Discourse Models for Collaboratively Edited Corpora

by

Erdong Chen

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

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2008

© Massachusetts Institute of Technology 2008. All rights reserved.

Author .....
Department of Electrical Engineering and Computer Science
May 23, 2008

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

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUL 0 1 2008
LIBRARIES

ARCHIVES

Discourse Models for Collaboratively Edited Corpora

by

Erdong Chen

Submitted to the Department of Electrical Engineering and Computer Science
on May 23, 2008, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

This thesis focuses on computational discourse models for collaboratively edited corpora. Due to the exponential growth rate and significant stylistic and content variations of collaboratively edited corpora, models based on professionally edited texts are incapable of processing the new data effectively.

For these methods to succeed, one challenge is to preserve the local coherence as well as global consistence. We explore two corpus-based methods for processing collaboratively edited corpora, which effectively model and optimize the consistence of user generated text. The first method addresses the task of inserting new information into existing texts. In particular, we wish to determine the best location in a text for a given piece of new information. We present an online ranking model which exploits this hierarchical structure – representationally in its features and algorithmically in its learning procedure. When tested on a corpus of Wikipedia articles, our hierarchically informed model predicts the correct insertion paragraph more accurately than baseline methods. The second method concerns inducing a common structure across multiple articles in similar domains to aid cross document collaborative editing. A graphical model is designed to induce section topics and to learn topic clusters. Some preliminary experiments showed that the proposed method is comparable to baseline methods.

Thesis Supervisor: Regina Barzilay
Title: Associate Professor

Acknowledgments

I would like to gratefully acknowledge my supervisor, Professor Regina Barzilay, for her support and guidance. I have been greatly impressed by her enthusiasm to scientific research and comprehensive knowledge, and benefited from her patience and carefulness. Thanks to Professor Michael Collins and Professor Samuel Madden for their helpful inspiration. Some of the data used in this work was collected and processed by Christina Sauper, and I am also grateful to her generous help. Many thanks are given to to Serdar Balci, S.R.K. Branavan, Eugene Charniak, Harr Chen, Michael Collins, Pawan Deshpande, Micha Elsner, Jacob Eisenstein, Dina Katabi, Igor Malioutov, Alvin Raj, Christina Sauper, Benjamin Synder, Jenny Yuen, and Luke Zettlemoyer. Lastly, I want to give my special thanks to people in Natural Language Processing Group and Spoken Language Systems Group in MIT who offered me help for this thesis.

Bibliographic Notes

Portions of this thesis are based on the following paper(s):

"Incremental Text Structuring with Online Hierarchical Ranking", by Erdong Chen, Benjamin Snyder, and Regina Barzilay, the paper appeared in *Joint Meeting of Conference on Empirical Methods in Natural Language Processing and Conference on Computational Natural Language Learning(EMNLP-CoNLL)* 2007.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

> *Barack Obama is a Democratic politician from Illinois. He is currently running for the United States Senate, which would be the highest elected office he has held thus far.*
>
> ***Biography***
>
> *Obama's father is Kenyan; his mother is from Kansas. He himself was born in Hawaii, where his mother and father met at the University of Hawaii. Obama's father left his family early on, and Obama was raised in Hawaii by his mother.*
>
> – Wikipedia article about Barack Obama on March 18, 2004

The first Wikipedia[1] article about Barack Obama appeared on March 18, 2004. In June 2007, when he was running for U.S. president, this article grew to more than 400 sentences after more than five thousand revisions. The huge amount of collaborative editing efforts on Obama's Wikpedia entry is not unique, but one of more than 9,000,000 articles collaboratively contributed by more than 75,000 active contributors. For instance, an average English article on Wikipedia has 38

[1]Wikipedia is one of world's largest online multilingual encyclopedia, which is also collaboratively maintained by millions of volunteers.

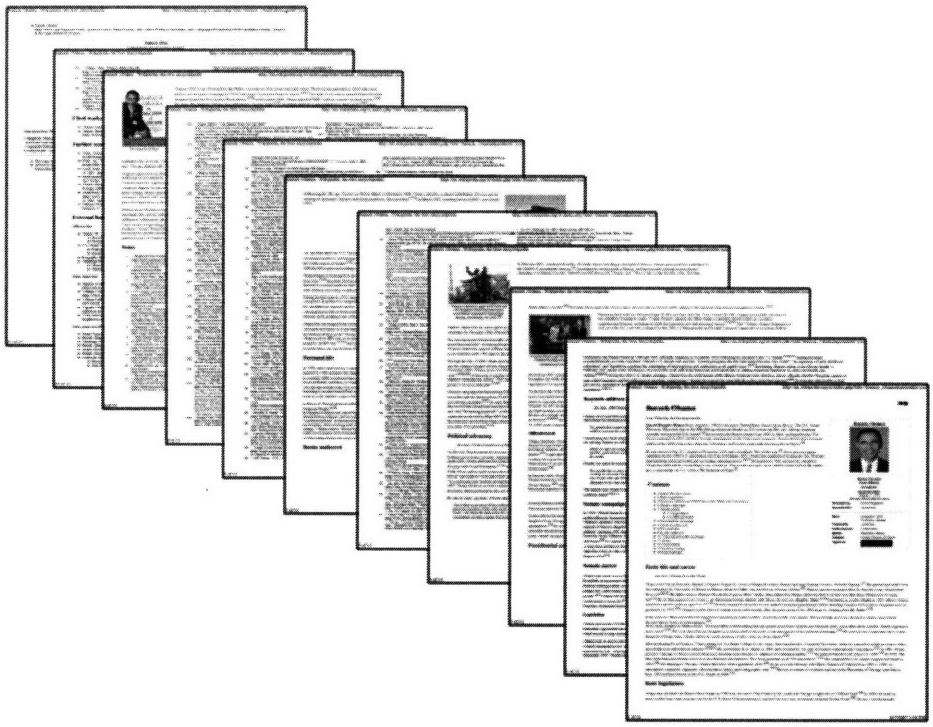


Figure 1-1: A Wikipedia article about Barack Obama in June 2007.

edits, and the English language version of Wikipedia averaged over 3 million edits[2] per month in 2006.

Moreover, in today's Web 2.0 age, collaborative editing has become an integral part of our daily life ; merely a few years ago, to finish a simple joint project would likely require many emails among co-editors , whereas today, they can easily finish an article together by using Wiki-like service(e.g. , MediaWiki[3] and Google Documents[4]). However, many emerging applications for collaborative editing require documents to be repeatedly updated. Such documents include newsfeeds, webpages, and shared community resources such as Wikipedia. Obviously, some tools that aid collaborative updating or automatically perform it could drastically decrease maintenance efforts and improve document quality.

[2] http://stats.wikimedia.org/EN/TablesWikipediaEN.htm
[3] http://wikimedia.org/
[4] http://docs.google.com/

In this thesis, we focus on statistical methods for text generation. These methods are particularly effective for processing collaboratively edited corpora. These methods are designed to aid collaborative updating or perform it automatically. Due to the variations and styles within the collaboratively edited corpora and the complexity of editing tasks, designing this new set of natural language processing algorithms requires some improvements over the existent text generation solutions:

- **Move from batch-mode generation to incremental generation.** Human editors rarely rewrite the whole document when new facts about stories are updated. Viewing text structuring as a process unfolding over time grants us the power to preserve the continuity and coherence of the original text. The closest relevant text structuring technique in natural language generation is the work on sentence ordering, in which a complete reordering of the text is undertaken. These methods are suboptimal for this new insertion task. Therefore, new algorithms are required that are able to take advantage of existing text structure.

- **Move beyond local continuity and coherence towards global consistence.** When an editor updates a sentence in an article from a document collection, he/she needs to ensure the local continuity and coherence within that paragraph as well as the global consistence within the article and across multiple articles. The local continuity and coherence of the original text may be maintained by examining sentences adjacent to the updating point. However, a local sentence comparison method such as this may fail to account for global document coherence (e.g. by allowing the mention of some fact in an inappropriate section or using inappropriate and different term). This problem is especially acute in the case of lengthy, real-world, collaboratively edited texts. Internally, these documents are commonly organized hierarchically into sections and paragraphs to aid reader comprehension. Furthermore, documents in a collection are always organized into categories based on some attributes (e.g. topics and authors). Our goal is to overcome local limitation by exploring methods which are capable of modeling hierarchical structure within documents and categorial organization across multiple documents.

The text generation methods discussed in this thesis are designed to address these issues, and

are particularly effective in collaboratively edited corpora. In particular, we explore two types of generation problems which address issues in single-document and cross-document collaboration respectively.

The first method is an incremental text structuring algorithm for inserting new sentences into existing, coherent texts. The main challenge is to maintain the continuity and coherence of the original text. We intend to preserve both the local and global coherence by exploring text structure. Rather than ignoring the inherent hierarchical structure of these texts, we desire to directly model such hierarchies and use them to our advantage – both representationally in our features and algorithmically in our learning procedure. By using the hierarchical representation, our method is able to analyze a text both in rough outline form and with focus on specific locations in more detail.

The second method is to induce a common structure across multiple articles in similar domains. The flexibility of editing unstructured text leads to different wordings of section titles that mean the same thing (for example musical equipment is referred to both as "Gear" and "Equipment" in different articles). One factor that contributes to the success of Wikipedia is that it does not enforce any fixed structure on its articles. The consequence of such a design means that documents are stored as an entire body of text, and indexing is only available at the document level. Anyone interested in a particular section (e.g., early life of politicians) across documents would have a hard time and essentially have to perform a sequential scan across all articles to find that section. For these reasons, we believe it is advantageous to induce a common structure of the corpus that allows readers to browse and edit documents easily when editing processes involve multiple documents. Words in both section titles and section text bodies indicates the topics presented in these sections. Equally important, a topical organization should be shared across the articles in similar domains. Therefore, both local lexical features and structural features across articles are crucial to induct a common organization of a document collection. Our method provides a solution to this problem, which is a graphical topic model which captures both lexical and structural dependencies.

In both cases, the methods are used to help users edit documents collaboratively. The incremental text structuring method is implemented and evaluated based on large scale corpora. It is ready

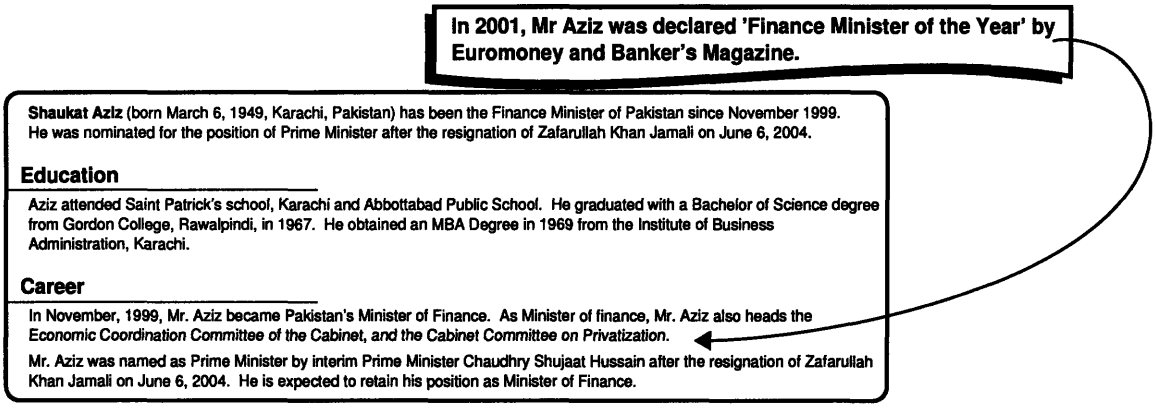


Figure 1-2: An example of Wikipedia insertion.

to be integrated into any practical text generation system. A preliminary experiment on common structure induction algorithms is conducted, while further improvement and validation is required before practical implementation.

1.2 Incremental Text Structuring

The first algorithm addresses the problem of incremental text structuring. For instance, we may want to preserve the coherence of a text after insert a new sentence into the original text. We are interested in automatically learning to perform such insertions based on the historical insertion log provided by the documents' contributors[5]. Figure 1-2 shows an example insertion on Shaukat Aziz's Wikipedia entry.

One challenge of this task is to preserve the local coherence as well as global coherence. These documents are commonly organized hierarchically into sections and paragraphs to aid comprehension. [6] Introducing a hierarchical representation into method for the task has both linguistic and practical motivations. This structure enables us to guide the ordering process with global features which have been identified as important in discourse analysis. From the practical viewpoint, this

[5]Figure 1-3 shows the screenshot of an insertion between two consecutive revisions on a Wikipedia article about Chef Tony.

[6]For documents where hierarchical information is not explicitly provided, such as automatic speech transcripts, we can use automatic segmentation methods to induce such a structure [22].

Chef Tony

From Wikipedia, the free encyclopedia

(Difference between revisions)

• *Interested in contributing to Wikipedia?* •

Jump to: navigation, search

Revision as of 07:06, 3 April 2006 (edit) Wvoutlaw2002 (Talk | contribs) ← Older edit	**Revision as of 07:09, 3 April 2006 (edit) (undo)** Wvoutlaw2002 (Talk | contribs) (→Competition) Newer edit →
Line 12:	**Line 12:**
The [[Jack Lalanne]] miracle juicer also butts heads with the Ultimate Chopper but the price gap between the two is considerable enough to target different market segments so the competition stays relatively friendly.	The [[Jack Lalanne]] miracle juicer also butts heads with the Ultimate Chopper but the price gap between the two is considerable enough to target different market segments so the competition stays relatively friendly.
	+
	+ **In addition, the Ultimate Chopper is in competition with [[Magic Bullet|The Original Magic Bullet]], which is pitched by fellow infomercial pitchman [[Mick Hastie]].**
==External link==	==External link==

Revision as of 07:09, 3 April 2006

Tony Notaro, aka "Chef Tony" is a successful infomercial pitchman. There is a dearth of knowledge available on the internet about his real name and personal life. He appears to be of Italian ancestry.

Products

He is most noted for selling his patented Chef Tony's Miracle Blades. "Over 12 million Miracle Blade knives have been sold since 1989" claims the website but relevant statistical evidence is not given that supports this claim. These are stamped-blade knives with curved handles apparently designed to allow greater clearance for the fingers beneath the handle. They are sold in sets only, with the main culinary blades being the Rock and Chop (resembles a santoku with a highly curved edge) and the Chop and Scoop (resembles a cleaver). Reports vary on the quality of the knives; common complaints are that they rust easily and dull quickly, though reviewers have praised the Rock and Chop in particular.

Tony also endorses The Ultimate Chopper with 750 watts of mega power that on TV has the power to powderize concrete and Smartware bakeware, which is made of non-stick Temperflex silicone.

Competition

There seems to be a competition between Tony and the competing Ronco blades that target the same audience but are slightly less expensive.

The Jack Lalanne miracle juicer also butts heads with the Ultimate Chopper but the price gap between the two is considerable enough to target different market segments so the competition stays relatively friendly.

In addition, the Ultimate Chopper is in competition with The Original Magic Bullet, which is pitched by fellow infomercial pitchman Mick Hastie.

External link

- His Miracle Blade Site
- His Ultimate Chopper Site

Retrieved from "http://en.wikipedia.org/wiki/Chef_Tony"

method can readily handle full-length documents. The method is particularly useful for long documents, guiding insertion process along the document hierarchy. The hierarchical representation allows us to analyze a text both in rough outline form and with focus on specific sections in more detail when necessary. Thus, our approach complements previous sentence-ordering work which has focused on short, unstructured documents where a local view of coherence suffices. As proven by the results, this is an effective method.

1.3 Common Structure Induction

Systematic terminology, which deals with all of the terms in a specific subject domain, is particularly useful for readers to comprehend documents in that domain. In addition, consistent topic organization of articles in a specific domain help readers to effectively access and navigate through large amounts of content. To address these issues, the second method discussed in this thesis is to solve a problem of inducing common structure across documents in similar domains.

To handle this problem, we have to overcome two main challenges. The first challenge is different wording to express the meaning. Table 4.1 shows 42 different Wikipedia section titles conveying similar meaning to *Early life*. Anyone interested in comparing the early life of different politicians have to perform a sequential scan across all articles to find that section. The second challenge is the variations of document organizations within a particular domain. Our approach to the task of inducing common document structure directly attack both challenges. At first, section topics can be predicted based on the words in both the section titles and the section text bodies. Secondly, equally important, a rough topical structure is shared across the articles in similar domains. Therefore, we capture both local lexical features and structural features across articles to generate a common organization.

1.4 Key Contributions

The key contributions of this thesis are three-fold: firstly, the incremental text structuring proposed in the thesis presents a new perspective on text generation. Secondly, our experiments show that

Early life, education, and family	Early years, education, military	Personal life and education
Early Life and Education	Early years	Personal life and family
Personal life and career	Childhood and Education	Early life and childhood
Childhood	Early life, education, and early career	Early years and education
Early life	Early biography	Childhood and education
Earlier life	Youth	Early Life and Family
Early years and family	Family and education	Family and early life
Family Life	Career after football	Curriculum vitae
Family and Personal Life	Upbringing	Early life and family
Early Years	Early and private life	Early career
The Early Years	Birth and education	Early and personal life
Background and early life	Education and Family	Early life and education
Family and Education	Early Life	Early Life and Family
Background and family	Personal and family life	Family and childhood

Table 1.1: 42 different Wikipedia section titles similar to *Early life*.

hierarchical representation coupled with hierarchically sensitive training improves performance. Finally, we discuss a model to induce a common structure of the corpus to aid editing processes which involve multiple documents.

1.4.1 Incremental Text Structuring

The traditional approach for text generation is batch-mode. The text is generated from scratch, and is seldom modified afterwards. However, a different view on generation, where text creation is viewed as an incremental process, is a common scenario. Newsfeeds are consistently updated by journalists when new facts about stories arrive. In addition, we propose a hierarchical learning framework for incremental text structuring. For future work, we want to combine our work with other generation modules to automatically update Wikipedia web pages.

1.4.2 Hierarchically Sensitive Training

We evaluate our method using real-world data where multiple authors have revised preexisting documents over time. We obtain such a corpus from Wikipedia articles,[7] which are continuously

[7]Data and code used in this chapter are available at `http://people.csail.mit.edu/edc/emnlp07/`

updated by multiple authors. Logs of these updates are publicly available, and are used for training and testing of our algorithm. Figure 1-1 shows an example of a Wikipedia insertion. We believe this data will more closely mirror potential applications than synthetic collections used in previous work on text structuring. We proposed a method which integrates this idea of selective hierarchical updates with the simplicity of the perceptron algorithm and the flexibility of arbitrary feature sharing inherent in the ranking framework. Unlike previous work on multiclass hierarchical classification, which rests on the assumption that a predetermined set of atomic labels with a fixed hierarchy is given, the set of possible insertion points in our task is unique to each input document.

1.4.3 Common Structure Induction

One contribution of this work is a corpus-based model to aid collaborated editing across multiple documents to provide users with more flexibility and power over the data. A graphical model is designed to induce section topics and to learn topic clusters. The process elucidated in this thesis is very general and easily applicable to other collaboratively edited corpora as well. Some preliminary experiments show that the proposed method is comparable to various baseline methods.

1.4.4 NLP on Collaboratively Edited Corpora

Our work contributes to research in natural language processing in collaboratively edited corpora. This work provides basic tools (including incremental text structuring algorithm) to process collaboratively edited corpora effectively. Our work also contributes to the ongoing research on collaboratively edited corpora by making our data and code available to public.

1.5 Scope of the Thesis

The thesis contains two interconnected parts. The first part, Chapter 3, investigates incremental text structuring, while the second part, Chapter 4, is dedicated to common structure induction.

In Chapter 2, we discuss related work in the areas of Hierarchical Learning, Text Structuring, Bayesian Topic Modeling, and Semantic Wikipedia.

In Chapter 3, we focus on sentence ordering methods in the context of collaboratively edited corpora. We present a novel corpus-based method for this task. The main contribution of this chapter is the incorporation of a rich hierarchical text representation into a flexible learning approach for text structuring. Our learning approach makes key use of the hierarchy by selecting to update only the layer found responsible for the incorrect prediction. Empirical tests on a large collection of real-world insertion data confirm the advantage of this approach. Chapter 3 is organized as follows. First, we introduce the problem and the importance of application in text generation. Next, we provide an overview of existing work on text structuring and hierarchical learning. Then, we define the insertion task and introduce our hierarchical ranking approach to sentence insertion. Next, we present our experimental framework and data. We conclude the chapter by presenting and discussing our results.

In Chapter 4, we explore methods to induce a common structure across documents in similar domains. Clustering section titles to form a consistent title set provides us with a consistent representation of documents in Wikipedia. To handle that problem, a graphical model is designed to find text alignment and to learn topic clusters. Both lexical features and structural features across articles are captured in the model. The process elucidated in this paper is very general and easily applicable to other collaboratively edited corpora as well. Chapter 4 is organized as follows. We first discuss the cross-document collaboration's negative influence on corpora and the need to have a common structure across documents. Next, we provide an overview of existing work on topic models. Then, we define our task and introduce our generative process with learning and inference methods. Finally, we present our experimental framework. To conclude Chapter 4, we present and discuss our methods.

In Chapter 5, we conclude the thesis by discussing the main findings of this thesis. Some directions for future research are discussed.

Chapter 2

Related Work

Our work focuses on text generation problems of collaboratively edited corpora. The two specific tasks we analyze require very different text generation. While the incremental text structuring algorithm addresses issues in single-document collaboration, the common structure induction method concerns cross-document cooperation. While the underlying theme is of text generation on collaboratively edited corpora, the two tasks need completely different algorithms, and derive from different streams of prior work. In the rest of this chapter, we describe the related work in Sentence Ordering, Symbolic Concept-to-text Generation, Hierarchical Learning, Bayesian Topic Modeling, and Semantic Wikipedia.

2.1 Incremental Text Structruing

The incremental text structuring problem can be viewed in terms of traditional generation architecture. Natural language generation (NLG)[39] is commonly decomposed into six subtasks: **content determination, document structuring, sentence aggregation, lexicalization , referring expression generation** , and **linguistic realization**.

- **Content determination** is the process of deciding what information should be expressed in the text. In our settings, information extraction[1] technology is able to extract summarized

[1]Information extraction is a type of information retrieval whose goal is to automatically extract structured information from semantically well-defined data, or from unstructured machine-readable documents.

sentences to update existent articles.

- **Document structuring** is the process of ordering a set of facts into a coherent text. This subtask is the focus of Chapter 3.

- **Sentence aggregation** is the process of summarizing structured facts into sentences. Aggregation is not necessary in NLG systems, because each fact can be viewed as a separate sentence.

- **Lexicalization, referring expression generation, and linguistic realization** are the processes of applying domain specific rules and grammar to produce a syntactically and logically correct text with selected words and phrases. These modules take the output of document structuring task as its input, and generate final results.

Traditionally, the methods in the prior work on document structuring [28, 2, 35, 37, 24] address the problem of finding the optimal full order of sentences, without assuming any partial order of the existent text. Also, these work focused on short, unstructured documents where a local view of coherence suffices. These document structuring methods are categorized into either statistical approaches or symbolic planners, which will be discussed in the following subsections.

2.1.1 Symbolic Concept-to-text Generation

Our approach is related to work on text planning in symbolic concept-to-text generation (see Reiter and Dale [39] for an overview). McKeown [35] proposed a schema-based discourse planning algorithm to implement a system called *TEXT*, which generates paragraph-length responses to questions in a database. Hovy [24] and Moore and Paris [37] proposed planning-based discourse planners which are based on Rhetorical Structure Theory or its modification.

In contrast to current ordering approaches (including the hierarchical learning algorithm proposed in this thesis), corpus-based approaches learn discourse relationships from training data. Text planners typically operate over a tree representation wherein leaves express the content being communicated and internal nodes indicate their discourse relations. The benefits of hierarchical

representation have been demonstrated in multiple symbolic generation systems: the ability of tree-based text planners to encode long-range discourse dependencies improves coherence of the generated text. This finding motivates our interest in hierarchical representations for statistical text structuring.

2.1.2 Statistical Methods on Sentence Ordering

The insertion task is closely related to the extensively studied problem of sentence ordering.[2] Most of the existing algorithms represent text structure as a linear sequence and are driven by local coherence constraints [28, 25, 38, 2, 8, 18]. These methods induce a total ordering based on pairwise relations between sentences. Researchers have shown that identifying precedence relations does not require deep semantic interpretation of input sentences: shallow distributional features are sufficient for accurate prediction. Our approach employs similar features to represent nodes at the lowest level of the hierarchy.

The key departure of our work from previous research is the incorporation of hierarchical structure into a corpus-based approach to ordering. While in symbolic generation and discourse analysis a text is typically analyzed as a tree-like structure [39], a linear view is prevalent in data-driven methods to text structuring.[3] Moving beyond a linear representation enables us to handle longer texts where a local view of coherence does not suffice. At the same time, our approach does not require any manual rules for handling tree insertions, in contrast to symbolic text planners.

2.1.3 Hierarchical Learning

There has been much recent research on multiclass hierarchical classification relevant to our hierarchical learning algorithm for the sentence insertion task. In this line of work, the set of possible labels is organized hierarchically, and each input must be assigned a node in the resulting tree. A prototype weight vector is learned for each node, and classification decisions are based on all the

[2] Independently and simultaneously with our work, Elsner and Charniak [18] have studied the sentence insertion task in a different setting.

[3] Though statistical methods have been used to induce such trees [42], they are not used for ordering and other text-structuring tasks.

weights along the path from node to root. The essence of this scheme is that the more ancestors two nodes have in common, the more parameters they are forced to share. Many learning methods have been proposed, including SVM-style optimization [9], incremental least squares estimation [11], and perceptron [15].

This previous work rests on the assumption that a predetermined set of atomic labels with a fixed hierarchy is given. In our task, however, the set of possible insertion points – along with their hierarchical organization – is unique to each input document. Furthermore, nodes exhibit rich internal feature structure and cannot be identified across documents, except insofar as their features overlap. As is commonly done in NLP tasks, we make use of a feature function which produces one feature vector for each possible insertion point. We then choose among these feature vectors using a single weight vector (casting the task as a *structured ranking* problem rather than a *classification* problem). In this framework, an explicit hierarchical view is no longer necessary to achieve parameter tying. In fact, each parameter will be shared by exactly those insertion points which exhibit the corresponding feature, both across documents and within a single document. Higher level parameters will thus naturally be shared by all paragraphs within a single section.

In fact, when the perceptron update rule of [15] – which modifies the weights of every divergent node along the predicted and true paths – is used in the ranking framework, it becomes virtually identical with the standard, flat, ranking perceptron of Collins [12].[4] In contrast, our approach shares the idea of [10] that "if a parent class has been predicted wrongly, then errors in the children should not be taken into account." We also view this as one of the key ideas of the incremental perceptron algorithm of [13], which searches through a complex decision space step-by-step and is immediately updated at the first wrong move.

Our work fuses this idea of selective hierarchical updates with the simplicity of the perceptron algorithm and the flexibility of arbitrary feature sharing inherent in the ranking framework.

[4]The main remaining difference is that Dekel et al. [15] use a passive-aggressive update rule [14] and in doing so enforce a margin based on tree distance.

2.2 Common Structure Induction

Having some meta-labels in these bodies of Wikipedia would contribute greatly to achieving the goal of producing the common structure that we envision for Wikipedia. Regrettably, Wikipedia does not yet have this meta-information. there has been much work to allow better programmatic access of information on Wikipedia or other large collaboratively edited datasets. In contrast to the prior work on semantic Wikipedia, an unsupervised Bayesian algorithm is proposed in this thesis to induce topics for documents without any domain-specific restriction.

The task of identifying the topics conveyed in a text is closely related to the extensively studied problem of Bayesian graphical models, where documents are modeled as a distribution of a sequence of topics, where each topic generates a sequence of words with Markov dependencies, such as the Latent Dirichlet Allocation (**LDA**) model [7, 19] and Correlated Topic Models (**CTM**)[4]. Our work extends the latent semantic framework to jointly model section words and document topic structure across multiple documents. At first, our model views topical properties as distributions over words in the sections. Furthermore, our approach is designed to favor the induced hidden topics toward frequent topical patterns embedded in the corpus. Bridging these two information sources improves the robustness of the hidden topics, thereby increasing the chance that the induced structure contains meaningful topical properties.

2.2.1 Bayesian Topic Modeling

One of the recent Bayesian approaches, Latent Dirichlet Allocation (LDA) [6], has been proposed to model the latent topics of a text collection. The LDA model can be viewed as a Bayesian extension of the unigram mixture model by representing topic weights as a latent Dirichlet random variable. Because words are associated with the topics, topics have distinct distributions over the words. The model estimates both the word distributions for each topic and the topic distributions for each article. Therefore, the model reduces the articles to a low dimensional representation over topics instead of a much higher one over words. Blei et. al. also showed that the LDA model improves over other latent variable models such as pLSI [23] on the tasks like collaborative filtering.

Griffiths et al. [19] extended LDA model to a composite model, in which the syntactic component is an HMM and the semantic component is a topic model. The composite model uses both short-term syntactic and long-term topical dependencies on words, as an effort to integrate semantics and syntax. On tasks like part-of-speech tagging and document classification, experimental results indicated that this HMM-LDA model is comparable to models that exclusively use short- and long-range dependencies respectively.

Li and McCallum [29] apply the HMM-LDA model to obtain word clusters. They overcome word ambiguity problem by allowing one word to probabilistically belong to multiple clusters. Experiments on part-of-speech tagging and Chinese word segmentation has shown improvements over a supervised learning baseline.

In addition to ignoring dependencies between words (compared with HMM-LDA), another shortcoming of the LDA model is that the topics are assumed to be independent of each other. This is obviously not true in general - for example, a Career section is highly likely to follow Education or Early Life section. Therefore, learning these correlations can potentially produce better topic models. To capture this point, Blei et al. proposed the Correlated Topic Model (CTM) [4], which extends the LDA model. The CTM model replace the Dirichlet prior of the LDA model with a logistic-normal prior. Correlation among topic proportions are modeled by first sampling a vector from a multivariate Gaussian distribution, followed by that mapping the vector to a vector of topic proportions through the logistic normal distribution. The logistic normal distribution's covariance matrix then models the correlation between the topics.

Rather than modeling pairwise topic correlation, recent work extended LDA to model the sequential dependencies of sentences in the document as a Markov chain [20]. The Hidden Markov Topic Model proposed by Amit Gruber et al. focuses on learning statistical pattern of topic transitions between sentences within a document.

In contrast to the prior work , our method is designed to capture the topical dependencies of high-level structure (e.g. *sections*) across multiple documents. Because the parameters of LDA are usually intractable , they are usually estimated by approximate inference algorithms, such as Markov Chain Monte Carlo (MCMC). Similar to most LDA tasks, a Gibbs sampling algorithm is

proposed to solve the common structure induction task.

2.2.2 Semantic Wikipedia

There are been many approaches towards the goal of using a Wikipedia for sematic knowledge database. Platypus Wiki and Rhizome Wiki are two examples showing how Semantic Web technologies are incorporating into Wikipedia. Tazzoli et. al.[40] described one of the first prototypes of a Semantic Wikipedia Web, Platypus Wiki, which utilizes the RDF[5] and OWL[6] to represent metadata and relations between Wikipedia pages. Inspired by Platypus Wiki, Rhizome is designed by Souzis[43] to represent collaboratively edited content in a semantically rich format, in order to handle inconsistence and to make Semantic Web technologys ease of use. After analyzed previous systems, Volker et. al. and proposed a system to add more formalized structure to Wikipedia via the addition of semantic fields [46]. Semantic and standardized fields allow us to assign meaning to text within Wikipedia pages. This allows a user the ability to find "all the movies produced in the 1960s by Italian directors". Krtzsch et. al. used typed links as an unobtrusive way for adding machine-readable links in Wikipedia in the same spirit as Volker [27].

Tagging systems also enable users to add semantics to collaboratively edited corpora. Marlow et. al. gave a survey on web-based tagging systems and describe the advantages of structured information such as the ability to improve search, spam detection, reputation systems, and personal organization [34].

2.2.3 Summary

Significant stylistic and content variations across texts in collaboratively edited articles distinguishes our corpus from more stylistically collections used in the prior text generation research. In contrast to the prior research in text structuring, our first method creates a direction by viewing text structuring as a incremental process unfolding over time. It grants us the power to take advantage of existing text structure. While the previous work on Semantic Wikipedia focuses on Semantic

[5] http://www.w3.org/RDF/

[6] http://www.w3.org/TR/owl-features/

Web technologies, our second method discussed in this thesis is an unsupervised Bayesian corpus-based algorithm to induce semantics on Wikipedia.

Chapter 3

Incremental Text Structuring

3.1 Overview

Many emerging applications require documents to be repeatedly updated. For instance, newsfeed articles are continuously revised by editors as new information emerges, and personal webpages are modified as the status of the individual changes. This revision strategy has become even more prevalent with the advent of community edited web resources, the most notable example being Wikipedia. At present this process involves massive human effort. For instance, the English language version of Wikipedia averaged over 3 million edits[1] per month in 2006. Even so, many articles quickly become outdated. A system that performs such updates automatically could drastically decrease maintenance efforts and potentially improve document quality.

Currently there is no effective way to automatically update documents as new information becomes available. The closest relevant text structuring technique is the work on sentence ordering, in which a complete reordering of the text is undertaken. Predictably these methods are suboptimal for this new task because they cannot take advantage of existing text structure.

We introduce an alternative vision of text structuring as a process unfolding over time. Instead of ordering sentences all at once, we start with a well-formed draft and add new information at each stage, while preserving document coherence. The basic operation of incremental text structuring is

[1] http://stats.wikimedia.org/EN/TablesWikipediaEN.htm

the insertion of new information. To automate this process, we develop a method for determining the best location in a text for a given piece of new information.

The main challenge is to maintain the continuity and coherence of the original text. These properties may be maintained by examining sentences adjacent to each potential insertion point. However, a local sentence comparison method such as this may fail to account for global document coherence (e.g. by allowing the mention of some fact in an inappropriate section). This problem is especially acute in the case of lengthy, real-world texts such as books, technical reports, and web pages. These documents are commonly organized hierarchically into sections and paragraphs to aid reader comprehension. For documents where hierarchical information is not explicitly provided, such as automatic speech transcripts, we can use automatic segmentation methods to induce such a structure [22]. Rather than ignoring the inherent hierarchical structure of these texts, we desire to directly model such hierarchies and use them to our advantage – both representationally in our features and algorithmically in our learning procedure.

To achieve this goal, we introduce a novel method for sentence insertion that operates over a hierarchical structure. Our document representation includes features for each layer of the hierarchy. For example, the word overlap between the inserted sentence and a section header would be included as an upper-level section feature, whereas a comparison of the sentence with all the words in a paragraph would be a lower-level paragraph feature. We propose a linear model which simultaneously considers the features of every layer when making insertion decisions. We develop a novel update mechanism in the online learning framework which exploits the hierarchical decomposition of features. This mechanism limits model updates to those features found at the highest incorrectly predicted layer, without unnecessarily disturbing the parameter values for the lower reaches of the tree. This conservative update approach maintains as much knowledge as possible from previously encountered training examples.

We evaluate our method using real-world data where multiple authors have revised preexisting documents over time. We obtain such a corpus from Wikipedia articles,[2] which are continuously updated by multiple authors. Logs of these updates are publicly available (see Figure 3-4), and are used for training and testing of our algorithm. Figure 3-1 shows an example of a Wikipedia inser-

[2]Data and code used in this chapter are available at `http://people.csail.mit.edu/edc/emnlp07/`

Raymond Chen

From Wikipedia, the free encyclopedia

(Difference between revisions)
Jump to: navigation, search

Revision as of 03:46, 26 April 2006 (edit)	**Revision as of 17:21, 1 May 2006 (edit) (undo)**
Warren (Talk | contribs)	JenKilmer (Talk | contribs)
m (+ Category:Windows people)	(→Hobbies)
← Older edit	Newer edit →
Line 9:	**Line 9:**
==Hobbies==	==Hobbies==
His computer-unrelated hobbies, as described in the blog, include knitting, cooking, classical music, bicycling, and learning multiple foreign languages ([[Swedish language|Swedish]], [[German language|German]], and [[Standard Mandarin|Mandarin Chinese]]).	His computer-unrelated hobbies, as described in the blog, include knitting, cooking, classical music, bicycling, and learning multiple foreign languages ([[Swedish language|Swedish]], [[German language|German]], and [[Standard Mandarin|Mandarin Chinese]]). **Chen grew up speaking English and some of the Taiwanese dialect.**
Before his career at Microsoft and lasting even into 1995, Raymond Chen identified himself as "just another [[Linux]] hacker" in his [[Usenet]] [[Signature block|sig]]. He is listed in the Linux kernel CREDITS file as "Author of Configure script". It is unknown if Chen is the only person who has contributed code to both Windows and Linux.	Before his career at Microsoft and lasting even into 1995, Raymond Chen identified himself as "just another [[Linux]] hacker" in his [[Usenet]] [[Signature block|sig]]. He is listed in the Linux kernel CREDITS file as "Author of Configure script". It is unknown if Chen is the only person who has contributed code to both Windows and Linux.

Revision as of 17:21, 1 May 2006

Raymond Juimong Chen is a well-known developer on the Windows Shell team at Microsoft. Chen joined Microsoft in 1992. He has worked on OS/2, Windows 95, DirectX, and later versions of Windows. He has also spoken at Microsoft PDCs and other conferences.

Raymond is known for his programming abilities, his dry, pithy comments, and his custom of wearing suits at work.

Writings

Chen writes a blog, popular among software developers, called The Old New Thing, which focuses on the history of Windows and his own experience in ensuring its backwards compatibility. He is noted for his 'Psychic Debugging' (example) articles, as well as two useful types of thought experiments in software design: "Imagine if this were possible" and "What if two programs did this?" *(see also his presentation at PDC05)* Chen also contributed the essay *Why Not Just Block the Apps That Rely on Undocumented Behavior?* to the book The Best Software Writing I, edited by Joel Spolsky.

Hobbies

His computer-unrelated hobbies, as described in the blog, include knitting, cooking, classical music, bicycling, and learning multiple foreign languages (Swedish, German, and Mandarin Chinese). **Chen grew up speaking English and some of the Taiwanese dialect.**

Before his career at Microsoft and lasting even into 1995, Raymond Chen identified himself as "just another Linux hacker" in his Usenet sig. He is listed in the Linux kernel CREDITS file as "Author of Configure script". It is unknown if Chen is the only person who has contributed code to both Windows and Linux.

This biographical article relating to a computer specialist is a stub. You can help Wikipedia by expanding it.

Retrieved from "http://en.wikipedia.org/wiki/Raymond_Chen"
Categories: Computer specialist stubs | Microsoft employees | Living people | American bloggers | Windows people

tion. We believe this data will more closely mirror potential applications than synthetic collections used in previous work on text structuring.

Our hierarchical training method yields significant improvement when compared to a similar non-hierarchical model which instead uses the standard perceptron update of Collins [12]. We also report human performance on the insertion task in order to provide a reasonable upper-bound on machine performance. An analysis of these results shows that our method closes the gap between machine and human performance substantially.

In the following section, we provide an overview of existing work on text structuring and hierarchical learning. Then, we define the insertion task and introduce our hierarchical ranking approach to sentence insertion. Next, we present our experimental framework and data. We conclude the chapter by presenting and discussing our results.

3.2 The Algorithm

In this section, we present our sentence insertion model and a method for parameter estimation. Given a hierarchically structured text composed of sections and paragraphs, the sentence insertion model determines the best paragraph within which to place the new sentence. To identify the exact location of the sentence within the chosen paragraph, local ordering methods such as [28] could be used. We formalize the insertion task as a structured ranking problem, and our model is trained using an online algorithm. The distinguishing feature of the algorithm is a selective correction mechanism that focuses the model update on the relevant layer of the document's feature hierarchy.

The algorithm described below can be applied to any hierarchical ranking problem. For concreteness, we use the terminology of the sentence insertion task, where a hierarchy corresponds to a document with sections and paragraphs.

3.2.1 Problem Formulation

In a sentence insertion problem, we are given a training sequence of instances $(s^1, \mathcal{T}^1, \ell^1), \ldots, (s^m, \mathcal{T}^m, \ell^m)$. Each instance contains a sentence s, a hierarchically structured document $\mathcal{T}$, and a node ℓ repre-

senting the correct insertion point of s into $\mathcal{T}$. Although ℓ can generally be any node in the tree, in our problem we need only consider leaf nodes. We cast this problem in the ranking framework, where a feature vector is associated with each sentence-node pair. For example, the feature vector of an internal, section-level node may consider the word overlap between the inserted sentence and the section title. At the leaf level, features may include an analysis of the overlap between the corresponding text and sentence. In practice, we use disjoint feature sets for different layers of the hierarchy, though in theory they could be shared.

Our goal then is to choose a leaf node by taking into account its feature vector as well as feature vectors of all its ancestors in the tree.

More formally, for each sentence s and hierarchically structured document $\mathcal{T}$, we are given a set of feature vectors, with one for each node: $\{\phi(s,n) : n \in \mathcal{T}\}$. We denote the set of leaf nodes by $\mathcal{L}(\mathcal{T})$ and the path from the root of the tree to a node n by $\mathcal{P}(n)$. Our model must choose one leaf node among the set $\mathcal{L}(\mathcal{T})$ by examining its feature vector $\phi(s,\ell)$ as well as all the feature vectors along its path: $\{\phi(s,n) : n \in \mathcal{P}(\ell)\}$.

3.2.2 The Model

Our model consists of a weight vector $\mathbf{w}$, each weight corresponding to a single feature. The features of a leaf are aggregated with the features of all its ancestors in the tree. The leaf score is then computed by taking the inner product of this aggregate feature vector with the weights $\mathbf{w}$. The leaf with the highest score is then selected.

More specifically, we define the *aggregate feature vector* of a leaf ℓ to be the sum of all features found along the path to the root:

$$\Phi(s,\ell) = \sum_{n \in \mathcal{P}(\ell)} \phi(s,n) \tag{3.1}$$

This has the effect of *stacking together* features found in a single layer, and *adding* the values of features found at more than one layer.

Input : $(s^1, \mathcal{T}^1, \ell^1), \ldots, (s^m, \mathcal{T}^m, \ell^m)$.
Initialize : Set $\mathbf{w}^1 = 0$
Loop : For $t = 1, 2, ..., N$:

1. Get a new instance $s^t, \mathcal{T}^t$.
2. Predict $\hat{\ell}^t = \arg\max_{\ell \in \mathcal{L}(\mathcal{T})} \mathbf{w}^t \cdot \Phi(s^t, \ell)$.
3. Get the new label ℓ^t.
4. If $\hat{\ell}^t = \ell^t$:
   $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t$
   Else:
   $i^* \leftarrow \max\{i : \mathcal{P}(\ell^t)^i = \mathcal{P}(\hat{\ell}^t)^i\}$
   $a \leftarrow \mathcal{P}(\ell^t)^{i^*+1}$
   $b \leftarrow \mathcal{P}(\hat{\ell}^t)^{i^*+1}$
   $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t + \phi(s, a) - \phi(s, b)$

Output : $\mathbf{w}^{N+1}$.

Figure 3-2: Training algorithm for the hierarchical ranking model.

Our model then outputs the leaf with the highest scoring aggregate feature vector:

$$\arg \max_{\ell \in \mathcal{L}(\mathcal{T})} \mathbf{w} \cdot \Phi(s, \ell) \tag{3.2}$$

Note that by using this criterion, our decoding method is equivalent to that of the standard linear ranking model. The novelty of our approach lies in our training algorithm which uses the hierarchical feature decomposition of Equation 3.1 to pinpoint its updates along the path in the tree.

3.2.3 Training

Our training procedure is implemented in the online learning framework. The model receives each training instance, and predicts a leaf node according to its current parameters. If an incorrect leaf node is predicted, the weights are updated based on the divergence between the predicted path and the true path. We trace the paths down the tree, and only update the weights of the features found at the split point. Updates for shared nodes along the paths would of course cancel out. In contrast to the standard ranking perceptron as well as the hierarchical perceptron of [15], no

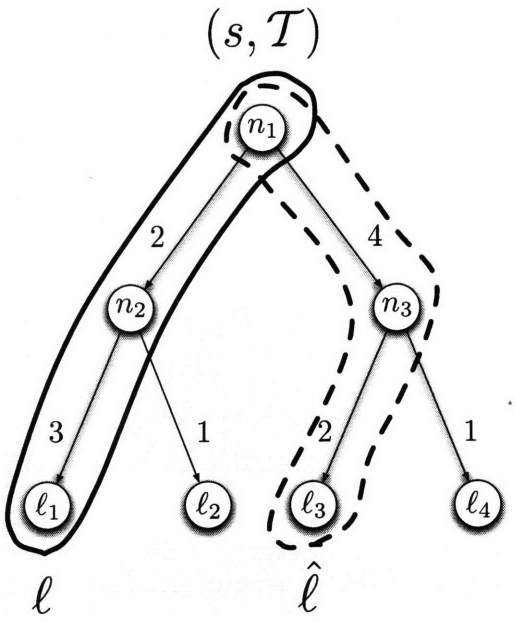


Figure 3-3: An example of a tree with the corresponding model scores. The path surrounded by solid lines leads to the correct node ℓ_1. The path surrounded by dotted lines leads to ℓ_3, the predicted output based on the current model.

features further down the divergent paths are incorporated in the update. For example, if the model incorrectly predicts the section, then only the weights of the section features are updated whereas the paragraph feature weights remain untouched.

More formally, let $\hat{\ell}$ be the predicted leaf node and let $\ell \neq \hat{\ell}$ be the true leaf node. Denote by $\mathcal{P}(\ell)^i$ the i^{th} node on the path from the root to ℓ. Let i^* be the depth of the lowest common ancestor of ℓ and $\hat{\ell}$ (i.e., $i^* = \max\{i : \mathcal{P}(\ell)^i = \mathcal{P}(\hat{\ell})^i\}$). Then the update rule for this round is:

$$\mathbf{w} \leftarrow \mathbf{w} + \phi\left(s, \mathcal{P}(\ell)^{i^*+1}\right) - \phi\left(s, \mathcal{P}(\hat{\ell})^{i^*+1}\right) \tag{3.3}$$

Full pseudo-code for our hierarchical online training algorithm is shown in Figure 4-2.

We illustrate the selective update mechanism on the simple example shown on Figure 3-3. The correct prediction is the node ℓ_1 with an aggregate path score of 5, but ℓ_3 with the higher score of 6 is predicted. In this case, both the section and the paragraph are incorrectly predicted.

In response to this mistake, the features associated with the correct section, n_2, are added to the weights, and the features of the incorrectly predicted section, n_3, are subtracted from the weights. An alternative update strategy would be to continue to update the feature weights of the leaf nodes, ℓ_1 and ℓ_3. However, by identifying the exact source of path divergence we preserve the previously learned balance between leaf node features.

3.3 Features

Features used in our experiments are inspired by previous work on corpus-based approaches for discourse analysis [33, 28, 17]. We consider three types of features: lexical, positional, and temporal. This section gives a general overview of these features and a list of sample features in Table 3.1 (see code for further details.)

3.3.1 Lexical Features

Lexical features have been shown to provide strong cues for sentence positioning. To preserve text cohesion, an inserted sentence has to be topically close to its surrounding sentences. At the paragraph level, we measure topical overlap using the TF*IDF weighted cosine similarity between an inserted sentence and a paragraph. We also use a more linguistically refined similarity measure that computes overlap considering only subjects and objects. Syntactic analysis is performed using the MINIPAR parser [30].

The overlap features are computed at the section level in a similar way. We also introduce an additional section-level overlap feature that computes the cosine similarity between an inserted sentence and the first sentence in a section. In our corpus, the opening sentence of a section is typically strongly indicative of its topic, thus providing valuable cues for section level insertions.

In addition to overlap, we use lexical features that capture word co-occurrence patterns in coherent texts. This measure was first introduced in the context of sentence ordering by Lapata [28]. Given a collection of documents in a specific domain, we compute the likelihood that a pair of words co-occur in adjacent sentences. From these counts, we induce the likelihood that two sen-

tences are adjacent to each other. For a given paragraph and an inserted sentence, the highest adjacency probability between the inserted sentence and paragraph sentences is recorded. This feature is also computed at the section level.

3.3.2 Positional Features

These features aim to capture user preferences when positioning new information into the body of a document. For instance, in the Wikipedia data, insertions are more likely to appear at the end of a document than at its beginning. We track positional information at the section and paragraph level. At the section level, we record whether a section is the first or last of the document. At the paragraph level, there are four positional features which indicate the paragraph's position (i.e., start or end) within its individual section and within the document as a whole.

3.3.3 Temporal Features

The text organization may be influenced by temporal relations between underlying events. In temporally coherent text, events that happen in the same time frame are likely to be described in the same segment. Our computation of temporal features does not require full fledged temporal interpretation. Instead, we extract these features based on two categories of temporal cues: verb tense and date information. The verb tense feature captures whether a paragraph contains at least one sentence using the same tense as the inserted sentence. For instance, this feature would occur for the inserted sentence in Figure 1-1 since both the sentence and chosen paragraph employ the past tense.

Another set of features takes into account the relation between the dates in a paragraph and those in an inserted sentence. We extract temporal expressions using the **TIMEX2** tagger [32], and compute the time interval for a paragraph bounded by its earliest and latest dates. We record the degree of overlap between the paragraph time interval and insertion sentence time interval.

Paragraph level features
The number of sentences in p which shared non-stop-words/nouns/proper nouns/verbs with sen
Whether p is the i-th paragraph in the section
Whether p is the first paragraph in the section
Whether p is the last paragraph in the section
TF score between p and sen based on non-stop-words/nouns/proper nouns/verbs
TF-IDF score between p and sen based on non-stop-words/nouns/proper nouns/verbs
The ratio that the subjects/objects of sen appear in the sentences of p
The ratio that p's sentences whose subjects/objects appear in sen
Both sen and p have present tense verbs
sen doesn't and p has present tense verbs
Average bayesian score introduced by Lapata [28] by comparing each sentence in p with sen
Top bayesian score introduced by Lapata [28] by comparing each sentence in p with sen
Both sen and p have dates
sen has dates, but p doesn't have dates
sen doesn't have dates, but p has dates
The date of sen is between the date of p's previous paragrah and that of p's next paragraph
The date of sen contradicts with that of p's previous paragraph or p's next paragraph
The date of sen is after that of the last paragraph of an article
The date of sen overlaps with that of the last paragraph of an article
The date of sen is after that of p's previous paragraph
Section level features
The number of sentences in sec which shared non-stop-words/nouns/proper nouns/verbs with sen
Whether sec is the first section in the section
Whether sec is the last section in the section
TF score between sec and sen based on non-stop-words/nouns/proper nouns/verbs
TF-IDF score between sec and sen based on non-stop-words/nouns/proper nouns/verbs
The ratio that the subjects/objects of sen appear in the sentences of sec
Average bayesian score introduced by Lapata [28] by comparing each sentence in sec with sen
Top bayesian score introduced by Lapata [28] by comparing each sentence in sec with sen

Table 3.1: List of sample features (given an insertion sentence sen, and a paragraph p from a section sec of a document d.)

		Section	Paragraph	Tree Dist
T1	J1	0.575	0.5	1.85
	J2	0.7	0.525	1.55
T2	J3	0.675	0.55	1.55
	J4	0.725	0.55	1.45

Table 3.2: Accuracy of human insertions compared against gold standard from Wikipedia's update log. T1 is a subset of the data annotated by judges J1 and J2, while T2 is annotated by J3 and J4.

3.4 Experimental Set-Up

3.4.1 Corpus

Our corpus consists of Wikipedia articles that belong to the category "Living People." We focus on this category because these articles are commonly updated: when new facts about a person are featured in the media, a corresponding entry in Wikipedia is likely to be modified. Unlike entries in a professionally edited encyclopedia, these articles are collaboratively written by multiple users, resulting in significant stylistic and content variations across texts in our corpus. This property distinguishes our corpus from more stylistically homogeneous collections of biographies used in text generation research [16].

We obtain data on insertions[3] from the update log that accompanies every Wikipedia entry. For each change in the article's history, the log records an article before and after the change. Figure 3-4 shows a part of Barack Obamba's Wikipedia article log. From this information, we can identify the location of every inserted sentence. In cases where multiple insertions occur over time to the same article, they are treated independently of each other. To eliminate spam, we place constraints on inserted sentences: (1) a sentence has at least 8 tokens and at most 120 tokens; (2) the MINIPAR parser [30] can identify a subject or an object in a sentence.

This process yields 4051 insertion/article pairs, from which 3240 pairs are used for training and 811 pairs for testing. These insertions are derived from 1503 Wikipedia articles. Table 3.3 shows a list of sample insertion sentences from our dataset. Relative to other corpora used in text structuring research [3, 28, 25], texts in our collection are long: an average article has 32.9

[3]Insertion is only one type of recorded update, others include deletions and sentence rewriting.

Make a donation to Wikipedia and give the gift of knowledge!

Revision history of Barack Obama

From Wikipedia, the free encyclopedia
View logs for this page
(Latest | Earliest) View (newer 40) (older 40) (20 | 50 | 100 | 250 | 500)
For any version listed below, click on its date to view it. For more help, see Help:Page history and Help:Edit summary.
(cur) = difference from current version, (last) = difference from preceding version, **m** = minor edit, → = section edit, ← = automatic edit summary
Compare selected versions

- (cur) (last) 05:12, 18 May 2008 Tvoz (Talk | contribs) (130,629 bytes) (remove subheading which gives undue weight to this section -)
- (cur) (last) 04:46, 18 May 2008 Arsenic99 (Talk | contribs) (130,669 bytes) (Fixed Grammar!)
- (cur) (last) 04:36, 18 May 2008 Fovean Author (Talk | contribs) (130,670 bytes) (Undid revision 213185598 by Tvoz (talk)If you want to make changes, get a consensus)
- (cur) (last) 04:33, 18 May 2008 Tvoz (Talk | contribs) (127,112 bytes) (Undid revision 213176598 by Fovean Author (talk) revert)
- (cur) (last) 04:28, 18 May 2008 Kossack4Truth (Talk | contribs) (130,670 bytes) (RVV)
- (cur) (last) 04:26, 18 May 2008 Kossack4Truth (Talk | contribs) (127,112 bytes) (RVV)
- (cur) (last) 04:23, 18 May 2008 Nykoreankid (Talk | contribs) (131,319 bytes)
- (cur) (last) 04:22, 18 May 2008 Nykoreankid (Talk | contribs) (131,290 bytes)
- (cur) (last) 04:18, 18 May 2008 Lordajay (Talk | contribs) (131,285 bytes) (→Presidential campaign)
- (cur) (last) 04:17, 18 May 2008 Lordajay (Talk | contribs) (130,970 bytes)
- (cur) (last) 04:15, 18 May 2008 Lordajay (Talk | contribs) (130,847 bytes)
- (cur) (last) 03:29, 18 May 2008 Fovean Author (Talk | contribs) (130,670 bytes) (Undid revision 213171020 by Newross (talk)If it isn't, it should be)
- (cur) (last) 02:53, 18 May 2008 Newross (Talk | contribs) (127,112 bytes) (Revert; absolutely not "consensus version")
- (cur) (last) 02:21, 18 May 2008 Kossack4Truth (Talk | contribs) (130,670 bytes) (Lulu, this is the consensus version. Please do not abbreviate it again. Such an edit might reasonably be construed as vandalism.)
- (cur) (last) 02:07, 18 May 2008 Avanatt (Talk | contribs) (127,112 bytes)
- (cur) (last) 02:04, 18 May 2008 Avanatt (Talk | contribs) (127,110 bytes)
- (cur) (last) 02:01, 18 May 2008 Modocc (Talk | contribs) (127,286 bytes) (Undid revision 213162048 by Nunh-huh (talk) Not so, check history before DianeFinn's edits)
- (cur) (last) 01:59, 18 May 2008 Tribulation725 (Talk | contribs) (127,294 bytes)
- (cur) (last) 01:54, 18 May 2008 Nunh-huh (Talk | contribs) (127,046 bytes) (→Early life: you're right, there was no consensus to change from the original, uncomplicated presentation of the accurate name.)
- (cur) (last) 01:44, 18 May 2008 Modocc (Talk | contribs) (127,038 bytes) (Policy aside, this is overly detailed, inconsistent, and ignors self-identity. There is no concensus to change name either.)
- (cur) (last) 00:36, 18 May 2008 Sli723 (Talk | contribs) (127,063 bytes) (rv edits--there is no proof)
- (cur) (last) 00:33, 18 May 2008 Foxcloud (Talk | contribs) (127,249 bytes) (→Early life)
- (cur) (last) 00:25, 18 May 2008 Foxcloud (Talk | contribs) (127,250 bytes) (→External links)
- (cur) (last) 00:24, 18 May 2008 Foxcloud (Talk | contribs) (127,213 bytes) (→External links)
- (cur) (last) 00:23, 18 May 2008 Nunh-huh (Talk | contribs) (127,064 bytes) (→Early life and career: restore accurate name; there's no "style" issue here, and the accurate name should be used. No valid reason to excise this information.)
- (cur) (last) 00:19, 18 May 2008 Foxcloud (Talk | contribs) (127,038 bytes) (→External links)
- (cur) (last) 23:46, 17 May 2008 Lulu of the Lotus-Eaters (Talk | contribs) (126,758 bytes) (rm unnecessary subheading)
- (cur) (last) 23:26, 17 May 2008 Lulu of the Lotus-Eaters (Talk | contribs) (126,795 bytes) (Soapboxing and WP:UNDUE weight continue to contradict WP style and policy)
- (cur) (last) 23:22, 17 May 2008 Lulu of the Lotus-Eaters (Talk | contribs) (129,473 bytes) (use WP style on names)
- (cur) (last) 22:21, 17 May 2008 Ndgp (Talk | contribs) (129,498 bytes)
- (cur) (last) 21:26, 17 May 2008 AlnoktaBOT (Talk | contribs) m (129,488 bytes) (robot Modifying: zh:贝拉克·奥巴马)
- (cur) (last) 21:00, 17 May 2008 DianeFinn (Talk | contribs) (129,487 bytes) (futher compromise (getting rid of parenthesis), also discussed at length in talk page, merely correcting mother's name)
- (cur) (last) 20:15, 17 May 2008 Kossack4Truth (Talk | contribs) (129,462 bytes) (→Rev. Wright and later primaries)
- (cur) (last) 20:08, 17 May 2008 Kossack4Truth (Talk | contribs) (129,025 bytes) (Undid revision 213101427 by Lulu of the Lotus-Eaters (talk))
- (cur) (last) 19:53, 17 May 2008 Lulu of the Lotus-Eaters (Talk | contribs) (126,744 bytes) (The soapboxing by Kossak4Truth is starting to verge on vandalism... there ARE other articles where a long discussion belongs)
- (cur) (last) 19:47, 17 May 2008 Kossack4Truth (Talk | contribs) (129,025 bytes) (→Presidential campaign)
- (cur) (last) 19:43, 17 May 2008 Kossack4Truth (Talk | contribs) (128,948 bytes) (Undid revision 213000894 by Lulu of the Lotus-Eaters (talk))
- (cur) (last) 19:41, 17 May 2008 Northwesterner1 (Talk | contribs) (126,744 bytes) (→Presidential campaign: why is this article using the popular vote map but not the (more relevant) delegate map?)
- (cur) (last) 19:16, 17 May 2008 Lulu of the Lotus-Eaters (Talk | contribs) (126,747 bytes) (WP policy links on MOST COMMONLY USED version of name. Explain name changes and extra names in articles so linked)
- (cur) (last) 18:49, 17 May 2008 DianeFinn (Talk | contribs) m (126,768 bytes) (→Early life: compromise: Stanley Ann Dunham (Ann Dunham) instead of just Ann Dunham)

Compare selected versions
(Latest | Earliest) View (newer 40) (older 40) (20 | 50 | 100 | 250 | 500)
Retrieved from "http://en.wikipedia.org/wiki/Barack_Obama"

Figure 3-4: A selected part of the revision history on a Wikipedia article about Barack Obama.

sentences, organized in 3.61 sections and 10.9 paragraphs. Our corpus only includes articles that have more than one section. When sentences are inserted between paragraphs, by convention we treat them as part of the previous paragraph.

3.4.2 Evaluation Measures

We evaluate our model using *insertion accuracy* at the section and paragraph level. This measure computes the percentage of matches between the predicted location of the insertion and the true placement. We also report the *tree distance* between the predicted position and the true location of an inserted sentence. *Tree distance* is defined as the length of the path through the tree which connects the predicted and the true paragraph positions. This measure captures section level errors (which raise the connecting path higher up the tree) as well as paragraph level errors (which widen the path across the tree).

3.4.3 Baselines

Our first three baselines correspond to naive insertion strategies. The RANDOMINS method randomly selects a paragraph for a new sentence, while FIRSTINS and LASTINS insert a sentence into the first and the last paragraph, respectively.

We also compare our HIERARCHICAL method against two competitive baselines, PIPELINE and FLAT. The PIPELINE method separately trains two rankers, one for section selection and one for paragraph selection. During decoding, the PIPELINE method first chooses the best section according to the section-layer ranker, and then selects the best paragraph within the chosen section according to the paragraph-layer ranker. The FLAT method uses the same decoding criterion as our model (Equation 3.2), thus making use of all the same features. However, FLAT is trained with the standard ranking perceptron update, without making use of the hierarchical decomposition of features in Equation 3.1.

Tone plans to release her second album in late 2006.
He started the 2006 season with AAA Iowa.
On November 16, 2006, he was traded to the Chicago White Sox for pitcher Neal Cotts.
He has won the George Orwell Prize for Political Journalismin 1998 and 2001.
He has also been accused by the Muslim Community for being an Islamophbic Journalist out to attack Pro Democractic Muslim Organisations such as MPACUK.
The ludologists are contrasted by the so-called "narrativists" such as Janet Murray.
In addition, he regularly writes columns for the Jewish Chronicle.
His successor is Willibrord van Beek member of parliament since 1998 .
Van Aartsen will not be a candidate for the VVD in the 2006 general election.
Jozias van Aartsen is the son of Jan van Aartsen, himself a minister of the Netherlands from 1958 to 1965.
Similarly, Danny, Tito, and team punishment member Matt Hammill are featured in a parody of the Ultimate Fighter called "The Gay Ultimate Fighter" which features excerpts of the show to the music of "Hot Stuff".
Mindi also sings, as evidenced by her vocal cover of the Eagle-Eye Cherry hit "Save Tonight" on her debut album.
His bestselling books, according to Christian Bookseller's Association listings, include "Harry Potter and the Bible" and "The Truth Behind the DaVinci Code.".
DNA testing later revealed, however, that Abbott was not O'Connor's biological father.
Abbey is a father of two childeren and also co-owns a small buisness along with his wife.
He is the the grandson of Dr. Joseph Ephraim Weinman, Professor Emeritus, and one of the founding fathers of the School of Anatomy in the College of Veterinary Medicine in the University of Missouri.
His Father lives in Morocco with a part of his family memebrs and is member of the Royal Advisory Council for Saharan Affairs (CORCAS).
Kareem Appeared in one of the Apple PowerBook commercials in the early 90's.
Jabar appeared as a cameo guest star in the television show, "Full House" as the referee of a charity basketball game.

Table 3.3: A list of sample insertion sentences from our Wikipedia dataset

3.4.4 Human Performance

To estimate the difficulty of sentence insertion, we conducted experiments that evaluate human performance on the task. Four judges collectively processed 80 sentence/article pairs which were randomly extracted from the test set. Each insertion was processed by two annotators.

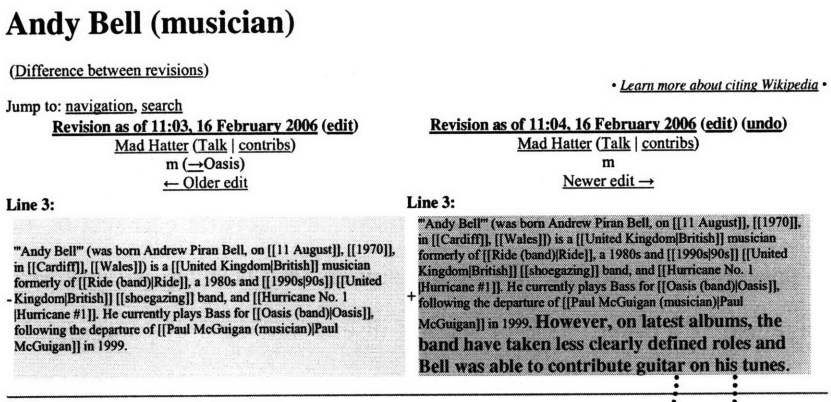
Table 3.2 shows the insertion accuracy for each judge when compared against the Wikipedia gold standard. On average, the annotators achieve 66% accuracy in section placement and 53% accuracy in paragraph placement. We obtain similar results when we compare the agreement of the judges against each other: 65% of section inserts and 48% of paragraph inserts are identical between two annotators. The degree of variability observed in this experiment is consistent with human performance on other text structuring tasks such as sentence ordering [1, 28].

Another conclusion from the analysis of human performance is that for many insertions more than one location is legitimate. Our evaluation is overly strict since it penalizes any variations from the gold standard. To provide another view of our model's performance, we manually analyze 50 insertions which did not match the gold standard. In 16 of these cases, the algorithm insertion was judged equally correct as the true insertion point. One typical pattern occurs when a sentence is inserted into an initial summary paragraph of the article, and instead our algorithm chooses a more specific location within a body of the article (or vice versa). See Figure 3-5 for an example where the inserted sentence would be appropriate in both the article summary as well as the more specific "Career" section.

3.5 Results

Table 3.4 shows the insertion performance of our model and the baselines in terms of accuracy and tree distance error. The two evaluation measures are consistent in that they yield roughly identical rankings of the systems. Assessment of statistical significance is performed using a Fisher Sign Test. We apply this test to compare the accuracy of the HIERARCHICAL model against each of the baselines.

The results in Table 3.4 indicate that the naive insertion baselines (RANDOMINS, FIRSTINS,



Andy Bell (musician)

(Difference between revisions)

• *Learn more about citing Wikipedia* •

Jump to: navigation, search

Revision as of 11:03, 16 February 2006 (edit)
Mad Hatter (Talk | contribs)
m (→Oasis)
← Older edit

Revision as of 11:04, 16 February 2006 (edit) (undo)
Mad Hatter (Talk | contribs)
m
Newer edit →

Line 3:

"Andy Bell" (was born Andrew Piran Bell, on [[11 August]], [[1970]], in [[Cardiff]], [[Wales]]) is a [[United Kingdom|British]] musician formerly of [[Ride (band)|Ride]], a 1980s and [[1990s|90s]] [[United Kingdom|British]] [[shoegazing]] band, and [[Hurricane No. 1 |Hurricane #1]]. He currently plays Bass for [[Oasis (band)|Oasis]], following the departure of [[Paul McGuigan (musician)|Paul McGuigan]] in 1999.

Line 3:

"Andy Bell" (was born Andrew Piran Bell, on [[11 August]], [[1970]], in [[Cardiff]], [[Wales]]) is a [[United Kingdom|British]] musician formerly of [[Ride (band)|Ride]], a 1980s and [[1990s|90s]] [[United Kingdom|British]] [[shoegazing]] band, and [[Hurricane No. 1 |Hurricane #1]]. He currently plays Bass for [[Oasis (band)|Oasis]], following the departure of [[Paul McGuigan (musician)|Paul McGuigan]] in 1999. **However, on latest albums, the band have taken less clearly defined roles and Bell was able to contribute guitar on his tunes.**

Revision as of 11:04, 16 February 2006

Andy Bell (was born Andrew Piran Bell, on 11 August 1970, in Cardiff, Wales) is a British musician formerly of Ride, a 1980s and 90s British shoegazing band, and Hurricane #1. He currently plays Bass for Oasis, following the departure of Paul McGuigan in 1999. **However, on latest albums, the band have taken less clearly defined roles and Bell was able to contribute guitar on his tunes.**

Ride

Bell formed Ride with Mark Gardener (Guitar), who he met at Cheney School in Oxford and Laurence Colbert (Drums) and Steve Queralt (Bass), who he met doing Foundation Studies in Art and Design at Banbury in 1988. While still at Banbury the band produced a tape demo including the tracks "Chelsea Girl" and "Drive Blind". In February 1989 "Ride" were asked to stand in for a cancelled student union gig at Oxford Poly that brought them to the attention of Alan McGee. After supporting The Soup Dragons in 1989 McGee signed them to Creation Records.

With Ride, Bell released three EPs between January and September 1990, entitled "Ride", "Play" and "Fall". While the EP's were not a chart successes, enough critical praise was received to make Ride the "darlings" of music journalists. The first two EPs were eventually released together as *Smile* in 1992, while the "Fall" EP was incorporated into their first LP, *Nowhere*, released in October 1990, which was hailed as a critical success and the media dubbed Ride "The brightest hope" for 1991. This was followed in March 1992 with *Going Blank Again*. The twin rhythm guitars of Bell and Gardener, both distorted, both using Wah-wah pedals and both feeding back on each other was seen as the highlight of the album's critical and chart success.

Despite having a solid fanbase and some mainstream success, the lack of a breakthrough contributed to inter-band tension, especially between Gardener and Bell. Their third LP, *Carnival of Light*, was released in 1994, after shoegazing had given way to Britpop. *Carnival of Light* was oriented towards this new sound, but sales were sluggish and the shift in musical tastes devastated much of their original audience. 1995 saw the dissolution of the band while recording fourth album *Tarantula* due to creative and personal tensions between Gardener and Bell. The track listing of *Carnival of Light* gives an indication of the tension that was mounting between the two guitarists, with the first half of the album being songs written by Gardener and the last half of the album being songs written by Bell - one or both had refused to let their songs be interspersed with pieces written by the other. Bell penned most of the songs for *Tarantula*, one of which - "Castle on the Hill" - was a lament for the band's situation and contains references to Gardener's self imposed exile from the group. The album was withdrawn from sales one week after release.

Since the break-up, both Bell and Gardener have been able to be more reflective on the reasons why the group disintegrated, with Bell especially admitting his own part in the process. It appears that they had just been too young and too stubborn and had no real idea of where the band was heading when they changed their style.

Oasis

After the split, Bell formed a new band called Hurricane #1 but this project was permanently dissolved when he was asked to play Bass for Oasis. Bell had never played the bass before, but Noel Gallagher was confident that he would make a suitable replacement for Paul McGuigan. Bell was obliged to learn playing bass and the entire Oasis catalogue before his first Oasis gig at the last minute.

Bell is also a member of Oasis' songwriting team, contributing *Heathen Chemistry's* instrumental "(A Quick) Peep", *Don't Believe the Truth*'s "Turn Up The Sun", and "Keep the Dream Alive", as well as b-side "Thank You for the Good Times", which appeared on the "Stop Crying Your Heart Out"-single.

As the token southerner, Bell is the butt of many jokes. On stage, in response to the arguing chants of *"Noel"* and *"Liam"*, Noel Gallagher appealed for the crowds to "cut out the *"Noel"* and *"Liam"* shit. Let's have a bit of *Who the fuck is Andy Bell?*".

Bell is married to Swedish singer Idha. Together, they have a daughter named Leia. He splits his time between Sweden and London.

Figure 3-5: Screenshot of an example insertion with two legitimate insertion points. An insertion sentence is shown in red.

	Section	Paragraph	Tree Dist
RANDOMINS	0.318*	0.134*	3.10*
FIRSTINS	0.250*	0.136*	3.23*
LASTINS	0.305*	0.215*	2.96*
PIPELINE	0.579	0.314*	2.21*
FLAT	0.593	0.313*	2.19*
HIERARCHY	**0.598**	**0.383**	**2.04**

Table 3.4: Accuracy of automatic insertion methods compared against the gold standard from Wikipedia's update log. The third column gives tree distance, where a lower score corresponds to better performance. Diacritic * ($p < 0.01$) indicates whether differences in accuracy between the given model and the Hierarchical model is significant (using a Fisher Sign Test).

LASTINS) fall substantially behind the more sophisticated, trainable strategies (PIPELINE, FLAT, HIERARCHICAL). Within the latter group, our HIERARCHICAL model slightly outperforms the others based on the coarse measure of accuracy at the section level. However, in the final paragraph-level analysis, the performance gain of our model over its counterparts is quite significant. Moreover, according to tree distance error, which incorporates error at both the section and the paragraph level, the performance of the HIERARCHICAL method is clearly superior. This result confirms the benefit of our selective update mechanism as well as the overall importance of joint learning.

Viewing human performance as an upper bound for machine performance highlights the gains of our algorithm. We observe that the gap between our method and human performance at the paragraph level is 32% smaller than that between the PIPELINE model and human performance, as well as the FLAT model and human performance.

Another conclusion from the analysis of human performance is that for many insertions more than one location is legitimate. Our evaluation is overly strict since it penalizes any variations from the gold standard. To provide another view of our model's performance, we manually analyze 50 insertions which did not match the gold standard. In 16 of these cases, the algorithm insertion was judged equally correct as the true insertion point. One typical pattern occurs when a sentence is inserted into an initial summary paragraph of the article, and instead our algorithm chooses a more specific location within a body of the article (or vice versa). See Figure 1-1 for an example where the inserted sentence would be appropriate in both the article summary as well as the more specific

"Career" section.

3.5.1 Sentence-level Evaluation

Until this point, we have evaluated the accuracy of insertions at the paragraph level, remaining agnostic as to the specific placement within the predicted paragraph. We perform one final evaluation to test whether the global hierarchical view of our algorithm helps in determining the *exact* insertion point. To make sentence-level insertion decisions, we use a local model in line with previous sentence-ordering work [28, 8]. This model examines the two surrounding sentences of each possible insertion point and extracts a feature vector that includes lexical, positional, and temporal properties. The model weights are trained using the standard ranking perceptron [12].

We apply this local insertion model in two different scenarios. In the first, we ignore the global hierarchical structure of the document and apply the local insertion model to every possible sentence pair. Using this strategy, we recover 24% of correct insertion points. The second strategy takes advantage of global document structure by first applying our hierarchical paragraph selection method and only then applying the local insertion to pairs of sentences within the selected paragraph. This approach yields 35% of the correct insertion points. This statistically significant difference in performance indicates that purely local methods are insufficient when applied to complete real-world documents.

3.6 Conclusion

We have introduced the problem of sentence insertion and presented a novel corpus-based method for this task. The main contribution of our work is the incorporation of a rich hierarchical text representation into a flexible learning approach for text structuring. Our learning approach makes key use of the hierarchy by selecting to update only the layer found responsible for the incorrect prediction. Empirical tests on a large collection of real-world insertion data confirm the advantage of this approach.

Chapter 4

Common Structure Induction

4.1 Overview

In the last decade, corpus-based approaches have been proven in the natural language generation (NLG) research community. These corpus-based approaches increased the scope and accuracy of automatic processing technology, but they typically require large amount of training data. Meanwhile, recent years have seen unprecedented interests in user generated content sites (e.g., review sites and shared community sites such as Wikipedia), where terabytes of annotated text data are generated. Experiments showed that NLG tools can enhance users' experiences with such sites, and thus the development and research on natural language generation systems has become not only research, but a very practical challenge. In this chapter, we will show a corpus-based generation method that can leverage collaboratively edited corpora to automatically induce common organization across documents.

Wikipedia – "The free encyclopedia that anyone can edit" – is the largest and most dominant general reference work currently available on the Internet. The power given to any users to edit and add unstructured data to Wikipedia leads to different wordings of section titles that mean the same thing. When a professional editor revises an encyclopedia, one needs to ensure that domain specific terms are consistent across all sections. This requires the information be analyzed and generated in a global way - to avoid inconsistent wording of section titles. Nevertheless, most

Wikipedia editors and existing NLG approaches make decisions about each document in isolation, suffering the consequent inconsistence across documents. Therefore new algorithms are required that are able to make global decisions.

The new algorithm has potential to benefit collaborative editing. When editing processes involve multiple documents in similar domains, it is advantageous to induce a common structure of the corpus that allows readers to browse and edit documents easily . To meet these needs, an automated method is necessary to infer common structures among multiple documents in similar domains.

The primary experimental focus of this chapter is to explain a method that creates a common organization of articles from Wikipedia where a document consists of a sequence of sections with various titles. We analyzed global models that cluster section titles into a representative set, alleviating the user's responsibility of upholding a consistent set of section titles as well as aligning sections across all documents. From these models, we design a graphical model to capture three kinds of important dependencies: similarity between section titles, co-occurrence of words, and statistical patterns of section-level topic structure. Our model assumes that section title transition probabilities are distributed according to a Hidden Markov Model (HMM) Topic model, and the transition probability distribution is estimated from the data.

In addition to experiments on mutually labeled data, we will examine this method on Wikipedia. Because we have the mappings from old section titles to new standardized section titles, we can upload these changes to Wikipedia in an effort to present a standard view of its articles. The model is designed to induce section topics and to learn topic clusters simultaneously. Viewing Wikipedia in this manner provides the users with more flexibility and power over the data. We implement the method, and perform initial experiments to validate our algorithms. We need some further experiments to prove the performance of the proposed method.

In the following sections, we define our task and formulate the problem. Then, we introduce our generative process, including learning and inference methods. Next, we present our experimental framework and preliminary result analysis. We conclude the chapter by presenting and discussing our methods.

4.2 Problem Formulation

In our task, we are given a number of documents from similar domain, where each document consists of several sections with title and text in each of these sections. Each section title may be divided into phrase units, which express semantical topics embedded in the text. Across the corpus, title phrase units may express the same topic. After being trained by accessing both section titles and section texts, our method is to predict the section topics for given section texts.

4.3 Model

The input of our model are nd documents $(d_1, d_2, \ldots, d_{dn})$, where the i-th document d_i has nt_i section titles $(y_1, y_2, ..., y_{nt_i})$ and section content $(w_1, w_2, ..., w_{nt_i})$. With that as input, we want to produce a representative set of k topics, where k is a parameter of the model. Each section is represented as a distribution of topics. The graphical model we designed captures three kinds of important dependencies: similarity between section titles, co-occurrence of words, and structural patterns of document topics. Our model assumes that section title transition probabilities are distributed according to a hidden Markov model, and the transition probability distribution is estimated from the data.

Our method trains the model on documents with both section titles and section texts. During training, the method learns a hidden topic model from the texts, and each section consists of a few topics reflected by section titles. At test time, the method takes documents without section titles as input. The hidden topic model of texts and hidden Markov model of section titles are used to determine the topics that each section in a document represents. Topics with top t highest likelihood in each section are predicted to present its semantic meaning.

4.4 Cross-document Structure Analysis

Our section level text analysis is through a probabilistic model based on the LDA model[6]. The mixture of topics per section in the LDA model is generated from a Dirichlet prior mutual to all

David Steel	**Megawati Sukarnoputri**	**Robert Wexler**	**Dingiri Wijetunga**	**Herbert London**
Early life	Early life	Early life	Early life	Early life
Political career	Political career	Political Career	Prime Minister	Professional life
Retirement	Presidency	Election results	Election History	Political campaigns
			President	

Table 4.1: Section titles in sample documents of our corpus.

documents in the corpus. This model captures correlations between section content words and section title words via the topics. Thus, rather than identifying a single topic for a section, the LDA model predicts a distribution over topics for each section. However, the assumption that the general order of sections in documents can be ignored is incorrect. We can see that *Early life* is followed by *Political career* with high likelihood from Table 4.1, which shows the section titles in a sample set of Wikipedia articles about politicians. We manually cluster the section titles of a different document collection sample. The Figure 4-1 shows that transitions between section titles among the whole corpus have some statistical patterns, which may help us detect document structure and align sections across multiple documents.

4.4.1 Generative Process

We assume that we cluster nd documents in to K topics. The model ties parameters between different documents by drawing θ of all documents from a common Dirichlet prior parameterized by α. We assume that some general patterns exist for topic transitions between consecutive sections across multiple documents. The parameters θ is estimated to model the transitions between section titles among the whole corpus. It also ties parameters between topics by drawing the vector ϕ_z of all topics from a common Dirichlet prior parameterized by β. In words, each topic z has a different language model parameterized by ϕ_z. The probabilistic model is given using plate notion in Figure 4-3.

The precise form of the model is as follows:

$$\theta \sim \text{Dirichlet}(\alpha) \tag{4.1}$$

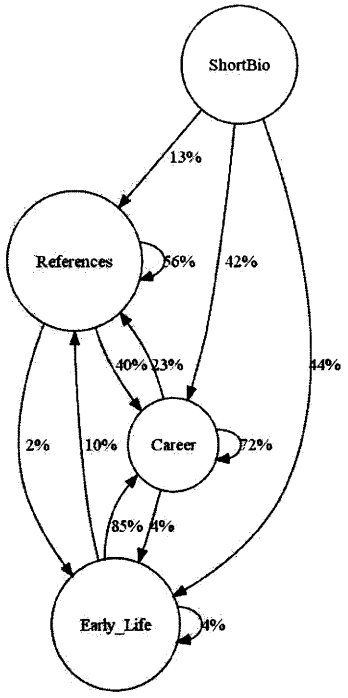


Figure 4-1: Topic transition of section titles (Mutually-clustered section titles).

For $z = 1 \ldots K$,
    Draw $\theta_z \sim Dirichlet(\alpha)$

For $d = 1 \ldots nd$,
    For $i = 1 \ldots nt_d$,
        Draw z_i from $z_i \mid z_{i-1} \sim$ Multinominals(θ)
        For $n = 1 \ldots nw$,
            Draw $w_n \sim Multinomial(\phi_{z_i})$.

Figure 4-2: Training algorithm for the graphical model.

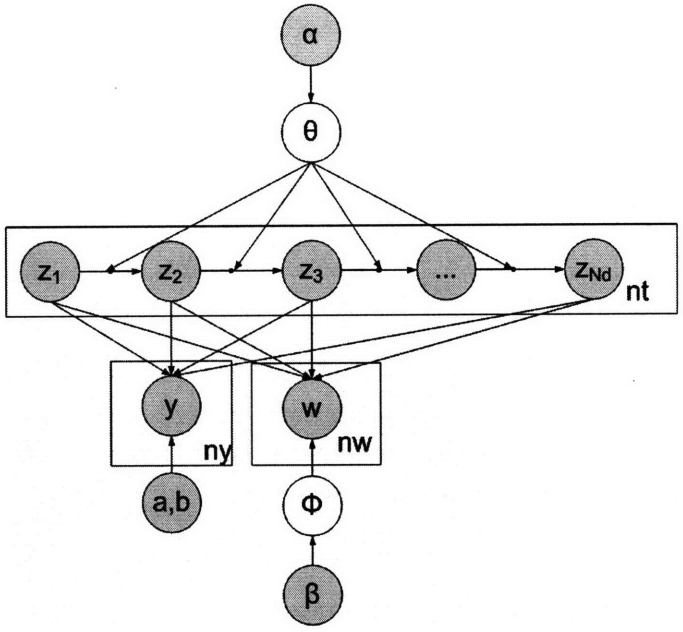


Figure 4-3: Graphical Model.

$$z_i \mid z_{i-1} = z \sim \text{Multinomial}(\theta) \tag{4.2}$$

$$sim_{i,j} = \begin{cases} \text{Beta}(1,2) & \text{if } y_i = y_j \ , \\ \text{Beta}(2,1) & \text{if } y_i = y_j \ . \end{cases} \tag{4.3}$$

$$\phi \sim \text{Dirichlet}(\beta) \tag{4.4}$$

$$w_i \sim \text{Multinomial}(\phi_{z_i}) \tag{4.5}$$

4.4.2 Learning & Inference

In general, the formula above is intractable between θ and ϕ. We use Gibbs Sampling to estimate parameters. Because Dirichlet priors are conjugate to multinomials, the conditional distribution

for y_i is given below. For each observation y_i in turn, we re-sample its state z_i conditioned on the states z_{-i} of the other observations, and the distribution of topic sequences converges to the desired posterior eventually.

The first step is to sample z_i according to the multinominals θ:

$$P(z_i \mid z_{-i}, y, \theta) = \frac{P(y \mid z_i, z_{-i}, y, \theta)P(z_i \mid z_{-i}, \theta)}{P(y \mid z_{-i}, \theta)}$$

$$\propto P(y \mid z, y, \theta)P(z_i \mid z_{-i}, \theta)$$

$$= P_{\text{beta}}(y; a_z, b_z)P(z_i \mid z_{-i}, \theta)$$

$$P(z_i \mid z_{-i}, \theta) = P(z_i \mid z_{-i}, \alpha)$$

$$= (\frac{n_{z_i, z_{i-1}+\alpha}}{n_{z_{i-1}]} + s\alpha})(\frac{n_{z_{i+1}, z_i} + I(z_{i-1} = z_i = z_{i+1}) + \alpha}{n_{z_i} + I(z_{i-1} = z_i)})$$

The second step is to sample w_z according to the multinominals ϕ, and ϕ is estimated as follows:

$$P(\phi \mid w, z, \beta) = \frac{P(w \mid z, \phi, \beta)P(\phi \mid z, \beta)P(w \mid z, \beta)}{P(w \mid z, \beta)}$$

$$\propto P(w \mid z, \phi, \beta)P(\phi \mid z, \beta)P(w \mid z, \beta)$$

$$= P_{\text{multinominal}}(w, \phi_z)P_{\text{Dirichlet}}(\phi_z; \beta)$$

$$\sim P_{\text{Dirichlet}}(\phi_z; \beta + \Sigma_{w \in z} w)$$

The final step is to re-estimate θ given the current topic states z:

$$P(\theta_i \mid z, \alpha) \propto P(z_i \mid z_{-i}, \theta_i)P(\theta_i \mid \alpha)$$

$$\sim P_{\text{Dirichlet}}(\alpha + n_{z_{i-1}, z_i})$$

4.5 Experiments

Corpus We evaluated our method based on a Wikipedia corpus. Our corpus consists of Wikipedia articles that belong to the category 'Living People.' Finally, we tested our method on the politician domain. At this domain, 120 documents are processed into two parts, with two thirds as training data and one third as testing data. Training data contain 814 sections, while testing data consist of 298 sections.

4.6 Data

A typical Wikipedia article has a document title, introductory text, content table, category information, and a document hierarchy with sections, subsections, paragraphs, and sentences. For parsing purposes, we simplify this, assuming a common layout across all articles. For simplicity, we remove sections with only tables, figures, or links. All the documents in the politician domain are then processed by our method.

4.6.1 Evaluation Measures

One of our goals is to cluster sections, such that each cluster corresponds to a well-defined topic. Based on the clustering results, further evaluation on common structure induction can be conducted via cross document validation. To infer the ground true for clustering, two judges collectively annotated 154 distinct section titles (see the appendix), which were extracted from all the documents in our corpus. Each judge was asked to cluster those documents into a few topics, assuming that sections with the same section titles are in the same topic. Preliminary evaluation shows that human agreement is comparatively low. To obtain a gold standard to evaluate various methods, a subset of the titles on which two judges agree are extracted. Finally, various methods are evaluated on seven agreed clusters (see Table 4.2).

Since only partial section title clusters are agreed across judges as a gold standard, a ground truth on document common organization can not be inferred. Instead of using common structure induction task, we evaluate our model based on a section alignment task. Two sections are aligned

Later career	**Honors**
Post-Presidency	Awards
Career after politics	Recognition
Life after politics	Decorations
Later life	Honours
Retirement	Leadership Achievements
Later years	
	Personal background
Early career	History
Early political career	Biography
Career before Politics	Background
Life before politics	Overview
Entry into politics	Personal life
	Personal
Family	Private Life
Family background	
	Childhood
Early life	Birth
Early years	Youth

Table 4.2: Seven topic clusters that are agreed between judges.

if at least a phrase from a section is from the same annotated cluster of a phrase from the other section. F-measure (the weighted harmonic mean of precision and recall) is used to evaluate various methods and baselines.

4.6.2 Baselines

We have two baseline methods to compare with. The first one is to compare with title generation method proposed by Branavan et. al[44] . The second one is to compare with unsupervised title clustering method.

We used the default settings of Branavan's model. The topic smooth prior, cluster prior, topic prior, and language model prior were equal to 0.1. All experiments were based on $5,000$ iterations. The number of topic clusters k were experimented from 10 to 40. The results were stable while the number of clusters k is increasing. We reached the best result when k equaled 30.

We used the most common unsupervised clustering method K-means as the first baseline. Doc-

	Precision	Recall	F-score
K-means (K = 1)	0.22	1	0.36
K-means (K = 3)	0.24	0.62	0.25
K-means (K = 10)	0.28	0.25	0.27
Branavan et al.	0.22 (K = 30)	0.98	0.36
CDHM-LDA (K = 27)	**0.26**	**0.85**	**0.38**

Table 4.3: Performance of various methods and baselines against a gold standard from human annotated clusters.

uments consist of a collection of independent section titles that generate some bodies of text which is modeled as a collection of independent words (i.e. bag of words). We measured the similarity between two bodies of texts as the cosine of the angle between the multidimensional feature vectors that represent the histograms of word counts of those texts. The k-means implementation in the Cluto[1] toolkit were used in our module.

4.6.3 Preliminary results

Some preliminary experiments were performed, and results are presented at Table 4.3. These numbers show that the Cross-document Hidden Markov and Latent Dirichlet Allocation model (CDHM-LDA) proposed in this thesis is comparable to various baseline methods. However, the CDHM-LDA method failed to generate improvements over the baselines. We need some further experiments to validate our proposed method.

[1] `http://glaros.dtc.umn.edu/gkhome/views/cluto/`

Chapter 5

Conclusions and Future Work

In this thesis, we introduced several problems arisen in the context of collaboratively editing. We explored two computational discourse models to solve two problems respectively. The first method is a hierarchical learning learning method based on Perceptron learning framework, while the second method incorporates both Bayesian topic model and Hidden Markov model.

The first model is a corpus-based model for incremental text structuring task. The hierarchical learning approach makes key use of the hierarchy by selecting to update only the layer found responsible for the incorrect prediction. The results of our experiments showed that our hierarchical update strategy produced more accurate predictions than two other hierarchical update strategies and all other baselines.

Next, we explained a model to induce a common structure across multiple collaboratively edited articles. The model bridges the advantage of a local Bayesian topic model and that of a global hidden Markov topic model, to improve the robustness of the hidden topics. We discussed the model's direct application in converting a collection of the articles in Wikipedia into a common structure with standardized section titles. Viewing Wikipedia in this manner provides the user with more flexibility and power over the data. We manually labeled the data, showed that our method reproduces as many original section alignments as baselines. However, further experiments are necessary to validate our method.

Sentence ordering algorithms too are likely to benefit from a hierarchical representation of text.

However, accounting for long-range discourse dependencies in the unconstrained ordering framework is challenging since these dependencies only appear when a particular ordering (or partial ordering) is considered. An appealing future direction lies in simultaneously inducing hierarchical and linear structure on the input sentences. In such a model, tree structure could be a hidden variable that is influenced by the observed linear order. We are also interested in further developing our system for automatic update of Wikipedia pages. Currently, our system is trained on insertions in which the sentences of the original text are not modified. However, in some cases additional text revisions are required to guarantee coherence of the generated text. Further research is required to automatically identify and handle such complex insertions. On anther direction, for those texts without hierarchical structure, we can first pre-process it by a structure induction process.

We have only run our preliminary experiments for the common structure induction task. We would like to evaluate its performance on a larger set, such as the entire Wikipedia. Furthermore, a new evaluation metric would be necessary to better evaluate performances of various methods on this task. We are interested in further development of our system to update Wikipedia by uploading the results of the standard section titles. Further research is required to improve the performance to handle more challenging cases on Wikipedia. We could also explore the area of human-assisted labeling and extend our current work by making Wikipedia collaborators help accepting or rejecting the changes. This common structure induction algorithm has a potential to identify short units of text that contain similar information even though they are written in different languages. Documents describing the same entity in different languages are linked on Wikipedia. The similar method can be utilized to analyze document structure of documents cross languages. First, given Chinese and English documents that describe the same event, our system can present sections that are supported by both the documents. Second, the system diversify both the Chinese and English documents by adding sections that are exclusively in only one language.

Appendix A

Examples of Sentence Insertions on Wikipedia

Given below are a few examples of Wikipedia articles and associated sentence insertions at various locations.

Santiago Calatrava

(Difference between revisions)

Revision as of 23:20, 21 October 2006 (edit)
80.103.196.233 (Talk)
← Older edit

Line 21:

Calatrava was born in [[Valencia (city in Spain)|Valencia]], [[Spain]], where he pursued undergraduate studies at the Architecture School and Arts and Crafts School. Following graduation in 1975, he enrolled in the [[Swiss Federal Institute of Technology]] (ETH) in [[Zürich|Zürich, Switzerland]] for graduate work in [[civil engineering]]. Calatrava was - influenced by the French/Swiss architect [[Le Corbusier]], whose [[Notre Dame du Haut]] chapel caused Calatrava to examine how complex form could be understood and generated in architecture. In 1981 after completing his doctoral [[thesis]], "On the Foldability of [[space frame|Space Frames]]", he started his architecture and engineering practice.

Revision as of 06:46, 22 October 2006 (edit) (undo)
69.167.58.163 (Talk)
Newer edit →

Line 21:

Calatrava was born in [[Valencia (city in Spain)|Valencia]], [[Spain]], where he pursued undergraduate studies at the Architecture School and Arts and Crafts School. Following graduation in 1975, he enrolled in the [[Swiss Federal Institute of Technology]] (ETH) in [[Zürich|Zürich, Switzerland]] for graduate work in [[civil engineering]]. Calatrava was influenced by the French/Swiss architect [[Le Corbusier]], whose [[Notre Dame du Haut]] chapel caused Calatrava to examine how complex form could be understood and generated in architecture.
+ **Many of his structural forms (parabolic arches, branching columns, ruled surfaces) appear to be inspired by fellow countryman [[Gaudi]] with obvious homages paid in his Bodegas Ysios and his design for the completion of St. John the Divine in Brooklyn.** In 1981 after completing his doctoral [[thesis]], "On the Foldability of [[space frame|Space Frames]]", he started his architecture and engineering practice.

Revision as of 06:46, 22 October 2006

Santiago Calatrava Valls (born July 28, 1951) is a Spanish architect whose work has become increasingly popular worldwide.

Calatrava was born in Valencia, Spain, where he pursued undergraduate studies at the Architecture School and Arts and Crafts School. Following graduation in 1975, he enrolled in the Swiss Federal Institute of Technology (ETH) in Zürich, Switzerland for graduate work in civil engineering. Calatrava was influenced by the French/Swiss architect Le Corbusier, whose Notre Dame du Haut chapel caused Calatrava to examine how complex form could be understood and generated in architecture. Many of his structural forms (parabolic arches, branching columns, ruled surfaces) appear to be inspired by fellow countryman Gaudi with obvious homages paid in his Bodegas Ysios and his design for the completion of St. John the Divine in Brooklyn. In 1981 after completing his doctoral thesis, "On the Foldability of Space Frames", he started his architecture and engineering practice.

Calatrava's unique, creative, and highly influential style combines a striking visual architectural style that interacts harmoniously with the rigid principles of engineering. His work often draws on form and structure found in the natural world, and can be described as anthropomorphic. His works have elevated the design of some civil engineering projects such as bridges to new heights. He has designed numerous train stations, heralded for their bright, open, and easily-traveled spaces.

While he is primarily known as an architect, Calatrava is also a prolific sculptor and painter, claiming that the practice of architecture combines all the arts into one.

Calatrava's first United States work was the Quadracci Pavilion addition to the Milwaukee Art Museum.

One of his newest projects is a residential skyscraper named "80 South Street" after its own address, composed of 10 townhouses in the shape of cubes stacked on top of one another. The townhouses move up a main beam and follow a ladder-like pattern, providing each townhouse with its own roof. The "townhouse in the sky" design has attracted a high profile clientele, willing to pay the hefty US$30 million for each cube. It will be built in New York City's financial district facing the East River.

He has also designed a skyscraper for 400 North Lake Shore Drive in Chicago, formerly known as the Fordham Spire. Originally commissioned by Chicagoan Christopher Carley, Irish developer Garrett Kelleher purchased the building site for the project in July of 2006 when Carley's financing plans fell through. Kelleher is currently in negotiations with Carley and Calatrava to purchase Calatrava's design for the building. Kelleher's close working relationship with the Anglo Irish Bank, and his own wealth which will allow him to personally finance 100 percent of the equity in the project, will make it easier for Kelleher to build this project than it was for Carley. Kelleher plans to begin construction of the building in Spring of 2007 for completion in 2010. When completed, 440 North Lakeshore Drive will, at 2,000 feet tall, be the tallest building in North America.

Calatrava has also designed three bridges that will eventually span the Trinity River in Dallas, the first of which will commence construction in December, 2005. When completed (target date 2010), Dallas will join the Dutch county of Haarlemmermeer in having three Calatrava bridges.

Calatrava was awarded the Eugene McDermott Award in the Arts. The Award is among the country's most esteemed arts awards. Established to honor Eugene McDermott, founder of Texas Instruments and long-time friend and benefactor to MIT, the award was created by the Council for the Arts at MIT in 1974, and further endowed by Eugene's wife, Margaret. Since its inception, the Council has bestowed the award upon 31 individuals producing creative work in the performing, visual and media arts, as well as authors, art historians and patrons of the arts.

His nephew Alex Calatrava is a professional tennis player.

Figure A-1: Screenshot of an insertion between two consecutive revisions on a Wikipedia article on *Santiago Calatrava*. An insertion sentence is shown in red.

Amanda Bynes

Revision as of 18:04, 19 June 2006 (edit) Jack O'Lantern (Talk | contribs) m ← Older edit	Revision as of 23:31, 19 June 2006 (edit) (undo) 69.196.252.112 (Talk) Newer edit →
Line 14:	**Line 14:**
|}'''Amanda Laura Bynes''' (born [[April 3]], [[1986]]) is an [[United States|American]] [[actor|actress]] and former show host on [[Nickelodeon (TV channel)|Nickelodeon]]. After appearing in several successful television series on Nickelodeon in the late [[1990s]] and early [[2000s]], Bynes has moved into a film career, starring in several films aimed at teenage audiences, including her latest, "[[She's the Man]]".	|}'''Amanda Laura Bynes''' (born [[April 3]], [[1986]]) is an [[United States|American]] [[actor|actress]] and former show host on [[Nickelodeon (TV channel)|Nickelodeon]]. After appearing in several successful television series on Nickelodeon in the late [[1990s]] and early [[2000s]], Bynes has moved into a film career, starring in several films aimed at teenage audiences, including her latest, "[[She's the Man]]". **Bynes is now working on the movie called "hairspary" set to be released in the summer of 2007**

Revision as of 23:31, 19 June 2006

Amanda Laura Bynes (born April 3, 1986) is an American actress and former show host on Nickelodeon. After appearing in several successful television series on Nickelodeon in the late 1990s and early 2000s, Bynes has moved into a film career, starring in several films aimed at teenage audiences, including her latest, *She's the Man*. **Bynes is now working on the movie called "hairspary" set to be released in the summer of 2007**

Bynes has been described as having an "Everygirl" appeal, embodying "both everything her teen fans dream of being and everything they know they really are, and they love her for it."[1] In 2006, she was named one of Teen People's "25 Hottest Stars Under 25".[2]

Biography

Early life

Bynes was born in Thousand Oaks, California to Richard Bynes (a dentist who also practiced stand-up comedy) and Lynn Organ (a dental assistant). She has two older siblings, Tommy, a chiropractor, and Jillian, who has a psychology degree from UCLA. Bynes grew up "half Catholic and half Jewish"[3] and identifies herself as Jewish.[4]

Career

Bynes, who was trained as an actress by Arsenio Hall and Richard Pryor at a comedy camp, started acting at the age of seven, appearing in a television advertisement for Buncha Crunch candies. After taking acting classes, she began her acting career as a regular cast member of the show *All That* in 1996 and later became the star of *The Amanda Show*, both on the Nickelodeon channel. In 2002, Bynes also voiced a character in five episodes of the Nickelodeon series *Rugrats*.

Bynes made her film debut in 2002's mild box office success, *Big Fat Liar*, where she played opposite Frankie Muniz. Her first solo leading role was in 2003's *What a Girl Wants*, which also performed fairly at the box office with her co-star Oliver James. Subsequently, Bynes starred in the WB Network's sitcom *What I Like About You* and had a voice part in 2005's CGI animated comedy, *Robots*.

Bynes came into the public eye during the time period when popular teen actresses Lindsay Lohan and Hilary Duff became well-known. Although she is often compared with them, Bynes has commented that "It's like being the hot girl at the high school party. I was never that girl. I grew up with terrible acne and feeling insecure. I was tall and skinny. I didn't feel pretty at all, and guys didn't even like me. That's why I got into comedy."[5] Bynes has also said that her relatability to teenage audiences stems from the fact that she is "more similar to them than some... socialite or whatever".[1]

Her newest film is *She's the Man*, a comedy based on William Shakespeare's *Twelfth Night*. In the film, which opened in March 2006, Bynes disguises herself as her brother in order to join an elite boarding school. Producers had originally wanted to cast singer Jesse McCartney as Bynes' brother, noting a physical resemblance between McCartney and Bynes disguised as a boy, but McCartney was unavailable.[6]

Bynes has completed filming on another romantic comedy *Lovewrecked*, which was shot before *She's the Man* but will be released after it, at some point in 2006. She was also recently cast as Penny Pingleton in *Hairspray*, a new film adaptation of the Broadway musical of the same name.

Bynes has commented that she wishes to start appearing in more mature roles and believes that she is still developing her acting skills and maturing as an actress, saying that she is "getting better" as she matures.[1]

Personal life

Bynes, who has a dog named Midge and drives a white Lexus LS430, graduated Thousand Oaks High School's independent study program and has expressed a desire to attend New York University in the near future. She briefly moved into an apartment in Hollywood, California, but has since returned to her family home.

Bynes is interested in drawing (she once painted a portrait of David Letterman as a gift to him) and fashion design, having commented that "I'm the girl whose biggest nightmare would be to lose my makeup bag while traveling".[5]

Figure A-2: Screenshot of an insertion between two consecutive revisions on a Wikipedia article on *Amanda Bynes*. An insertion sentence is shown in red.

Chen Lu

From Wikipedia, the free encyclopedia

(Difference between revisions)

• *Interested in contributing to Wikipedia?* •

Jump to: navigation, search

Revision as of 08:37, 29 December 2005 (edit)	**Revision as of 08:43, 29 December 2005 (edit) (undo)**
67.161.180.79 (Talk)	67.161.180.79 (Talk)
(→To the Top of the World)	(→Her Struggles)
← Older edit	Newer edit →
Line 34:	**Line 34:**
Lu struggled after her win at the World Championships. Although she managed top two finishes at all three events she entered in the fall of 1995, she skated inconsistently. For example, at a competition in France she finished 7th in the technical program and 1st in the free skate. Moreover, Michelle Kwan of the United States was attracting a lot of attention and praise, had won three events during the inaugural Gran Prix Series (then known as the Champion Series), and had defeated Lu Chen at Skate America. The low point of Lu's season came at the Champion Series Final, where she led going into the free skate but dropped to 4th overall after struggling with her jumps. Therefore, many doubted Lu could repeat as the World Champion.	Lu struggled after her win at the World Championships. Although she managed top two finishes at all three events she entered in the fall of 1995, she skated inconsistently. For example, at a competition in France she finished 7th in the technical program and 1st in the free skate. Moreover, Michelle Kwan of the United States was attracting a lot of attention and praise, had won three events during the inaugural Gran Prix Series (then known as the Champion Series), and had defeated Lu Chen at Skate America. The low point of Lu's season came at the Champion Series Final, where she led going into the free skate but dropped to 4th overall after struggling with her jumps. Therefore, many doubted Lu could repeat as the World Champion.
At the 1996 World Championships, Lu Chen skated very well--better than she had skated all season--but she finished 2nd overall to Michelle Kwan of the USA. Both skated well and both garnered two perfect marks of 6.0 for Presentation, but Kwan had edge on the technical scores and won by a vote of 6 judges to 3. Although it remains a hotly debated result, Kwan landed seven triple jumps to Lu's 6 and had harder and more varied spins, which may have been the basis for Kwan's higher technical scores.	**Lu Chen won the 1996 Winter Asian Games, a prestigious competition that occurs every four years before the World Championships.** At the 1996 World Championships, Lu Chen skated very well--better than she had skated all season--but she finished 2nd overall to Michelle Kwan of the USA. Both skated well and both garnered two perfect marks of 6.0 for Presentation, but Kwan had edge on the technical scores and won by a vote of 6 judges to 3. Although it remains a hotly debated result, Kwan landed seven triple jumps to Lu's 6 and had harder and more varied spins, which may have been the basis for Kwan's higher technical scores.

Revision as of 08:43, 29 December 2005

Lu Chen (Simplified Chinese: 陈露, pinyin: Chén Lù) (born 24 November 1976 in Changchun, China) is a Chinese figure skater. Her mother was a table tennis player and her father was an ice hockey coach. She was coached by Li Minzhu. She is China's most successful woman figure skater. She is called "Butterfly on Ice" by the fans and media in China for her popular performance to *Butterfly Lovers' Violin Concerto*.

Amateur Career

Early Success

Lu Chen became one of the most decorated figure skaters of the 1990s winning two Olympic and four World medals. Her success brought attention to Chinese figure skating and spurred more Chinese success.

As a young skater in the early 90's, Lu demonstrated her tremendous athletic abilities. For example, she landed seven triple jumps, including a triple Toe Loop/triple Toe Loop combination at the 1991 World Championships held in Munich, Germany. In fact, in the free skating portion of the event, she landed more triple jumps than any of the top 5 finishers. But, Lu also demonstrated great artistic potential and her skating was praised by such American commentators as Scott Hamilton and Sandra Bezic.

Later that year, she became the first Chinese figure skater to compete in the United States when she finished 4th at the Skate America competition held in Oakland, California. This finish was particularly strong considering that the competition also included Tonya Harding and Kristi Yamaguchi, the world's top two skaters at the time.

In 1992, she had even greater successes, winning the bronze medal at the Junior World Championships. She then went on to shock the skating world with a surprising 6th place finish at the Olympics. At those Olympics, she was one of a few skaters that attempted a triple Lutz combination in the technical program. Although she landed the difficult combination, she had problems executing other required elements and was ranked 11th after the opening phase of the competition. In the longer Free Skate, she landed six triple jumps, more than any of the skaters that finished ahead of her. Also, Lu's performance was remarkable in that she was the only top 6 skater that did not fall on a jump.

To the Top of the World

She followed her Olympic success with bronze medals at the 92 and 93 World Championships, the first two won by a Chinese figure skater. In 1994 she became the first Chinese figure skater to medal at the Olympic games, winning the bronze medal for a performance that included five triple jumps skated to the soundtrack from Nausicaa by Joe Hisaishi. This success was largely overshadowed by the Kerrigan/Harding controversy surrounding the Olympics.

After the 1994 Olympics, Nancy Kerrigan and Oksana Baiul (the Olympic Silver and Gold medalists, respectively) retired from amateur competition and Lu became the favorite to win the World title in 94. However, a stress fracture injury kept her out of the competition and jeopardized her career. She made a successful comeback in the fall winning the NHK Trophy in Japan. In 1995, she became the World Champion (another first for a Chinese skater) over Surya Bonaly of France and younger competitors from the USA.

Figure A-3: Screenshot of an insertion between two consecutive revisions on a Wikipedia article on *Chen Lu* (part 1). An insertion sentence is shown in red.

Her Struggles

Lu struggled after her win at the World Championships. Although she managed top two finishes at all three events she entered in the fall of 1995, she skated inconsistently. For example, at a competition in France she finished 7th in the technical program and 1st in the free skate. Moreover, Michelle Kwan of the United States was attracting a lot of attention and praise, had won three events during the inaugural Gran Prix Series (then known as the Champion Series), and had defeated Lu Chen at Skate America. The low point of Lu's season came at the Champion Series Final, where she led going into the free skate but dropped to 4th overall after struggling with her jumps. Therefore, many doubted Lu could repeat as the World Champion.

Lu Chen won the 1996 Winter Asian Games, a prestigious competition that occurs every four years before the World Championships. At the 1996 World Championships, Lu Chen skated very well--better than she had skated all season--but she finished 2nd overall to Michelle Kwan of the USA. Both skated well and both garnered two perfect marks of 6.0 for Presentation, but Kwan had edge on the technical scores and won by a vote of 6 judges to 3. Although it remains a hotly debated result, Kwan landed seven triple jumps to Lu's 6 and had harder and more varied spins, which may have been the basis for Kwan's higher technical scores.

The second place finish at the World Championships was not what Lu had wanted. Her skating deteriorated further as she struggled with injury and conflict with both her long-time coach and her skating federation. She withdrew from competitions in the fall of 1996, citing injury and was ill-prepared for the 1997 World Championships. There, she finished only 25th in the World and did not qualify for the final free skate. Nor did her finish qualify China for the figure skating competition at the 1998 Olympic Games.

Thus, in the fall of 1997, Lu, working with a new coach, had to qualify for the Olympics. She did this by winning an event in Vienna and by finishing 4th and 3rd at events in France and Japan, respectively. Still, she had not regained the form that had won her the World title and many doubted she could win a medal. This was because Michelle Kwan and Tara Lipinski (the top two skaters in the world at the time) seemed likely to occupy the top two spots on the podium and there were many other skaters that could challenge for the bronze medal.

Her Comeback and Farewell

At the Olympics, Lu announced her intention to retire from amateur skating after the Olympic games. Thus, her performances, took on a special significance as a comeback and as a farewell. She performed well to "Adios Noninos" in her technical program and to "Butterfly Lovers" in the free skate. Although she had struggled before the Olympics and had to fight to land her triple jumps during the competition, she was able to complete the two programs well enough to compete for a medal. But, she faced intense competition from two Russian competitors, Maria Butyrskaya and Irina Slutskaya. They too skated well, but, like Lu, made mistakes. The final placements were very close and far from unanimous. Lu beat Irina Slutskaya by the vote of 6 judges to 3 and beat Maria Butyrskara 5 judges to 4. In fact, most of the judges had Lu in 4th place, away from a medal. But, the votes for 3rd were split unevenly and because Irina and Maria both received many 4th and 5th place ordinals as well and each received few 3rd place votes, it was enough for Lu to win the Bronze medal.

Her performance was regarded as one of the great comebacks of the Olympic games and is memorable for the emotion she displayed both during and after her free skate. Immediately after the free skate, she bowed to the audience and to her coaches. She then retired from amateur skating and turned professional.

Professional life

She toured with Stars on Ice for two seasons. In July 2005 she married Denis Petrov, a Russian and 1992 Winter Olympics pairs skating silver medalist. Chen is now chief director of an ice skating club named after her in Shenzhen. She likes the job very much and hopes to train more skating athletes for the city and China.

Figure A-4: Screenshot of an insertion between two consecutive revisions on a Wikipedia article on *Chen Lu* (part 2). An insertion sentence is shown in red.

CM Punk

(Difference between revisions)

Revision as of 11:46, 2 July 2006 (view source)
Lid (Talk | contribs)
m (→Total Nonstop Action Wrestling and Ring of Honor)
← Older edit

Line 25:

During his early career Punk cracked his skull when he attempted one of his signature moves, a [[Professional wrestling aerial techniques#Flying neckbreaker|springboard corkscrew flipping neckbreaker]], during a match against his opponent Reckless Youth. Youth was one step too far in and Punk didn't correctly complete the corkscrew leading to Youth's head ending up on top of Punk's head rather than next to it cracking his skull when they impacted on the mat. He refused [[Analgesic|pain medication]] due to his Straight Edge lifestyle. Doctors told him not to strain himself for a year. However, Punk returned to wrestling only a few months later.

==Total Nonstop Action Wrestling and Ring of Honor==

Revision as of 12:13, 2 July 2006 (view source)
Lid (Talk | contribs)
(→Early career)
Newer edit →

Line 25:

During his early career Punk cracked his skull when he attempted one of his signature moves, a [[Professional wrestling aerial techniques#Flying neckbreaker|springboard corkscrew flipping neckbreaker]], during a match against his opponent Reckless Youth. Youth was one step too far in and Punk didn't correctly complete the corkscrew leading to Youth's head ending up on top of Punk's head rather than next to it cracking his skull when they impacted on the mat. **Despite the injury Punk managed to complete the rest of the match before going to hospital**. He refused [[Analgesic|pain medication]] due to his Straight Edge lifestyle. Doctors told him not to strain himself for a year. However, Punk returned to wrestling only a few months later.

==Total Nonstop Action Wrestling and Ring of Honor==

Revision as of 12:13, 2 July 2006

Phil Brooks (born October 26, 1978 in Chicago, Illinois), better known by his ring name **CM Punk**, is an American professional wrestler currently working for World Wrestling Entertainment (WWE) in its ECW brand and developmental territory Ohio Valley Wrestling. He is the current reigning OVW Heavyweight Champion.
His straight edge gimmick reflects his lifestyle, taken to a more extreme level in the ring. As CM Punk says to his opponents, "*Straight Edge means I'm drug-free, alcohol-free, and better than you!*" As an inside joke to Punk's fans, asking him what "C.M." stands for yields a different answer each time. The most widely-believed theory is that Brooks was in a tag team early in his backyard wrestling years called the Chick Magnets[1], although Dave Prazak has once said it stood for Chuck Mosley. Another common belief is that the C.M stands for "Clean Made" due to his straight-edge gimmick and actual lifestyle mixed with the typicial stereotype of a "punk".

Career

Early career

Following a stint in a backyard wrestling federation called the Lunatic Wrestling Federation in the mid-late 90's, Brooks trained as a wrestler at the *Steel Domain*. He was mostly trained by Ace Steel. It was in the Steel Domain that he met Scott Colton, who later adopted the stage name Colt Cabana. Brooks befriended Colton and spent most of the time working in the same independent promotions as opponents or allies. Punk's home promotion for his early career could be considered IWA Mid-South. He, along with Colt Cabana, Paul/Chuck E. Smoothe, and manager Dave Prazak, formed the Gold Bond Mafia. Punk, Cabana, and Smoothe (along with Chris Hero, among others) became some of IWA Mid-South's hottest young wrestlers. CM Punk also feuded with both Cabana and Hero. Punk's matches with Cabana led him to getting a job in Ring of Honor. Punk's matches in IWA Mid-South are not only considered to have given Punk recognition, but are also credited for putting IWA Mid-South on the map. In his last match in IWA Mid-South, Punk faced Delirious in a 60-minute draw. Half-way through the match, Punk exposed his butt in memory of Chris Candido.
During his early career Punk cracked his skull when he attempted one of his signature moves, a springboard corkscrew flipping neckbreaker, during a match against his opponent Reckless Youth. Youth was one step too far in and Punk didn't correctly complete the corkscrew leading to Youth's head ending up on top of Punk's head rather than next to it cracking his skull when they impacted on the mat. Despite the injury Punk managed to complete the rest of the match before going to hospital. He refused pain medication due to his Straight Edge lifestyle. Doctors told him not to strain himself for a year. However, Punk returned to wrestling only a few months later.

Total Nonstop Action Wrestling and Ring of Honor

Image:Punkcabana.jpg
Punk with Colt Cabana as the ROH Tag Team Champions.

Punk joined Ring of Honor and started climbing the ranks, winning the ROH Tag Team Championship twice with Colt Cabana as the **Second City Saints**. Punk also joined NWA:Total Nonstop Action, where he was soon paired with Julio Dinero as sidekicks for Raven. Punk's feud with Raven in RoH was a big success, and soon they were feuding in TNA. Raven and Punk were originally slated to face off in a hair versus hair match in RoH. TNA objected and insisted the match to be had in their promotion with Raven facing Shane Douglas instead of Punk. ROH ended up arranging a steel cage match between them instead. Punk's time in TNA ended when he had a scuffle with Teddy Hart outside of a restaurant shortly after a TNA show. Punk was still under contract, but was not used, though, according to Punk, TNA was intending to use him again. However, when TNA wanted him to stop working ROH shows, he refused, and thus never returned. The contract ran out almost a year later.
Several months after being released from TNA, Punk faced off against ROH World Heavyweight Champion Samoa Joe for the championship in a three match series in what is considered one of the best trilogies in wrestling.
The first match, on June 12, 2004 in Dayton, Ohio, resulted in a 60 minute time-limit draw. The show was named *World Title Classic*. The second encounter was scheduled for Punk's hometown of Chicago, Illinois. At *Joe vs. Punk II* on October 16, they wrestled to a second 60 minute draw once again. In addition to becoming Ring of Honor's best selling DVD at that point, the

Figure A-5: Screenshot of an insertion between two consecutive revisions on a Wikipedia article on *CM Punk* (part 1). An insertion sentence is shown in red.

match received the first 5-star rating by Dave Meltzer's Wrestling Observer newsletter for a match in North America since 1997. Joe finally won the series in the third match that December. Those matches were very highly rated, and resulted in much stir and raise in ROH's publicity. It is generally thought that the company would not have survived its slump without Punk. Punk had been laid off from his previous full-time job as a laboratory technician for Underwriters Laboratories. Soon after the lay off, he became the head trainer for ROH's wrestling school. Punk has stated that he never had the credentials for the laboratory work, and only worked there so he could afford wrestling.
On February 25, 2005 Punk began a heated feud with Jimmy Rave of The Embassy. The feud stemmed from a match between Spanky and Punk from February 19 in which Prince Nana invited Spanky into the Embassy which Spanky rejected. After which Nana and Punk got into an argument after which Punk beat up both Nana and the Outcast Killas. On February 25 Punk had a match with Alex Shelley after which he was attacked by Jimmy Rave and Fast Eddie Vegas leading to a grudge match between Punk and Rave on February 26.

Image:Punkrohchampion.jpg
Punk as the ROH World Heavyweight Champion.

The grudge match ended after Rave blinded Punk with flyspray while a member of the Embassy distracted the referee. After the match Rave proceeded to give Punk's valet, Traci Brooks, the Rave Clash. Later, after a tag match with Colt Cabana against Rave and Fast Eddie Vegas which they lost after significant interference on the part of the Embassy, The Embassy held down Punk and Rave attempted to remove Punk's straight edge tattoo on his stomach with a cheese grater. On April 16 Punk wrestled against Mike Kruel who was subbing for the absent Jimmy Rave who Prince Nana claimed was injured after "falling off an elephant in Africa". Punk defeated Kruel but was soon attacked by Jimmy Rave who ran through the crowd and jumped Punk which allowed the Embassy to hang Punk with a steel chain from the top rope to the outside of the ring. Following that incident Rave defeated Punk for a third time in a dog collar match at *Manhattan Mayhem* after five chairshots to the head. Punk finally defeated Rave in a steel cage match after a suplex off the top of the cage.
In June of 2005, CM Punk accepted a deal with World Wrestling Entertainment, after wrestling try-out matches on its *Sunday Night HEAT* show. Even though he had accepted the deal, CM Punk went on to defeat Austin Aries to win the ROH World Heavyweight Championship at *Death Before Dishonor III*.
This started a critically acclaimed angle where Punk threatened to bring the ROH title to the WWE with him. For weeks, Punk teased the ROH locker room and the ROH fans, proving his great versatility as a performer by garnering a great amount of heel heat for a popular wrestler. At each ROH show, it seemed like Punk was set to lose the title, but would continue to hold on to the belt for the next show. A notable part of this angle was Mick Foley making several ROH appearances, attempting to convince Punk to do the right thing and defend the title on his way out.
On August 12, 2005 CM Punk lost his ROH World Title to James Gibson in Dayton, Ohio in a four corner elimination match consisting of himself, Gibson, Samoa Joe and Christopher Daniels. Austin Aries replaced Punk as head trainer at the ROH Wrestling School, leaving him no commitments and making him available to leave ROH. His final match in Ring of Honor took place on August 13, 2005 against his good friend Colt Cabana at *Punk:The Final Chapter*.
He reappeared on February 11, 2006, due to a severe snowstorm which prevented several ROH wrestlers from attending. Punk asked for permission from WWE official Tommy Dreamer to appear so he could help out ROH. Dreamer approved and Punk appeared during the night to fill gaps where others were supposed to appear. In the main event, he teamed with Bryan Danielson as a replacement for Low Ki to wrestle Jimmy Rave and Adam Pearce.

Ohio Valley Wrestling

Punk was assigned to Ohio Valley Wrestling, a WWE developmental territory. On September 26, 2005 in his OVW television debut, Punk suffered a ruptured eardrum and broken nose at the hands of Danny Inferno, after he was hit by a overly stiff right hand. Despite the injury, Punk finished the match and quickly recovered.

Image:Punkwinsovw.jpg
Promotional banner of Punk winning the OVW Heavyweight Championship

On November 9, 2005, Punk became the OVW Television Champion after defeating Ken Doane. This led immediately to a feud between Punk and Brent Albright, who had previously been feuding with Doane for the television championship and had lost his chance to wrestle Doane because Punk had hit him with a chair so he himself could wrestle Doane. This led to a series of matches, including one which ended in overtime with Albright having Punk submit to the Crowbar. However, Punk was able to keep the title as he didn't agree to the extra time. On January 4, 2006, Punk lost the TV Title in a rematch during a three way dance between him, Albright and Doane. Doane was injured halfway through the match and could not continue. Aaron Stevens then came in the match to replace Doane. Punk then submitted to Albright's Crowbar, and after interference by Punk, Stevens was able to get the pin on Albright to become the new OVW Television Champion.
The feud continued after a short lapse where Albright and Punk became a tag team, but that all changed when Albright wanted the respect of Punk who would never give it to him and instead proceeded to "punk out" Albright repeatedly. This continued for weeks with Punk always getting the better of Albright until a double turn occurred on February 1, 2006 when Albright turned heel during a tag match allowing The Spirit Squad to destroy Punk and, in doing so, turning Punk face.
During this time CM Punk had a minor appearance at WrestleMania 22 on April 2, 2006 as one of the gangsters who rode a 1930s era car to the ring before John Cena's entrance.[2]
When Matt Cappotelli vacated the OVW Heavyweight Championship due to brain cancer, a tournament was held to crown a new champion. The finals were Brent Albright vs CM Punk with Albright defeating Punk to become the new champion. Punk and Albright continued their feud with Albright becoming more and more unstable and paranoid about maintaining his championship after several close call matches against Punk, resulting in acts such as threatening Maria. On May 3, 2006, Punk finally defeated Brent Albright in a strap match to win the OVW Heavyweight Championship. Punk also had some minor feuds, retaining the title in matches against Shad Gaspard & Ken Kennedy.

Extreme Championship Wrestling

On June 24, 2006 Punk debuted on ECW during a house show at the ECW arena, defeating Stevie Richards.

Figure A-6: Screenshot of an insertion between two consecutive revisions on a Wikipedia article on *CM Punk* (part 2). An insertion sentence is shown in red.

Chris Chelios

Revision as of 21:29, 11 February 2006 (edit)
136.176.185.118 (Talk)
(→International Play)
← Older edit

Line 4:

==Playing Career==

Chelios was drafted by the [[Montreal Canadiens]] in the [[1981]] [[NHL Entry Draft]]. Prior to being drafted, he played for the [[Moose Jaw Canucks]] of the [[SJHL]]. He played for two years at the [[University of Wisconsin-Madison|University of Wisconsin]] after being drafted. In [[1983]], he was named to the [[All-Tournament Team]] and the Second [[WCHA All-Star Team]]. He made his debut for the Canadiens during the following season, playing twelve games in the regular season and 15 in the playoffs.

Revision as of 02:23, 13 February 2006 (edit) (undo)
69.164.55.213 (Talk)
Newer edit →

Line 4:

==Playing Career==

Chelios was raised in [[Southern California]] and was a standout youth hockey player. Chelios was drafted by the [[Montreal Canadiens]] in the [[1981]] [[NHL Entry Draft]]. Prior to being drafted, he played for the [[Moose Jaw Canucks]] of the [[SJHL]]. He played for two years at the [[University of Wisconsin-Madison|University of Wisconsin]] after being drafted. In [[1983]], he was named to the [[All-Tournament Team]] and the Second [[WCHA All-Star Team]]. He made his debut for the Canadiens during the following season, playing twelve games in the regular season and 15 in the playoffs.

Revision as of 02:23, 13 February 2006

Christos K. Chelios (born January 25, 1962, of Greek origin in Chicago, Illinois) is a defenseman for the Detroit Red Wings of the NHL. He has earned many awards during his long career, and is considered one of the best Americans to ever play in the NHL. With the retirement of Mark Messier, Chelios has become the oldest active player in the NHL.

Playing Career

Chelios was raised in Southern California and was a standout youth hockey player. Chelios was drafted by the Montreal Canadiens in the 1981 NHL Entry Draft. Prior to being drafted, he played for the Moose Jaw Canucks of the SJHL. He played for two years at the University of Wisconsin after being drafted. In 1983, he was named to the All-Tournament Team and the Second WCHA All-Star Team. He made his debut for the Canadiens during the following season, playing twelve games in the regular season and 15 in the playoffs.

In 1984, he made the team for good, and distinguished himself with his play. He earned a trip to the National Hockey League All-Star Game and was named to the 1985 NHL All-Rookie Team. He scored 64 points in 74 games, a high total for a defenseman, even in the higher-scoring 1980s. In the playoffs that year, he scored 10 points in games, with a +17 plus/minus. Although he only played 41 games in the 1985-1986 season, he won his first Stanley Cup, playing in front of Conn Smythe Trophy winner Patrick Roy.

Following two more good seasons, Chelios really broke out in the 1988-1989 season. He scored 73 points in 80 games at +35, was named to the All-Star First-Team, and won the James Norris Memorial Trophy. After he only played 53 games in the next season, on June 29, 1990, he was traded to the Chicago Blackhawks with a 2nd-round draft pick for Denis Savard, who is now in the Hockey Hall of Fame.

In his first season with Chicago, he continued to score at his usual rate, tallying 64 points, and earned a spot on the Second NHL All-Star Team. After a slightly less offensively impressive season (although he had a very good playoffs), Chelios was in top form for the 1992-1993 season. He scored 73 points and won another Norris Trophy. In 1996, he would win it again. All told, he won three Norris Trophies, was named to 3 First All-Star Teams and 2 Second All-Star Teams, and played in 6 All-Star Games as a Blackhawk. He was captain of the Blackhawks from 1995 to 1999.

By 1999, though, Chelios was starting to show signs of age. At 37, his career was clearly in decline, and he was no longer the offensive and defensive force he had once been. However, even if he did not have much to offer the Blackhawks, he could still help teams with his veteran leadership and his largely-remaining talent. On March 23, he was traded to the Detroit Red Wings for Anders Eriksson and two first-round draft picks.

The move to Detroit, where he had fewer responsibilities and more skilled teammates, helped keep Chelios playing at close to his peak level. In 2002, his +40 plus/minus led the league, and he was again named to the First All-Star Team. He also led the United States hockey team to a silver medal in the 2002 Winter Olympics, and was named to the Tournament's All-Star Team. His season culminated in the Red Wings' victory over the Carolina Hurricanes in the Stanley Cup Finals, giving Chelios his second Stanley Cup.

In 2004, due to the cancellation of the NHL season, Chelios, along with fellow Red Wing teammates Derian Hatcher and Kris Draper, decided to play hockey for the Motor City Mechanics, a UHL team based out of Fraser. In October 2004 he trained with the U.S. bobsled federation in a bid to compete for the Greek bobsled team at the 2006 Winter Olympics.

August 4, 2005, the 43-year-old defenseman re-signed with the Red Wings for a one-year contract.

On February 1, 2006, Chelios was again named captain of the US Olympic Hockey Team. Chelios was also captain in the 1998 Nagano games and of the silver-medal-winning team in the 2002 Salt Lake City games.

Chelios is rumored to be retiring soon, since at the age of 44 he is the oldest player in the league. If he does retire, he will surely be elected to the Hall of Fame when he becomes eligible. His 19-year career has shown that he can both score and play defense. He also plays with an edge to his game, as demonstrated by his 2695 penalty minutes. In his prime he combined his offensive skills with his physical edge to win 3 Norris Trophies. Over his career, he played in 11 All-Star games and was named to 7 NHL First or Second All-Star Teams. Even with his career tapering off, he has proved that he can play an important role for a Stanley Cup-winning team. All in all, Chelios has secured a legacy as one of the most decorated Americans to ever play in the NHL, and is considered by some to be the greatest American ever to play hockey.

Trivia

In 2004 Chris and surfer Laird Hamilton trained with the US bobsled team, and hope to form the first Greek bobsled team at the 2006 Winter Olympics.

Figure A-7: Screenshot of an insertion between two consecutive revisions on a Wikipedia article on *Chris Chelios*. An insertion sentence is shown in red.

Appendix B

Examples of Document Structure and Section Titles on Wikipedia

The following tables show some sample documents with their section titles (Table B.1, B.2, and B.3) and all distinct titles (Table B.4 and B.5)in our corpus described in Chapter 4.

Michael Ahern	**Richard Arrington**
Early life	Childhood
Parliamentary Career	Academic career
Premier	Political career
Further reading	References
Carlos Romero Barcel	**Ernest Benach Pascual**
Education	Education and professional background
Political career	Civic background
Legacy	Political background
Publications	Institutional background

Table B.1: Section titles in sample documents of our corpus (part 1).

Belisario Betancur	**Mario Cuomo**
Early years	Early life
Presidency	Political career
Post-Presidency	Views
Honors	Personal life
Vuk Dra	**Harold Ford**
Early life and career	Early life
Career in politics	House of Representatives
Personal	2006 Senate campaign
Quotations	Controversies
	References
Bas de Gaay Fortman	**John Jay Hooker**
Career before politics	Early life
Political career	Legal career
Career after politics	Political career
Political Views	Later life
Miscellaneous facts	
Quotes	
John Ikenberry	**Ed Koch**
Positions	Early life
Publications	Political career
Views	Later life
References	Political stance
	Legacy
	Books
Mary Landrieu	**Harry Lee**
Senate career	Background
Gang of 14	Legal career
Hurricane Katrina	Political career
Election History	Hurricane Katrina
Mohammad Naseem	**Joni Madraiwiwi**
History	Career
Leadership	Opinions
Controversy	Achievements
Politics	Personal life

Table B.2: Section titles in sample documents of our corpus (part 2).

Jim Moran
Early life
Congress
Controversies
2006 election

Ren Gill Pratt
Political career
Controversies
Education
Election history

Anna Sim Castell
Education and professional background
Civic background
Political background
Institutional background

Wilebaldo Solano
Youth
Civil War
Exile
Later years

Alejandro Toledo
Early years
Professional career
Political career
The presidency

Malcolm Turnbull
Early life
Career
Politics
Family
Writing

Ibrahim Nasir
Genealogy
Political career
Criticism
Succession

Volen Siderov
Early life
Rise in politics
Controversy
Presidential election

Louise McIntosh Slaughter
Personal Background
Early Political Career
Congressional career
Election results

Richard Tarrant
Early life
Business career
Philanthropy
2006 campaign for U.S. Senate

Haunani-Kay Trask
Background
Education
Activism
Resources

Dingiri Banda Wijetunga
Early life
Political Career
Prime Minister
President

Table B.3: Section titles in sample documents of our corpus (part 3).

Early life	Childhood
Family and early life	Birth and education
Early years	Youth
Family	Personal
Private life	Personal life
Family and personal life	Private Life
Personal background	Family background
Life	Personal Background
Homosexuality	Genealogy
Health	Education
Academic career	Education and professional background
Education and early career	Academic life
Education and Early Career	Early life, career, and family
Life before politics	Retirement
Post-Presidency	Later years
Life after politics	Later life
Platform	Political Views
Political views	Opinions
Political Activities	Philanthropy
Civil War	Activism
Publications	Published works
Books	Writing
Current activities	Recent developments
Recent activities	Legacy
Further reading	Trivia
Biography	References
Contact	Notes
Sources	Views
Quotations	Miscellaneous facts
Background	Quotes
Resources	External sources
Filmography	Issues
History	Overview
Honors	Awards and recognition
Awards and decorations	Achievements
Awards	Honours
Recognition	Indictment

Table B.4: Distinct section titles in our corpus (part 1).

Charges	Trial
Controversies	Controversy
Criticism	Exile
Parliamentary Career	Parliamentary career
Political career	Political life
Life and career	Civic background
Political background	Institutional background
Early career	Presidency
Premier	Prime Minister
Professional career	Timeline of career
Early political career	Term as Mayor
Later career	Early life and career
Prosecutor	Minister
Career	Career in politics
Early life and entry into politics	Parliament
House of Representatives	2006 Senate campaign
State Legislature	Business career
Political Career	Election History
Campaign for Lieutenant Governor	Legal career
Positions	Politics
Appointment	Medical career
Political stance	Senate career
Electoral history	2004 elections
United States Senate	Professional life
Political campaigns	Leadership
Legislative Career	2006 campaign for U.S. Senate
Political Life	2006 Election
The presidency	Committee assignments and caucus memberships
Mayoralty	President
Entry into politics	Career before politics
Career after politics	Congress
2006 election	2006 campaign for the U.S. House of Representatives
Early Political Career	Congressional career
Rise in politics	Presidential election
2006 Gubernatorial Race	Public service
Career before Politics	Electoral History
Election history	Election results
Hurricane Katrina	Public perception
Succession	Politician
Administration	Gang of 14

Table B.5: Distinct section titles in our corpus (part 2).

References

[1] Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. Inferring strategies for sentence ordering in multidocument news summarization. *JAIR*, 17:35–55, 2002.

[2] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. In *Proceedings of the ACL*, pages 141–148, 2005.

[3] Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of HLT-NAACL*, pages 113–120, 2004.

[4] David M. Blei and John D. Lafferty. Correlated topic models. In *NIPS*, 2005.

[5] David M. Blei and J. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems*, 2007.

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. A bottom-up approach to sentence ordering for multi-document summarization. In *Proceedings of the COLING/ACL*, pages 385–392, 2006.

[9] Lijuan Cai and Thomas Hofmann. Hierarchical document categorization with support vector machines. In *Proceedings of the CIKM*, pages 78–87, 2004.

[10] Nicolò Cesa-Bianchi, Claudio Gentile, and Luca Zaniboni. Hierarchical classification: Combining bayes with SVM. In *Proceedings of the ICML*, pages 177–184, 2006.

[11] Nicolò Cesa-Bianchi, Claudio Gentile, and Luca Zaniboni. Incremental algorithms for hierarchical classification. *JMLR*, 7:31–54, 2006.

[12] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the EMNLP*, pages 1–8, 2002.

[13] Michael Collins and Brian Roark. Incremental parsing with the perceptron algorithm. In *Proceedings of the ACL*, pages 111–118, 2004.

[14] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *JMLR*, 7:551–585, 2006.

[15] Ofer Dekel, Joseph Keshet, and Yoram Singer. Large margin hierarchical classification. In *Proceedings of the ICML*, pages 209–216, 2004.

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

[17] Micha Elsner, Joseph Austerweil, and Eugene Charniak. A unified local and global model for discourse coherence. In *Proceedings of the HLT-NAACL*, pages 436–443, 2007.

[18] Micha Elsner and Eugene Charniak. A generative discourse-new model for text coherence. Technical Report CS-07-04, Brown University, 2007.

[19] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17*, Cambridge, MA, 2005. MIT Press.

[20] A. Gruber, M. Rosen-Zvi, and Y. Weiss. Hidden topic markov models. In *Artificial Intelligence and Statistics (AISTATS)*, 2007.

[21] Donna Harman. Overview of the first text retrieval conference (trec-1). In *TREC*, pages 1–20, 1992.

[22] Marti Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the ACL*, pages 9–16, 1994.

[23] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA, 1999. ACM.

[24] Eduard Hovy. Automated discourse generation using discourse structure relations. 63:341–385, 1993.

[25] Nikiforos Karamanis, Massimo Poesio, Chris Mellish, and Jon Oberlander. Evaluating centering-based metrics of coherence for text structuring using a reliably annotated corpus. In *Proceedings of the ACL*, pages 391–398, 2004.

[26] Richard Kittredge, Tanya Korelsky, and Owen Rambow. On the need for domain communication knowledge. *Computational Intelligence*, 7(4):305–314, November 1991.

[27] Markus Krotzsch, Denny Vrandecic, and Max Volkel. Wikipedia and the semantic web - the missing links. In *Proceedings of Wikimania 2005 The First International Wikimedia Conference*. Wikimedia Foundation, JUL 2005.

[28] Mirella Lapata. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the ACL*, pages 545–552, 2003.

[29] Wei Li and Andrew McCallum. Semi-supervised sequence modeling with syntactic topic models. In *AAAI*, pages 813–818, 2005.

[30] Dekang Lin. Dependency-based evaluation of minipar. In *Proceedings of the Workshop on the Evaluation of Parsing Systems, LREC*, pages 48–56, 1998.

[31] Dekang Lin. LaTaT: Language and text analysis tools. In *Proceedings of the 1st International Conference on Human Language Technology Research*, pages 1–6, 2001.

[32] Inderjeet Mani and George Wilson. Robust temporal processing of news. In *Proceedings of the ACL*, pages 69–76, 2000.

[33] Daniel Marcu and Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the ACL*, pages 368–375, 2002.

[34] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM.

[35] K. R. McKeown. Discourse strategies for generating natural language text. *Artificial Intelligence*, 27(1):1–41, 1985.

[36] Chris Mellish, Mick O'Donnell, Jon Oberlander, and Alistair Knott. Experiments using stochastic search for text planning. In *Proceedings of International Conference on Natural Language Generation*, pages 98–107, 1998.

[37] Johanna D. Moore and Cecile L. Paris. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19(4):651–694, 1993.

[38] Naoaki Okazaki, Yutaka Matsuo, and Mitsuru Ishizuka. Improving chronological sentence ordering by precedence relation. In *Proceedings of the COLING*, pages 750–756, 2004.

[39] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, 1990.

[40] Paolo Castagna Roberto Tazzoli and Stefano Emilio Campanini. Towards a semantic wiki wiki web. In *In Demo Session of ISWC2004*, 1999.

[41] Jacques Robin and Kathleen McKeown. Empirically designing and evaluating a new revision-based model for summary generation. *Artificial Intelligence*, pages 135–179, 1996.

[42] Radu Soricut and Daniel Marcu. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the HLT-NAACL*, pages 149–156, 2003.

[43] Adam Souzis. Building a semantic wiki. *IEEE Intelligent Systems*, 20(5):87–91, 2005.

[44] Jacob Eisenstein S.R.K. Branavan, Harr Chen and Regina Barzilay. Learning document-level semantic properties from free-text annotations. In *(To appear) Proceedings of the ACL*, 2008.

[45] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[46] M. Voelkel, M. Kroetzsch, D. Vrandecic, H. Haller, and R. Studer. Semantic Wikipedia. In *Proceedings of the International World Wide Web Conference (WWW'06)*, pages 585–594, Edinburgh, Scotland, 2006. ACM.