asp?durki=21485. Retrieved on 2008-06-28.

be observed closely for signs of recurrence whenever treatment is discontinued. [20]

A.8 Hypophosphatasia

Diagnosis

In general, the earlier the diagnosis is made the more severe the skeletal manifestations. Cases detected in the womb or with severe deformity at birth almost always have a lethal outcome within days or weeks. When the diagnosis is made before six months of age, some infants have a downhill and fatal course, others survive and may even do well. When diagnosed during childhood, there can by presence or absence of skeletal deformity from underlying rickets, but premature loss of teeth (less than five years of age) is the most common manifestation. Adults may be troubled by recurrent fractures in their feet and painful, partial fractures in their thigh bones. [23]

Causes

The symptoms of Hypophosphatasia develop due to abnormally low levels of the enzyme alkaline phosphatase (ALP). The infantile and childhood forms of the disorder are inherited as autosomal recessive genetic traits. The adult form is inherited as an autosomal dominant genetic trait. In recessive disorders, the condition does not appear unless a person inherits the same defective gene for the same trait from each parent. If an individual receives one normal gene and one gene for the disease, the person will be a carrier for the disease, but usually will not show symptoms. The risk of transmitting the disease to the children of a couple, both of whom are carriers for a recessive disorder, is 25 percent. Fifty percent of their children risk being carriers of the disease, but generally will not show symptoms of the disorder. Twenty-five percent of their children may receive both normal genes, one from each parent, and will be genetically normal (for that particular trait). The risk is the same for each pregnancy. In dominant disorders, a single copy of the disease gene (received from either the mother or father) will be expressed "dominating" the other normal gene and resulting in the appearance of the disease. The risk of transmitting the disorder

[23] "Hypophosphatasia", http://www.magicfoundation.org/www/docs/175/. Retrieved on 2008-06-28.

from affected parent to offspring is 50 percent for each pregnancy regardless of the sex of the resulting child. [24]

Symptoms

The signs and symptoms of hypophosphatasia vary widely and can appear anywhere from before birth to adulthood The most severe forms of the disorder tend to occur before birth and in early infancy. Hypophosphatasia weakens and softens the bones, causing skeletal abnormalities similar to another childhood bone disorder called rickets. Affected infants are born with short limbs, an abnormally shaped chest, and soft skull bones. Additional complications in infancy include poor feeding and a failure to gain weight, respiratory problems, and high levels of calcium in the blood (hypercalcemia), which can lead to recurrent vomiting and kidney problems. These complications are life-threatening in some cases. [25]

Treatment

Genetic counseling is important for all families who have affected children. A pedigree is essential, especially for the childhood, adult, or odontohypophosphatasic forms, which can have either autosomal dominant or recessive forms. Options for future pregnancies, such as prenatal testing for the perinatal form, should be discussed with parents. [26]

[24] "Hypophosphatasia", `http://www.cigna.com/healthinfo/nord518.html`. Retrieved on 2008-06-28.

[25] "Hypophosphatasia", `http://ghr.nlm.nih.gov/condition=hypophosphatasia`. Retrieved on 2008-06-28.

[26] Horacio Plotkin. "Hypophosphatasia", `http://www.emedicine.com/ped/topic1126.htm`. Retrieved on 2008-06-28.

A.9 Macular Hole

Diagnosis

Macular degeneration is a condition affecting the tissues lying under the retina, while a macular hole involves damage from within the eye, at the junction between the vitreous and the retina itself. There is no relationship between the two diseases. Depending upon the degree of attachment or traction between the vitreous and the retina, there may be risk of developing a macular hole in the other eye. Your eye care provider can determine the status of the vitreous jell and its degree of traction on the retinal surface in the uninvolved eye. In those cases where the vitreous has already become separated from the retinal surface, there is very little chance of developing a macular hole in the other eye. On the other hand, when the vitreous remains adherent and pulling on the macular region in both eyes, then there may be a greater risk of developing a hole in the second eye. In very rare instances, trauma or other conditions lead to the development of a macular hole. In the vast majority of cases, however, macular holes develop spontaneously. As a result, there is no known way to prevent their development through any nutritional or chemical means, nor is there any way to know who is at risk for developing a hole prior to its appearance in one or both eyes. [27]

Causes

The eye contains a jelly-like substance called the vitreous. Shrinking of the vitreous usually causes the hole. As a person ages, the vitreous becomes thicker and stringier and begins to pull away from the retina. If the vitreous is firmly attached to the retina when it pulls away, a hole can result. [28]

[27] "Macular Hole", http://www.avclinic.com/macular_hole.htm. Retrieved on 2008-06-28.

[28] "Eye Disorders – Macular Hole", http://www.columbiaeye.org/coc/Eye_Dissorders_5.asp. Retrieved on 2008-06-28.

Symptoms

However, if the vitreous is firmly attached to the retina when it pulls away, it can tear the retina and create a macular hole. Also, once the vitreous has pulled away from the surface of the retina, some of the fibers can remain on the retinal surface and can contract. This increases tension on the retina and can lead to a macular hole. In either case, the fluid that has replaced the shrunken vitreous can then seep through the hole onto the macula , blurring and distorting central vision. [29]

Treatment

One of the newest strategies that holds promise is the use of new drugs that stop blood vessels in wet AMD and can cause existing blood vessels to regress. This new class of drugs is known as anti-angiogenic agents. At Bascom Palmer Eye Institute, we are currently investigating six such drugs. Three drugs are being investigated for the treatment of AMD and three drugs are being investigated for the treatment of diabetes. [30]

[29] "Facts About Macular Hole", `http://www.nei.nih.gov/health/macularhole/index.asp`. Retrieved on 2008-06-28.

[30] "Macular Disease", `http://www.bpei.med.miami.edu/site/disease/disease_macular.asp`. Retrieved on 2008-06-28.

A.10 Vestibular neuronitis

Diagnosis

In large part, the process involves ascertaining that the entire situation can be explained by a lesion in one or the other vestibular nerve. It is not possible on clinical examination to be absolutely certain that the picture of "vestibular neuritis" is not actually caused by a brainstem or cerebellar stroke, so mistakes are possible. Nevertheless, this happens so rarely that it is not necessary to perform MRI scans or the like very often. Signs of vestibular neuritis include spontaneous nystagmus and unsteadiness. One may notice that vision is disturbed or jumpy on looking to a particular side. This usually means that the opposite ear is affected – it is called "Alexander's Law" and is due to asymmetric gaze evoked nystagmus . Occasionally other ocular disturbances will occur such as vertical double vision – skew deviation. However if symptoms persist beyond one month, reoccur periodically, or evolve with time, testing may be proposed. In this situation, nearly all patients will be asked to undergo an audiogram and an ENG. An audiogram is a hearing test needed to distinguish between vestibular neuritis and other possible diagnoses such as Meniere's disease and Migraine. The ENG test is essential to document the characteristic reduced responses to motion of one ear. [31]

Causes

Vestibular Neuronitis is felt to be caused by a viral infection of the balance nerve that runs from the inner ear to the brain. We do not know which virus in particular causes this problem, and in fact, many different viruses may be capable of infecting the balance nerve. Some patients will report having an upper respiratory infection (common cold) or a flu prior to the onset of the symptoms of vestibular neuronitis, others will have no viral symptoms prior to the vertigo attack. [32]

[31] "Vestibular neuritis and labyrinthitis", http://www.dizziness-and-balance.com/disorders/unilat/vneurit.html. Retrieved on 2008-06-28.

[32] "Vestibular Neuronitis and Migrainous Vertigo", http://www.uphs.upenn.edu/pennorl/bal_ves.htm. Retrieved on 2008-06-28.

Symptoms

The main symptom of vestibular neuronitis is vertigo, which appears suddenly, often with nausea and vomiting. Vertigo usually lasts for several days or weeks. In rare cases it can take months to go away entirely. Vestibular neuronitis does not lead to loss of hearing. [33]

Treatment

Viral infection of the vestibular nerve and/or labyrinth is believed to be the most common cause of vestibular neuronitis. Acute localized ischemia of these structures also may be an important cause. Especially in children, vestibular neuritis may be preceded by symptoms of a common cold. However, the causative mechanism remains uncertain. [34]

[33] "Vestibular Neuronitis", `http://www.cigna.com/healthinfo/aa75303.html`. Retrieved on 2008-06-28.

[34] Keith A. Marill. "Vestibular Neuronitis", `http://www.emedicine.com/emerg/TOPIC637.HTM`. Retrieved on 2008-06-28.

Bibliography

[1] Eugene Agichtein, Steve Lawrence, and Luis Gravano. Learning search engine specific query transformations for question answering. In *Proceedings of the WWW*, pages 169–178, 2001.

[2] Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain, 1997.

[3] Regina Barzilay and Mirella Lapata. Collective content selection for concept-to-text generation. In *In Proceeding of the EMNLP*, pages 331–338, 2005.

[4] Regina Barzilay and Kathleen R. McKeown. Sentence fusion for multidocument news summarization. *Comput. Linguist.*, 31(3):297–328, 2005.

[5] Fadi Biadsy, Julia Hirschberg, and Elena Filatova. An unsupervised approach to biography production using wikipedia. In *Proceedings of ACL/HLT*, pages 807–815, 2008.

[6] Sasha Blair-Goldensohn, Kathleen R. McKeown, and Andrew Hazen Schlaikjer. Answering definitional questions: A hybrid approach. In Mark Maybury, editor, *New Directions In Question Answering*. AAAI Press, 2004.

[7] Eric Brill, Jimmy Lin, Michele Banko, Susan T. Dumais, and Andrew Y. Ng. Data-intensive question answering. In *Text REtrieval Conference*, 2001.

[8] James Clarke and Mirella Lapata. Modelling compression with discourse constraints. In *Proceedings of the 2007 Joint Conference on Empirical Methods*

in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 1–11, 2007.

[9] James Clarke and Mirella Lapata. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:399–429, 2008.

[10] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Intoduction to Algorithms*. The MIT Press, 1992.

[11] Mark Dras. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. PhD, Macquarie University, Australia, 1999.

[12] Pablo A. Duboue and Kathleen R. McKeown. Statistical acquisition of content selection rules for natural language generation. In *Proceedings of the EMNLP*, pages 121–128, 2003.

[13] Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. Web question answering: is more always better? In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 291–298, New York, NY, USA, 2002. ACM.

[14] Julian M. Kupiec, Jan Pedersen, and Francine Chen. A trainable document summarizer. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73, Seattle, Washington, 1995.

[15] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, 2004.

[16] Inderjeet Mani and Mark T. Maybury. *Advances in Automatic Text Summarization*. The MIT Press, 1999.

[17] Tomasz Marciniak and Michael Strube. Beyond the pipeline: Discrete optimization in NLP. In *Proceedings of the Annual Conference on Computational Natural Language Learning*, pages 136–143, Ann Arbor, MI, 2005.

[18] Kathleen McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. Towards multidocument summarization by reformulation: Progress and prospects. In *AAAI/IAAI*, pages 453–460, 1999.

[19] Kathleen McKeown and Dragomir R. Radev. Generating summaries of multiple news articles. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82, New York, NY, USA, 1995. ACM.

[20] Kathleen R. McKeown. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text.* Cambridge University Press, 1985.

[21] Dan I. Moldovan, Marius Pasca, Sanda M. Harabagiu, and Mihai Surdeanu. Performance issues and error analysis in an open-domain question answering system. In *ACL*, pages 33–40, 2002.

[22] D. Molla and S. Wan. Macquarie university at duc 2006: Question answering for summarisation. In *Proceedings of DUC*, 2006.

[23] Manabu Okumura, Takahiro Fukusima, and Hidetsugu Nanba. Text summarization challenge 2: text summarization evaluation at ntcir workshop 3. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop*, pages 49–56, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[24] M. F. Porter. An algorithm for suffix stripping. In *Readings in information retrieval*, pages 313–316, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[25] Vasin Punyakanok, Dan Roth, Wen tau Yih, and Dav Zimak. Semantic role labeling via integer linear programming inference. In *COLING '04: Proceedings*

of the 20th international conference on Computational Linguistics, page 1346, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[26] Dragomir Radev, Weiguo Fan, Hong Qi, Harris Wu, and Amardeep Grewal. Probabilistic question answering on the web: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 56(6):571–583, 2005.

[27] Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *NAACL-ANLP 2000 Workshop on Automatic summarization*, pages 21–30, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[28] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems.* Cambridge University Press, Cambridge, 2000.

[29] Dan Roth. Learning to resolve natural language ambiguities: a unified approach. In *Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence*, pages 806–813, Madison, US, 1998. AAAI Press, Menlo Park, US.

[30] Somayajulu G. Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. A two-stage model for content determination. In *Proceedings of the ACL-ENLG*, pages 3–10, 2001.

[31] Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445, 2002.