NLP Driven Models for Automatically Generating Survey Articles for Scientific Topics

by

Rahul Kumar Jha

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in The University of Michigan
2015

Doctoral Committee:

Professor Dragomir R. Radev, Chair
Professor Qiaozhu Mei
Professor Rada Mihalcea
Professor Emily Kaplan Mower Provost

ACKNOWLEDGEMENTS

First of all, I would like to thank Prof. Dragomir Radev for taking me as a PhD student and guiding my research. Drago's commitment to quality research is inspirational and I found his grounded approach to research a very good fit for me. Over the past few years, I have learnt a lot from him about doing good research as well as presenting it, and I am very grateful to him for helping me achieve the goals I set for myself.

Next, many thanks to Prof. Rada Mihalcea, Prof. Emily Mower Provost, and Prof. Qiaozhu Mei for being on my thesis committee. I was very happy to get a chance to discuss a number of research ideas with Rada and always found her full of new ideas and great advice. Emily brought a different perspective to the table and helped me see many aspects of my thesis work that I would have missed. Finally, Qiaozhu with his attention to detail helped me notice several finer points of my work and improve them. I feel very fortunate to have such a great committee help me shape and refine my thesis work.

I had the pleasure to work with a number of great people as part of the Clair lab during my time here including Ben King, Reed Coke, Catherine Finegan-Dollak, Amjad Abu Jbara, Vahed Qazvinian, and Ahmed Hassan. Thanks to them for giving me the chance to collaborate on various projects over the years.

Thanks to the AI faculty at University of Michigan for providing a great environment for doing research. In particular, I would like to thank Prof. John Laird, for making me feel welcome during the initial time of my PhD and Mike Wellman, whose

great instruction for EECS 592 taught me a lot of things about research.

During my PhD, I was fortunate enough to collaborate with many researchers from other institutions as part of the IARPA FUSE program. It was a great learning experience for me, and I'd like to thank all the professors I had a chance to work with including Kathleen McKeown, Hal Daume, Owen Rambow, Simone Teufel, Luis Gravano, Kenneth Fleischmann, Fei Xia, and Mike Collins as well as their students including Kapil Thadani, Ioannis, Paparrizos, Pablo Gonzalez, Diarmuid O Seaghdha, Snigdha Chaturvedi, Taesun Moon, Or Biran, and Clay Templeton. Additionally, thanks to Suvarna Bothe, Kevin McInerney, and Arfath Pasha for their help during this project.

Thanks to all my friends at Michigan who have supported me during my time here including Shiwali Mohan, Vineet Raichur, Gagan Mand, Komal Kampasi, Girish Kulkarni, Ayush Shah, and Yash Adhia. Finally, I would never be able to do any of this without the constant love and support from my parents, Bishwa Mohan Jha and Poonam Jha and my brother, Saurabh Jha. They were always there for me at every point of the PhD journey and cheered me on as I stumbled along on the tortuous path of doctoral research.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

NLP Driven Models for Automatically Generating Survey Articles for Scientific Topics

by

Rahul Kumar Jha

Chair: Dragomir Radev

This thesis presents new methods that use natural language processing (NLP) driven models for summarizing research in scientific fields. Given a topic query in the form of a text string, we present methods for finding research articles relevant to the topic as well as summarization algorithms that use lexical and discourse information present in the text of these articles to generate coherent and readable extractive summaries of past research on the topic. In addition to summarizing prior research, good survey articles should also forecast future trends. With this motivation, we present work on forecasting future impact of scientific publications using NLP driven features.

CHAPTER I

Introduction

The exponential growth of scientific output has created new challenges for researchers. New researchers often find themselves struggling with the massive amount of prior literature in their chosen fields that they must understand before being able to make new contributions of their own. Experts in any research area also face the challenge of keeping up with the progress of their rapidly growing fields. Additionally, scientific research today tends to be highly interdisciplinary, which means researchers are expected to understand many related fields in addition to their own field in order to pursue interesting research projects.

The dissemination of scientific literature through electronic means goes some way towards solving this problem. Traditional publishers such as Elsevier and ACM are increasingly offering electronic access to their publications through online portals. Easy access to electronic versions of scientific publications has created opportunities and demand for tools that allow researchers to quickly find the relevant research in their area. Two major commercial search engines catering to this demand are Google Scholar and Microsoft Academic Search. In addition to offering a simple search interface to look for the most relevant papers in any area, they offer additional tools to help researchers. For example, Figure 1.1 shows the output of Microsoft Academic Search for the query *question answering*. The output, in addition to the relevant

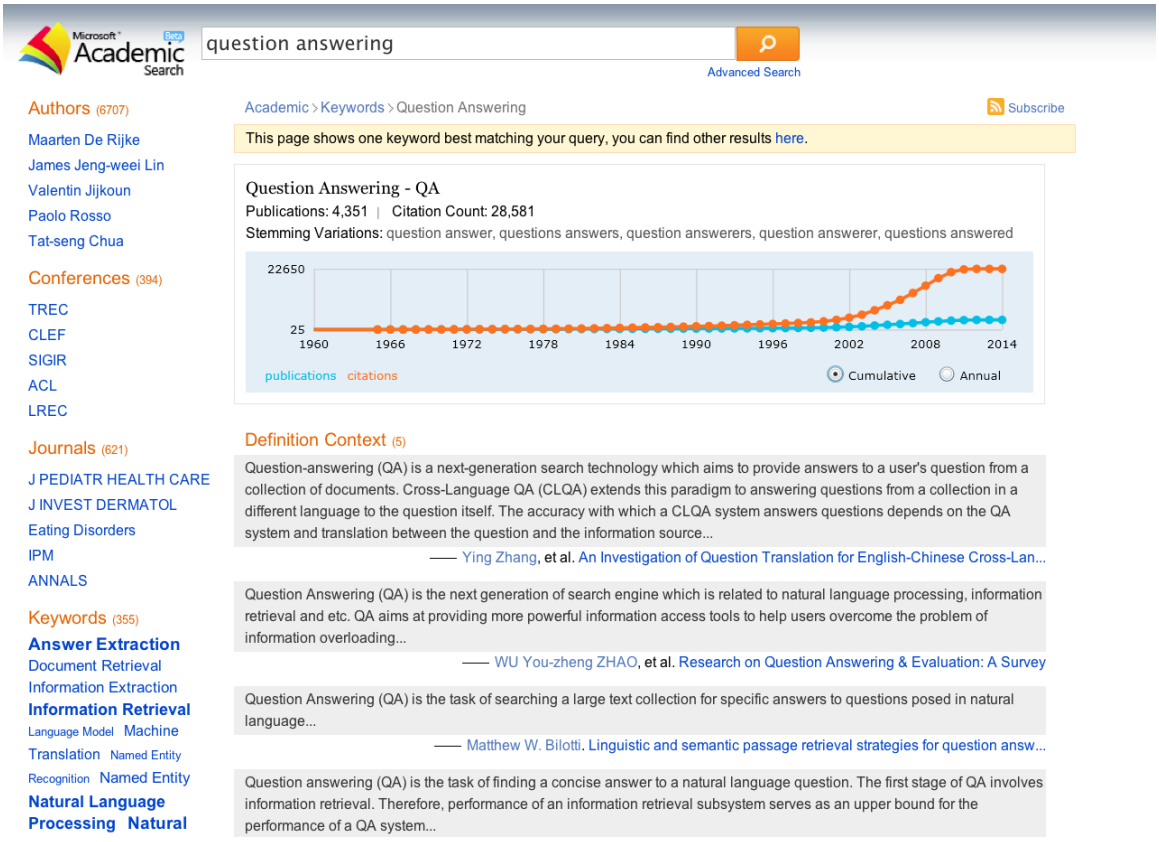

Figure 1.1: Results displayed by Microsoft Academic Search for the query *Question Answering*.

papers in the area, shows some of the top authors and main venues. It also shows topic definitions that it extracts from relevant papers, and shows a plot of citations to papers that belong to the topic in the past few years. All of this is potentially useful information for someone looking to learn more about a scientific topic.

These simple retrieval technologies, though useful, are not enough. Researchers need tools that can summarize and present relevant literature in a way that makes it easy for them to understand the main sub-problems in the area, their relationship to each other and the contributions of the main papers. The traditional means of acquiring this kind of knowledge about a scientific field is a survey article written by experts in the area. However, the rapid growth of publications, especially in technological fields, has meant that the supply of such survey articles has not kept up with the demand. Good human written survey articles do not exist for many

The recent interest in computational lexicography has fueled a large body of recent work on this 40-year-old problem ,e.g., Black (1988) , Brown et al. (1991) , Choueka nd Lusignan (1985) , Clear (1989) , Dagan et al. (1991) , Gale et al. (to appear) , Hearst (1991) , Lesk (1986) , Smadja and McKeown (1990) , Walker (1987) , Veronis and Ide (1990) , Yarowsky (1992) , Zemik (1990, 1991).
Word-sense disambiguation is a long-standing problem in computational linguistics (e.g., Kaplan (1950) , Yngve (1955) , Bar-I-Iillel (1960) , Masterson (1967)), with important implications for a number of practical applications including text-to-speech (TI'S) , machine translation (MT), information retrieval (IR), and many others.
Word sense disambiguation (WSD) has been found useful in many natural language processing (NLP) applications ,including information retrieval (Krovetz and Croft 1992; McRoy 1992), machine translation (Brown et al. 1991; Dagan, Itai, and Schwall 1991; Dagan and Itai 1994), and speech synthesis (Yarowsky 1992).
WSD has received increasing attention in recent literature on computational linguistics (Lesk 1986; Schi.itze 1992; Gale , Church , and Yarowsky 1992; Yarowsky 1992, 1995; Bruce and Wiebe 1995; Luk 1995; Ng and Lee 1996; Chang et al. 1996).
Given a polysemous word in running text, the task of WSD involves examining contextual information to determine the intended sense from a set of predetermined candidates.
It is a nontrivial task to divide the senses of a word and determine this set, for word sense is an abstract concept frequently based on subjective and subtle distinctions in topic, register, dialect, collocation, part of speech, and valency.

Figure 1.2: Part of the summary for the topic of *word sense disambiguation* generated using our current system. The system finds the most relevant sentences that should be included in the summary using a relevance model and uses a discourse model for presenting the sentences as a coherent summary with appropriate context for each sentence.

scientific fields and the ones that do exist get outdated quickly.

In this thesis, we present Natural Language Processing (NLP) based methods that aim to fill this gap. We explore algorithms that in addition to retrieving documents relevant to a scientific topic, also generate natural language summaries similar to human written survey articles. For example, in response to the query *word sense disambiguation*, our system can generate the summary shown in Figure 1.2.

In addition to enabling better search technologies, trying to solve the above problems allows us to investigate the process of scientific research as a social process.

Scientists in any field discuss research problems, come up with shared terminology that evolves over long periods of time, and record their arguments as text narratives. This is a process of social "meaning making" that has traditionally been studied under *social semiotics* (Leeuwen, 2004). We refer to the linguistic data that emerges out of such social processes as *collective discourse*. In this thesis, we look at scientific discourse as a special form of collective discourse.

We now present an outline of the components a system needs in order to generate a summary given a query provided by a user and a corpus of scientific publications. Our corpus is the ACL Anthology Network[1] (Radev *et al.* , 2013). This outline also serves as a roadmap to the chapters in this thesis.

1. **Process Input Query and Build a Target Document Collection:** The first task is to process the input query and build a collection of documents that we need to summarize. In Chapter III, we compare several methods for building document collections in response to queries and find that a simple model we propose called *Restricted Expansion* produces reasonable results. A description of this method was first published in Jha *et al.* (2013). We also look at the problem of predicting future importance of recently published papers on a topic that have not had a chance to accrue citations. Such papers may not be deemed important by a simple citation counts based retrieval model, but might be included by a good survey due to high expected impact in the future.

2. **Aggregate Relevant Text:** Once we have a target document collection, we need to find the text that will be used to summarize these documents. Previous research has shown that in the domain of scientific publications, documents citing the target documents provide useful information for summarizing them (Mohammad *et al.* , 2009). In Chapter IV and Chapter V, we explore several ways of operationalize this idea and present ways of aggregating text that are

[1] http://clair.eecs.umich.edu/aan/index.php

suitable to build the kind of summaries that we aim to produce. Some of these results were previously published in Jha *et al.* (2013).

3. **Build a Content Model for Input Text:** Once the input text segments for the summarization system are available, we need to create a data structure to represent the information present in these segments that allows us to build a summary. The main output of such a data structure is a relevance score assigned to the elementary discourse content units (sentences in our case). In Chapter IV, we explore existing network based models for this problem (Erkan & Radev, 2004; Qazvinian & Radev, 2008a). We also experiment with Bayesian content models (Daumé & Marcu, 2006; Haghighi & Vanderwende, 2009) and present a new way to combine the information from these models with network based models. In Chapter V, we present a set of new content models that exploits the network structure between citing and cited documents in the input set. Some of these results have been previously published in Jha *et al.* (2015b) and Jha *et al.* (2015a).

4. **Generate Coherent Output Text:** Using the above components, we can find relevant sentences in the input set and generate summaries, but the output summaries generated in this way are usually not very readable. To generate coherent summaries, we need to identify the main topics we would like our summary to talk about and infer the right order of presenting these topics. In addition, since the sentences are used outside their original context, we also need to make sure we include appropriate context necessary to understand each sentence in the summary. We discuss all of these issues in Chapter VI where we present a complete pipeline for generating coherent surveys for scientific topics. Many of these results were published earlier in Jha *et al.* (2015b).

We first begin by a description of the related work in Chapter II. This is followed

by algorithms for various components in Chapter III, IV, V and VI as described above. Finally, Chapter VII concludes the thesis along with pointers to future work.

CHAPTER II

Related Work

The work in this thesis draws upon and makes contributions to several different fields that have traditionally worked independently from one another. Thus, the related work is quite large. We will make an attempt to summarize the work most relevant to the material in this thesis, pointing the readers to comprehensive surveys when possible. The related work is broadly divided into four main sections.

- **Summarization:** Summarization has been studied as a problem in computational linguistics since the early work of Luhn in 1958 though a lot of major techniques appeared during and after the 1990's, driven by the TREC, DUC, and MUC conferences. I will summarize relevant work in salience models, coherence models, discourse oriented summarization, post processing strategies, and automated coherence scoring.

- **Scientometrics:** In this section, I will discuss and relate the work done in the field of Scientometrics on measuring scientific impact of entities (papers, authors, institutions etc.), modeling the history and dynamics of scientific fields, forecasting future impact of authors and papers, and models for literature search.

- **Modeling scientific text:** In this part, I will focus on the work done in modeling the text of scientific articles. Specifically, I will look at two main

threads of research: 1) analysis of citation text and 2) analysis of discourse structure of scientific text.

- **Repurposing scientific text:** In this section, I describe research that draws upon ideas discussed in the previous sections to use the text in scientific papers for new interesting applications. This includes work on finding new science using literature mining, scientific article summarization, survey article generation for scientific topics, and scientific indexing and retrieval.

2.1 Summarization

The problem of automatic text summarization is now one of the standard tasks in the field of natural language processing. Starting from the seminal work of Luhn (1958), research in automatic summarization has made great progress with many different variants of the problem and several classes of methods explored in the community. The Document Understanding Conferences (DUC) and now the Text Analytics Conference (TAC) have provided the main thrust of research in this area by creating standardized data and evaluation methods.

Summarization methods can be divided into two major classes: single document summarization and multi-document summarization. Single document summarization deals with creating short summaries of a single text document. Multi-document summarization is the problem of summarizing a collection of documents that are topically related in some way. Other classifications of summarization methods includes indicative vs informative and extractive vs abstractive. A large number of recent summarization methods are extractive, which means they pick the most relevant sentences from the original document and aggregate them in some order to produce the output summary.

Given this, the first task is to assign sentences importance scores. Most early

document summarization methods used the word distribution statistics to first find the most relevant words, and then pick sentences that contain these words (Luhn, 1958; Baxendale, 1958). In later work, more sophisticated methods were explored that range from using external knowledge (Barzilay & Elhadad, 1997), supervised machine learning based methods (Kupiec *et al.* , 1995), methods based on discourse properties of input text (Marcu, 1995; Mani *et al.* , 1998) and network based methods (Erkan & Radev, 2004).

After sentence relevance has been determined, a good summarization system must process the output summary to form a coherent and readable summary by reordering sentences (Barzilay *et al.* , 2001a), fusing multiple sentences (Barzilay & McKeown, 2005) and other operations (Jing & McKeown, 2000). It also needs to make sure that the output does not contain redundant information (Carbonell & Goldstein, 1998).

In the following sections, we will discuss different aspects of summarization systems including content models, discourse based summarization, summary post processing, and automatic coherence scoring in more detail. For a more thorough treatment of the research in summarization, we refer the readers to the excellent survey by Nenkova & McKeown (2011).

2.1.1 Content Models

The goal of content models in the context of summarization is to extract a representation from input text that can help in identifying important sentences that should be in the output summary. We will organize our discussion of content models across five categories: term based models, network based models, probabilistic models, and discourse based models. In this section, we discuss the first three types of models and cover discourse based methods in detail in the next subsection.

2.1.1.1 Term Based Models

Term based content models depend on assigning importance scores to sentences based on statistic measures of word distributions. One of the earliest term based methods (and one of the earliest summarization methods in general) would be that presented by Luhn (1958). They first identify significant words in the text based on the frequency of words, with higher frequency words assigned higher significant (after accounting for stop words). Sentences are then assigned a significance or relevance score based on the number and proximity of significant words in the sentence. In later work, Edmundson (1969) presented a summarization system that used a linear weighted combination of scores based on cue words, title and heading words, and sentence location in addition to significant words to rank sentences. Kupiec *et al.* (1995) used similar features but approached summarization as a classification problem. They train a Naive-Bayes classifier with features corresponding to sentence length, cue phrases, sentence position, word frequency, and capitalization. The training data consists of scientific articles and manual abstractive summaries. The sentences in the abstractive summaries are aligned with the sentences in the original documents to create training data for extractive summarization.

Lin & Hovy (2000) presented a summarization system called SUMMARIST based on topic signatures. Topic signatures are defined as a family of related terms that are highly correlated with a target concept. These terms acquired and ranked from a pre-classified corpus using the likelihood-ratio test. The score of a sentence is simply the sum of all the scores of content-bearing terms in the sentence.

Another data structure for modelling significant terms for multi-document summarization is the centroid (Radev *et al.* , 2004b). A centroid is a set of words that are statistically important to a cluster of documents. In the centroid based method, relative documents are first grouped into clusters. A centroid is generated by starting with the first document in the cluster. As new documents are processed, their

TF*IDF values are compared with the centroid using cosine similarity and if it's within a threshold, the new document is included in the cluster and its terms are added to the centroid. The summarization algorithm then assigns an importance score to each sentence based on its centroid score (computed as the sum of the centroid values of all words in the sentence) as well as its position and overlap with the first sentence.

2.1.1.2 Network Based Models

Sentences in any text document are related to each other. Term based models do not model these relationships between sentences. An obvious choice for modelling such data is networks. In recent years, there has been a tremendous amount of progress in the field of network theory (Newman, 2010). Network based content models (Erkan & Radev, 2004; Mihalcea & Tarau, 2004) work by converting the input sentences into a network. Each sentence is represented by a node in the network and the edges between sentences are given weights based on the similarities of sentences. We then run Pagerank on this network and sentences are selected based on their pagerank in the network. For computing the pagerank, the network can either be pruned by removing edges that have weights less than a certain threshold, or a weighted version of pagerank can be run on the network. The method can be modified for doing query focused summarization as well (Otterbacher *et al.* , 2009). C-Lexrank (Qazvinian & Radev, 2008a) modifies Lexrank by first running a clustering algorithm on the network to partition the network into different communities and then selecting sentences from each community by running Lexrank on the sub-network within each community.

2.1.1.3 Probabilistic Models

Probabilistic methods for multi-document summarization depend on Bayesian modeling of word distributions in the input documents. One of the first proba-

bilistic content models seems to be BayeSum (Daumé & Marcu, 2006), designed for query focused summarization. BayeSum models a set of document collections using a hierarchical LDA style model. Each word in a sentence can be generated using one of the three language models: 1) a general english language model that captures english filler or background knowledge, 2) a document specific language model, and 3) a query language model. These language models are inferred using expectation propagation and the sentences are ranked based on their likelihood of being generated from the query language model. A similar model for general multidocument summarization called TopicSum was proposed by Haghighi & Vanderwende (2009), where the query language model is replaced by a document collection specific language model; thus sentences are selected based on how likely they are to contain information that summarizes the entire document collection instead of information pertaining to individual documents or background knowledge. They also introduce a more sophisticated content model called HierSum, that further divides the document collection specific language model into multiple content distributions representing various topics a document might talk about.

Barzilay & Lee (2004) present a Hidden Markov Model (HMM) based content model where the hidden states of the HMM represent the topics in the text. The transition probabilities are learnt through Viterbi decoding. They show that the HMM model can be used for both re-ordering of sentences for coherence as well as discriminative scoring of sentences for extractive summarization. Fung & Ngai (2006) present a similar HMM based model for multi-document summarization (Fung *et al.* , 2003).

Jiang & Zhai (2005) presented an HMM based system for extracting coherent relevant passages as part of an information retrieval system. In their model, they treat the document as a sequence of words that is generated from two language models: a relevant language model and a background language model. A different stochas-

tic process determines when the language model switches from one to the other. The background language model is estimated using maximum likelihood trained on the entire document collection. The relevance language model is estimated using a maximum likelihood estimator on one of three different samples: 1) the original query words, 2) a within document pseudo-feedback corpora created by extracting from the document a short passage highly likely to be relevant to the query, and 3) a cross-document pseudo feedback corpora created by extracting starting passages from different documents that are relevant to the same query. Once all output probabilities have been computed, the transition probabilities can be trained. The HMM trained with cross-document feedback performed best.

2.1.1.4 Other Approaches

Boguraev & Kennedy (1997) present a summarization system that generates a summary for a document in the form of a list of representative phrases. Their system works by first identifying the set of all nominal expressions in the text. This set is then reduced to a smaller set of expressions which uniquely identify the objects referred to in the text using an anaphora resolution procedure. These referents are then assigned a salience score as a function of how a candidate satisfies a set of grammatical, syntactic and contextual parameters such as if any term in the coreference class for the referent is a subject, a direct object, complement of a preposition etc. This salience score allows the system to determine the prominence of an expression in a local segment of discourse. To generate a summary for the entire of text, TextTiling (Hearst, 1994) is used to divide the text into discourse segments and the most salient phrases for each discourse segment are presented in the order of the original sequence of the discourse segments in the original text. Bateman (1993) briefly describe a system that aggregates knowledge from various source articles to generate biographical summaries of individuals.

Kraaij *et al.* (2001) present a multi-document summarization system that scores sentences based on a combination of a mixture language model for text and a Naive Bayes model for non text features such as sentence position, sentence length and cue phrases. Their mixture language model consists of three unigram language models: corresponding to the document, corresponding to the cluster and corresponding to background. Each of the unigram models are learnt using maximum likelihood procedures. Their hypothesis is that "a 'good' sentence is both salient for a document and for the corresponding cluster" and their final model is the geometric mean of the log likelihood ratio of a sentence given the mixture model and given the background model. After this, they manually scored each sentence for salience on a score ranging from -2 to 2. They then did feature selection using this ground truth data and picked cue phrase, sentence length and first sentence as useful features. A Naive Bayes classifier was then trained on these features. They then combined the mixture model and Naive Bayes model by using the posterior log-odds of the Naive Bayes classifier and interpolating the two values. They used this relevance model along with diversity based reranking using MMR.

Conroy & O'leary (2001) present two extractive summarization methods. Their first method is based on identifying the terms in the document using named entity recognition. A term document matrix is then formed and the summarization system works by choosing a subset of sentences that cover the main terms in the document using QR decomposition with partial pivoting. This method iteratively chooses sentence vectors with the largest weight and updates the weights of remaining vectors to avoid redundancy similar to MMR. The second method is based on an HMM that contains two states corresponding to summary sentences and non-summary sentences. The features used for the HMM are sentence position within the document and paragraph, number of terms and their likelihood. Given this HMM, two methods of summary sentence selection are presented: the first method chooses sentences with

the maximum posterior probability of being a summary sentence, the second method uses the QR decomposition to remove any redundant sentence that might be included by the HMM maximum posterior probability method. In Conroy *et al.* (2001), they also present an additional logistic regression model for summarization trained on four features: the number of unique query terms in the sentence, the number of content words in a sentence, distance of the sentence from one with a query term, and the position of sentence in a document. The query terms used in these features were automatically extracted from the input documents.

Afantenos *et al.* (2004) present a multi-document summarization method based on establishing cross document structural relationships inspired by the Cross Document Structure Theory (CST) (Radev, 2000). CST describes RST (Rhetorical Structure Theory) style relationships that can exist across multiple documents to words, phrases, sentences, paragraphs or even entire documents. The authors however, argue that CST as derived from RST might not be suitable for modelling cross document relationships because of the lack of a coherent discourse across multiple documents. They propose a methodology for establishing cross document relationships based on the creation of a topic specific ontology, message types and relations. They present a case study of football match descriptions for which they extract each of the above. The method seems to be highly domain dependent.

2.1.2 Discourse Based Summarization

Modeling discourse has been a topic of interest within the computational linguistics community for a number of years. Discourse processing involves several different subproblems such as how entities are introduced and discussed in a coherent text (Grosz & Sidner, 1986) as well as abstract data structures for modelling how different text segment in a piece of text relate to each other (Mann & Thompson, 1988; Prasad *et al.* , 2008; Hahn, 1990; Wolf & Gibson, 2005). For a detailed overview of

the research in discourse processing, we refer the readers to Stede (2012). Before a discussion of the related work on discourse based summarization is presented, it is necessary to differentiate between two important linguistic concepts related to discourse: cohesion and coherence. These two concepts provide a good way of characterizing and understanding the prior research in this area.

The concept of cohesion is dealt with in great detail in an influential book (Halliday & Hasan, 1976). In this book that inspired several important papers in discourse based summarization, the authors described the concept of cohesion in terms of the concept of text. A text is defined as being constituted as "any piece of language that is operational, functioning as a unity in some context of situation" and in this way is distinguished from just a string of sentences. A text is not thought of as a grammatical unit like a supersentence, but rather as a semantic unit. The unity in the above definition "is a unity of meaning in context [...]". Cohesion helps to create text and "occurs where the interpretation of some element in the discourse is dependent on that of another". It is distinguished from discourse structure, which is the concern of coherence as discussed later. The concept of cohesion is "setup to account for relations in discourse, but [...] without the implication that there is some structural unit that is above the sentence. Cohesion refers to the range of possibilities that exist for linking something with what has gone before." Halliday & Hasan (1976) discuss how grammatical cohesion is achieved by reference, substitution, ellipsis, conjunction as well as how lexical cohesion is achieved by reiteration and collocation.

Coherence, as opposed to cohesion "has to do with macro-level, deliberative structuring of a multi-sentence text in terms of relations between sentences and clauses." Mani *et al.* (1998). If we define sense as "knowledge that actually is conveyed by expressions occurring in a text" (Beaugrande & Dressler, 1981), a piece of text "makes sense" because there is a continuity of senses among the knowledge activated by the expressions of the text. This continuity of senses is the foundation of coherence. Co-

hesion in this sense, supports coherence. Coherence models thus aim to represent the overall structure of a multi-sentence text in terms of macro-level relations between clauses or sentences.

To summarize, "coherence relation is a relation among clauses or sentences, such as elaboration, support cause or exemplification [...] cohesion relations are relations among elements in a text: reference, ellipsis, substitution, conjunction and lexical cohesion [...]" (Morris & Hirst, 1991).

2.1.2.1 Coherence Based Models

There has been a steady stream of research focused on using coherence based discourse models for summarization, we now summarize some of this work. Note that this section does not focus on methods that *produce* coherent summaries, but methods that try to exploit features of input text that are present due to its coherence to assign importance score to sentences.

Two early works explored the cognitive structures used by humans for summarization. Van Dijk (1979) did some experimental work on building an understanding of the cognitive nature of human summarization process. Their main theory was that the processing of complex discourses requires a macro-structural component that specifies the global structures of the discourse, and the mapping rules relating these with the sequence of propositions of the text. These macro-rules are operations of semantic information reduction (deletion, generalization, combination). It is assumed that what is 'best' stored in memory of a longer discourse is essentially its macro-structure. The macro-rules predict which sentences are recalled or forgotten and used or not used in summaries. They describe the linguistic and cognitive basis of their hypothesis and then present seven experiments for illustration.

Hovy (1993) discussed discourse structure and relations from the perspective of text planning. After arguing that an understanding of discourse structure is essential

for text generation, they outlined the work in building text generation systems, and identified discourse structure relations that arise regardless of the discourse theory employed and their effects on sentence planning and text formatting.

Ono *et al.* (1994) present a summarization system for Japanese expository writings that is based on extracting rhetorical structure. Their discourse model represents rhetorical structure in two layers: intra-paragraph and inter-paragraph structures. The rhetorical structure is represented by a binary tree whose sub-trees form argumentative constituents. For summarization, their system calculates the importance of each sentence in the original text based on the relative importance of rhetorical relations. It then imposes penalties on both nodes for each rhetorical relation according to its relative importance. Finally, the system recursively cuts out the nodes from the terminal nodes which are imposed the highest penalty. The list of terminal nodes of the final structure becomes an abstract for the original document.

Marcu's work (Marcu, 1995, 1997) showed that the concept of discourse structure and nuclearity from RST can be effectively used to assign importance scores to textual units for a summarization system. Their RST parser determines the discourse markers and the elementary units that make up that text, uses the information derived from corpus analysis to hypothesize rhetorical relations among elementary units, and applies a constraint-satisfaction procedure to assign the best RS-tree to the text. Given the RS-tree, each node is assigned salient units that are computed recursively: each leaf is associated with the leaf itself, and each internal node is associated with the salient units of the nuclei of the rhetorical relation corresponding to that node. A salience score is then assigned to each textual unit depending on the depth in the tree where it occurs as a salient unit. Thus, the textual units that are salient units of the top nodes in a tree get a higher score than those that are salient units of the nodes found at the bottom of the tree. The summarization program selects the textual units that are most salient in the text based on this score.

Marcu (2001) present discourse based summarization system that they used to create single-document and multi-document summaries for DUC 2001. The single document summarization system starts by extracting the discourse structure of the document and extracting important sentences from the discourse structure using method described in Marcu (2000). They then used a syntactic parser to identify all the noun constructs as well as find all co-references. They then used a set of post-processing steps adding sentences to the pool so as to avoid dangling discourse relations to improve the coherence and compactness of the summary. In the final generation step, the sentences from the pool of important sentences are printed in the order of their occurrence in the text while making to replace third person pronouns with the associated referring expression when needed. Their multi-document summarization uses 100 word single-document summaries for each document as an input. It then calculates the similarity between every pair of documents and between every sentence pair in all single document summaries. Sentences are then assigned an importance score based on their average similarity scores. They define a number of heuristic scores (lying between 0 and 1) for an output summary. Some heuristics can be dependent on sentence pairs, e.g. summaries that maintain order of sentences from individual documents are better, sentences that produces sentences in chronological order are better and summaries that present sentences from documents with high average similarity scores before documents with low average similarity scores etc. There are also sentence specific heuristics based on sentence length, position of the sentence in the original document etc. In order to build a multi-document summary, they use a very simple (and apparently inefficient) optimization procedure. If there are n documents in the collection, they start with n active summaries created by taking the first sentence from the single-document summary for each document. After this, the system iterates over all possible summaries of length two that can be created by appending one sentence to a summary from the list of n active summaries.

The top 100 highest scoring summaries are kept from this pool, and the system then follows the same process to generate summaries of length three and so on.

Carlson *et al.* (2001) criticized the existing work of Ono *et al.* (1994) and Marcu on the grounds that it was restricted to short scientific texts, and that their attempts to extend summarization based on rhetorical structure trees to longer texts did not yield good results. They explain this by suggesting that the heuristic used by both these systems, namely that the importance of textual units is determined by their distance to the root of the corresponding rhetorical structure is reasonable but limited. Two other shortcomings of these systems are that they are un-localized and insensitive to the semantics of rhetorical relations. For their own experiments, the authors manually annotated 380 articles with rhetorical structures based on RST.

Daumé & Marcu (2002) present a discourse based summarization system that first derives the syntactic structure of each sentence and the discourse structure of the text. The discourse structure is derived using an RST parser. It then uses a statistical hierarchical model to drop non-important syntactic and discourse units to generate coherent and grammatical document compressions of arbitrary length. Their compression model is similar to the noisy channel model for sentence compression used by Knight & Marcu (2000), but it adds discourse units to the model. The source model assigns to the string the probability that the summary sentence is good english ($P(S)$), the channel model assigns to a document summary pair the probability that the document is a good expansion of the summary ($P(D|S)$), and the decoder searches through all possible summaries of a document D for the summary that maximizes the posterior probability ($P(D|S)P(S)$). The source model assigns scores to compressions based on bigram probability, context-free syntactic probabilities and context-free discourse probabilities. The channel model is allowed to add syntactic constituents or discourse units and is trained on the RST corpus of 385 Wall Street Journal articles as well as 150 documents from the same corpus paired with

extractive summaries. The decoder produces a list of possible compressions, and the best compression is chosen based on the log-probability of a compression normalized by length.

Bosma (2005) create a question answering system that generates detailed answers to user questions by using an RST based summarization method. Their approach consist of first parsing the text into an RS-Tree and then transforming it into a graph representation. A vertex is created for each sentence in the RS-tree and for each directed relation, an edge is created from each of the sentences of the nucleus to each of the sentences of the satellite of the relation, resulting in an acyclic directed graph. They define a weight for every edge in this graph calculated based on features from rhetorical structure as well as features of the sentence corresponding to the vertex which is targeted by the edge. The second step exploits a graph search algorithm in order to extract the most salient sentences from the graph. The starting node of the search is the node representing the answer sentence. The generated summary consists of the most salient sentences, given the answer sentence, where the salience is defined by the shortest path from the sentence to the answer sentence in the weighted graph.

Farzindar & Lapalme (2004) present a system for legal text summarization based on "the identification of thematic structures of the document and the determination of semantic roles of the textual units in the judgment. They analyzed a corpus of legal judgements to identify the organizational structure of a typical judgement by comparing model summaries written by humans with the texts of original judgements. The textual units considered as important by the professional abstractors were aligned manually with one or more elements of the source text. They observed that the original texts are organized according to a macro-structure with sentences belonging to 5 themes: decision data, introduction, context, juridical analysis, and conclusion. Based on this, their summarization system proceeds in four phases. In the first phase, the text is segmented into themes by using several heuristics such as the presence of

significant section titles, absolute and relative positions of a segment, certain linguistic markers etc. The second phase is a filtering phase that eliminates less important sentences such as citations. Third phase is a selection phase where each sentence is assigned a score based on heuristic functions related to position of the paragraphs in the document, position of the paragraphs in the thematic segment, position of the sentences in the paragraph, distribution of the words in document and corpus, and some cue phrases and linguistic markers. In the final phase, a summary of size 10% of the original document is generated and presented in a tabular format.

Louis *et al.* (2010) analyze the aspects of discourse that provide the strongest indication of text importance. They compare the utility of discourse features for single-document summarization from three frameworks: Rhetorical Structure Theory, GraphBank and Penn Discourse Treebank (PDTB). Discourse relations in these frameworks can provide two kinds of information: 1) structural information in terms of a tree or graph structure that relates the sentences, and 2) semantic information in terms of the specific discourse relations between the sentences. They test both these features. They tested the predictive powers of both of these types of features against gold-standard human summaries and found that structure information is the most robust indicator of importance and outperform semantic features by a large margin. They also found that the performance of the classifier is substantially improved when both types of features are used. They also compared the results of using discourse based structure with methods based on simple lexical overlap similarity such as Lexrank (Erkan & Radev, 2004) and found that similarity graph representation is even more helpful than RST or GraphBank. They conclude that "for use in content selection, lexical overlap information appears to be a good proxy for building text structure in place of discourse relations" and "for content selection in summarization, current systems can make use of simple lexical structures to obtain similar performance as discourse features."

2.1.2.2 Cohesion Based Models

Methods based on cohesion can be primarily divided into lexical cohesion and coreference. Lexical cohesion is created by using semantically related words (Barzilay & Elhadad, 1997). For modeling lexical cohesion, lexical chains is a popular method that has been used for producing summarization systems (Morris & Hirst, 1991). Lexical cohesion can occur not just between pairs of words but over a number of nearby related words spanning a topical unit of the text. These sequences of related words are defined as lexical chains and they can delineate portions of text that have a strong unity of meaning.

Mani & Bloedorn (1997) presented a system for multi-document summarization that uses a graph based representation based on entities and relations between entities. The driving principle behind this work is to use the relations that fall under cohesion as defined by Halliday & Hasan (1976) to assign salience to textual units that are then used to generate the summary. The graph is generated by creating a node for each word occurrence and linking nodes based on different cohesion links such as adjacency and coreference links as well as knowledge based links extracted from NetOwl and Wordnet. Words and phrases are assigned a salience score using a tf*idf score. The system takes a parameter specifying a topic with respect to which the summary is generated. Document nodes whose strings are equivalent to the topic terms act as entry points into the graph and a spreading activation algorithm is used to find nodes that are linked to the activated nodes till a system-defined threshold on the number of output nodes is met. To create the final summary, the system first finds nodes that are common and different between the documents to highlight similarities and differences between the set of documents. It then greedily selects sentences based on the average activated weight of the covered words. In Mani *et al.* (1998), the authors explored methods based on both cohesion and coherence for summarization. Cohesion is modeled by the graph structure as described earlier, while coherence is

modelled by Rhetorical Structure Theory (RST). They use the same salience metric for coherence as described by Marcu (1997), which we describe now.

Barzilay & Elhadad (1997) present an algorithm that produces summaries of a text by relying on a model of topic progression in the text derived from lexical chains. They computes lexical chains using Wordnet similar to Morris & Hirst (1991), but use a non-greedy method for disambiguating word senses. They then present a method for scoring lexical chains based on strength which is determined based on chain length and homogeneity. A summary representation is created by simply selecting chains that have a higher strength than a threshold. Sentences corresponding to these chains can then be extracted to build the summary. They present three different methods for extracting sentences corresponding to each chain: 1) choose the sentence that contains the first appearance of a chain member, 2) choose the sentences that contains the first appearance of a "representative" chain member (representative words have a frequency in the chain no less than the average word frequency in the chain), and 3) choose the sentence with the first chain appearance in a central text unit, defined as a text unit where the chain is highly concentrated. Silber & McCoy (2002) later presented a more efficient linear time model for lexical chain computation for summarization.

Brunn *et al.* (2001) also used lexical chains to develop a summarization system for the DUC 2001 summarization task. Their summarizer first segments the original text based on topic. This is followed by modules for part of speech tagging and syntactic parsing. This is followed by a noun filtering component that removes "noisy" nouns based on the heuristic that "nouns contained within subordinate clauses are less useful for topic detection than those contained within main clauses." Following this, they compute lexical chains by first selecting the set of candidate words that come from open class of words that are noun phrase or proper names. These candidate words are then exploded into senses, and semantic relatedness between two words exists

if their senses have an intersection. A strength is assigned to semantic relatedness based on "the length of the path taken in the matching with respect to the levels of the two compared sets.". The longest chains are retained based on a set of preference criterion. The sentence extractor first assigns a score to each segment using a tf*idf type score computed using chains and then the top n segments with the highest scores are selected for sentence extraction. Each sentence is then assigned a score by "summing the number of shared chain members over the sentence.". The final summary consists of a ranked list of top-scoring sentences.

Doran *et al.* (2004) describe a summarizer based on lexical chains and compare it with several earlier methods based on lexical chains. Unlike previous chaining approaches, their algorithm produces two disjoint sets of chains: noun chains and proper noun chains. Proper noun chains are created by using a fuzzy string matching function to find repetition relationships between proper nouns phrases like "George Bush" and "President Bush". They compared e five different chain scoring metrics, three based on semantic relationships between the words of the chain, one based on corpus statistics, and one that assigns the same score to each chain. The differences between the first three lie in the way relations in the chain are handled. They evaluate these systems using an extrinsic task: for each distinct set of summaries generated, summary quality is evaluated by observing whether the system can correctly detect if two documents are about the same event by looking at the two summaries.

We now move to coreference based systems. Baldwin & Morton (1998) presented a query-sensitive based summarization system based on dynamic coreference. They first compute noun phrase coreference relations between the tokens in query and the document. Event references are captured by associating verbs or nominalizations in the query with those in the document. These associations are then used to rank sentences from the document. Coreference chains are built by adding every token in the document that is associated with the same token in the query or headline to the

same chain. A number of scores are computed for each sentence based on coreference chains and these are used to sort the sentence. These scores are based on the overlap of the sentences with the coreference chains, e.g. "the number of coreference chains from the query which are covered by the sentence and haven't been covered by a previously ticketed sentence" and "the number of noun coreference chains from the query which are covered by the sentence and the number of verbal terms in the sentence which are chained to the query. " The selected sentences are then presented in the order in which they occurred in the document.

For DUC 2003, Bergler *et al.* (2003) used a method based on coreference chains to create 10-word indicative summaries of text. The approach is to simply order the entities in the text based on the number of times it is referred to in the text calculated as the length of its noun phrase coreference chain. Their coreference chain computation engine uses a knowledge poor noun phrase coreference system that models the certainty of the heuristics using fuzzy set theory.

Alonso i Alemany & Fuentes Fort (2003) tried to integrate both cohesion and coherence based cues for summarization. They start with lexical chains as computed by Barzilay & Elhadad (1997). The lexical chains in this method are scored exclusively based on lexical information. They augment this information by incorporating rhetorical and argumentative relations and found an improvement in scores compared to only using lexical information.

Karamuftuoglu (2002) presented an approach for summarization using lexical bonds in DUC 2002, where a lexical bond exists between two sentences if they share two or more word stems. The first sentence of the summary was the first sentence in the main body of the document that has at least one forward lexical bond. A summary was formed by following lexical bonds one by one from the source sentence to the one lexically bonded with it, and from that sentence to the other which has a bond with it, and so on.

Chan *et al.* (2000) reported on efforts to automatically identify and classify discourse markers in chinese texts using heuristic-based and corpus-based data-mining methods as part of automatic text summarization via rhetorical structure.

Finally, in his thesis (Bosma, 2008), Bosma discusses the role of discourse for summarization in detail and presents a graph based framework for automatic discourse oriented text summarization.

2.1.3 Summary Post-Processing

Summaries generated by extractive systems suffer from a lack of coherence – especially for multi-document summarization systems – because sentences that are picked from different documents or different parts of the same document may not fit well together. One way to alleviate this problem is to do some post-processing or "smoothing" on generated extractive summaries in order to improve the coherence and cohesion. Researchers in summarization recognized this problem early on and there have been several papers that have looked at strategies for smoothing out summaries that include sentence rewriting, sentence fusion and sentence ordering.

2.1.3.1 Information Fusion

McKeown *et al.* (1999) describe a multi-document summarization system for news that works by identifying and synthesizing similarities across a set of related document. Their systems consists of three main components that identify similar paragraphs or "themes", find intersection of similar phrases within paragraphs, and formulate a summary using language generation. For grouping paragraphs into themes, they extract a set of linguistic and positional features based on word co-occurence, matching noun phrases, wordnet synonyms and common semantic classes for verbs. These features are then used to train a supervised system on a manually labelled dataset of 8,225 paragraphs. Given a set of paragraphs, the classifier can yield pair-

wise similarity scores, which are then fed into a clustering algorithm that identifies the themes. The algorithms for finding theme intersection and summary formulation are described in greater detail in Barzilay *et al.* (1999). They try to go beyond extractive summarization by using a comparison of extracted similar sentences from a theme to select the phrases that should be included in the summary and sentence generation to reformulate them as new text. Given a set of sentences that form a theme, the first stage of their system, content planning, tries to find phrases that represent common information of the theme. This is done by first parsing the sentences to a dependency representation and then comparing the dependency trees for phrases in pairs of sentences (a phrase being defined as a verb with at least two constituents). This comparison can deal with some variation in the surface realization of phrases by using a set of paraphrasing rules. The common phrases found are filtered based on a frequency of occurrence threshold. In the second phase, these set of phrases are used as the input to the sentence generator, which is a variant of the FUF/SURGE system that the authors modified to work with phrases with input features derived from shallow analysis during content planning. The generation component is able to revise phrases by attaching new constituents as well. Finally, the ordering of sentences is guided by timestamps extracted from the sentences belonging to the theme being summarized.

Daume III & Marcu (2004) reported on a set of human evaluations for a restricted version of sentence fusion task where two humans are "given two sentences and asked to produce a single coherent sentence that contains only the important information from the original two." They did their analysis on a corpus of 50 pairs of sentence fusions extracted from a larger corpus of computer product reviews from the Ziff-Davis corporation. Using a series of evaluations, they showed that there is very little measurable agreement between humans regarding what information should be considered important.

2.1.3.2 Revision Strategies

Mani *et al.* (1999) present a system that tries to improve extractive summaries by revising them. Their approach to revision involves constructing an initial draft summary and then improving it by combining and excising information in the draft based on revision rules involving aggregation and elimination operations. Aggregation gathers and draws in relevant background information in the form of descriptions of discourse entities from different parts of the source while elimination increases the amount of compression. Their revision approach is based on representing input sentences as syntactic trees with nodes annotated with coreference information. The initial draft is generated by selecting salient sentences from the input; revisions are then done by selecting highly weighted sentences and applying a rule from a sequence of revision rules until it can no longer be applied. The rules can be unary rules applied to a single sentence or a binary rule applied to a pair of sentences. Revision rules carry out three types of operations: 1) elimination operations eliminate constituents such as parentheticals, sentence-initial PP's, and adverbial phrases, 2) aggregation operations combine constituents from two sentences (one of which can be a sentence not currently in the draft) based on referential identity (two sentences are candidates if they have NPs that are coreferential) , and 3) smoothing operations that apply to a single sentence to arrive at a more compact sentence by simplifying coordinated constituents and applying reference adjustment rules to "fix" the results of other revision operations.

Nanba & Okumura (2000) investigate factors that make extractive summaries hard to read using human evaluation and divide them into five classes, most of which are related to cohesion. 12 graduate students were asked to produce extracts of 25 news articles and then asked to revise them building a dataset of 343 revisions. The revisions were classified into the following five categories: 1) lack of conjunctive expressions/presence of extraneous conjunctive expressions, 2) syntactic complexity, 3) re-

dundant repetition, 4) lack of information, and 5) lack of adverbial particles/presence of extraneous adverbial particles. Compared with Mani *et al.* (1999), Category 4 is related to their aggregation operation, categories 1 and 5 are related to their elimination operation, while 2 and 3 are related to smoothing operations. They came up with revision rules for factors 1, 3 and 4. For 1, they created a list of 52 conjunctive expressions that are deleted whenever the extract does not include the related sentence (which are identified by a partial RST style discourse analysis). For 3, if subjects of adjacent sentences in an extract were the same, the repeated expressions were deleted. For 4, there are two rules: sentences with anaphors that don't have the sentence with its antecedent in the summary are deleted, and if a subject in a sentence in an extract is omitted, it's supplemented by the subject from the nearest preceding sentence whose subject is not omitted in the original text.

Jing & McKeown (2000) present a cut-and-paste summarizer that edits extractive summaries using operations derived from analysis of human abstracts. They wrote an HMM based decomposition method (Jing, 2002) that automatically identifies the most likely document position for each word in the human written summary. Based on their analysis, they define six operations based on a manual analysis of human abstracts for 30 articles: 1) sentence reduction that removes extraneous information from a selected sentence, 2) sentence combination that merges content from several sentences, 3) syntactic transformation such as moving the position of subject, 4) lexical paraphrasing, 5) generalization or specification that involves replacing phrases or clauses with more general or specific descriptions, and 6) reordering of extracted sentences. Based on this, they created modules for sentence reduction and sentence combination. The sentence reduction module removes extraneous phrases by identifying the obligatory components of a sentence important for its grammaticality, identifying important contents based on local context of a sentence, and identifying components that are likely to be removed using corpus statistics. The sentence

combination module includes operations for adding descriptions for named entities, aggregation (e.g. extract common subjects/objects), substituting incoherent phrases (e.g. dangling anaphora and noun phrases) and substituting phrases with more general or specific information. Rules for when to apply combination operations were manually written based on corpus analysis. The combination operations involved joining two parse trees, substituting a subtree with another, or adding additional nodes implemented using a formalism based on Tree Adjoining Grammar.

Otterbacher *et al.* (2002) investigate problems with cohesion in extractive multi-document summaries by analyzing a set of 15 news summaries. They first revised each summary manually resulting in 160 revisions. Based on this, they identified five major categories of concerns related to text cohesion in multi-document summaries: 1) discourse based concerns about the relationships between sentences in a summary, 2) entity based concerns involving resolution of referential expressions, 3) temporal concerns relating to establishment of the correct temporal relationships between events, 4) grammatical errors stemming from juxtaposition of sentences and previous revisions, and 5) location or setting based concerns that involve establishing where each event in a summary takes place. They found that 82% of the revisions fall in the first three categories. The last two categories were not so prominent because sentences in extractive summaries are typically grammatical and given the corpus, the place or setting of an event was typically obvious in the summary. Within discourse based revisions, the majority of the revisions were due to abrupt topic shifts (45%) and purpose of a sentence (33%). In all these cases, the revision involved adding a phrase or a sentence that provided a transition or motivation for the problematic sentence. Entity based concerns typically involved underspecified entities (38%), where the revision involved finding the missing information in the source documents so that an entity being introduced for the first time has the associated description. Rewrites for temporal concerns mostly involved issues with temporal

ordering of the extracted sentences (89%). The other errors in the temporal category were issues with time of event, event repetition, synchrony, and anachronism. The majority of the grammatical concerns resulted from previous revisions and included run on sentences, mismatched verbs, missing punctuation, and awkward syntax. Location/setting type of revisions were least frequent and typically involved sentences that retain the place/time stamp at the beginning of the article and missing location information. For each of these categories, they suggest operations that can be used to fix the issues. These operations were arranged in a taxonomy ordered by complexity "ranging from concrete repairs that require only knowledge of the surface structures of sentences, to knowledge-intensive repairs that cannot be implemented without a discourse model."

2.1.3.3 Sentence Ordering

Barzilay *et al.* (2001b) analyze ordering strategies for output of multi-document summarization for achieving coherence. Their summarization system clusters sentences from input documents into themes. They describe two naive ordering strategies. The majority ordering algorithm works by first creating pairwise counts of how often sentences from one theme occur before sentences from another theme. This gives them a directed graph between themes where the weights are the counts between the themes. Order in the summary is then calculated by looking for a path with maximal weight. The problem with this method is that it does not provide enough constraints to determine one optimal ordering and creates several orderings with the same weight. The chronological ordering algorithm works by assigning a date to each theme based on the first time the theme was reported in the input set and ordering sentences from themes in chronological order based on these dates. In their evaluation, assessors were asked to rate the information ordering in each summary as poor, fair or good. The majority algorithm produces a small number of good summaries but

most of the summaries are graded as fair. They found that majority algorithm does not work well when the input text have very different orderings. The chronological ordering produces a similar number of good summaries and a larger number of poor summaries. They found that badly ordered sentences resulting from chronological ordering cannot be ordered based on their temporal occurence. They also observed that poor summaries typically contain abrupt switches of topics and general forms of incoherence. To come up with a better ordering algorithm, they asked 10 human subjects to order 10 sets of sentences, obtaining 100 orderings. They noticed that majority of orderings were different but some sentences always appear together. They clustered sentences into blocks based on these adjacency relationships in the human orderings and found that these blocks correspond to clusters of topically related sentences that deal with the same subject and exhibit cohesive properties. Based on this, they create an augmented ordering algorithm that groups themes into blocks of topically related themes based on the occurence of themes in the same segment of text in original documents. Each block is then assigned a timestamp based on the earliest timestamp of the themes it contains. Themes inside each block are ordered using chronological ordering. This algorithm ensures that cohesively related themes appear together in generated summary and avoids abrupt topic switches. In their evaluation, their augmented algorithm showed a significant improvement in the number of good summary orderings.

Lapata (2003) describe an unsupervised probabilistic model that learns ordering constraints from a large corpus. Their method represents sentences as a set of informative features and learns which sequence of features are likely to co-occur and makes predictions concerning preferred orderings. Their model uses a markov assumption where the probability of a given sentence depends only on its previous sentence. They extract feature representations for each sentence in the corpus and compute the probability of a pair of features occurring one after the other in the corpus. The features

extracted from sentences include verbs, nouns, and dependencies. Given a new set of sentences, the features are first extracted from each sentence and the sentence are represented as a graph where the set of vertices are sentences and the edges represent the probability of the vertices appearing after each other. Finding an optimal ordering through this directed graph is an NP complete problem, but a simple greedy algorithm that starts with the vertex with the highest probability and then selects the next set of nodes based on the conditional probability given previously selected nodes provides an approximate solution.

Bollegala *et al.* (2010) present a bottom up approach for arranging sentences in extractive multi-document summaries. In their method, segments represent a sequence of ordered sentences. Starting with a set of segments with one sentence each, the algorithm works by combining segments with the strongest association recursively. The direction and strength of association between two segments is based on four criteria: 1) chronology criterion based on arranging sentences chronological order of publication date, 2) topical-closeness criterion which deals with association based on topical similarity (computed using a cosine similarity of vectors consisting of nouns and verbs in sentences), 3) precedence criterion which measures the substitutability of the presuppositional information of a segment as a segment that appears before it, and succession criterion which assesses the coverage of the succedent information for a segment by the segment that appears after it. They combine these criteria using an SVM trained on human-ordered extracts.

2.1.3.4 Other Work

Knight & Marcu (2002) argue that a good summarization system should be able to compress sentences while keeping the most important information. They present two data-driven approach to sentence compression: the first is a probabilistic noisy-channel model, the second is a decision-based, deterministic model. In the noisy

channel framework, a long string is assumed to be originally a short string to which some additional, optional text was added. Given this, compression is a matter of recovering the original short string. The noisy channel model has three parts: the source model which assigns to every sentence s a probability $P(s)$ of being generated from an original short string, the channel model which assigns to every pair of sentences (s, t) a probability $P(t|s)$ for the short string s being expanded to the long string t, and a decoder which given a long string t, looks for the short s that maximizes $P(s|t)$. The probabilities are actually assigned to syntactic trees instead of strings. For training the model, they used 1067 sentence pairs automatically from article/abstract pairs in the Ziff-Davis corpus that contains newspaper articles. The decision based model is based on rewriting a parse tree into a smaller tree using a series of shift-reduce-drop actions. It is more flexible than the noisy channel model by allowing the derivation of trees whose skeleton can differ quite drastically from that of the tree given as input. A learning case is assigned to each configuration of the shift-reduce-drop rewriting model; 46383 learning cases were derived from the 1067 pairs of sentences and a set of 99 features were associated with each learning example. Given this, the decision-based compression module learns decision trees that specify how large syntactic trees can be compressed into shorter trees.

Miller (2003) describe a system that aims to produce coherent extractive summaries by adding linking material between semantically dissimilar sentences. Given a document, the sentences are first divided into topics. A representative sentence is chosen from each topic by finding the sentence that has the highest semantic similarity to all the sentences in the topic segment (the similarity is computed using LSA based vectors computed from the term-sentence co-occurence matrix). The coherence of the resulting summary is then improved by considering the glue sentences for each pair of sentences in the original extract. A glue sentence is a sentence which occurs between the two sentences in the source document and is semantically similar to both.

The glue sentences are selected based on a metric that rewards glue sentences which are similar to their boundary sentences, but penalizes glue sentences that are too biased in favour of only one of the boundary sentences. The gluing process continues recursively that stops upon encountering one of four stopping criteria.

Nenkova (2005) discussed the problem of generating appropriate referring expressions to entities in extractive summaries. They conduct a corpus study to identify syntactic properties of first and subsequent mentions to people in news data and created a statistical model of the flow of referring expressions which suggests a set of rewrite rules that can transform an extractive summary back to a more natural and readable text. Their model is based on the idea that the syntactic form of a reference depends on the syntactic forms of previous mentions. Their model was trained on a corpus of news stories and the highest probability paths in the model were coded as rewrite rules. In addition to these, they performed a small manual evaluation for the task of annotating hearer-old vs hearer-new status of entities and found substantial agreement between evaluators, suggesting that this is a doable task. An SVM trained for this classification achieved 76% accuracy. With this component, a summarization system can omit descriptions of hearer-old entities.

2.1.4 Automatic Coherence Metrics

Lapata (2005) explores linguistically rich quantitative models for automatic evaluation of text coherence. They focus on local coherence, which "captures text organization at the level of sentence to sentence transitions". They explore two broad classes of coherence models. models based on syntactic aspects of coherence characterize the transitions of entity mentions in different syntactic positions across adjacent sentence. Models based on semantic aspects of coherence quantify local coherence as the degree of connectivity across text sentences in terms of semantic relatedness. The syntactic models are built around the idea of the entity-grid, which was further developed in

a separate paper summarized below. The semantic models are based on modelling similarity using word-based, distributional, and taxonomy based measures. In their experiments, they found that a model that fuses the syntactic and the semantic views yields substantial improvement over any single model.

Barzilay & Lapata (2005) describes in detail the entity-grid model that was previously mentioned. Entity-grid is a representation of discourse that captures patterns of entity distributions in a text and is based on Centering Theory (Grosz *et al.* , 1995). Entity-grid represents a piece of text as a two-dimensional array where the rows correspond to sentences and the columns correspond to discourse entities (where discourse entities refer to a class of coreferent noun phrases). For each discourse entity, the corresponding cell contains information about its presence in the set of sentences as well as the syntactic role in each sentence as being either subject, object, or neither. This representation makes it possible to extract features based on entity transitions, e.g. subject followed by object, and encode a piece of text as entity transition sequences. The probability of various sequences in coherent texts can be learnt from data and these can be uses to evaluate the local coherence of a new text. The paper presents experimental results that show that entity-grid performs well on the tasks of automated sentence ordering, summary coherence rating, and readability assessment.

Finally, Christensen *et al.* (2013) present a joint model for optimizing salience and coherence in multi-document summarization. Their system called G-Flow models coherence using a discourse graph that represents pairwise ordering constraints on input sentences. Edges between sentences are formed based on deverbal noun references, event/entity continuation, discourse markers, cross-document inferred edges based on semantic overlap, and coreferent mentions. Each edge is assigned a weight based on how many of the above indicators participate in that edge, with negative edges added for unfulfilled deverbal noun references, discourse markers, and co-reference mentions. Given this, the coherence score for a piece of text is estimated as the sum of edge

weights between successive summary sentences. The salience of a sentence is estimated using a linear regression classifier trained on ROUGE scores over the DUC 2003 datset, and the salience of a piece of text is the sum of salience of individual sentences. Redundancy between two sentences in G-Flow is captured by overlap of relational tuples extracted using an open information extractor. Their objective function combines the salience, coherence and redundancy scores. The summarization process then becomes an optimization process that starts from an initial summary and uses stochastic hill climbing with random restarts to find a local optimal summary.

2.2 Scientometrics

2.2.1 Measuring Scientific Impact

2.2.1.1 Citation Based Metrics

Citation based measures depend on using a citation to a paper as a vote for the paper and use aggregate statistics on citation counts to estimate research impact. Perhaps the most widely known citation based metric for evaluating the impact of scholars is the h-index (Hirsch, 2005) which is defined as the number of papers with citation number greater than or equal to h. In a later paper (Hirsch, 2010), Hirsch acknowledged the shortcoming of the h-index not taking into account the number of coauthors of a paper and proposed a variant of h-index that accounts for multiple coauthorship.

Several studies have suggested other problems with the h-index. Gaudry (2006) studied the resistance of h-index to malicious users, e.g. a researcher who wants to artificially increase his h-index. They describe several problems with the reliability of h-numbers and suggest some countermeasures. Bletsas & Sahalos (2009) show that expected value and standard deviation constitute the minimum information required for meaningful h-index ranking campaigns and propose scaling of the h-index based on

its probability distribution which is calculated for any underlying citation distribution. Dehghani *et al.* (2011) use a dataset of 999 authors from three fields in the Scopus dataset to analyze how the h-index of authors are affected by citations from co-authors as well as self citations. Ferrara & Romero (2013) propose a method to discount the effect of self-citations in computing h-index and similar measures. Schreiber (2013a) cast doubts on the predictive power of the h-index by reporting results from a study that shows "the increase of the h-index with time after a given point of time is not necessarily related to the scientific achievements after this date."

Havemann & Larsen (2014) compare 16 bibliometric indicators for a sample of 29 astrophysics researchers who later co-authored highly cited papers before their first landmark paper with the distributions of these indicators over a random control group of 74 young authors in astronomy and astrophysics. The indicators that were tested were based on productivity, influence, and collaboration. They found that indicators of total influence based on citation numbers normalised with expected citation numbers are the only indicators among a total of 16 which show significant differences between later stars and random authors. Specifically, the h-index performs much worse than indicators of total influence.

Other researchers have proposed variants of the h-index that alleviate some of its problems. The e-index (Zhang, 2009) complements the h-index by taking into account the excess citations that are ignored by the h-index. Cormode *et al.* (2012b) propose a variant of h-index called the social h-index that redistribute the h-index score for collaborative papers in order to reflect an individual's impact on the research community. Another variant of h-index is the g-index Egghe (2006), which is calculated as the largest number such that top g articles received at least g^2 citations. The authors claim that the g-index better takes into account the citation scores of the top articles. They also proposed a normalized version of g-index in a later paper (Egghe, 2014).

Zhu *et al.* (2015) present a variant of h-index called the hip-index (influence

primed h-index). They posit that not all citations that are referenced by a paper are equally important and create a new dataset consisting of papers and references that were influential for these papers. The ground-truth dataset was created by asking authors of 100 papers to identify the essential references in their paper. They then trained a SVM model using five classes of features based on counts, similarity, context, position and miscellaneous. They found that the number of times a reference was cited in a paper was the most predictive feature for academic influence. Given this, the influenced primed citation count for a paper is computed by summing the square of in-paper citation count for the paper in each of the citing papers. The hip-index is computed similarly to h-index except that the citation counts used are the influenced primed citation counts. They found that hip-index does a better job of finding ACL fellows. Alonso *et al.* (2009) is a comprehensive review of the h-index and its related indicators as well as their main advantages, drawbacks, and applications.

Giles & Councill (2004) suggest that acknowledgements can be used as parallel metric to evaluate academic contribution and present methods for automatic acknowledgement extraction. They propose that acknowledgment analysis can be combined with citation indexing to measure the average impact of the research influenced of individuals as well as sponsors of scientific research.

Sidiropoulos *et al.* (2014) propose new indicators to measure performance of a researcher that take into account the distributions of citations per article and categorize researchers into two categories: influentials that produce articles with high impact and mass producers that produce a lot of articles with moderate or zero impact.

Mimno & McCallum (2007) present a method for finding the most influential authors on a particular topic using a probabilistic model that combines citation pagerank scores with lexical information from paper abstracts.

Ding *et al.* (2009) experiment with different damping factors for Pagerank algorithm to rank authors in author co-citation networks. They also experiment with

weighted Pagerank algorithms on the co-citation network where the weight for a node can be the count of publications or citations of the author. They select 108 most highly cited authors in the field of information retrieval for their experiments. They present a detailed analysis of how the rankings of various authors changes as the damping factor in the Pagerank algorithm is varied as well as the effects of using the two weighting factors. They also compare their measure with other network based measures such as degree centrality, betweenness centrality, and closeness centrality and found that citation rank is highly correlated with PageRank with different damping factors but citation rank and PageRank are not significantly correlated with centrality measures. They also found that h-index is not highly correlated with the rest of the measures.

Pepe & Kurtz (2012) propose two metrics for research productivity: 1) total research impact (tori) that quantifies for a researcher the total amount of scholarly work that others have devoted to his/her work, and 2) research impact quotient (riq), an age independent measure of an individual's research ability. They show that these measures are comparatively less vulnerable to temporal debasement and cross-disciplinary bias.

Similar citation based metrics have also been proposed for journals and research groups. One of the most widely used metric for ranking journals is the journal impact factor (Garfield, 2006b) but h-index type metrics have also been proposed (Braun *et al.* , 2006). Eigenfactor (Bergstrom *et al.* , 2008) is a Pagerank style metric to measure the value and prestige of scholarly journals. Mryglod *et al.* (2012) proposed citation based measures for research excellence of research groups and found that citation-based indicators are "highly correlated with peer-evaluated measures of group strength but are poorly correlated with group quality."

2.2.1.2 Criticism of Citation Based metrics

Several researchers have cast doubts about the use of citation based measures for scientific impact.

Moravcsik & Murugesan (1975) do a quantitative analysis about the use of citation counts for measuring impact. They selected 30 articles from physical review and classified each reference in these articles into several categories. One of their main results was that about two-fifths of the references were perfunctory, raising serious doubts about the use of citations as a quality measure.

Radicchi (2012) did empirical analysis using data from thirteen major publication outlets in science and found that commented papers are more cited than non commented papers, despite the fact that the goal of comments is to correct or criticize previously published papers.

Waltman *et al.* (2013) propose a model in which the number of citations to a publication depends on a set of random factors in addition to the scientific impact of the publication. Their analysis suggests that there might be systematic consequences of random effects in citation counts and more may not always be better in terms of using citation counts to assess impact of publications. They propose a modified highly-cited-publications (HCPs) indicator. However Schreiber (2013b) reported some inconsistencies with this indicator as well.

Petersen *et al.* (2013) explore the relationship between reputation and career growth of researchers through a longitudinal analysis on the careers of 450 highly-cited scientists. They found that a publication can gain a significant early advantage corresponding to roughly a 66% increase in the citation rate for each tenfold increase in reputation (measured by the total citations of the author).

Brembs & Munafo (2013) perform a similar analysis for the reputation of journals and conclude that "journal rank is a weak to moderate predictor of utility and perceived importance" and, disturbingly, that "journal rank is a moderate to strong

predictor of both intentional and unintentional scientific unreliability".

Amancio *et al.* (2010) use formalisms of complex networks for two datasets of papers from the arXiv repository to show that authors often don't follow important principles while preparing related work sections of their scientific manuscripts which can hamper fair assessment of authors usually done on the basis of citations.

Other methods of measuring impact of papers that do not depend solely on citations have been suggested. Dietz *et al.* (2007) present an unsupervised probabilistic topic model that model the influence of citations in paper collections. The model's ability to predict the strength of influence of citations was evaluated against manually rated citations. Recent work has also shown that taking into account the number of times a paper is cited in the citing paper often does a better job of measuring the impact of the cited paper (Wan & Liu, 2014b; Hou *et al.* , 2011). Wan & Liu (2014a) present a regression method for automatically estimating the strength value of each citation and show the estimated values can achieve good correlation with human-labeled values. Thelwall *et al.* (2013) evaluate the use of measurements derived from social web sites such as twitter as early indicators of the impact of scientific articles. Their results show that such metrics associate with citation counts, but do not provide conclusive evidence about the magnitude of correlation. Additionally, the coverage of all the altmetrics seems to be very low and thus cannot be used as universal sources of evidence.

2.2.2 Social Dynamics of Science

Several researchers have discussed the nature of science and scientific discovery. Contrary to the general public view of the scientist as lonely researchers toiling at insurmountable problems in the lab, the process of science should be regarded as a social process with its own social norms. We first give a overview of two books that look at the social aspects of scientific development, followed by a literature review of

research that tries to quantify the nature of social dynamics of science.

In his 1968 essay (Ziman, 1968), Ziman defines science as "Public Knowledge". According to the essay, science is not merely published knowledge or information[1], but comprises facts and theories that survive a period of critical study and testing and are found persuasive enough to be universally accepted. Thus the goal of science is a consensus of rational opinion over the widest possible field, and scientific research in this sense should be regarded as a social activity. This argument is elaborated throughout the 150 page essay with discussion of topics ranging from what distinguishes science from other fields such as law, philosophy, technology etc., the nature of scientific method and argument, the problems in teaching science, and the roles of individual scientist, community, institutions and authorities.

A counterpoint to Ziman's discussion is provided by Kuhn's theory about the structure of scientific revolutions (Kuhn, 1970). In his book, Kuhn first introduces normal science characterized by a set of shared beliefs or paradigms generally accepted by the scientific community that students study in order to become members of the scientific community. Scientific research in normal science is seen as puzzle solving: an attempt to force nature into the conceptual boxes supplied by professional education. This is disrupted by the emergence of a "crisis", a failure of existing theory to solve problems or explain facts, leading to new theories that create a paradigm-shift. These new theories compete with each other until a satisfactory paradigm is adopted by the scientific community that replaces the existing paradigm. This is the idea of scientific "revolution" at the core of Kuhn's theory.

2.2.2.1 Models of Scientific Growth

Lehman (1947), measured the rate of growth of publications in several areas of sci-

[1]Interestingly, he supports this by noting that "anyone can make an observation, or conceive a hypothesis, and, if he has the financial means, get it printed and distributed". This is even more pertinent today with easy and cheap means of publication through the world wide web available to most individuals.

ence including chemistry, genetics, geology, mathematics, pathology, economics, and philosophy starting from about A.D. 1500 till early 20th century and found that the rate is exponential in nature. Four years later, de Solla Price (1951) analyzed growth of publications for two fields: a general field, physics and a specialized field, determinants and matrices. He found that the during normal times a general field such as physics increases exponentially, while a specialized field increases exponentially to a certain point at which the growth changes to linear. He hypothesized that the switch from exponential to linear growth occurs as a result of researchers being attracted initially to a new and growing field, and "as the research front moves forward, recruiting to some fields slows and stops [...] leaving a constant body of workers in those particular fields." In later work (de Solla Price, 1970), Price examined journals from different fields. Price's classic 1963 book (reprinted with later writings in de Solla Price (1974)) examined the transformation in the structure of science from Little Science consisting of individual researchers largely working on their own to Big Science characterized by large-scale collaborations and expenditures of man-power and money.

More recently, Bornmann & Mutz (2014) do a detailed analysis of the stages of growth in science since the mid 1600's using the Web of Science corpus. Redner (2004) analyzed the citations for all publications in the Physical Review journal from 1983 to 2003. Some of their main findings are that the growth of citations follows a linear preferential attachment model, citations from a publication have an exponentially decaying age distribution, and citations to a publication follow a power law age distribution.

2.2.2.2 Models of Author Productivity and Collaboration

Zuckerman (1967) analyzed the productivity patterns of Nobel laureates and found that they begin publishing earlier, publish for longer, and publish at a much higher

rate. They also found that laureates tend to collaborate with other distinguished and highly productive scientists. Finally, they found that the productivity of laureates tends to decline sharply in the five years following receipt of the prize, which can be attributed to high visibility following the prize leading to increased demands of social requests and invitations.

Wei *et al.* (2013) try to understand if researchers tend to follow hot topics. They found that new papers are more likely to be attracted by hot fields, but "there are qualitative differences among scientists from various countries, among research works regarding different number of authors, different number of affiliations and different number of references".

Viana *et al.* (2013) derive collaborative networks parametrized along time. They define affine groups as a set of authors that either co-author papers or belong to the same community and show that the average size of the affine groups grows exponentially, the number of authors increases as a power law, and larger affine groups tend to be less stable.

Velden *et al.* (2010) study patterns of collaboration between authors in three research fields in chemistry. They base their research on both the individual researcher as node in a co-author network and the observed modular structure of co-author networks. They find two main types of coauthor-linking patterns: 1) transfer-type connections due to career migrations or one-off services rendered, and 2) stronger, dedicated inter-group collaboration. They propose that the coauthorship network of a research area can be understood as the overlay of these two types of cooperative networks.

2.2.2.3 Interesting Case-Studies

Deutsch *et al.* (1971) tried to characterize major advances in the field of social science and found that they came from efforts in a small number of interdisciplinary

centers and had widespread acceptance in surprisingly short spans of times. Midorikawa (1983) compare 15 subfields in physics by investigating half-life, citation degree, form dispersion and title dispersion in articles from 74 physics journals. They conclude that the most significant communication medium for most physics subfields was the journal, but for subfields in which large experimental or observational devices are used, the use of reports and letter journals was prevalent as well.

Van Raan (2004) try to do a bibliometric analysis of the phenomenon of sleeping beauties: publications that go unnoticed for a long time and then suddenly attract a lot of attention. They model three main self-explanatory variables: 1) depth of sleep or the number of citations (0-2) that the paper receives during its sleep period, 2) length of sleep, and 3) awake intensity measured by the number of citations during four years after the sleep period. They used a corpus of 20,000,000 articles to measure the above values and derived a grand equation that gives the number of sleeping beauties for any sleeping time, sleep intensity and awake intensity. Based on their observations, they find several characteristics of sleeping beauties: probability of awakening after a deep sleep is smaller for longer sleeping periods, for a less deep sleep, the length of the sleeping period matters less for the probability of awakening and, the probability for higher awakening intensities decreases extremely rapidly. Later research in Costas *et al.* (2009) proposed a new methodology for the general analysis of ageing and "durability" of scientific papers that classifies documents into three general types: normal documents which have a typical distribution of citations over time, delayed documents which receive the main part of their citations later than normal documents (these include sleeping beauties), and flash in the pans which receive citations immediately after their publication but are not cited in the long term.

Martin *et al.* (2013) use publications from Physical Review over a span of 116 years from 1893 to 2009 to study a hybrid coauthorship/citation network. Their

main findings were that Physical Review is growing exponentially, the fraction of self-citations and citations among coauthors is more or less constant over time, authors tend to cite their own papers sooner after publication than do their coauthors (who in turn cite sooner than non-coauthors), a strong tendency towards reciprocal citations, and a small triadic closure effect (two researchers who share a common coauthor but have never collaborated themselves have only a small probability of collaborating in future).

Kuhn *et al.* (2014) analyze memes in scientific literature and find that they are governed by a surprisingly simple relationship between frequency of occurrence and the degree to which they propagate along the citation graph. They propose a formalization of this pattern and evaluate it using a datset of 50 million publications from Web of Science, PubMed Central, and American Physical Society.

Arbesman & Christakis (2011) introduce a discovery based approach to scientometrics called eurekometrics that is concerned with quantitatively examining scientific discoveries rather than examining the properties of scientific publications. They discuss the new resources that make such research possible, and analyze its potential and limitations.

Guerini *et al.* (2012) did a preliminary examination of the correlation of stylistic aspects of a paper with its popularity by analyzing the words in the abstracts of the papers. They used the Linguistic Inquiry and Word Count (LIWC) corpus (Pennebaker *et al.* , 2001) to tag the words in the abstracts based on the psycholinguistic class it belongs to. LIWC includes about 4,500 words and stems that are grouped into 65 categories based on their linguistic category (e.g. 1st person singular, Future tense) and psychological processes (e.g. Certainty, Negation).

2.2.2.4 Mapping History of Scientific Fields

Garfield *et al.* (1984) proposed the use of citation data in doing historical analysis

of science. In their study, they created two models of the history of the discovery of the DNA: one using Issac Asimov's book called "The Genetic Code", the second using bibliographic citation data contained in the documents which are the original published studies of events represented in the Asimov book. They found that 65% of historical dependencies in Asimov's book were confirmed by linkages established by citations while 31 new citation connections were found with no reported historical dependency reported in Asimov's book. Based on this, they conclude that "bibliographic citation data, if presented in the form of network diagrams and or citation indexes, reveal historical dependencies which can be easily overlooked by the historian.".

There has been a lot of interest in using maps to visualize the history of scientific fields. Fukuda *et al.* (2012) present a method for automatically creating a technical trend map from both research papers and patents by focusing on the elemental (underlying) technologies and their effects. Fried & Kobourov (2013) describe a different approach to visualize research fields by creating maps where words and phrases are cities in the map and countries are created based on word and phrase similarity, calculated using co-occurrence. Heatmaps can then be used to visualize profiles of conferences, journals, researchers and departments. Shahaf *et al.* (2012) present yet another visualization technique: building metro maps of scientific fields that can show the relationship between papers and the evolution of the research field.

Sim *et al.* (2012) propose a probabilistic model over the citations between authors and the words used to do the citing to discover latent factions in the computational linguistics community such as discourse and parsing. They analyze the relationships between different factions and their evolution over time. Anderson *et al.* (2012) have also presented a historical analysis of computational linguistics by identifying topics authors work in and how authors move between different topics. They also identify four epochs in the history of the field where topical overlaps are stable and use the flow

of authors between different fields to discover how some subfields flow into the next. Scientometrics studies of many other fields have been conducted including computer science (Guha *et al.* , 2013), quantum information processing (Winiarczyk *et al.* , 2012), high energy physics (Pia *et al.* , 2012), chemistry (Milard, 2014), and law (Liu *et al.* , 2014a), specific conferences such as CHI (Bartneck & Hu, 2009) and CSCW (Horn *et al.* , 2004) and interestingly enough, scientometrics itself (Szanto-Varnagy *et al.* , 2014).

Erosheva *et al.* (2004) presented a mixed-membership model classifying papers in biological science based on semantic decompositions of abstracts and bibliographies and found that the traditional discipline classifications correspond to a mixed distribution over the internal categories. In their later paper (Airoldi *et al.* , 2010), they explore ways to classify papers published in PNAS that capture the interdisciplinary nature of science. Serpa *et al.* (2012) present Statistical Common Author Networks (SCAN), a method to visualize the relatedness of scientific areas based on measuring the overlap of researchers between those areas.

2.2.2.5 Models of Evolution

Several models for evolution of citation relationships have been proposed. Liu & Rousseau (2014) characterize the citation graphs of articles written by Nobel Prize winners in physics and show that concave, convex, and straight curves represent different types of interactions between old ideas and new insights. Borner *et al.* (2004) presented a process model for the simultaneous evolution of coauthor and paper citation networks and validate the model against a 20-year dataset of PNAS articles. The model incorporates a partitioning of authors and papers into topics, a bias for authors to cite recent papers, and a tendency for authors to cite papers cited by papers that they have read. Peterson *et al.* (2010) presented a dual model for distribution of citations where an author cites an old paper directly or the author

finds an old paper via the reference list of a newer intermediary paper. Their results on the ISI database indicates that papers having fewer citations are mainly cited by the direct mechanism while classic papers are cited mainly by the indirect mechanism (where papers become classic in the ISI database after accumulating 25-40 citations depending on the database). They also found that the power law exponent is different between highly cited and less cited individuals.

A different set of models based on lexical analysis have been explored in the *natural language processing* community. Gerrish & Blei (2010) present a dynamic topic model that probabilistically models how topics change over time and propose a document influence model that measures the importance of a paper by measuring how its appearance changes the language of subsequent papers in the field. Hall *et al.* (2008) use LDA to divide ACL Anthology publications into topics and examine the popularity of various topics over time. They also define a metric for measuring the topic entropy of a conference and use it to analyze the diversity of various conferences in the ACL Anthology such as COLING, ACL and EMNLP.

Gupta & Manning (2011) present a novel approach for analyzing the dynamics of research. They first extract concepts corresponding to focus, technique and domain from the papers in the ACL Anthology using a bootstrapping algorithm. Once these concepts are extracted, the influence of communities (which are extracted using topic models) on each other is measured based on the number of times its focus, technique or domain have been used as a technique in other communities. Using this, they are able to observe trends such as the decline in influence of the speech recognition community in recent years and the increase in the influence of the statistical machine translation community.

Yun *et al.* (2014) propose a mechanistic model for scientific evolution using analysis of a corpus of digitized English texts between 1800 and 2008. Their results indicate that slowly-but-commonly adopted science and technology tend to have higher innate

strength than fast-and-commonly adopted ones.

Sun *et al.* (2012) presented an agent-based model that uses social interactions among agents representing scientists in a social network of collaborations to guide the evolution of disciplines. New disciplines emerge from splitting and merging of social communities in a collaboration network. They validated their model by showing a good fit against three datsets: NanoBank, Scholarometer, and Bibsonomy.

Recently, Wang *et al.* (2013) presented a mechanistic model for citations to scientific papers with three parameters: 1) the relative fitness or perceived novelty and importance of the paper, 2) immediacy that captures the time for the paper to reach its citation peak, and 3) longevity that captures the decay rate for the citations received by the paper. They showed that citation histories of the papers from Physical Review fit their model well and that the model has predictive power as well.

2.2.2.6 Reliability of Scientific Publications

A major issue in science has been reliability of results presented in scientific publications. Hamblin (1981) describe several cases of fake science that range from glaring experimental errors[2] to downright fraud and plagiarism. He suggests that one of the reasons for this behavior was the enormous pressure on scientists to consistently publish and come up with big breakthroughs. Ioannidis (2005a) designed a study to understand how often highly cited claims in clinical research are contradicted afterwards. They looked at studies published in 3 major journals between 1990-2003 and cited more than 1000 times. Of the 49 highly cited students, 16% were contradicted by subsequent studies and 16% found effects that were stronger than subsequent studies. Only 44% of these were replicated and 24% remained largely unchallenged. In another controversial paper published in the same year (Ioannidis, 2005b), the au-

[2]The article opens with an amusing case study: a tenfold overestimation of the amount of iron in spinach reported in 1890's that was used as a propaganda tool during the meatless days of the second world war. This claim was refuted in 1930's and ascribed to a misplaced decimal point by the original researchers.

thors show that for most study designs and settings, it is more likely for a research claim to be false than true and that for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias (Although this is perhaps not as shocking given Kuhn's characterization of science). In a later essay, Moonesinghe *et al.* (2007) extend this analysis to show that replication of previous studies can improve the predictive value of the truth of a research finding. Tatsioni *et al.* (2007) evaluate the citations for two highly cited epidemiological studies that have "continued being defended in the literature, despite strong contradicting evidence from large randomized clinical trials." They also examined the amount of supporting citations for highly cited but contradicted protective effects of beta-carotene on cancer and of estrogen on Alzheimer disease. They found the citing articles for these studies to be equivocal and consisting of "a range of counterarguments raised to defend effectiveness against contradicting evidence."

2.2.2.7 Interdisciplinarity

Stirling (2007) present a framework for analyzing diversity in science that is based on three properties: 1) variety, which answers how many types of thing do we have?, 2) balance, which answers how much of each type of thing do we have?, and 3) disparity, which answers how different from each other are the types of thing that we have? They propose a quantitative non-parametric diversity heuristic that allows for a systematic exploration of diversity under different perspectives.

Rafols & Meyer (2009) present a conceptual framework for interdisciplinarity based on diversity indicators that describe the heterogeneity of a bibliometric set viewed from predefined categories and coherence indicators that measure the intensity of similarity relations within a bibliometric set. Porter & Rafols (2009) compute bibliometric indicators as well as an interdisciplinarity index for publications in six research domains from 1975 to 2005. Their results show major increases in the number

of cited disciplines and references per article but only a modest increase in the interdisciplinarity index, which suggests that science is becoming more interdisciplinary but in small steps. Cordier (2012) propose a measure of similarity between journals based on co-authorship and measure of homogeneity of a publication list that can be used to characterize its interdisciplinarity.

2.2.2.8 Cross-disciplinary studies

Tang (2008) analyze citation distributions for 750 randomly selected scholarly monographs in religion, history, psychology, economics, mathematics, and physics. Their main findings were: in contrast with general assumption, the average book citation counts were lowest for religion and history and highest for psychology; monographs in psychology, mathematics, and physics have a statistically higher citation rate than those in religion, history, and economics; half-life for monographs in physics was found to be longest (13 years) while the shortest half-lives were history and psychology.

Van Zyl & Van Der Merwe (2012) study citation statistics across various fields including chemistry, economics, mathematics, biology, and physics and conclude that there is a large difference in how research is cited across various disciplines and "any attempt to rate researchers or journals using a single measure or unified benchmark system across subject fields is thus inherently biased towards fields with a natural culture of high numbers of citations per document.".

2.2.2.9 Role of Gender and Geography

West *et al.* (2012) quantitatively assess the role of gender in scholarly authorship using a dataset of over eight million papers across the natural sciences, social sciences, and humanities and note that even though women are increasingly represented as authors of scientific publications, gender inequities persist in subtle ways.

They find that "in certain fields, men predominate in the prestigious first and last author positions [...] women are significantly underrepresented as authors of single-authored papers". Mihalcea & Welch (2015) try to identify differences in the topics of interest across genders in the field of computer science by using data mining on publications in the ACM digital library. They found that there were areas in CSE (e.g. human-centered computing and applied computing) that were clearly preferred by women.

Pan *et al.* (2012) analyzed publications between 2003 and 2010 in the ISI Web of Science dataset to understand the role of geography played in the dynamics of science. They report that citation flows and collaboration flows between cities decrease with distance and follow gravity laws, and the total research impact of a country grows linearly with the amount of national funding for research and development but is also governed by a threshold effect.

Birnholtz *et al.* (2013) conduct a case study to understand the nature of collaboration across two campuses belonging to the same institution but located at two different geographic locations. Their results suggest that cross-campus collaboration is increasingly common but is mainly accounted for by a small number of departments and researchers.

2.2.3 Literature Search

This subsection focuses on retrieval models that can be used to find documents related to a topic input by the user.

Strohman *et al.* (2007) present a literature search system that takes the manuscript of a paper as a query. It then finds relevant document by using textual similarity and citation network and then ranks them using 6 hand-engineered features. Nascimento *et al.* (2011) also present a system that can recommend papers provided a manuscript, their system generates several potential queries by using terms in that

paper, which are then submitted to existing Web information sources that hold research papers. Once a set of candidate papers for recommendation is generated, the framework applies content-based recommending algorithms to rank the candidates in order to recommend the ones most related to the input paper. In a different flavor of the problem, Bethard & Jurafsky (2010) present a retrieval model for finding which papers should be cited given the abstract of the paper. They learn a supervised model using several features like topical similarity, recency and similarity based on citing terms.

El-Arini & Guestrin (2011) instead use a few seed papers as a query and then find more papers by using an objective function based on a notion of influence between documents. The influence between two documents is based on the technical terms or informative phrases that are shared between these documents. He *et al.* (2010) presented a context aware citation recommendation system that can suggest citations at a given location in a manuscript based on the surrounding context of the location. Their system can also predict the citations to a paper. In a later paper (He *et al.* , 2011), they present a system that can take a manuscript as input and suggest the locations where citations are needed.

Huang *et al.* (2012) present a method for "translating" paper drafts into references. They use citations and their contexts from existing papers as parallel data written in two different languages and use a translation model to create a relationship between these two "vocabularies". In a more recent paper (Huang *et al.* , 2015), they present a neural probabilistic model that jointly learns the semantic representations of citation contexts and cited papers in order to predict a reference given a citation context.

In other work, (Sugiyama & Kan, 2013) attempt to alleviate the problem of sparsity in the citation network by identifying "potential citation papers" through the use of collaborative filtering. Küçüktunç *et al.* (2014) attack the problem of diversify-

ing the results of citation-based literature search. They present random-walk based diversification algorithms, enhance them with direction awareness to allow users to reach classic papers or more recent papers, and propose a set of new algorithms based on vertex selection and query refinement.

2.2.4 Forecasting Scientific Impact

Several recent papers have explored the problem of forecasting future scientific trends. The 2003 KDD Cup (Gehrke *et al.* , 2003) focused on the task of predicting the difference between the citations received by well cited papers between two consecutive 3 month spans. The winning team (Manjunatha *et al.* , 2003) used a time series based approach to learn citation patterns over time.

(Yan *et al.* , 2011) use a variety of features including topic models, diversity, and recency to predict the exact future citation counts of papers and report R^2 values of 0.75 and 0.79 for 5-year and 10-year predictions respectively. In later work (Yan *et al.* , 2012), they do more comprehensive analysis and report higher R^2 values of 0.87 and 0.92 for forecast periods of 5 and 10 years respectively. (Fu & Aliferis, 2008) build similar prediction models for biomedical articles.

Yogatama *et al.* (2011), try to predict the response of a scientific community to an article. One of the problems they looked at was the problem of predicting whether a paper will receive any citations within the first three years given data at publication time. They use a model based on logistic regression, but introduce time regularized models for scientific prediction, where they learn separate features for each year, but add a regularization term to force the coefficients across different years to change smoothly. They show that this leads to better results. This also allows us to see the trends in a community by looking at how the coefficients for different terms change over time.

Chen (2012) describe a method for discerning the potential of a paper using information available and derivable upon the publication of a scientific paper. They present their structural variation model which predicts the potential of a scientific publication in terms of the degree to which it alters the intellectual structure of the state of the art. The structural variation approach focuses on the novel connections introduced by a paper between previously disparate patches of knowledge. They present three metrics of structural variation. The first metric called Modularity Change Rate is based on how the new paper and its references affect the modularity of the network. The second metric called Cluster Linkage measures the overall structural change introduced by an article in terms of new connections added between clusters. Finally, the third metric called Centrality Divergence measures the structural variation caused by an article a in terms of the divergence of the distribution of betweenness centrality of nodes in the baseline network. They evaluate these metrics as predictive measures for global citations for papers in four research areas (terrorism, mass extinction, complex network analysis, and knowledge domain visualization) and show that an article that introduces novel connections between clusters of co-cited references is likely to subsequently become highly cited.

Sarigöl *et al.* (2014) analyse the extent to which the success of a scientific article can be attributed to social influence. They extract time-evolving coauthorship networks from a dataset of 100,000 publications and study to what extent the centrality measures of an author in these networks can be used to predict the success of research articles in terms of number of citations. The centrality measures used are degree centrality, eigenvector centrality, betweenness centrality and k-core centrality. They report a precision of 60% on the task of predicting whether an article will belong to the top 10% most cited articles after five years. The authors argue that "this result quantifies the existence of a social bias, manifesting itself in terms of visibility and attention [. . .]". Castillo *et al.* (2007) describe a method to estimate the number of

citations for a paper based on statistics of past papers written by the same author(s).

Dong *et al.* (2014) formalize a new impact prediction problem for papers: whether a paper will increase the author's h-index. They experiment on the ArnetMiner dataset consisting of 1,712,433 authors and 2,092,356 papers in computer science with several predictive features that can be divided into six categories: author, content, venue, social, reference and temporal. They found that the two most important factors were the author's authority on a topic and the level of the venue in which the paper is published. They also found that publishing on an academically "hot" but unfamiliar topic is unlikely to further one's scientific impact.

Wang *et al.* (2014) present a HITS like framework for simultaneous ranking of future impact of papers and authors called MRFRank. They extract two types of text features based on words and words co-occurrence in the same sentence titles and abstracts of papers. They then measure the innovativeness of text features by using a burst detection based method: a text feature is labelled innovative if its frequency increases remarkably in a specified time window. Papers and authors are then characterized as a set of text features. They then form five types of time-aware graphs (coauthor graph, paper citation graph, author-paper graph, author-text feature graph, and paper-text feature graph) using three types of nodes (authors, papers, and text features). The graphs are made time aware by using exponentially decaying weights. They then use their HITS style MRFRank algorithm to predict future authority scores for papers, authors and text features. Their experiments over the ArnetMiner dataset shows that their joint prediction method performs better than other baseline methods.

Acuna *et al.* (2012) present a formula for predicting the future h-index of scientists based on a small set of parameters: current h-index, number of articles written, years since publishing first article, number of distinct journals published in, and number of articles in top venues (e.g. Nature, Science). Penner *et al.* (2013) however, show

that the predictive power of this model decreases significantly when applied to early career years and advice caution in using this model to make early career decisions.

Klosik & Bornholdt (2013) propose a measure based on the shape and size of the wake of a paper within the citation network and find that it is able to detect a large fraction of seminal articles co-authored by Nobel prize laureates.

Hirsch (2007) conduct an empirical study to evaluate the predictive power of the h-index and found that it is better than the other indicators they considered, namely: total citation count, citations per paper, and total paper count.

Indirect means of predicting future impact of scientific publications and technologies has been studied as well. Brody *et al.* (2006) discuss how early web-usage statistics derived from download counts can be used to predict future citation impact. Using a dataset of papers extracted from ArXiv, they show a high correlation between downloads and citations (0.4). They also report that "if the baseline correlation for a field is significant and sufficiently large, the download data could be used after 6 months as a good predictor of citation impact after 2 years." Eysenbach (2011) explore the feasibility of measuring social impact of scholarly articles by analyzing buzz in social media and whether these metrics are sensitive and specific enough to predict highly cited articles. They report results that indicate tweets can predict highly cited articles within the first three days of article publication and propose the use of social impact measures based on tweets to complement traditional citation metrics. Erdi *et al.* (2012) use patent citation networks to detect new emerging recombinations of technologies and predict emerging new technology clusters.

2.3 Modeling Scientific Text

2.3.1 Citation Text Analysis

Studying citation patterns and referencing practices has interested researchers for many years (Garfield, 1972; Garfield *et al.* , 1984). White (2004) provides a good survey of the different research directions that study or use citations by reviewing contributions from the 1970s to the present in three major lines of research: citation classification, content analysis of citation contexts, and studies of citer motivations.

Early on, Leydesdorff & Wouters (1999) investigate the history of evolution of referencing and its implications. Several research efforts have focused on studying the different purposes for citing a paper (Garfield, 1964; Weinstock, 1971; Moravcsik & Murugesan, 1975; Chubin & Moitra, 1975; Bonzi, 1982). Chubin & Moitra (1975); Bonzi (1982) analyze citer motivations in 43 high energy physics papers by classifying them into one of six categories: basic, subsidiary, providing additional information, perfunctory, partial negation, and total negation. They found that negational references are rare (5%) and short lived. Full length papers published in highly visible journals are most recognized as useful contributions. Essential articles continue to be cited at a substantial level.

Bonzi (1982) studied the characteristics of citing and cited works that may aid in determining the relatedness between them. Garfield (1964) enumerated several reasons why authors cite other publications, including "alerting researchers to forthcoming work", paying homage to the leading scholars in the area, and citations which provide pointers to background readings. Weinstock (1971) adopted the same scheme that Garfield proposed in her study of citations.

Spiegel-Rösing (1977) proposed 13 categories for citation purpose based on her analysis of the first four volumes of Science Studies. Some of them are: cited source is the specific point of departure for the research question investigated, cited source

contains the concepts, definitions, interpretations used, cited source contains the data used by the citing paper. Nanba & Okumura (1999) came up with a simple schema composed of only three categories: *Basis*, *Comparison*, and other *Other*. They proposed a rule-based method that uses a set of statistically selected cue words to determine the category of a citation. They used this classification as a first step for scientific paper summarization. Teufel *et al.* (2006a), in their work on citation function classification, adopted 12 categories from Spiegel-Rosing's taxonomy. They trained an SVM classifier and used it to label each citing sentence with exactly one category. Further, they mapped the twelve categories to four top level categories namely: weakness, contrast (4 categories), positive (6 categories) and neutral.

The polarity (or sentiment) of a citation has also been studied previously. Previous work showed that positive and negative citations are common, although negative citations might be expressed indirectly or in an implicit way (Ziman, 1968; MacRoberts & MacRoberts, 1984; Thompson & Yiyun, 1991). Athar (2011) addressed the problem of identifying sentiment in citing sentences. They used a set of structure-based features to train a machine learning classifier using annotated data.

Citation purpose and relevance has been used for doing scientometric analysis in a number of different fields. Li *et al.* (2014) use citation motivation to study science linkage: a widely used patent bibliometric indicator to measure patent linkage to scientific research based on the frequency of citations to scientific papers within the patent. Liu *et al.* (2014a) use citation relevance based main path analysis for tracing main paths of legal opinions and show that relevancy information helps main path analysis uncover legal cases of higher importance. Cheang *et al.* (2014) use citation classification to do an evaluation of 39 selected management journals. Doboli *et al.* (2014) analyzed 30 publications in the area of high-frequency analog circuit design and defined two new measures to characterize the creativity (novelty and usefulness) of a publication based on its pattern of citations clustered by reason, place and citing

scientific group.

Bonzi & Snyder (1991) did a study to understand citation purpose in the context of self-citation in natural sciences. They investigated the citation motivation among 51 self-citing authors in several natural science disciplines by creating a survey that asked authors to mark the reasons for citing a set of 4 papers that they had cited in their own works from one of 14 reasons. The references were chosen in order to pair each self-citation with a citation to another work, selecting from the same paragraph if possible, or if not, from the same section of the article. Only three reasons for self-citations were found to be significantly different than as opposed to citation of others: 1) that of establishing the writer's authority in the field by citing their own work, 2) demonstrating knowledge of important work by citing work by other authors, and 3). identification of earlier research on which the work builds by self-citation. They found minimal difference in the words devoted to self-citation compared to other citations. This lack of substantive difference in citation behaviour was consistent across fields.

Wan & Liu (2014a) present a regression method for automatically estimating the strength value of each citation and show the estimated values can achieve good correlation with human-labeled values.

We now look at some work on linguistic analysis of citation text. Nakov *et al.* (2004) proposed the use of citation text as a tool for semantic interpretation of bioscience text and propose several applications: a source for unannotated comparable corpora, summarization of target papers, synonym identification and disambiguation, entity recognition and relation extraction, targets for curation, and improved citation indexes for information retrieval. They identify several issues for effective use of citing sentences as well: detecting the span of text for a citation, identifying the different reasons a given paper is cited for, and normalizing or paraphrasing citing sentences. They also propose a method for paraphrase extraction from citing sentences that works by using dependency paths between named entities in sentences.

Ding *et al.* (2014) introduced the notion of Citation Content Analysis (CCA) and discussion the nature and purposes of CCA along with potential procedures to conduct CCA. Halevi & Moed (2013) present a citation context analysis for the journal of infometrics. Zhao & Strotmann (2014) also analyze the feasibility, benefits, and limitations of in-text author citation analysis and test how well it works compared with traditional author citation analysis using citation databases. Angrosh *et al.* (2013) present a dataset for citation context sentences and present a model for citation context identification based on Conditional Random Fields (CRFs).

One of the important uses of citation context is for scientific summarization. Nanba & Okumura (1999) use the term *citing area* to refer to the same concept as citation context. In Nanba *et al.* (2004a), they use their algorithm to improve citation type classification and automatic survey generation. Kaplan *et al.* (2009) present a method for identifying citation contexts based on coreference analysis and their use for research paper summarization. Qazvinian & Radev (2010a) propose a method for finding implicit citing sentences using a method based on Markov Random Fields and show how adding implicit citing sentences improves the quality of a survey generation system.

Citation context has also been used for literature retrieval models for scientific domains. Liu *et al.* (2014b) designed a retrieval system for the PubMed Central database using citation contexts. Yin *et al.* (2011) similarly used citation context for the task literature retrieval in the biomedical domain.

Athar & Teufel (2012a) observed that taking the context into consideration when judging sentiment in citations increases the number of negative citations by a factor of 3. They also proposed two methods for utilizing the context. Their experiments surprisingly gave negative results and showed that classifying sentiment *without* considering the context achieves better results. They attributed this to the small size of their training data and to the noise that including the context text introduces to

the data. In Athar & Teufel (2012b), the authors present a method for automatically identifying all the mentions of the cited paper in the citing paper. They show that considering all the mentions improves the performance of detecting sentiment in citations.

Finally Abu-Jbara & Radev (2012) looked at the complementary task to citation context detection called reference scope extraction: given a citing sentence that cites multiple papers, the goal in this task is to extract the segments from the sentence that are relevant to a specific target paper. They experimented with several supervised methods for this task and found that a CRF based classifier performed best.

2.3.2 Discourse Structure of Scientific Text

Early work in the study of discourse in scientific articles seems to have been guided by the need to develop pedagogical material for non-native speakers of english who might need to read or write academic papers. Swales (1981) analyzed introductions of sixteen articles from physics, biology, and social sciences and identified four moves, each of which can be further subcategorized: 1) Establishing the field, 2) Summarize previous research, 3) Preparing for present research, and 4) Introducing present research. Crookes (1986) annotated a larger corpus of 96 articles with Swales's categories and found that the most common structures were 2-4 and 1-2-3-4.

Thompson & Yiyun (1991) describe the results of a project to identify the kinds of verbs used in citations in academic papers. They analyze reporting verbs under two main headings: denotation and evaluative potential. In their terminology, writer is the researcher who wrote the paper being analyzed and author refers to the person who wrote a paper being cited by the paper being analyzed. Under denotation, they found that most verbs belong to three groups of processes relating to author acts: textual (e.g. state, write, point out), mental (e.g. believe, think, consider), and research (e.g. measure, calculate, obtain) and two groups of processes relating to

writer acts: comparing (e.g. correspond to, accord with, anticipate), and theorizing (e.g. account for, explain, support). With respect to the evaluative potential, the reporting verbs are again broken down into three factors: 1) author's stance which can be positive (e.g. emphasize, hypothesize, invoke), negative (e.g. attack, challenge, dismiss), or neutral (e.g. assess, examine, evaluate), 2) writer's stance which can be factive (e.g. acknowledge, bring out, demonstrate), counter-factive (e.g. betray, confuse, disregard), or non-factive (e.g. advance, believe, claim), and 3) writer's interpretation which can be author's discourse interpretation (e.g. add, comment, continue), author's behaviour interpretation (e.g. admit, advocate, assert), status interpretation (e.g. account for, bringout, confirm), or non-interpretation (e.g. adopt, apply, calculate).

Liddy (1991) undertook an exploratory three phase study to determine whether information abstracts reporting on empirical work possess a predictable discourse-level structure and whether there are lexical clues that reveal this structure. The linguistic model was developed by asking the expert abstractors to list the components of information they believed constituted an abstract reporting on empericial work. The components were then analyzed manually and coded leading to a structured representation of each abstract showing how the components of the abstracters' model actually exhibited themselves in text and lists of the lexical ciues for each component. Finally a linguistic model was developed based on this structure and validated against the human written abstracts. Their results indicated that expert abstractors do possess an internalized structure of empirical abstracts. In later work (Liddy *et al.* , 1987), they also investigated the use of anaphoric references in scientific abstracts.

Paice & Jones (1993) argue that the main concepts discussed in technical papers fit into a predictable range of semantic roles conveyed by a variety of characteristic stylistic constructs and expressions and that these constructs provide evidence for the semantic roles of the concepts bound to them. They use a set of manually extracted

context patterns to identify important concepts in technical papers.

RAFI was a system (Lehmam, 1999) intended for generating indicative summaries of scientific texts in French. Their preliminary study indicated that the structure of scientific texts contain the following important parts: previous knowledge, content, method, and results. Their summarization system is based around assigning an importance score to each sentence through comparison with a base of pre-constituted knowledge (thesauras), eliminating sentences which do not obtain a threshold score and making sure that the summary contains at least one of the parts of the structure.

Teufel (1999) propose a new framework for rhetorical analysis of scientific text called Argumentative Zoning. In this framework, the rhetorical status of a sentence with respect to the communicative function of the paper is classified into one of seven categories: 1) BKG or general scientific background, 2) OTH, neutral descriptions of other people's work, 3) OWN, neutral descriptions of the own, new work 4) AIM, statements of the particular aim of the current paper 5) TXT, statements of textual organization of the current paper 6) CTR, contrastive or comparative statements about other work; explicit mention of weaknesses of other work 7) BAS, statements that own work is based on other work.

Grover *et al.* (2003) present a system for summarizing legal documents by first classifying sentences according to their argumentative role similar to the work of Teufel & Moens (2002). They define three high level rhetorical labels for legal judgements: Background sentences conveying generally acceptable background knowledge, Case sentences containing description of the case including the events leading up to legal proceedings, and Own sentences that can be attributed to the Lord speaking about the case (they use Judgments of the House of Lords for their study). Each of these is further divided into more detailed sub-categories. Using this annotation scheme, they annotated five randomly selected appeals cases. They then perform chunking on the sentences to find the main verb-group for each sentence and its tense. This is

motivated by their hypothesis that tense may be a useful feature in identifying the rhetorical structure and their initial experimentation confirms this hypothesis.

Cormode *et al.* (2012a) introduces the notion of scienceography: the study of how science is written. They analyze Latex source code of papers in computer science and mathematics to find broad patterns and trends regarding the use of comments, paper length, use of figures, distribution of theorems etc. Tan & Lee (2014) built a corpus of sentence level revisions in academic writing by comparing different versions of the same papers uploaded on ArXiv. They use Mechanical Turk to label sentence revisions as one of stronger, weaker, no strength change, or can't tell.

2.4 Repurposing Scientific Text

2.4.1 Single Paper Summarization

Most of the research in summarizing scientific literature has focused on producing extractive summaries of scientific documents. The oldest work in this direction seems to be the work of Kupiec *et al.* (1995). The goal of their work was to produce an indicative summary of a scientific article that can be used in the absence of a manually generated abstract. They used a supervised approach to the problem, given a manually generated extractive summary, they trained a Naive Bayes classifier that predicted the probability of a sentence being included in a summary using features such as sentence length, cue phrases, word frequency and sentence position in a paragraph. They reported an accuracy of 84%.

Elhadad & McKeown (2001) presented a system for summarizing medical articles that given a query finds and extracts results from multiple medical journals, filters results that match the patient and merges and orders the remaining facts for the summary.

Kan *et al.* (2002) use a corpus of 2000 annotated bibliographies for scientific papers as a first step towards a supervised summarization system. They found that summaries in their corpus were mostly single-document abstractive summaries that were both indicative and informative and were organized around a "theme," making them ideal for query-based summarization.

Teufel & Moens (2002) present a system for summarizing scientific articles that is based on the argumentative zoning roles discussed earlier. In addition to annotating sentences in 80 scientific papers with rhetorical status, they also annotate each sentence as relevant or irrelevant for the summarization of the article. For automatic classification of relevance, sentences classified as AIM, CONTRAST, and BASIS sentences are marked relevant directly since these categories are overall rare, while sentences classified as BACKGROUND are further classified into relevant or non-relevant using a separate classifier trained for relevance.

Da Cunha & Wanner (2005) explore integrating textual, lexical, discursive, informative, and syntactic features for automatic summarization of medical articles in Spanish. Zhang *et al.* (2013) explore methods for biomedical summarization by identifying cliques in a network of semantic predications extracted from citations. These cliques are then clustered and labeled to identify different points of view represented in the summary.

Most current publications however, contain manually written abstracts and therefore, obviate the need of generating such summaries. However, Elkiss *et al.* (2008) provided a new perspective on this problem. They suggest that using the citing sentences as defined sentences in later papers that talk about a particular paper might be useful for summarizing the actual contributions of the papers. Two later works in subsequent years used this hypothesis to improve scientific article summarization. Mei & Zhai (2008) used citing sentences to create *impact based summaries* of scientific articles, where sentences are still chosen from the text of the article to be

summarized, but the sentence selection is driven by the similarity of the sentences to citing sentences (where the similarity measure is negative KL-divergence between language models estimated from citing sentences and those of the candidate sentences). Qazvinian & Radev (2008a) abandoned the source sentences completely and used the citing sentences directly to generate extractive summaries of research papers. The intuition behind their method, C-Lexrank, is that the citing sentences can be clustered into groups that focus on specific aspects of the target paper. Once these groups of sentences are found, lexically central sentences in each of these groups can be used to summarize each aspect of the paper resulting in a diverse and informative summary of the paper. The sentences are selected based on the size of clusters and their lexical centrality in each cluster as determined by their Lexrank scores.

These works focused on the informativeness of the resulting summaries. In practice, however, the summaries were incoherent and difficult to read. Abu-Jbara & Radev (2011) introduced a method for coherent citation based summarization. The crux of their method is to first use a trained SVM to classify each input sentence into one of five functional categories: Background, Problem Statement, Method, Results, and Limitations. Once this is done, C-Lexrank is run over the sentences in each of the functional categories. The sentence order is chosen by the categories as listed above (for example, a Background sentence would proceed a Method sentence), as well as the sizes of the clusters and the lexical centrality of the sentences. In addition to these, they have a preprocessing step for sentences in which they remove parts of a sentence that are not relevant for the summarization of the target paper. This is important because citing sentences often talk about multiple papers. A post-processing step removes non-syntactic reference markers and adds pronouns for repeated entities. They show that their methods improve the coherence of the resulting summaries are significantly improved.

2.4.2 Scientific Topic Summarization

A summary of the main contributions of a individual paper might be useful, but more often, researchers are looking to understand all the related work in a given field of interest. For example, an example might be "dependency parsing". This is a multi-document summarization task where the summary should focus on the general trends in an area instead of focusing on the contributions of one or two papers. In our knowledge, the first work to investigate this problem was Nanba *et al.* (2004b). In this paper, they present a prototype system called PRESRI intended to assist users in generating review articles. They define citing areas in papers as the citing sentences as well as relevant context sentences around the citing sentences that are identified using 86 manually generated cue phrases. The main hypothesis of their work is that the citing areas from papers in a given subject domain can act as a review article if properly classified and organized. The workflow consists of two parts: classifying papers and summarizing them. For classifying papers, they use the citation types of citing areas, where citation types are based on reasons for citing a particular paper. Three citation types are defined in their work: type B corresponding to citations that show other work for theoretical basis, type C corresponds to citations that point problems or gaps in related work and type O corresponding to all the other citations (Teufel *et al.* (2006b) presented a more finer grained classification for citing sentences, but its usefulness for summarization has not been shown). An automatic classifier is used to classify each citing area into a citing type. This is then used to cluster papers topically by using a form of bibliographic coupling based on the citing type as a similarity measure. Their system then presents citing areas for the papers to the user, who can then use these to generate a review article.

Mohammad *et al.* (2009) attempted to completely automate the review article generation process. They focused on two topics in NLP, Dependency Parsing and Question Answering with 16 and 10 papers in the input document set respectively.

For each of these topics, they experimented with using abstracts, full text or citing sentences as input for a survey generation system. The algorithms used were C-Lexrank that was previously shown to have good performance in summarizing single scientific articles and Trimmer, an extractive summarizer based on sentence compression. 250 word surveys were generated from all these sources. The main conclusions from the paper were that citations contained unique survey worth information and useful surveys can be generated using these sources. In a follow up work, Qazvinian & Radev (2010b) presented a method based on Markov Random Fields (MRF) for finding relevant context sentences around the explicit citation sentences and showed that including these sentences in the input to the summarizer improved the pyramid scores of the resulting surveys.

Dunne *et al.* (2012) present a system called Action Science Explorer (ASE) for helping researchers in exploring a research field integrating "statistics, text analytics, and visualization in a multiple coordinated window environment that supports exploration".

More recently, Hoang & Kan (2010) present a method for automated related work generation. Their system takes as input a set of keywords arranged in a hierarchical fashion that describes a target paper's topic. They hypothesize that sentences in a related work provide either background information or specific contributions. They use two different models to extract these two kinds of sentences using the input tree and combines them to create the final output summary.

2.4.3 Other Applications

2.4.3.1 Indexing and Retrieval

Early on, Kostoff *et al.* (2001) introduce Citation Mining: a literature based approach that integrates text mining and bibliometrics for identifying ways in which research can impact other research, technology development, and applications. The

idea of citation indexes for scientific literature can be attributed to Garfield (Garfield, 1964, 2006a).

Berendt *et al.* (2010) presented an interactive retrieval system that supports active and constructive exploration of a domain. Their method uses data mining and interactivity to transform a typical search into a dialogue that involves *sense-making*, and the user constructs a bibliography and a domain model of the search terms. The use of citation context in literature retrieval has been covered in the Section 2.3.1.

Bradshaw (2003) introduced Reference Directed Indexing (RDI): an approach for scientific document retrieval that is based on building an index for a document using the terms used by authors in describing that document.

2.4.3.2 Finding New Science

Swanson proposed the idea of undiscovered public knowledge (Swanson, 1986): knowledge that consists of independently created fragments that are logically related but never retrieved, brought together, and interpreted. This is an even bigger problem today with huge number of publications that scholars need to keep pace with. In this paper, three examples are constructed and analyzed to illustrate the idea of undiscovered public knowledge. He developed these ideas further in a follow up paper (Swanson, 1990) where three examples are presented: 1) a link between two independent findings that reported that dietary fish oils lead to certain blood and vascular changes and that similar changes might benefit patients with Raynaud's syndrome that was later confirmed using clinical trials 2) an inference that magnesium deficiency might be a causal factor in migraine headache based on eleven indirect connections, and 3) implicit connections between arginine intake and blood levels of somatomedins. The paper also describes a method for aid in discovering such logically related findings. Ten years after the first paper on undiscovered public knowledge, a ten year update Swanson & Smalheiserf (1996) reported progress in

creating interactive software and database strategies for finding hidden connections in scientific literature as well seven examples of literature based knowledge synthesis.

Some other researchers have pursued this direction as well. Kuhn *et al.* (2008) review attempts to apply large-scale computational analyses to predict novel interactions of drugs and targets from molecular and cellular features. Frijters *et al.* (2010) describe a tool called CoPub Discovery that mines scientific literature for new relationships between biomedical concepts based on co-occurence. They used CoPub Discovery to find new relationships between genes, drugs, pathways and diseases that were validated using independent literature sources.

CHAPTER III

Building Document Collections in Response to Scientific Queries

In this chapter, we look at the task of building document collections in response to queries. There are two main subproblems here: first, we need to find the seminal historical papers that should be described in a summary of the query topic. We formulate this as an information retrieval problem and experiment with several algorithms. Secondly, we need to find recent important papers relevant to the topic that the user should know about. We formulate this second problem as a prediction problem, and experiment with features that allow us to predict the future prominence of papers with high confidence.

3.1 Finding Seminal Papers Relevant to Query

Given a query representing the topic to be summarized, the first task is to find the set of relevant documents from the corpus. The simplest way to do this for a corpus of scientific publications is to do a query search using exact match or a standard TF*IDF system such Lucene, rank the documents using either citation counts or pagerank in the bibliometric citation network, and select the top n documents. However, comparing the results of these techniques with the papers covered by gold

Document selection algorithm	CG_5	CG_{10}	CG_{20}
Title match sorted with citation count	1.82	2.75	3.29
Title match sorted with pagerank	1.77	2.55	3.34
Citation expansion sorted with citation count	0.53	1.20	2.29
Citation expansion sorted with pagerank	0.20	0.78	1.99
TF*IDF ranked	0.14	0.14	0.56
TF*IDF sorted with citation count	0.44	2.25	3.18
TF*IDF sorted with pagerank	1.54	2.22	2.85
Restricted Expansion	**2.52**	**3.91**	**6.01**

Table 3.1: Comparison of different methods for document selection by measuring the Cumulative Gain (CG) of top 5, 10 and 20 results.

standard surveys on a few topics, we found that some important papers are missed by these simple approaches. One reason for this is that early papers in a field might use non-standard terms in the absence of a stable, accepted terminology. Some early Word Sense Disambiguation papers, for example, refer to the problem as Lexical Ambiguity Resolution. Additionally, papers might use alternative forms or abbreviations of topics in their titles and abstracts, e.g. for input query "Semantic Role Labelling", papers such as (Dahlmeier et al., 2009) titled "Joint Learning of Preposition Senses and Semantic Roles of Prepositional Phrases" and (Che and Liu, 2010) titled "Jointly Modeling WSD and SRL with Markov Logic" might be missed.

To find these papers, we add a simple heuristic called *Restricted Expansion*. In this method, we first create a base set B, by finding papers with an exact match to the query. This is a high precision set since a paper with a title that contains the exact query phrase is very likely to be relevant to the topic. We then find additional papers by expanding in the citation network around B, that is, by finding all the papers that are cited by or cite the papers in B, to create an extended set E. From this combined set ($B \cup E$), we create a new set F by filtering out the set of papers that are not cited by or cite a minimum threshold t_{init} of papers in B. If the total number of papers is lower than f_{min} or higher than f_{max}, we iteratively increase or decrease t till $f_{min} \leq |F| \leq f_{max}$. This method allows us to increase our recall without losing

Document selection algorithm	Precision	Recall
Title match sorted with citation count	0.36	0.18
Title match sorted with pagerank	0.34	0.16
Citation expansion sorted with citation count	0.21	0.10
Citation expansion sorted with pagerank	0.21	0.10
TF*IDF ranked	0.12	0.06
TF*IDF sorted with citation count	0.30	0.15
TF*IDF sorted with pagerank	0.26	0.13
Restricted Expansion	**0.53**	**0.27**

Table 3.2: Comparison of different methods for document selection by measuring precision and recall for the top 50 documents. The improvement of restricted expansion over each of the other methods for both precision and recall is statistically significant with $p < 0.05$

precision. The values for our current experiments are: $t_{init} = 5, f_{min} = 150, f_{max} = 250$.

To evaluate different methods of candidate document selection, we use Cumulative Gain (CG), where the weight for each paper is estimated by the fraction of surveys it appears in. Table 3.1 shows the average Cumulative Gain of top 5, 10 and 20 documents for each of eight methods we tried. Restricted Expansion outperformed every other method. Once we obtain a set of papers to be summarized, we select the top n most cited papers in the document set as the papers to be summarized, and extract the set of citing sentences S from all the papers in the document set to these n papers. S is the input for our sentence selection algorithms. We also compare the precision and recall of each of the methods for the top 50 documents. These results are summarized in Table 3.2. Restricted Expansion outperforms every other method for this evaluation as well.

3.2 Forecasting Future Impact of Papers

We now focus on methods for forecasting the impact of papers that have been recently published and have not had time to accumulate citations. What makes a

new paper an important piece of work? How much is the reception of a paper in the scientific community affected by the prestige of the authors and the publication venue? What role is played by the novelty of the work and how it is positioned with respect to prior work? These are some of the research questions that we seek to answer in the work presented here. We seek to experimentally measure the role played by four sets of features in forecasting both the short term and long term impact of a paper: prestige, positioning, content and style. Prestige relates to the attributes of a paper that might play a role in its initial popularity such as the authors, their affiliations and the venue of publication. Positioning relates to how the contributions of the paper are linked to previous work by the authors. Content and style relate to what information is presented in the paper and how it is presented.

Our short term prediction task is similar to (Yogatama *et al.* , 2011), where the goal is to predict whether a paper will receive a citation within the first 3 years of its publication given information present at publication time. We extend their work by adding a richer set of features that capture several important aspects of a paper that might be useful in the prediction task.

However, the short term success of a paper might not be indicative of its long term impact. Figure 3.1 shows the cumulative citations for four of the papers published in 2002 over the next 10 years. Notice that until 2004, the citations counts for the different papers are similar. However, by 2011, the numbers have diverged greatly. Particularly, citations to (Papineni *et al.* , 2002) are substantially higher than the other papers [1]. Therefore, we also formulate a long term prediction task where we want to predict which papers would have a high impact in a longer horizon of 10 years, given data at publication time. We present experimental results on this prediction task using the above set of features.

[1]This is the paper that introduced BLEU, a popular method for evaluating machine translation systems.

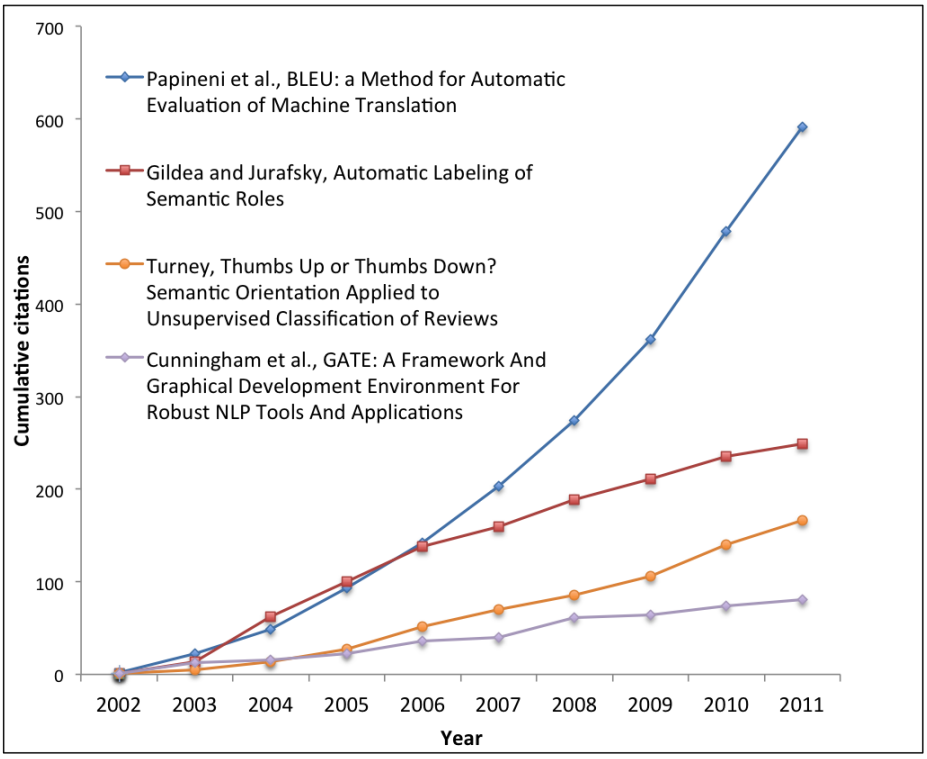

Figure 3.1: Cumulative citation counts for four papers published in 2002 with different long term citation patterns.

3.3 Features for Forecasting Impact

We now describe the main feature classes that we use for the task of predicting the future impact of a paper given data at publication time.

Prestige Intuitively, papers published by well known authors and appearing at more prestigious venues might have a higher chance of being cited. The prestige features are derived from these immediately visible aspects of a paper. (Yogatama *et al.* , 2011) used the author last names and venues as metadata features. We similarly model the prestige of a paper by using Boolean features for the venue of publication, each author and his/her affiliation. However, since author last names can be highly ambiguous (e.g. there are 13 authors in the ACL Anthology with the last name "Ng"), we use a manually disambiguated list of author names in our data set as the author features.

Word Class	Sample Words
Tentativeness	almost; apparently; appears; approximat*
Certainty	accura*; always; clear; confident
Discrepancy	could; hope; expect*; lack*
Achievement	achiev*; acquir*; better; improve*
Positive emotion	advantage*; benefits; efficien*; interest*

Table 3.3: Sample LIWC categories and a few example words for each of them.

Content This includes unigrams, bigrams and trigrams from the title and full text of the paper. We did feature pruning by removing any features that appear in less than 2% and more than 98% of the papers. In addition to these, we use the terms present in both the title and abstract of the paper as features. The terms were extracted from the titles and abstracts of the entire corpus (we describe our corpus later in Section 4.3) using the C-Value/NC-Value method (Frantzi *et al.* , 1998).

Positioning For a new piece of research to be properly interpreted, it's important to situate it with respect to the previous literature in the area. For this, we include features computed on citation network as well as different lexical networks derived from the corpus.

(Shi *et al.* , 2010) describe a set of network based metrics that measure how a new paper is related to the papers it cites. They build two networks: one is the citation projection of the papers that are cited by the current paper onto the global citation network, G_p. That is, G_p is a network built by including only edges from the global citation network for which both the nodes are in the set of papers cited by the current paper, but excluding the current paper. G_{p0} is a network built by including the citing paper in this network as well. Following (Shi *et al.* , 2010), we include the following features from these two networks: graph density, clustering coefficient, connectivity and maximum betweenness of G_p, along with the betwenness centrality of the current paper in G_{p0}.

We derive a second set of features from a projection of the corpus term co-

occurrence network. We start by building a global co-occurence network for terms (extracted as described above) from the titles and abstracts. The nodes in this network represent the terms and the edge weights represent the number of times these terms have co-occurred in either a title or an abstract. Given a new paper, we project the terms in its title and abstract onto this co-occurence network similarly to the citation network above. From this projected network, we extract as features the maximum, minimum and average edge weight along with its density.

In addition, we also compute the average, minimum and maximum cosine similarity of the title of this paper to the titles of its cited papers. We also create features based on the cosine similarity of the title of this paper with all the published papers in the same year, and also the papers published in the previous year. These values model how this paper is positioned with respect to the cited literature and its contemporary papers.

Style (Guerini *et al.* , 2012) did a preliminary examination of the correlation of stylistic aspects of a paper with its popularity by analyzing the words in the abstracts of the papers. They used the Linguistic Inquiry and Word Count (LIWC) corpus (Pennebaker *et al.* , 2001) to tag the words in the abstracts based on the psycholinguistic class it belongs to. LIWC includes about 4,500 words and stems that are grouped into 65 categories based on their linguistic category (e.g. 1st person singular, Future tense) and psychological processes (e.g. Certainty, Negation). Table 3.3 shows some sample categories and example words from each of the categories. The LIWC 2007 corpus contains certain classes that are irrelevant for our purposes, such as "Non-fluencies", "Money", and "Religion". We filtered a set of 36 categories for our experiments. For each paper, we use the counts of words found from each LIWC category as a feature.

Feature set	Accuracy			
	2004	**2005**	**2006**	**2001-2006**
Majority (baseline)	0.677	0.744	0.624	0.699
Pr	0.585	0.653	0.616	0.606
Pr + Po	0.663	0.688	0.685	0.687
Pr + Co	0.705	0.720	**0.700**	0.707
Pr + St	0.594	0.653	0.638	0.634
Pr + Po + Co	0.705	0.720	0.696	0.711
Pr + Co + St	0.705	0.712	0.698	0.702
Pr + Po + Co + St	**0.746**	**0.751**	**0.700**	**0.723**

Table 3.4: Results on the task of predicting whether a paper is cited within the first 3 years for three years, 2004, 2005 and 2006 and also averaged over 2001-2006. The abbreviations stand for Pr = Prestige, Po = Positioning, Co = Content, St = Style. The feature group Pr+Co corresponds to the feature set presented in (Yogatama *et al.* , 2011). The highest values in each column are highlighted. The improvement of all features over purely prestige features is statistically significant with $p < 0.01$ using a two-tailed t-test.

3.4 Experiments

We use the ACL Anthology Network (AAN) (Radev *et al.* , 2013) for our experiments. AAN is a corpus of publications that contains the metadata and full-text of papers from various conferences in the field of Computational Linguistics. We use the papers from 1980 till 2012 for our experiments, and only include the papers from ACL, EACL, HLT and NAACL following (Yogatama *et al.* , 2011). This restricted set contains 5,727 papers and is used as the corpus in our experiments.

3.4.1 Predicting Short Term Impact

Our first task is to predict whether a paper will be cited within the first 3 years of its publication. For testing on the papers in any given year y, we use the papers till $y-1$ as training data. The data till year $y-3$ is fully observable. However, for the papers in the years $y-2$ and $y-1$, we don't have enough data to assign the labels, since we are not allowed to look past year y; these papers fall in the forecast gap. We extrapolate the number of citations for such papers using the fully observable

training data. If the prediction is to be made for t years in future (t = 3 in our case), for any year t' in the range $t' \in [y-t, y]$, we first estimate the ratio r of the number citations in t' years to t years using the training data from year $[b, y-t]$ (where b is the epoch, 1980 in our case). The counts of citations in t' are then extrapolated by scaling the observed citations by r^{-1}.

Table 3.4 shows the results of our experiments for 2004, 2005 and 2006 and also the average over 2001-2006. The feature group Pr+Co corresponds to the feature set presented in (Yogatama *et al.* , 2011). The results show that our additional features lead to an improvement in accuracy. Specifically, the features in the positioning and content classes lead to the largest improvement. The style based features do not lead to major improvements by themselves, but improve results when used in conjunction with other features.

Feature set	**Precision**	**Recall**	**F-score**
Pr	0.283	0.215	0.226
Pr + Po	0.240	0.239	0.222
Pr + St	0.226	**0.271**	0.236
Pr + Co	**0.603**	0.206	**0.272**
Pr + Po + Co	0.389	0.211	0.248
Pr + Co + St	0.300	0.149	0.196
Pr + Po + Co + St	0.327	0.163	0.203

Table 3.5: Results on predicting whether a paper appears in the top 90 percentile at the end of 10 years averaged over results from 1995-1999. The baseline of assigning all to True has Precision = 0.1, Recall = 1 and F-score = 0.18. The abbreviations stand for Pr = Prestige, Po = Positioning, Co = Content, St = Style. For the improvement in precision using pr+co over pr, p is estimated at 0.06 using a two-tailed t-test.

3.4.2 Predicting Long Term Impact

In this section, we turn to the problem of predicting the prominence of papers on a large horizon of 10 years after publication. The exact counts of citations to papers are dependent on the rate of new publications, which increases over years and thus, the counts are not comparable for papers published in different years. To get

around this, we estimate the prominence of a paper by comparing it to the papers that are published in the same year. The intuition is that if some of these papers get many more citations than other papers published in the same year, then they are prominent. Let p_i refer to the publication year of a paper i. The forecast year $f > p_i$ is defined as the year for which we have to make a prediction for the paper ($f = 10$ for the current experiments). $PR_i(y)$ is defined as a function that takes as input a year y and outputs the percentile rank of the paper i in year y with respect to all the papers published in the same year as i. The percentile rank is computed with respect to the cumulative citations accumulated by i from p_i till y. Given this formulation, we define the class of papers with long term prominence, P_d, as the set of papers that are in the top 10 percentile at the end of the forecast year, $P_d := \{i : PR_i(f) > 0.9\}$. Given a test year y, the forecast gap is $[y - 10, y]$ for this set of experiments. We extrapolate the citation counts for papers in the forecast gap the same way as for the short term prediction experiments.

Table 3.5 shows the results of our experiments on test data spanning 1995-1999. The content features lead to the highest improvement in precision over the baseline and purely prestige based features. However, we see low recall values for all combinations of features. This points towards the fact that these sets of features might not be enough to find all the papers that will have a high impact in the long term and additional features might be needed to detect all such papers. Finding such papers might need also investigation into the phenomenon of "sleeping beauties" discussed in the scientometrics literature (Van Raan, 2004; Costas *et al.* , 2009). These are papers that do not show high impact early on in their life cycle but tend to be highly cited in later years.

3.4.3 Flash In the Pans and Sleeping Beauties

We want to investigate models for detecting papers known as "sleeping beauties" in the scientometrics literature Van Raan (2004); Glänzel *et al.* (2003). These are the papers that do not get a lot of recognition in early years, but become important papers in later years. An extreme example of this is Gregor Mendel's work, which was so ahead of it's time that it took the scientific community thirty-four years to catch up to it. However, as described later, we find less extreme examples of this in our dataset. A related class of papers are the "flash in the pans" Costas *et al.* (2009): these papers receive a lot of attention immediately after publication, but fail to create a long term impact. We'd like to see if our models can detect such papers, and what attributes of such papers might be helpful.

Let p_i refer to the publication year of a paper i. We define reference year $r_i > p_i$ as the time till which our system is allowed to look at the data. The forecast period $f_i > r_i > p_i$ is defined as the year for which we have to make a prediction for the paper. For example, for a paper published in 1990, $p_i = 1990$, two possible values of r_i and f_i are 1991 and 2000 respectively. $PR_i(y)$ is defined as a function that takes as input a year y and outputs the percentile rank of the paper i in year y with respect to all the papers published in the same year as i. $PR_i(y)$ is only defined for $y > p_i$ The percentile rank is computed with respect to the cumulative citations accumulated by i from p_i till y. It should be clear that $PR_i = PR_j$ if $p_i = p_j$.

Given this formulation, we define the class of papers with long term prominence, P_d, as the set of papers that are in the top 10 percentile at the end of the forecast year, $P_d := \{i : PR_i(f_i) > 0.9\}$. The class of papers with delayed recognition, or "sleeping beauties", P_s, are defined as the set of papers that are not in the top 10 percentile at reference year, but are present in the top 10 percentile in the forecast year, $P_s := \{i : PR_i(r_i) < 0.9, PR_i(f_i) > 0.9\}$. Similarly, papers are defined as being in class P_f, or as "flash in the pan", if they follow the opposite pattern, $P_s := \{i :$

$PR_i(r_i) > 0.9, PR_i(f_i) < 0.9\}$.

For experimentation, we generate ground truth data from the ACL Anthology Network Radev *et al.* (2013) or AAN. We take papers with publication years from 1980 to 2002 and generate classification data for each of the three sets for five different reference periods, $r_i \in (p_i, p_i + 5)$.

3.4.3.1 Approach

We adopt a supervised machine learning approach and experiment with several classes of features based on metadata and content. We describe our main feature classes below.

Metadata Features These include author based features, such as the average h-index of the authors, the number of citations received by the authors on their previous papers and their affiliations. We also include features based on the publication venue of the paper, such as the impact factor of the venue. Additionally, we include features based on the papers cited by this paper, eg. the pagerank of cited papers. We also include features based on papers and authors citing this paper in the reference period.

Features derived from Heterogeneous Networks An additional set of features is derived from a heterogeneous network that combines authors, papers, venues, institutions and terms used in titles into a single network. For each of the entities connected to the paper, we compute the pagerank of the entity in this heterogeneous network and use them as features. Additionally, we calculate the slope of the change in the pagerank of these features over the last 5 years, and use this as a trend feature for the entity. The intuition behind these features is that a paper that is connected to prominent entities has a higher chance of becoming prominent itself. For example, a paper that has a prominent term in its title might get noticed more.

Lexical Features We add several features that try to capture the diversity of the current paper. We compute the average, minimum and maximum similarity of the title of this paper with the titles of its cited papers. We also create features based on the similarity of the title of this paper with all the published in the same year, and also the papers published in the previous year. These values model how this paper is positioned with respect to the prior work and also with respect to its contemporary papers.

We derive an additional set of lexical features using terms. We extract a gold standard set of terms from the entire AAN by first extracting noun chunks and then manually labelling them. We compute a metric for the impact of this paper by measuring the ratio of the number of papers with this term in reference period to the number of such papers in a 5-year window prior to this paper's publication. We also compute a metric for novelty by creating a network from all the terms in the current paper. We add an edge in this network for any co-occurrence of these terms prior to the publication year of this paper. We use the maximum, minimum and average edge weight in this network along with the density of this network as a metric of the novelty of this paper.

3.4.3.2 Experiments

Reference Lag	**All features Classification F-score**	**Baseline (citation count)**
1	0.34	0.09
2	0.86	0.83
3	0.87	0.83
4	0.86	0.83
5	0.87	0.83

Table 3.6: Performance on the task of detecting prominent papers in a horizon of 10 years based on 1-5 years of evidence. The baseline values are derived from only using the number of citations to the papers in the reference period as a feature

We trained a logistic regression classifier for each of the problem formulations

	Delayed Recognition		Flash in the Pan	
Reference Lag	**Classification F-score**	**Number of Papers**	**Classification F-score**	**Number of Papers**
1	0.12	334	0.19	275
2	0.25	229	0.22	225
3	0.34	200	0.27	177
4	0.16	145	0.19	152
5	0.13	112	0.12	114

Table 3.7: Performance on the task of detecting papers with delayed recognition and flash in the pan in a horizon of 10 years based on 1-5 years of evidence. The count indicates the number of papers that are in the given class out of the 6,995 papers in the data set.

using the features discussed in the previous section. To compare our results with that presented in Yogatama *et al.* (2011), we extract features by setting the reference lag as 0 and setting a boolean label based on whether the paper is cited in the first 3 years or not. We then use the papers between 1980-2003 as training data and measure the accuracy of prediction from 2004-2006. We achieve an average accuracy of 0.70, which is close to the one reported in their paper.

Results on predicting the prominence of the papers are shown in Table 3.6. We can see that we can achieve an F-score of more than 86% on the task of predicting the papers that will be prominent in forecast year by looking at just two years of data. The drastic different between the accuracy with one and two years of data points can be attributed to the publication cycles of the conferences that are a part of AAN. Most of them publish proceedings annually, as opposed to monthly. Thus, it takes about one to two years of time before a paper starts accumulating citations. Even an important paper may not get enough visibility within the first year to start to receive attention from people.

Results on predicting whether a paper is going to be a "sleeping beauty" or a "flash in the pan" is shown in Table 3.7. We find that this is an extremely difficult problem, with maximum F-score of 34% on detecting "sleeping beauties" and an

F-score of 27% on detecting "flash in the pans".

We did feature analysis by running the Chi-square test on our training set. The top features are mostly features derived from the heterogeneous networks. Especially, the trend features that capture the properties of the current paper and its authors, and the authors of its citing papers are important. The most important lexical features are the term impact of abstracts, the average similarity of the paper to its cited papers, minimum term edge weight of the terms appearing in titles and the minimum similarity of the paper with all the papers published in the same year. The similarity features capture both the novelty of the paper compared to its cited papers and the papers published in the current year. The term edge weight feature measures the novelty of the paper in terms of new ideas it brings together, while the term impact measures how popular these terms become in the reference period after the current paper mentions them.

3.4.4 Combined Prediction

Recent work on the problem of citation prediction has focused on either predicting whether a paper will receive a certain number of citations within a short time span after publication Yogatama *et al.* (2011) or to predict the exact citation count at some point in the future Yan *et al.* (2012). Even though these are interesting formulations of this problem, they do not give us insight into the temporal nature of the citations. A recent model that provides such insight was presented by Wang *et al.* (2013), we'll refer to it as the WSB model henceforth. They present a mechanistic model for the citation dynamics of papers that can be used to fit and predict citation trajectories for papers. The model estimates a curve for a paper based on three parameters: μ which measures immediacy or how soon a paper grows in importance, σ which captures the decay rate for the citations of a paper, and λ which captures the relative inherent quality of the paper with respect to other papers.

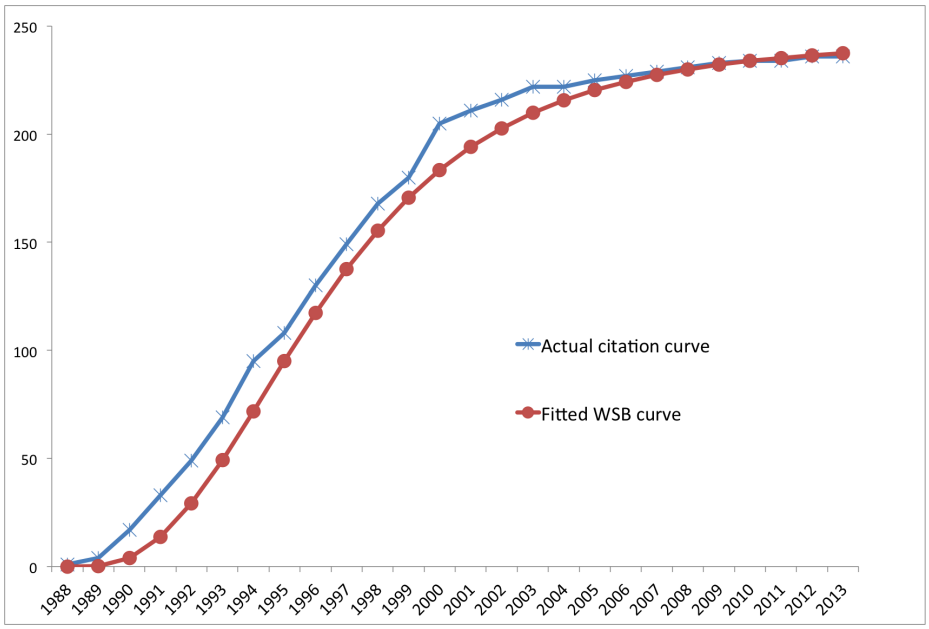

Figure 3.2: The cumulative citation graph for Church (1988). The blue line with cross markers represents the actual citation curve, while the red line with circle markers represents the output for each year from the fitted WSB model. Parameters learned for the WSB model for this curve are $\lambda = 2.21, \mu = 7.6, \sigma = 0.64$

The WSB model successfully captures three different factors in the citation dynamics of a paper: preferential attachment, long term decay and its inherent quality. We find that the papers in the ACL Anthology fit the model closely. Figure 3.2 shows the actual and fitted citation curve for Church, (1988). We first fit the WSB model to AAN papers using an existing implementation[2] and measured the goodness of fit. For measuring the goodness of fit, we used the Kolmogorov-Smirnov test similar to Wang *et al.* (2013). We found that for 69.78% of the papers, the null hypothesis that they fit the WSB model cannot be rejected at level 0.1.

Given this, if we can predict the values of λ, μ and σ for a new paper within reasonable error, we should be able to predict the future citation path for a paper with a good accuracy. Thus the fitted parameters of the WSB model can be used as response variables in a citation prediction setting. Using this formulation also allows us to measure the influence of different features on the immediacy, longetivity, and

[2]http://josiahneuberger.github.io/citation_prediction/

the perceived quality of paper.

Our initial set of experiments, however, showed that a simple regression model for predicting the number of citations outperforms the strategy of first predicting λ, μ and σ and then using it to predict the citation using the WSB curve. The RMSE for simple regression on papers in a single year (2005) was 7.12 while the RMSE for WSB strategy was 11.0. This can be attributed to the cumulative error accumulation of errors for each of the three variables in the factored WSB model. Despite this, we think that modeling features that correlate well with each of the parameters of the WSB model can be used to gain new insights into the dynamics of evolution of scientific fields, but leave this to future work.

3.5 Publications

Most of the work presented in this chapter was published previously in Jha *et al.* (2013). Some of the features for citation prediction were derived from heterogenous scholarly networks, which was first presented in King *et al.* (2014).

CHAPTER IV

Content Models for Extracting Survey-worthy Sentences

A number of network based content models have been proposed and evaluated for the task of scientific summarization (Qazvinian & Radev, 2008b; Mohammad *et al.* , 2009). The evaluation for these content models has been done using pyramid evaluation based factoids extracted from the citing sentences which form the input to the summarizer itself. Although pyramid evaluation is better than ROUGE for assessing the quality of these content models, since the factoids used are extracted from the input, the evaluation does not give us any indication of how these content models compare with human written surveys. In this chapter, we describe a dataset of factoids that we extracted from several human written surveys on seven topics. We use these factoids to conduct pyramid evaluation of existing content models.

We first describe the network based content models that we evaluated, followed by a description of the data and our experimental results. This is followed by a section on combining network based content models with bayesian content models.

4.1 Network Based Content Models

We experiment with three existing network based content models: Centroid, Lexrank and C-Lexrank, each of which is described below.

4.1.1 Centroid

The centroid of a set of documents is a set of words that are statistically important to the cluster of documents. Centroid based summarization of a document set involves first creating the centroid of the documents, and then judging the salience of each document based on its similarity to the centroid of the document set. In our case, the input citing sentences represent the documents from which we extract the centroid. We use the centroid implementation from the publicly available summarization toolkit, MEAD (Radev *et al.* , 2004a).

4.1.2 Lexrank

LexRank (Erkan & Radev, 2004) is a network based content selection algorithm that works by first building a graph of all the documents in a cluster. The edges between corresponding nodes represent the cosine similarity between them. Once the network is built, the algorithm computes the salience of sentences in this graph based on their eigenvector centrality in the network.

4.1.3 C-Lexrank

C-Lexrank is another network based content selection algorithm that focuses on diversity (Qazvinian & Radev, 2008a). Given a set of sentences, it first creates a network using these sentences and then runs a clustering algorithm to partition the network into smaller clusters that represent different aspects of the paper. The motivation behind the clustering is to include more diverse facts in the summary.

Many corpus based methods have been proposed to deal with the sense disambiguation problem when given de nition for each possible sense of a target word or a tagged corpus with the instances of each possible sense, e.g., supervised sense disambiguation (Leacock et al. , 1998), and semi-supervised sense disambiguation (Yarowsky, 1995).
Most researchers working on word sense disambiguation (WSD) use manually sense tagged data such as SemCor (Miller et al. , 1993) to train statistical classifiers, but also use the information in SemCor on the overall sense distribution for each word as a backoff model.
Yarowsky (1995) has proposed a bootstrapping method for word sense disambiguation.
Training of WSD Classifier Much research has been done on the best supervised learning approach for WSD (Florian and Yarowsky, 2002; Lee and Ng, 2002; Mihalcea and Moldovan, 2001; Yarowsky et al. , 2001).
For example, the use of parallel corpora for sense tagging can help with word sense disambiguation (Brown et al. , 1991; Dagan, 1991; Dagan and Itai, 1994; Ide, 2000; Resnik and Yarowsky, 1999).

Figure 4.1: A sample output survey of our system on the topic of "Word Sense Disambiguation" produced by paper selection using Restricted Expansion and sentence selection using Lexrank. In our evaluations, this survey achieved a pyramid score of 0.82 and Unnormalized RU score of 0.31.

4.2 Evaluation Data for Network Based Models

We use the ACL Anthology Network (AAN) as the corpus for our experiments (Radev *et al.* , 2013). We built a factoid inventory for seven topics in NLP based on manual written surveys in the following way. For each topic, we found at least 3 recent tutorials and 3 recent surveys on the topic and extracted the factoids that are covered in each of them. Table 4.1 shows the complete list of material collected for the topic of "Word Sense Disambiguation". We found around 80 factoids per topic on an average. Once the factoids were extracted, each factoid was assigned a weight based on the number of documents it appears in, and any factoids with weight one were removed. Table 4.2 shows the top ten factoids in the topic of Word Sense Disambiguation along with their distribution across the different surveys and tutorials and final weight.

Authors	**Year**	**Size**
Surveys		
ACL Wiki	2012	4
Roberto Navigli	2009	68
Eneko Agirre; Philip Edmonds	2006	28
Xiaohua Zhou; Hyoil Han	2005	6
Nancy Ide; Jean Vronis	1998	41
Tutorials		
Sanda Harabagiu	2011	45
Diana McCarthy	2011	120
Philipp Koehn	2008	17
Rada Mihalcea	2005	186

Table 4.1: The set of surveys and tutorials collected for the topic of "Word Sense Disambiguation". Sizes for surveys are expressed in number of pages, sizes for tutorials are expressed in number of slides.

For each of the topics, we used the method described earlier to create a candidate document set and extracted the candidate citing sentences to be used as the input for the content selection component. Each sentence in each topic was then annotated by a human judge against the factoid list for that topic. A sentence is allowed to

Factoid	S1	S2	S3	S4	S5	T1	T2	T3	T4	Weight
definition of wsd	X	X	X	X	X	X	X	X	X	9
wordnet	X	X	X		X	X	X	X	X	8
knowledge based wsd		X	X	X	X	X		X	X	7
supervised wsd	X	X	X	X	X	X		X		7
senseval	X	X	X			X	X	X	X	7
definition of word senses	X		X	X		X		X	X	7
machine readable dictionaries		X	X		X	X		X	X	6
unsupervised wsd		X	X	X		X	X	X		6
bootstrapping algorithms	X	X	X			X	X	X		6
supervised wsd using decision lists	X	X	X	X	X			X		6

Table 4.2: Top 10 factoids for the topic of "Word Sense Disambiguation" and their distribution across various data sources. The last column shows the factoid weight for each factoid.

have zero or more than one factoid. The human assessors were graduate students in Computer Science who have taken a basic "Natural Language Processing" course or an equivalent course. On an average, 375 citing sentences were annotated for each topic, with 2,625 sentences being annotated in total. We present all our experimental results on this large annotated corpora which is also available for download [1].

4.3 Experiments

To do an evaluation of our different content selection methods, we first select the documents using our Restricted Expansion method, and then pick the citing sentences to be used as the input to the summarization module. Given this input, we generate 500 word summaries for each of the seven topics using the four methods: Centroid, Lexrank, C-Lexrank and a random baseline.

For each summary, we compute two evaluation metrics. The first is the Pyramid

[1] http://clair.si.umich.edu/corpora/survey_data/

score (Nenkova & Passonneau, 2004) computed by treating the factoids as Summary Content Units (SCU's). The second metric is an Unnormalized Relative Utility score (Radev & Tam, 2003), computed using the factoid scores of sentences based on the method presented in (Qazvinian, 2012). We call this Unnormalized RU since we are not able to normalize the scores with human generated gold summaries. The parameter α is the RU penalty for including a redundant sentence subsumed by an earlier sentence. If the summary chooses a sentence s_i with score w_{orig} that is subsumed by an earlier summary sentence, the score is reduced as $w_{subsumed} = (\alpha * w_{orig})$. We approximate subsumption by marking a sentence s_j as being subsumed by s_i if $F_j \subset F_i$, where F_i and F_j are sets of factoids covered in each sentence. We now describe the two evaluation metrics and compare them.

4.3.1 Relative Utility

In Relative Utility (RU) (Radev & Tam, 2003), a number of judges are asked to assign utility scores to each sentence in the input set. Let there be N judges in total, n input sentences, and e number of sentences in the extract to be evaluated. The sentence utility vector of judge i over all n input sentences is defined as:

$$\vec{U_i} = \{u_{i,1}, u_{i,2}, ..., u_{i,n}\}$$

The total extractive self-utility of all N judges is calculated using:

$$U' = \sum_{j=1}^{n} \epsilon_j \cdot u_{i,j}$$

ϵ_j is the multi-judge summary characteristic function and is 1 for the top e sentences according to the sum of utility scores from all judges. U' is the maximum utility that any system can achieve at a given summary length e.

Now, given an extract of length e, we can calculate its utility by adding the scores

given to each of its sentences by each of the judges. The Relative Utility of the summary is the ratio of this divided by the maximum possible performance. Thus, Relative Utility of a summary is judged based on its utility relative to the maximum possible against the set of judges, and is given by:

$$S = \frac{\sum_{j=1}^{n} \xi_{s,j} \cdot \sum_{i=1}^{N} u_{i,j}}{U'}$$

$\xi_{s,j}$ is equal to 1 for the top e sentences extracted by the system to be evaluated. To use relative utility in our experiments, we follow (Qazvinian, 2012) in assigning utility scores to each sentence. We assume each factoid source (tutorial or survey) represents a judge and assigns a utility to any sentence based on the number of its factoids it contains. Thus for source S_i, if a sentence contains two factoids present in the source, it gets a utility score of 2.

Relative Utility also contains a mechanism for evaluating redundancy in the summary content through conditional sentence utility values. This is incorporated using the idea of subsumption. Sentence s_i subsumes sentence s_j if all the information present in s_j is also present in s_i. This means that once we include s_i in our summary, the utility of s_j should be dropped. This is done by penalizing the addition of sentences already subsumed by the existing summary sentences by a parameter α, the utility of such sentences becomes:

$$U_{subsumed} = \alpha * U_{original}$$

In the original setup, subsumption is identified by human judges. In our setting, we can use the following approximation: if all the factoids in a new summary sentence already appear in the existing summary, we mark the sentence as subsumed by the current summary and apply the penalized score for this sentence. The paramter α, which takes a value from 0 to 1, determines the amount of penalty for subsumed

sentences. We use an α of 0.5 in our experiments.

4.3.2 Pyramid Evaluation

For pyramid evaluation (Nenkova & Passonneau, 2004), we first organize our factoids in a pyramid of order n. The top tier in this pyramid contains the highest weighted factoids, the next tier contains the second highest weighted factoids and so on. The score assigned to a summary is the ratio of the sum of the weights of the factoids it contains to the sum of weights of an optimal summary with the same number of factoids.

Factoid ID	G_1	G_2	Total Count
f1	X	X	2
f2	X	X	2
f3	X	X	2
f4	X	X	2
f5		X	1
f6		X	1
f7	X		1
f8	X		1

(a) Factoid distribution over two sources, 'X' indicates that the factoid is present in the source represented by the column

Sentence id	Factoids
s1	f5
s2	f6
s3	f5, f6
s4	f7, f8

(b) Factoids in each of the 4 sentences

Table 4.3: Example illustrating difference between Pyramid and Relative Utility

4.3.3 Comparing Pyramid Evaluation with Relative Utility

We illustrate the difference between the two metrics using a simple example. Assume two gold factoid sources, G_1 and G_2 and 8 factoids, with the distribution shown in Table 4.3(a). Assume that the input set contains four sentences s1, s2, s3 and s4 with the factoid distribution as shown in Table 4.3(b). Consider two output summaries: {s1, s2} and {s3, s4}.

The pyramid score for {s1,s2} is computed in the following way. It contains 2 factoids each with weight 1, so it has a factoid weight of 2. The ideal weight for a

Topic	Rand	Cent	LR	C-LR
Summarization	0.68	0.61	0.91	0.82
Question Answering	0.52	0.50	0.65	0.56
Word Sense Disambiguation	0.78	0.73	0.82	0.76
Named Entity Recognition	0.90	0.90	0.94	0.94
Sentiment Analysis	0.75	0.78	0.77	0.78
Semantic Role Labeling	0.78	0.79	0.88	0.94
Dependency Parsing	0.67	0.38	0.71	0.53
Average	0.72	0.68	0.81*	0.76

Table 4.4: Results of pyramid evaluation for each of the three methods and the random baseline on each topic.

Topic	Rand	Cent	LR	C-LR
Summarization	0.16	0.57	0.29	0.17
Question Answering	0.32	0.39	0.48	0.30
Word Sense Disambiguation	0.28	0.33	0.31	0.30
Named Entity Recognition	0.36	0.38	0.34	0.31
Sentiment Analysis	0.23	0.34	0.48	0.33
Semantic Role Labeling	0.11	0.17	0.16	0.21
Dependency Parsing	0.16	0.05	0.30	0.15
Average	0.23	0.32	0.34*	0.25

Table 4.5: Results of Unnormalized Relative Utility evaluation for the three methods and random baseline using $\alpha = 0.5$.

summary with 2 factoids is 4. Thus the pyramid score is $2/4 = 0.5$. Similarly, the weight for {s3,s4} is 0.5.

Now in Relative Utility, the self utility U' for all judges at summary length 2 is 4 ({s3,s4}). Now, summary {s1,s2} gets a score of 2 from the judge G_2 and no score from G_1. It's relative utility is given by $2/4 = 0.5$. {s3,s4} gets a score of 2 from judge G_1 and a score of 2 from judge G_2, giving it a relative utility score of $4/4 = 1$.

Thus Relative Utility gives higher score to {s3,s4} as compared to {s1,s2}, which makes intuitive sense since {s3,s4} covers more factoids than {s1,s2}. This difference arises because pyramid score measures how well a summary does against an ideal summary containing the same number of *factoids*, while RU score measures how well a summary does against an ideal summary containing the same number of *sentences*.

We report both of these metrics for our experiments.

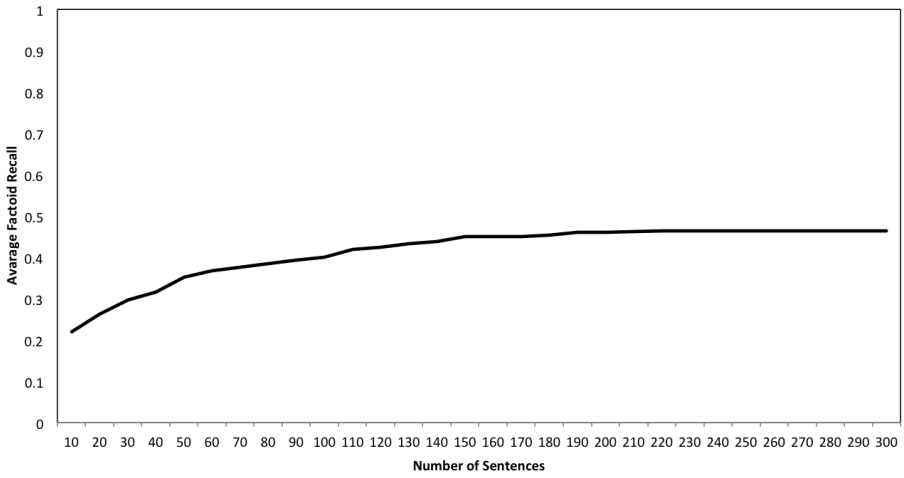

Figure 4.2: Factoid distribution in the gold standard data for the different topics

4.3.4 Results

The pyramid scores for each summary are shown in Table 4.4 and the Unnormalized RU scores are shown in Table 4.5.

The reason for the relatively high scores for the random baseline is that our process to select the initial set of sentences eliminates many bad sentences. For example, for a subset of 5 topics, the total input set contains 1508 sentences, out of which 922 of the sentences (60%) have at least one factoid. This makes it highly likely to pick good content sentences even when we are picking sentences at random.

We find that the Lexrank method outperforms other sentence selection methods on both evaluation metrics. The higher performance of Lexrank compared to Centroid is consistent with earlier published results (Erkan & Radev, 2004). The reason for the low performance of C-Lexrank as compared to Lexrank on this data set can be attributed to the fact that the input sentence set is derived from a much more diverse set of papers which can have a high diversity in lexical choice when describing the same factoid. Thus simple lexical similarity is not enough to find good clusters in this sentence set.

In recent years, conditional random fields (CRFs) (Lafferty et al. , 2001) have shown success on a number of natural language processing (NLP) tasks, including shallow parsing (Sha and Pereira, 2003), named entity recognition (McCallum and Li, 2003) and information extraction from research papers (Peng and McCallum, 2004).
In natural language processing, two aspects of CRFs have been investigated sufficiently: one is to apply it to new tasks, such as named entity recognition (McCallum and Li, 2003; Li and McCallum, 2003; Settles, 2004), part-of-speech tagging (Lafferty et al., 2001), shallow parsing (Sha and Pereira, 2003), and language modeling (Roark et al., 2004); the other is to exploit new training methods for CRFs, such as improved iterative scaling (Lafferty et al., 2001), L-BFGS (McCallum, 2003) and gradient tree boosting (Dietterich et al., 2004)
NP chunks are very similar to the ones of Ramshaw and Marcus (1995).
CRFs have shown empirical successes recently in POS tagging (Lafferty et al. , 2001), noun phrase segmentation (Sha and Pereira, 2003) and Chinese word segmentation (McCallum and Feng, 2003)
CRFs have been successfully applied to a number of real-world tasks, including NP chunking (Sha and Pereira, 2003), Chinese word segmentation (Peng et al., 2004), information extraction (Pinto et al., 2003; Peng and McCallum, 2004), named entity identification (McCallum and Li, 2003; Settles, 2004), and many others.

Figure 4.3: A sample output survey produced by our system on the topic of "Conditional Random Fields" using Restricted Expansion and Lexrank.

The lower Unnormalized RU scores compared to Pyramid scores indicate that we are selecting sentences containing highly weighted factoids, but we do not select the most informative sentences that contain a large number of factoids. This also shows that we select some redundant factoids, since Unnormalized RU contains a penalty for redundancy. This is again, explained by the fact that the simple lexical diversity based model in C-Lexrank is not able to detect the same factoids being present in two sentences. Despite these shortcomings, our system works quite well in terms of content selection for unseen topics, Figure 4.3 shows the top 5 sentences for the query "Conditional Random Fields".

Finally, Figure 4.2 shows an interesting trend, the recall of the system does not increase beyond a 0.4 as we keep increasing the output length of the summary. This shows that the citing sentences themselves do not contain all the information that is needed to summarize a scientific topic. In the next chapter, we focus on other methods of finding the missing information.

4.4 Combining Network and Bayesian Models for Content Selection

In this section, we explore two different recent approaches for multi-document summarization: probabilistic content models and lexical network models and present a new joint model that uses information from both of these models.

Given a document set with documents relevant to a topic, Bayesian content models (Daumé & Marcu, 2006; Haghighi & Vanderwende, 2009) learn a word distribution for the topic and assign importance to sentences based on this word distribution. Network based models (Erkan & Radev, 2004; Qazvinian & Radev, 2008a) on the other hand, model the input as a network of sentences and assign importance to sentences based on their centrality in this network.

Word distribution	
srl	0.034
semantic	0.029
role	0.025
arguments	0.015
argument	0.013
verb	0.013
labeling	0.012
propbank	0.012
roles	0.012
predicate	0.011
palmer	0.010
shared	0.009

(a)

Kl-div	Sentence text
9.09	The identification of event frames may potentially benefit many natural language processing (NLP) applications, such as information extraction (Surdeanu et al. 2003), question answering (Narayanan and Harabagiu 2004), summarization (Melli et al. 2005), and machine translation (Boas 2002).
9.21	The benefit of semantic roles has already been demonstrated for a number of tasks, among others for machine translation (Boas, 2002), information extraction (Surdeanu et al., 2003), and question answering (Narayanan and Harabagiu, 2004).

(b)

Centrality	Sentence text
0.00352	The SRL task is to identify semantic roles (or arguments) of each predicate and then label them with their functional tags, such as 'Arg0' and 'ArgM' in PropBank (Palmer et al., 2005), or 'Agent' and 'Patient' in FrameNet (Baker et al., 1998).
0.00332	Semantic Role Labeling (SRL) aims to identify and label all the arguments for each predicate in a sentence.

(c)

Figure 4.4: An example showing the different sentences selected for topic of *semantic role labeling* by Bayesian content models and network based models. (a) shows the topic word distribution learnt by the Bayesian model and (b) shows the top two sentences based on their KL-divergence score with the topic word distribution (lower is better). (c) shows the top two sentences by their pagerank centrality in the lexical network.

As an example, Figure 4.4(a) shows the top few words in the word distribution learnt for the topic of *semantic role labeling* and Figure 4.4(b) shows the top two sentences selected based on their score using the Bayesian content model (the score for a sentence is its KL-divergence with topic word distribution, so lower is better). Figure 4.4(c) shows the top sentences selected on the same input using Lexrank, a centrality score computed using the lexical network of sentences. Both methods tend to select useful, but different sentences.

As described in Section 4.6, Bayesian content models capture the hierarchical structure of the data, but not the relationships between sentences. Network based models ignore the hierarchical nature of the data, effectively treating the data as a "bag of sentences", but capture inter-sentential relationships in the data. Thus, it would be useful to have a joint model that can use the information from both these views of the data.

The main contribution of this chapter is a joint model for combining Bayesian and network models for summarization. We describe our approach and present empirical evidence that it can improve content selection in multi-document summarization.

4.5 Data Preparation

We use the ACL Anthology Network (AAN) as a corpus for our experiments (Radev *et al.* , 2013). AAN provides the full text of papers published in most of the venues in the field of *natural language processing* and provides additional useful data such as a manually curated citation network between all the papers in the corpus. For our experiments, we picked 15 topics, the list of topics appears later in Table 5.4.

For each of these topics, we use an adaptation of the algorithm described in (Jha *et al.* , 2013) for building the document set. We first found a core set C of documents highly relevant to the topic in the following way. At least three published surveys were found for each topic. The bibliographies of all these surveys were processed using Parscit (Luong *et al.* , 2010). Any document that appeared in the bibliography of more than one survey was added to C. On average, we found only 33% of the documents in C of any topic to be in AAN. Since the citation network for AAN contains only citations within AAN documents, we implemented a heuristic record matching algorithm to find all the papers in AAN that cite any arbitrary document outside AAN. We then used this enhanced citation network to find all the papers in AAN citing papers in C. The citing documents are ordered based on the number of papers

in C that they cite and the top n documents are then selected ($n = 20$ for current experiments). This is the set of documents that cite a number of important papers on the topic and thus contain useful sentences for summarizing the topic. Based on preliminary data analysis, introduction sections of papers in AAN contain a number of background sentences useful for summarizing the topic because the introduction is usually the place where authors describe the background of their field and how it relates to the new work being presented. Therefore, we extract the introductions from each of these n papers, which form the input document set for each topic.

4.6 Methodology for Combining Models

We now describe three models: TopicSum, which is a probabilistic content model; Lexrank, which is a network centrality model; and TSLR, which is our joint model that combines information from both these models.

4.6.1 TopicSum

TopicSum is a probabilistic content model presented in Haghighi & Vanderwende (2009) and is very similar to an earlier model called BayeSum proposed by Daumé & Marcu (2006). It is a hierarchical, LDA (Latent Dirichlet Allocation) style model that is based on the following generative story[2]: words in any sentence in the corpus can come from one of three word distributions: a background word distribution ϕ_B that flexibly models stop words, a content word distribution ϕ_C for each document set that models content relevant to the entire document set, and a document specific word distribution ϕ_D. The generative model for TopicSum is reproduced in Figure 4.5. The word distributions are learnt using Gibbs sampling. Given n document sets each with k documents, we get n content word distributions and $n * k$ document specific

[2]To avoid confusion in use of the term "topic", we call topics in the LDA sense word distributions. "Topic" in this paper refer to the natural language processing topics such as *question answering*, *word sense disambiguation*, etc.

distributions leading to a total of $1 + n + n * k$ word distributions.

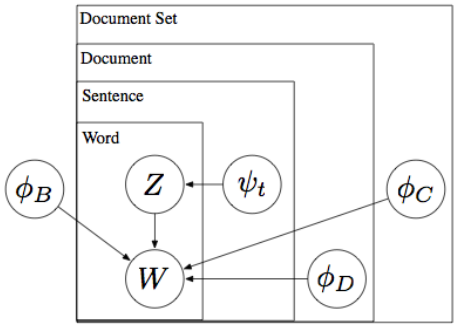

Figure 4.5: Graphical model for TopicSum from (Haghighi & Vanderwende, 2009).

To illustrate the kind of distributions TopicSum learns, Figure 4.6 shows the top words along with their probabilities from the background word distribution, a content word distribution and a document specific word distribution. We see that the model effectively captures general content words for each topic. $\phi_{C/QA}$ is the word distribution for the topic of *question answering* while $\phi_{D/J07-1005}$ is the document specific word distribution for a specific paper in the document set for *question answering* [3] that focuses on clinical question answering. The word distribution $\phi_{D/J07-1005}$ contains words that are relevant to the specific subtopic in the paper, while $phi_{C/QA}$ contains content words relevant to the general topic of *question answering*.

These topics, learnt using Gibbs sampling, can be used to select sentences for a summary in the following way. To summarize a document set, we greedily select sentences that minimize the KL-divergence of our summary to the document set specific topic. Thus the score for each sentence s is $KL(phi_C||P_s)$ where P_s is the sentence word distribution with add-one smoothing applied to both distributions. Using this objective, sentences that contain words from the content word distribution with high probability are more likely to be selected in the generated summary.

[3]Dina Demner-Fushman and Jimmy Lin. 2007. *Answering Clinical Questions with Knowledge-Based and Statistical Techniques.* Computational Linguistics.

ϕ_B		$\phi_{C/QA}$		$\phi_{D/J07-1005}$	
the	0.06643	question	0.04368	metathesaurus	0.00032
of	0.03964	questions	0.03793	umls	0.00032
and	0.03427	answer	0.02845	biomedical	0.00024
a	0.02887	answering	0.02236	relevance	0.00024
in	0.02745	qa	0.02067	citation	0.00024
to	0.02718	answers	0.01695	wykoff	0.00024
is	0.01737	2001	0.01600	bringing	0.00016
for	0.01449	system	0.01086	appropriately	0.00016
that	0.01200	trec	0.00815	organized	0.00016
we	0.01137	factoid	0.00782	foundation	0.00016

Figure 4.6: Top words from three different word distributions learnt by TopicSum on our input document set of 15 topics. ϕ_B is the background word distribution that captures stop words. $\phi_{C/QA}$ is the word distributions for the topic of *question answering*. $\phi_{D/J07-1005}$ is the document specific word distribution for a single paper in *question answering* that focuses on clinical question answering.

4.6.2 Lexrank

Lexrank is a network based content selection algorithm. Given a set input of sentences, it first creates a network using these sentences where each node represents a sentence and each edge represents the tf-idf cosine similarity between the sentences. Two methods for creating the network are possible. First, we can remove all edges that are lower than a certain threshold of similarity (generally set to 0.1). The Lexrank value for a node $p(u)$ in this case is calculated as:

$$\frac{1-d}{N} + d \sum_{v \in adj[u]} \frac{p(v)}{deg(v)}$$

Where N is the total number of sentences, d is the damping factor (usually set to 0.85), $deg(v)$ is the degree of the node v, and $adj[u]$ is the set of nodes connected to the node u. A different way of creating the network is to treat the sentence similarities as edge weights and use the adjacency matrix as a transition matrix after normalizing

the rows, the formula then becomes:

$$\frac{1-d}{N} + d \sum_{v \in adj[u]} \frac{cos(u,v)}{TotalCos_v} p(v)$$

Where $cos(u,v)$ gives the tf-idf cosine similarity between sentence u and v and $TotalCos_v = \sum_{z \in adj[v]} cos(u,v)$. In our experiments, we employ this second formulation. The above equation can be solved efficiently using the power method (**?**) to obtain the $p(u)$ for each node, which is then used as the score for ordering the sentences. The final lexrank values $p(u)$ for a node represent the stationary distribution of the Markov chain represented by the transition matrix.

Network based algorithms have been shown to perform well in summarization experiments (Erkan & Radev, 2004; Qazvinian & Radev, 2008a). However, these models do not have a rich representation of word distribution and do not take into account the hierarchical structure of the data. This suggests combining network models with the probabilistic content models in a way that leverages the strengths of both representations.

4.6.3 TSLR

The damping factor in the Lexrank calculations is introduced in order to provide a certain default centrality to all the nodes regardless of the number of edges the node has. In terms of the random walk view of the network, this makes the random walk jump with a certain probability to any node of the network. However, this random jump can also be used to encode prior belief about the importance of each node. By making the probability of jumping to a node non-uniform, we can bias the random walk towards certain nodes that we know to be important through other means. This idea has been exploited earlier in biased Lexrank (Otterbacher *et al.* , 2009) for query focused summarization where the probability of jumping to any random node is made

to depend on the similarity of the sentence to the give query. In our case, we can use the information provided by TopicSum to bias the scores. This allows sentences that have high centrality but low TopicSum scores to obtain higher scores. Similarly, sentences that have a low centrality, but high TopicSum scores obtain a higher overall score by having a higher probability of a random jump transitioning to them. The integrated score, which we call $TSLR(u)$, is given by:

$$(1-d)KL_{norm}(u) + d \sum_{v \in adj[u]} \frac{cos(u,v)}{TotalCos_v} p(v)$$

where $KL_{norm}(u)$ is a normalized version of the TopicSum score for the sentence u and is calculated as:

$$KL_{norm}(u) = \frac{KL_{max} - KL(phi_c||u)}{\sum_{z \in C} KL_{max} - KL(phi_c||z)}.$$

KL_{max} can be set to a value so that $KL_{max} - KL(phi_c||u)$ for any u would be positive.

In TSLR, The contribution of TopicSum scores and Lexrank centrality score to the final score can be controlled by changing the damping factor d. We experiment with different values of the damping factors and report the results in Section 5.1.3.

4.7 Experiments and Results

For evaluating our models, we used Lexrank, TopicSum and TSLR with different values of the damping factor to generate 2000 character summaries for each topic. Since all the methods provide a score to each sentence in the input set, this can easily done by ordering the sentences by their respective scores and cutting off the summary when the desired length is reached.

For generating reference summaries for ROUGE evaluation, we asked two assessors

Topic	LR	TSLR-0.7	TSLR-0.4	TSLR-0.1	TS
coreference resolution	0.53	0.50	0.41	0.34	0.34
question answering	0.42	0.47	0.51	0.45	0.56
sentiment analysis	0.32	0.45	0.47	0.54	0.53
dependency parsing	0.39	0.40	0.38	0.43	0.37
semi supervised learning	0.48	0.51	0.42	0.44	0.46
grammar induction	0.33	0.35	0.35	0.34	0.31
information extraction	0.41	0.41	0.43	0.38	0.37
information retrieval	0.42	0.41	0.41	0.41	0.38
machine translation	0.36	0.39	0.38	0.38	0.37
named entity recognition	0.26	0.25	0.24	0.31	0.32
semantic role labeling	0.52	0.51	0.51	0.51	0.49
speech recognition	0.36	0.36	0.38	0.34	0.36
summarization	0.43	0.41	0.39	0.34	0.35
topic models	0.59	0.58	0.55	0.51	0.46
word sense disambiguation	0.34	0.34	0.35	0.41	0.40
Average	**0.411**	**0.423**	**0.412**	**0.409**	**0.405**

Table 4.6: ROUGE-1 score for each topic for the different methods. We show scores for Lexrank (LR), TopicSum (TS), and TSLR. TSLR scores are shown with three values of the damping factor: 0.7, 0.4, 0.1.

to manually generate a 2000 character summary for each of the 15 topics in our dataset. Thus, two reference summaries were created for each topic which were used for the ROUGE evaluation. Additionally, we did a manual evaluation by showing pairs of summaries generated by different systems to one of three different assessors and asked them to mark which summary they preferred in terms of content selection or mark "indifferent" if no summary could be deemed better than the other. The summaries were randomly assigned and ordered so the assessors could not figure out the original systems that produced the summaries.

The final ROUGE-1 scores for each of the topics are shown in Figure 5.4 for TopicSum, Lexrank and TSLR with three values of the damping factor (0.1, 0.4 and 0.7). On an average, TSLR achieves better ROUGE-1 than either TopicSum or Lexrank. Higher values of the damping factor seem to improve the ROUGE-1 score and TSLR-0.7 gives the best scores among all the variants.

The individual scores for the different topics provide more insight into the working

of TSLR. Comparing just TopicSum and Lexrank, we notice that for some topics such as *coreference resolution* and *topic models*, Lexrank achieves a better ROUGE-1 score compared to TopicSum. On the other hand, for some other topics such as *question answering* and *sentiment analysis*, TopicSum yields better scores. In each case, however, combining the two methods using TSLR-0.7 improves the score over the lower performing model by using information from the better performing model. For some topics such as *dependency parsing*, the combined model performs better than either Lexrank or TopicSum. In only one case, *named entity recognition*, TSLR-0.7 does worse than either model. This shows that combining the two models allows us to leverage the information from each model to produce better summaries on an average.

Since TSLR-0.7 achieved the best ROUGE-1 scores, we chose this variant for our human evaluations. We showed different assessors three pairs of summaries: Lexrank vs TopicSum, Lexrank vs TSLR-0.7 and TopicSum vs TSLR-0.7 for each topic. Between Lexrank and TopicSum, our assessors preferred TopicSum 66% of the time while preferring Lexrank 33% of the time. Between Lexrank and TSLR-0.7, assessors preferred TSLR-0.7 20% of the time while being indifferent 80% of the time. Between TopicSum and TSLR-0.7, the assessors preferred TSLR-0.7 47% of the time and TopicSum 53% of the time. These results indicate that in terms of human evaluation, compared to Lexrank, TSLR-0.7 produces summaries that are either as good or better. Compared to TopicSum, even though TSLR-0.7 summaries do not perform better in all the cases, they are preferred in several cases. These results are consistent with our observations from the ROUGE scores, which shows that TSLR-0.7 does not perform better than either model all the time but, for most topics, produces summaries competitive with the best model and much better than the worst model.

4.8 Publications

The work on extracting factoids and pyramid evaluation results were presented previously in Jha *et al.* (2013).

CHAPTER V

Content Models Based on Linking Citing and Source Text

Earlier chapters have investigated content models based on lexical networks (Mohammad *et al.* , 2009; Qazvinian & Radev, 2008a). These models take as input citing sentences that describe important papers on the topic and assign them a salience score based on centrality in a lexical network formed by the input citing sentences. In this section, we propose a new content model based on network structure previously unexplored for this task that exploits the lexical relationship between citing sentences and the sentences from the original papers that they cite. Our new formulation of the lexical network structure fits nicely with the hubs and authorities model for identifying important nodes in a network (Kleinberg, 1999), leading to a new content model called HitSum. We also explore supervised methods for aligning citing sentences with source sentences.

5.1 Alignment Based Content Models

For the task of evaluating various content models discussed in this paper, we have annotated a total of 3,425 sentences across 7 topics in the field of *natural language processing* with factoids from each of the topics. The factoids we use were extracted

Factoid	Weight
Question Answering	
answer extraction	6
question classification	6
definition of question answering	5
TREC QA track	5
information retrieval	5
Dependency Parsing	
non-projective dependency structures / trees	6
projectivity / projective dependency trees	6
deterministic parsing approaches: Nivre's algorithm	5
terminology: head - dependent	4
grammar driven approaches for dependency parsing	4

Figure 5.1: Sample factoids from the topics of *question answering* and *dependency parsing* along with their factoid weights.

from existing survey articles and tutorials on each topic (Jha *et al.* , 2013), and thus represent information that must be captured by a survey article on the corresponding topic. Each of the factoids is assigned a weight based on its frequency in the surveys/tutorials, which allows us to do pyramid evaluation of our content models. Some sample factoids are shown in Figure 5.1. Evaluation using factoids extracted from existing survey articles can help us understand the limits of automated survey article generation and how well these systems can be expected to perform. For example, if certain kinds of factoids are missing consistently from our input sentences, improvements in content models are unlikely to get us closer to the goal of generating survey articles that match those generated by humans, and effort must be directed to extracting text from other sources that will contain the missing information. On the other hand, if most of the factoids exist in the input sentences but important factoids are not found by the content models, we can think of strategies for improving these models by doing error analysis.

The main contributions presented in this section are:

- HitSum, a new HITS-based content model for automatic survey generation for

Topic	# Sentences
dependency parsing	487
named entity recognition	383
question answering	452
semantic role labeling	466
sentiment analysis	613
summarization	507
word sense disambiguation	425

Table 5.1: List of seven NLP topics used in our experiments along with input size.

Input sentence	Factoids
According to [1] , the corpus based supervised machine learning methods are the most successful approaches to WSD where contextual features have been used mainly to distinguish ambiguous words in these methods.	supervised wsd, corpus based wsd
Compared with supervised methods, unsupervised methods do not require tagged corpus, but the precision is usually lower than that of the supervised methods.	supervised wsd, unsupervised wsd
Word sense disambiguation (WSD) has been a hot topic in natural language processing, which is to determine the sense of an ambiguous word in a specific context.	definition of word sense disambiguation
Improvement in the accuracy of identifying the correct word sense will result in better machine translation systems, information retrieval systems, etc.	wsd for machine translation, wsd for information retrieval
The SENSEVAL evaluation framework (Kilgarriff 1998) was a DARPA-style competition designed to bring some conformity to the field of WSD, although it has yet to achieve that aim completely.	senseval

Table 5.2: Sample input sentences from the topic of *word sense disambiguation* annotated with factoids.

scientific topics.

- A new dataset of 3,425 factoid-annotated sentences for scientific articles in 7 topics.
- Experimental results for pyramid evaluation comparing three existing content models (Lexrank, C-Lexrank, TopicSum) with HitSum.

The rest of this section is organized as follows. Section 5.1.1 describes the dataset used in our experiment and the factoid annotation process. Section 5.1.2 describes each of the content models used in our experiments including HitSum. Section 5.1.3 describes our experiments and Section 5.1.4 summarizes the results.

5.1.1 Data

Prior research in automatic survey generation has explored using text from different parts of scientific papers. Some of the recent work has treated survey generation as a direct extension of single paper summarization (Qazvinian & Radev, 2008a) and used citing sentences to a set of relevant papers as the input for the summarizer (Mohammad *et al.* , 2009). However, other researchers have observed that it's difficult to generate coherent and readable summaries using just citing sentences and have proposed the use of sentences from introductory texts of papers that cite a number of important papers on a topic (Jha *et al.* , 2015b) . The use of full text allows for the use of discourse structure of these documents in framing coherent and readable surveys. Since the content models we explore are meant to be part of a larger system that should be able to generate coherent and readable survey articles, we use the introduction sentences for our experiments as well.

The corpus we used for extracting our experimental data was the ACL Anthology Network, a comprehensive bibliographic dataset that contains full text and citations for papers in most of the important venues in *natural language processing* (Radev

et al. , 2013). An oracle method is used for selecting the initial set of papers for each topic. For each topic, the bibliographies of at least three human-written surveys were extracted, and any papers that appeared in more than one survey were added to the target document set for the topic.

The text for summarization is extracted from introductory sections of papers that cite papers in the target document set. The intuition behind this is that the introductory sections of papers that cite these target document summarize the research in papers from the target document set as well as the relationships between these papers. Thus, these introductions can be thought of as mini-surveys for specific aspects of the topic; combining text from these introductory sections should allow us to generate good comprehensive survey articles for the topic. For our experiments, we sort the citing papers based on the number of papers they cite in the target document set, pick the top 20 papers, and extract sentences from their introductions to form the input text for the summarizer. The seven topics used in our experiments and input size for each topic are shown in Table 6.3.

Once the input text for each topic has been extracted, we annotate the sentences in the input text with factoids for that topic. Some annotated sentences in the topic of *word sense disambiguation* are shown in Table 5.2. Given this new annotated data, we can compare how the factoids are distributed across different citing sentences (as annotated by Jha *et al.* (2013)) and introduction sentences that we have annotated. For this, we divide the factoids into five categories: definitions, venue, resources, methodology, and applications. The fractional distribution of factoids in these categories is shown in Table 5.3. We can see that the distribution of factoids relating to venues, methodology and applications is similar for the two datasets. However, factoids related to definitional sentences are almost completely missing in the citing sentences data. This lack of background information in citing sentences is one of the motivations for using introduction sentences for survey article generation as opposed

A dictionary such as the LDOCE has broad coverage of word senses, useful for WSD .
This paper describes a program that disambiguates English word senses in unrestricted text using statistical models of the major Roget's Thesaurus categories.
Our technique offers benefits both for online semantic processing and for the challenging task of mapping word senses across multiple MRDs in creating a merged lexical database.
The words in the sentences may be any of the 28,000 headwords in Longman's Dictionary of Contemporary English (LDOCE) and are disambiguated relative to the senses given in LDOCE.
This paper describes a heuristic approach to automatically identifying which senses of a machine- readable dictionary (MRD) headword are semantically related versus those which correspond to fundamentally different senses of the word.

Figure 5.2: A sentence from P_{citing} with a high hub score (bolded) and some of sentences from P_{cited} that it links to (italicised). The sentence from P_{citing} obtain a high hub score by being connected to the sentences with high authority scores.

Factoid category	**% Citing**	**% Intro**
definitions	0	4
venue	6	6
resources	18	2
methodology	70	83
applications	6	5

Table 5.3: Fractional distribution of factoids across various categories in citing sentences vs introduction sentences.

to previous work.

5.1.2 HitSum

Lexrank, C-Lexrank and TopicSum have been described in earlier chapters. In this section, we describe our new algorithm HitSum.

The input set of sentences in our data come from introductory sections of papers that cite important papers on a topic. We'll refer to the set of citing papers that provide the input text for the summarizer as P_{citing} and the set of important papers that represent the research we are trying to summarize as P_{cited}. Both Lexrank and C-Lexrank work by finding central sentences in a network formed by the input sentences

and thus, only use the lexical information present in P_{citing}, while ignoring additional lexical information from the papers in P_{cited}. We now present a formulation that uses the network structure that exists between the sentences in the two sets of papers to incorporate additional lexical information into the summarization system. This system is based on the hubs and authorities or the HITS model (Kleinberg, 1999) and hence is called HitSum.

HitSum, in addition to the sentences from the introductory sections of papers in P_{citing}, also uses sentences from the abstracts of P_{cited}. It starts by computing the tf-idf cosine similarity between the sentences of each paper $p_i \in P_{citing}$ with the sentences in the abstracts of each paper $p_j \in P_{cited}$ that is directly cited by p_i. A directed edge is created between every sentence s_i in p_i and s_j in p_j if $sim(s_i, s_j) > s_{min}$, where s_{min} is a similarity threshold (set to 0.1 for our experiments). Once this process has been completed for all papers in P_{citing}, we end up with a bipartite graph between sentences from P_{citing} and P_{cited}.

In this bipartite graph, sentences in P_{cited} that have a lot of incoming edges represent sentences that presented important contributions in the field. Similarly, sentences in P_{citing} that have a lot of outgoing edges represent sentences that summarize a number of important contributions in the field. This suggests using the HITS algorithm, which, given a network, assigns hubs and authorities scores to each node in the network in a mutually reinforcing way. Thus, nodes with high authority scores are those that are pointed to by a number of good hubs, and nodes with high hub scores are those that point to a number of good authorities. This can be formalized with the following equation for the hub score of a node:

$$h(v) = \sum_{u \in successors(v)} a(u)$$

Where $h(v)$ is the hub score for node v, $successors(v)$ is the set of all nodes that v

Topic	Lexrank	C-Lexrank	TopicSum	HitSum
dependency parsing	0.47	0.76	0.62	1.00*
named entity recognition	0.80	0.89	0.90*	0.80
question answering	0.65	0.67	0.65	0.76*
sentiment analysis	0.64	0.62	0.75*	0.63
semantic role labeling	0.75*	0.67	0.65	0.69
summarization	0.52	0.75*	0.57	0.68
word sense disambiguation	0.78	0.66	0.67	0.79*
Average	**0.66**	**0.72**	**0.69**	**0.76***

Table 5.4: Pyramid scores obtained by different content models for each topic along with average scores for each model across all topics. For each topic as well as the average, the best performing method has been highlighted with a *.

has an edge to, and $a(u)$ is the authority score for node u. Similarly, the authority score for each node is computed as:

$$a(v) = \sum_{u \in predecessors(v)} h(u)$$

Where $predecessors(v)$ is the set of all nodes that have an edge to v. The hub and authority score for each node can be computed using the power method that starts with an initial value and iteratively updates the scores for each node based on the above equations until the hub and authority scores for each node converge to within a tolerance value (set to 1E-08 for our experiments).

In our bipartite lexical network, we expect sentences in P_{cited} receiving high authority scores to be the ones reporting important contributions and sentences in P_{citing} that receive high hub scores to be sentences summarizing important contributions. Figure 5.2 shows an example of a sentence with a high hub score from the topic of *word sense disambiguation*, along with some of the sentences that it points to. HitSum computes the hub and authority score for each sentence in the lexical network and then uses the hub scores for sentences in P_{citing} as their relevance score. Sentences from P_{cited} are part of the lexical network, but are not used in the output summary.

5.1.3 Experiments

For evaluating our content models, we generated 2,000-character-long summaries using each of the systems (Lexrank, C-Lexrank, HitSum, and TopicSum) for each of the topics. The summaries are generated by ranking the input sentences using each content model and picking the top sentences till the budget of 2,000 characters is reached. Each of these summaries is then given a pyramid score (Nenkova & Passonneau, 2004) computed using the factoids assigned to each sentence.

For the pyramid evaluation, the factoids are organized in a pyramid of order n. The top tier in this pyramid contains the highest weighted factoids, the next tier contains the second highest weighted factoids, and so on. The score assigned to a summary is the ratio of the sum of the weights of the factoids it contains to the sum of weights of an optimal summary with the same number of factoids. Pyramid evaluation allows us to capture how each content model performs in terms of selecting sentences with the most highly weighted factoids. Since the factoids have been extracted from human-written surveys and tutorials on each of the topics, the pyramid score gives us an idea of the survey-worthiness of the sentences selected by each content model.

5.1.4 Results and Discussion

The results of pyramid evaluation are summarized in Table 5.4. It shows the pyramid score obtained by each system on each of the topics as well as the average score. The highest performing system on average is HitSum with an average performance of 76%. HitSum does especially well for the topics of *dependency parsing*, *question answering*, and *word sense disambiguation*. The second best performing system is C-Lexrank, which is not surprising because it was developed specifically for the task of scientific paper summarization. However, HitSum outperforms C-Lexrank on several topics and by 4% on average.

TopicSum does well on the topics of *named entity recognition* and *sentiment*

analysis, but does not do well on average. This can be attributed to the fact that it was developed as a content model for the domain of news summarization and does not translate well to our domain. All systems outperform Lexrank, which achieves the lowest average score. This result is also intuitive, because every other system in our evaluation uses additional information not used by Lexrank: C-Lexrank exploits the community structure in the input set of sentences, HitSum exploits the lexical information from cited sentences, and TopicSum exploits information about global word distribution across all topics.

The different systems we tried in our evaluation depend on using different lexical information and seem to perform well for different topics. This suggests that further gains can be made by combining these systems. For example, C-Lexrank and HitSum can be combined by utilizing both the network formed by citing sentences and the network between the citing sentences and the cited sentences into a larger lexical network. TopicSum scores can be combined with these network-based system by using the TopicSum scores as a prior for each node, and then running either Pagerank or HITS on top of it. We leave exploration of such hybrid systems to future work.

5.2 Aligning Citing Sentences with Source Sentences

Summarizing research papers based on only citing sentences ignores the fact that in some cases, the source sentence might be a good summary of the contribution. Thus, even though we still need to look at the citing sentence to know what piece of information is important about a given paper, a good linguistic summary might instead lie in the source paper itself. On the other hand, a large number of sentences in the source paper are concerned with specific details of the paper and are not suitable for summarization. Therefore, we need to find the small subset of source sentences that contain the information that the paper is being cited about. This motivates the task of citation source alignment, which we now formally define. Given a source

Citing Text	Aligned Source Text
Current approaches have used clustering ... to identify sense-specic subgraphs	To detect the different areas of meaning in our local graphs, we use a cluster algorithm for graphs (Markov clustering, MCL) developed by van Dongen
sentence retrieval for question answering (Otterbacher et al., 2005)	Our goal is to build a question-focused sentence retrieval mechanism
in fact, Pedersen (2001) found that bigrams alone can be effective features for word sense disambiguation	This paper shows that the combination of a simple feature set made up of bigrams and a standard decision tree learning algorithm results in accurate word sense disambiguation

Table 5.5: Examples of citing text along with aligned source text from the cited paper

paper S comprising of n sentences $s_1, s_2 \ldots, s_n$ and a paper C citing this source paper, consider the sentence c_i in source paper. The task of citation source alignment is to find a subset of source sentences $S_c \subset S$ such that S_c contain the information that is being cited by c_i. Table 5.5 shows some examples of citing sentences along with aligned source sentences from the original paper.

5.2.1 Data

For our experiments, we use the SciSumm Corpus [1]. The SciSumm corpus contains annotated data for 10 sets of source papers. For each source paper, upto 10 citing papers are found. Each citing paper is first annotated to extract the text segments that explicitly cite the source paper. Each citing text segment is then matched one or more text spans in the source papers. In total, there are 140 annotations.

We pre-processed the SciSumm corpus in the following way. We first sentence segmented the source paper text for each of the 10 papers provided in the original SciSumm corpus. We then matched each of these source sentences to the SciSumm annotation files. This provided us a with a fixed set of source sentences from the

[1] https://github.com/WING-NUS/scisumm-corpus

original files, a subset of which were matched to each citing sentence. In this way given, a citing sentence, we can compare the matching sentences from the source paper returned by our system to the gold standard sentences matched from the source paper and compute precision/recall.

The average number of source sentences matched for each citing sentence is 1.28 (with standard deviation 1.92). The maximum number of source sentences matched for a citing sentence is 7. Given that the total number of source sentences for papers ranges between from 100 to 600, this makes it a very challenging classification problem.

5.2.2 System Description

5.2.2.1 Features

Lexical Features We use two lexical features. The first feature is based on TF*IDF cosine similarity. The IDF's were computed over all sentences for each source paper, thus the IDF values differed across each of the 10 source papers. For any citing sentence, we computed the TF*IDF cosine similarity with all the sentences in the source paper and use them as a feature. The second lexical feature is based on the LCS (Longest Common Subsequence) between the citing sentence (C) and source sentence S and is computed as:

$$\frac{|LCS|}{min(|C|, |S|)}$$

Knowledge Based Features We also compute a set of features based on Wordnet similarity. We use six wordnet based word similarity measures and combine these word similarities to obtain six knowledge based sentence similarity features using the method proposed in (Banea *et al.* , 2012). The wordnet based word similarity measures we use are path similarity, WUP similarity (Wu & Palmer, 1994) , LCH sim-

ilarity (Leacock & Chodorow, 1998), Resnik similarity (Resnik, 1995), Jiang-Conrath similarity (Jiang & Conrath, 1997), and Lin similarity (Lin, 1998).

Given each of these similarity measures, the similarities between two sentences is computed by first creating a set of senses for each of the words in each of the sentences. Given these two sets of senses, the similarity score between citing sentence C and source sentence S is calculated as follows:

$$sim_{wn}(C, S) = \frac{(\omega + \sum_{i=1}^{|\phi|} \phi_i) * (2|C||S|)}{|C| + |S|}$$

Here ω is the number of shared senses between C and S. The list ϕ contains the similarities of non-shared words in the shorter text, ϕ_i is the highest similarity score of the ith word among all the words of the lower text (Tiantian & Lan, 2013).

Syntactic Features We compute an additional feature based on similarity of dependency structures using the method described in (Tiantian & Lan, 2013) . We use the Stanford parser to obtain dependency parse all the citing sentences and source sentences. Given a candidate sentence pair, two syntactic dependencies are considered equal if they have the same dependency type, govering lemma, and dependent lemma. If R_c and R_s are the set of all dependency relations in C and S, the dependency overlap score is computed using the formula:

$$sim_{dep}(C, S) = \frac{2 * |R_c \cap R_s| * |R_c||R_s|}{|R_c| + |R_s|}$$

We extracted all these features and trained a logistic regression classifier.

5.2.3 Results and Discussion

Since we had a limited amount of data, we evaluated our results using 10-fold cross validation. We report the precision, recall, F1-score, and F2-score for different feature

Feature set	Precision	Recall	F1-score	F2-score
only lex	0.011	0.168	0.021	0.044
lex+wn	0.010	0.187	0.019	0.041
lex+wn+dep	0.011	0.192	0.022	0.045

Table 5.6: 10-fold cross validation results for citing sentence alignment

combinations in Table 5.6. The recall seems to increase slightly as we augment simple lexical features with wordnet and dependency features. However, since the precision does not improve correspondingly, the gains in F scores are not much.

A number of errors made by the system are due to source sentences that match the words but differ slightly in their information content. Here is an example.

Citing text: use the BNC to build a co-occurrence graph for nouns, based on a co-occurrence frequency threshold

True Positives:

- Following the method in (Widdows and Dorow, 2002), we build a graph in which each node represents a noun and two nodes have an edge between them if they co-occur in lists more than a given number of times 1.

False positives:

- Based on the intuition that nouns which co-occur in a list are often semantically related, we extract contexts of the form Noun, Noun,... and/or Noun, e.g. "genomic DNA from rat, mouse and dog".

- To detect the different areas of meaning in our local graphs, we use a cluster algorithm for graphs (Markov clustering, MCL) developed by van Dongen (2000).

- The algorithm is based on a graph model representing words and relationships between them.

Even though the false positive sentences contain the same lexical items (nouns, co-occurrence, graph), they differ slightly in the facts presented. Detection of such subtle differences in meaning might be challenging for an automated system.

Another set of difficult sentences is when the citing sentence says something that is implied by the sentence in the source paper. For example:

Citing text: The line of our argument below follows a proof provided in ... for the maximum likelihood estimator based on nite tree distributions

False negatives:

- We will show that in both cases the estimated probability is tight.

Here, the citing text mentions a proof from source paper, but to match the sentence in the source paper, the system needs to understand that the act of showing something in a scientific paper constitutes a proof. The conclusion of this research was that a simple tf*idf cosine baseline can be used as a good proxy for this alignment as it provides competitive results compared to the supervised methods we explored.

5.3 Publications

The work on HitSum and other content models as well as their experimental evaluation will appear in a forthcoming publcation Jha *et al.* (2015a). The results presented here for aligning citing sentences with source sentences for the SciSumm corpus were previously published as part of Jha *et al.* (2015b).

CHAPTER VI

Generating Coherent Surveys

This chapter is about generating coherent summaries of scientific topics. Given a set of input papers that are relevant to a specific topic such as *question answering*, our system called Surveyor extracts and organizes text segments from these papers into a coherent and readable survey of the topic. There are many applications for automated surveys thus generated. Human surveys do not exist for all topics and quickly become outdated in rapidly growing fields like computer science. Therefore, an automated system for this task can be very useful for new graduate students and cross-disciplinary researchers who need to quickly familiarize themselves with a new topic.

Our work builds on previous work on summarization of scientific literature (Mohammad *et al.* , 2009; Qazvinian & Radev, 2008a). Prior systems for generating survey articles for scientific topics such as C-Lexrank have focused on building informative summaries but no attempt has been made to ensure the coherence and readability of the output summaries. Surveyor on the other hand focuses on generating survey articles that contain well defined subtopics presented in a coherent order. Figure 6.1 shows part of the output of Surveyor for the topic of *question answering*.

Our experimental results on a corpus of computational linguistics topics show that Surveyor produces survey articles that are substantially more coherent and readable

Traditional Information Retrieval (IR) focuses on searching and ranking a list of documents in response to a user's question.
However, in many cases, a user has a specific question and want for IR systems to return the answer itself rather than a list of documents (Voorhees and Tice (2000)).
To satisfy this need, the concept of Question Answering (QA) comes up, and a lot of researches have been carried out, as shown in the proceedings of AAAI and TREC (Text REtrieval Conference).
Li and Roth (2002) used a Sparse Network ofWinnows (SNoW) (Khardon et al., 1999).

Question classification is a crucial component of modern question answering system.
It classifies questions into several semantic categories which indicate the expected semantic type of answers to the questions.

The Question Answering (QA) task has received a great deal of attention from the Computational Linguistics research community in the last few years (e.g., Text Retrieval Conference TREC 2001,2003) .

Figure 6.1: Example output of Surveyor for the topic of *question answering*. The survey contains three distinct subtopics illustrated by different colors and separated by dashed lines.

compared to previous work.

6.1 Overview of Summarization Approach

We first describe the two main components of our system and then describe our summarization algorithm that is built on top of them.

6.1.1 Content Model

Given a set of research papers relevant to a scientific topic, each of them focuses on a specific aspect of the problem. For example, a paper on supervised word sense disambiguation might describe the background on word sense disambiguation followed by a review of supervised methods for the problem. Similarly, a paper on unsupervised word sense disambiguation may give some general overview of the field, then briefly describe supervised approaches followed by a more detailed overview of unsupervised methods. We capture these subtopics in the input documents and their transitions

subtopic 1
BB constructs classifiers for English-to-Chinese translation disambiguation by repeating the following two steps: (1) Construct a classifier for each of the languages on the basis of classified data in both languages, and (2) use the constructed classifier for each language to classify unclassified data, which are then added to the classified data of the language.
In translation from English to Chinese, for example, BB makes use of unclassified data from both languages.
subtopic 2
Word sense disambiguation (WSD) is the problem of assigning a sense to an ambiguous word, using its context.
The task of Word Sense Disambiguation (WSD) is to identify the correct sense of a word in context.
subtopic 3
We extend previously reported work in a number of different directions: We evaluate the method on all parts of speech (PoS) on SemCor.
Previous experiments evaluated only nouns on SemCor, or all PoS but only on the Senseval2 and Senseval3 data.

Figure 6.2: Example sentences from three subtopics learnt by the HMM for *word sense disambiguation*.

using a Hidden Markov Model (HMM) where the states of the HMM correspond to subtopics. Given the set of k subtopics $S = (s_1 \cdots s_k)$, the state transitions of the HMM are defined as:

$$p(s_j|s_i) = \frac{Count(s_i, s_j) + \delta}{Count(s_i) + \delta * m}$$

Where $Count(s_i, s_j)$ is the number of times a sentence from subtopic s_j appears immediately after a sentence from subtopic s_i in the input document collection and $Count(s_i)$ is the total number of times the subtopic s_i appears in the input document set. δ is a smoothing parameter and m is the number of sentences in s_i.

To initialize the states of the HMM, we use a network based clustering approach. We build a lexical network where the sentences represent the nodes of the network and the edge weights are the tf*idf similarity between each pair of sentences [1]. Given

[1]The idfs are computed over the entire input corpus.

	subtopic 1	**subtopic 2**	**subtopic 3**
start	0.35	0.50	0
subtopic 1	0.49	0.22	0
subtopic 2	0.24	0.41	0.02
subtopic 3	0.25	0.25	0.50

Table 6.1: A partial table of transition probabilities between three subtopics for *word sense disambiguation*. The probabilities do not add up to 1 because the table only shows a few states from a larger transition matrix.

this lexical network, we use the Louvain clustering method (De Meo *et al.* , 2011) to partition the lexical network into clusters. Each cluster in the network is then initialized to a sub-topic. Louvain is a hierarchical clustering algorithm that does not need the number of output clusters as a parameter. The HMM is then learned through Viterbi decoding. Our HMM model is similar to (Barzilay & Lee, 2004), but we take the novel step of using the transition matrix to guide the summarization output, as described below.

Figure 6.2 shows sentences from three of the subtopics learned for the topic of *word sense disambiguation*. In a coherent summary, subtopic 2 containing background sentences should appear before subtopic 1 that contains details about a specific method. We use the transition matrix of the learned HMM to model these subtopic transitions in the original documents and use it to guide the summarizer output. As an example, a partial table of transition probabilities learned for the subtopics in Figure 6.2 is shown in Table 6.1, where *start* is a pseudo-state representing the beginning of the document. The highest outgoing probability from *start* is to subtopic 2, which allows the summarizer to include background information about the topic at the beginning followed by sentences from more specific subtopics represented by subtopic 1 and subtopic 3.

s1	Opinion words are words that convey positive or negative polarities.
s2	They are critical for opinion mining (Pang et al., 2002; Turney, 2002; Hu and Liu, 2004; Wilson et al., 2004; Popescu and Etzioni, 2005; Gamon et al., 2005; Ku et al., 2006; Breck et al., 2007; Kobayashi et al., 2007; Ding et al., 2008; Titov and McDonald, 2008; Pang and Lee, 2008; Lu et al., 2009).
s3	The key difficulty in finding such words is that opinions expressed by many of them are domain or context dependent.
s4	Several researchers have studied the problem of finding opinion words (Liu, 2010).
s5	The approaches can be grouped into corpus-based approaches (Hatzivassiloglou and McKeown, 1997; Wiebe, 2000; Kanayama and Nasukawa, 2006; Qiu et al., 2009) and dictionary-based approaches (Hu and Liu 2004; Kim and Hovy, 2004; Kamps et al., 2004; Esuli and Sebastiani, 2005; Takamura et al., 2005; Andreevskaia and Bergler, 2006; Dragut et al., 2010).

$$
\begin{aligned}
midc(s1) &= \emptyset \\
midc(s2) &= \{s1\} \\
midc(s3) &= \{s1, s2\} \\
midc(s4) &= \emptyset \\
midc(s5) &= \{s4\}
\end{aligned}
$$

Figure 6.3: A paragraph from an input paper on the topic of *opinion mining* along with the *midc* for each sentence on the right.

Discourse relationship	Dependency rule
Coreference	Add a dependency between s_i and s_j if they belong to a coreference chain.
Discourse Transition	Add a dependency between s_{i-1} and s_i if s_i contains an explicit discourse marker.
Entity Transition	Add a dependency between s_i and s_j if they both share a prominent entity.

Table 6.2: Discourse rules used to create *minimum independent discourse contexts*

6.1.2 Discourse Model

A common problem with extractive summaries is that the sentences used from the original input documents may not be understandable when pulled out of their original context. To avoid such problems, we introduce the idea of *Minimum Independent Discourse Contexts (MIDC).*

Definition. Given a text segment T containing n sentences $(s_1 \cdots s_n)$, the minimum independent discourse context (midc) of a sentence s_i is defined as the minimum set of j sentences $midc(s_i) = (s_{i-j} \cdots s_i)$ such that given $midc(s_i)$, s_i can be interpreted independently of the other sentences in T.

Figure 6.3 shows how this idea works in practice. Sentences $s1$ and $s4$ can be included in a summary without requiring additional context sentences. Sentences $s2$, $s3$ and $s4$ on the other hand, depend on a set of previous sentences in order to be understandable. A summary that includes sentence $s3$, for example, must include sentences $s1$ and $s2$ for it to be understood outside of its original text.

To calculate the *midc*s for each sentence, we use discourse rules that are triggered by coreference dependencies, explicit discourse dependencies and entity links between sentences. These rules are summarized in Table 6.2. Every time a discourse rule is triggered, a dependency is added between two sentences. The *midc* for a sentence s_i is all the sentences preceding s_i in the input document to which it has a dependency edge. The coreference chains are found using the Stanford dependency parser

(de Marneffe *et al.* , 2006) and the discourse markers are obtained from the Penn Discourse TreeBank (Prasad *et al.* , 2008). The prominent entities used for creating entity links are nouns that appear in the syntactic role of subject or object in any sentence in the input.

6.1.3 Summarization Algorithm

We now describe how our summarization algorithm works given the output of these two components. The pseudocode for the algorithm is presented in Figure 6.4.

The algorithm accepts a set of input documents *docs* and a maximum summary length *maxlen*. It first learns the subtopics and their transition matrix by running HMM on the input document set. After initializing the first subtopic to the pseudo-subtopic *start*, it iteratively picks the next subtopic by using the HMM transition matrix. Given each subtopic, it runs a salience algorithm on all the sentences of the subtopic to find the most central sentence of the subtopic. In the current implementation, this is done using Lexrank (Erkan & Radev, 2004). Given the subtopic's most central sentence, it calculates the *midc* for this sentence and if the *midc* is valid, it is added to the output summary. An *midc* can be invalid if it exceeds a maximum threshold number of sentences [2] The *midc* is then removed from the subtopic so it will not be picked if we visit this subtopic again. This procedure continues till we obtain a summary of the desired length. Important subtopics in the input can get more than one *midc* in the summary because the transition matrix contains high probabilities for transitioning to these subtopics.

6.2 Experimental Setup

The main research questions that we want to answer using our experiments are:

[2]This constraint is added so that a highly salient sentence with a long *midc* does not dominate most of the output summary.

```
input  : docs, maxlen
output: summary of length maxlen
summary ← ∅;
transitionMatrix ← HMM(docs);
curSubtopic ← start;
while len(summary) < maxlen do
    nextSubtopic ← transitionMatrix.next(curSubtopic);
    nextSent ← getMostSalient(sents(nextSubtopic)));
    nextMidc ← midc(nextSent);
    if valid(nextMidc) then
        summary.add(nextMidc); sents(nextSubtopic).remove(nextMidc)
    end
    curSubtopic ← nextSubtopic;
end
return summary;
```

Figure 6.4: Summarization Algorithm

1. Are the summaries created using Surveyor more coherent than previous state-of-the-art methods for survey article generation?

2. What are the individual contributions of the content model and the discourse model?

3. How does Surveyor compare against state-of-the-art systems for coherent news summarization applied to the survey generation problem?

For Research question 1, we compare our system with C-Lexrank (Mohammad *et al.* , 2009), a state-of-the-art system for survey generation. For Research question 2, we measure the effects of HMM and MIDC models in isolation on the quality of output summaries. For Research question 3, we compare our system with G-Flow (Christensen *et al.* , 2013), a state-of-the-art system for coherent summarization of news articles. We now describe the data used in our experiments.

We used the ACL Anthology Network (AAN) (Radev *et al.* , 2013) as a corpus for our experiments and selected 15 established topics in computational linguistics for our evaluation. The input documents used for summarization of a research topic

Topic	# Sentences
coreference resolution	397
dependency parsing	487
grammar induction	407
information extraction	495
information retrieval	560
machine translation	552
named entity recognition	383
question answering	452
semantic role labeling	466
semi supervised learning	506
sentiment analysis	613
speech recognition	445
summarization	507
topic modeling	412
word sense disambiguation	425

Table 6.3: List of topics used in our experiments.

should be research papers that describe the most relevant research in the topic. Since the focus of this paper is on summarization, we used an oracle method for selecting the initial set of papers for each topic. We collected at least three human-written surveys on each topic. The bibliographies of all the surveys were processed using Parscit (Luong *et al.* , 2010) and any document that appeared in the bibliography of more than one survey was added to the initial document set D_i.

An ideal survey article on the topic should describe the research represented by D_i. These sentences are actually found in papers that cite papers in D_i and thus describe their contributions. Therefore to create the final document set D_f, we collect all the papers in AAN that cite the papers in D_i. [3] The citing documents are then ordered based on the number of papers in D_i that they cite and the top n documents are added to D_f. The text input for the summarization system is extracted from D_f. For our current experiments, the value of n is set to 20.

[3]On average, we found only 33% of the documents in D_i to be in AAN. Since the citation network for AAN contains only citations within AAN documents, we implemented a record matching algorithm to find all the papers in AAN that cite any arbitrary document outside AAN.

For the task of survey article generation, the most relevant text is found in the introduction sections of D_f since this is where researchers describe the prior work done by subsets of papers in D_i. Therefore, we extract the sentences in the introductions of each of the papers in D_f as the text input for our summarizer. Table 6.3 shows the set of 15 topics and size of summarizer input for each topic.

6.3 Experiments

6.3.1 Coherence Evaluation with C-Lexrank

For coherence evaluation, we generated fixed length 2000 character summaries using both C-Lexrank and Surveyor. Six assessors with background in computational linguistics manually evaluated pairs of output summaries that were assigned randomly to them. Given two summaries, the assessors were asked to mark which summary they preferred, or mark "indifferent" if they could not choose one against the other. Compared to C-Lexrank, the assessors preferred a summary generated by Surveyor 67% of the time and were indifferent 20% of the time.

Surveyor	**Indifferent**	**C-Lexrank**
67%	20%	13%

Additionally, the assessors were asked to rate each summary based on the standard DUC quality questions [4]. The DUC quality questions are a standard benchmark used for evaluating summaries on the aspects of overall coherence, avoiding useless text, avoiding repetitive information, avoiding bad referents and avoiding overly explicit referents. For each of the questions, the assessors can assign a score from 1 to 5 with higher being better.

As shown in Figure 6.5, the assessors also assigned much higher scores to summaries generated by Surveyor on an average compared to C-Lexrank on all the DUC

[4]http://duc.nist.gov/duc2004/quality.questions.txt

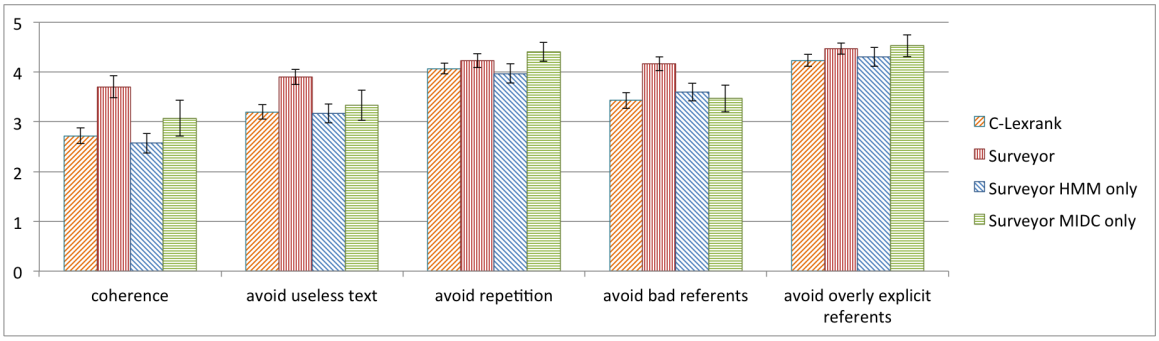

Figure 6.5: Average scores on the DUC quality questions for the different systems along with standard error.

quality questions. On the metric of coherence, the scores for Surveyor compared to C-Lexrank were higher by 36%. Both on the metrics of avoiding useless text and avoiding bad referents, the scores for Surveyor were higher by about 22%.

6.3.2 Contribution of Individual Components

To compare the contribution of the content model and the discourse model, we created two additional variants of our system. *Surveyor HMM Only* contains only the HMM component, but does not use the discourse component that adds the *midc*s for the output sentences. *Surveyor MIDC only* uses the discourse component, but instead of relying on the HMM transition matrix to generate the subtopic flow, chooses the subtopics based on their size, where size of a subtopic is the number of sentences assigned to the subtopic. It starts from the largest subtopic and goes through subtopics in order of their size.

We asked our assessors to compare summaries output by each system with the output of C-Lexrank as well rate summaries produced by each system on the DUC quality questions. The results of the direct comparison is summarized below and the average DUC ratings are reported in Table 6.5.

Even with just the HMM content model, the summaries from *Surveyor HMM Only* are preferred by assessors compared to C-Lexrank. *Surveyor MIDC Only* does

Surveyor HMM Only	Indifferent	C-Lexrank
53%	27%	20%
Surveyor MIDC Only	**Indifferent**	**C-Lexrank**
33%	27%	40%

not do as well, which suggests that without a coherent flow of subtopics, the addition of *midc*s to the output sentences does not improve performance. This shows the importance of the HMM content model and suggests that a summary that jumps between subtopics in an incoherent way will not be perceived as coherent even if the individual sentences in the summary have appropriate context. However, the scores for both of these systems on the DUC quality questions (Figure 6.5) show that the addition of *midc*s does affect the assessors' judgement of specific summary qualities and is an important component of the system. This explains why the combination of both the content model and the discourse model leads to much better results than either of them in isolation.

6.3.3 Informativeness Evaluation

We use ROUGE (Lin, 2004b) for informativeness evaluation. ROUGE is a standard evaluation metric for automatic evaluation of summaries that uses n-gram co-occurrences between automated summaries and human generated reference summaries to score the automated summaries. ROUGE has been shown to correlate well with human evaluations (Lin, 2004a).

For ROUGE evaluation, we asked two assessors to generate 2000 character long gold summaries using the input for each topic. We then did ROUGE evaluation of the summaries generated using C-Lexrank and Surveyor against these gold summaries. The average ROUGE-1 and ROUGE-2 scores are summarized below [5]. The improvement in ROUGE scores of Surveyor over C-Lexrank is statistically significant with $p < 0.05$. Thus Surveyor, in addition to producing more coherent summaries, also

[5]ROUGE-1 and ROUGE-2 correspond to unigram and bigram co-occurrence analysis respectively.

produces summaries that are more informative given the same input text.

System	ROUGE-1	ROUGE-2
C-Lexrank	0.40	0.05
Surveyor	0.44	0.19
Surveyor HMM Only	0.42	0.13
Surveyor MIDC only	0.42	0.13

We also used the factoid data created as part of the work presented in Chapter V to compute pyramid scores for each of these systems. Here are the average pyramid scores for the methods on the 7 topics in the factoid dataset:

System	Pyramid Score
C-Lexrank	0.72
Surveyor	0.68
Surveyor HMM Only	0.69
Surveyor MIDC only	0.73

In this evaluation, the Surveyor scores are slightly lower than the score for C-Lexrank. However, the average score of 68% for Surveyor is still high and given the high improvement in coherence, we hypothesize that this small trade-off in terms of informativeness would be acceptable for a potential user.

Previous evaluations for survey generation systems use citing sentences as input as opposed to sentences from the main text. There is no standard summarization evaluation that allows us to evaluate the informativeness of summaries generated using two different input sources. To compare summaries created using citing sentences and source sentences in terms of coherence, we ran C-Lexrank using both citing sentences and introduction sentences as summarizer input and did a coherence evaluation with our assessors. The assessors preferred summaries generated by using introductions as source 60% of the time while preferring summaries generated by using citing sentences as source only 27% of the time. Even though a direct comparison of informativeness is not possible, we posit that since our summaries include background information as part of the survey, our summaries would have to be slightly longer than those based

on citing sentences in order to be as informative. However, results from coherence evaluation show that using source sentences allows us to use topical and discourse information in the original papers to generate much more coherent summaries compared to citing sentences.

6.3.4 Evaluation with G-Flow

G-Flow (Christensen *et al.* , 2013) is a recent state of the art system for generating coherent summaries that has been evaluated on newswire data. We compared Surveyor with G-Flow by running the implementation of G-Flow obtained from the original authors on our evaluation data. The coherence evaluation with G-Flow was done in the same way as for C-Lexrank except the output summary length for both systems was limited to 1000 characters. This is because the optimization procedure implemented in G-Flow becomes intractable for output of 2000 characters [6].

In the coherence evaluation, assessors preferred Surveyor 47% of the time compared to 40% of the time for G-Flow.

Surveyor	**Indifferent**	**G-Flow**
47%	13%	40%

Surveyor also obtains higher scores than G-Flow on the DUC quality questions. The scores for Surveyor and G-Flow are summarized below separately because of the difference in the output length compared to the previous evaluation. The numbers are reported with standard error.

Quality question	**Surveyor**	**G-Flow**
coherence	3.53 ± 0.36	3.40 ± 0.25
avoid useless text	3.60 ± 0.36	3.47 ± 0.17
avoid repetition	4.93 ± 0.07	4.53 ± 0.19
avoid bad referents	3.93 ± 0.33	3.80 ± 0.22
avoid overly explicit referents	4.73 ± 0.12	4.47 ± 0.19

[6]Personal communication with Christensen *et al.* (2013)

In informativeness evaluation with ROUGE, the 1000 character summaries generated by Surveyor got an average ROUGE-1 score of 0.41 compared to a score of 0.36 obtained by G-Flow. The ROUGE-2 score of Surveyor was 0.13 compared to 0.07 for G-Flow. p-values for the ROUGE-1 and ROUGE-2 improvements of Surveyor over G-Flow are 0.12 and 0.11 respectively. Based on these results, Surveyor does slightly better than G-Flow in terms of coherence evaluation while producing much more informative summaries. This indicates that the HMM based content model does a better job of modeling the flow of subtopics in scientific articles compared to G-Flow which does not include such a component.

6.3.5 Introduction Sentences vs Citing Sentences

We also compared the coherence of summaries produced using introduction sentences and citing sentences as input. The summaries were produced using the same algorithm for both input sets, C-Lexrank. For each topic, an assessor was presented summaries produced by giving the two input sets separately to C-Lexrank. The assessor was asked to mark the summary they preferred and rate each summary on the DUC quality questions. For 15 topics, this generated 15 pairs of summaries for comparison. We found that between introduction and citing sentences, the assessors preferred the summaries generated using the introduction sentences as input 60% of the time, while preferring the ones generated using the citing sentences only 27% of the time. The ratings on the DUC quality questions also show that the summaries generated using introduction sentences obtained higher scores compared to citing sentences on coherence (2.72 vs 2.60), avoiding useless text (3.20 vs 2.93), avoiding bad referents (3.43 vs 3.40) and avoiding overly explicit referents (4.23 vs 4.13).

6.3.6 An Upper Baseline for Coherence

Finally, we add a system RandomPaper as an upper bound on coherence. For this method, given a topic, a random document from the input document set of the topic is picked and its first n characters are output as the summary (where n is the maximum summary length). The RandomPaper system is an upper bound for coherence, since it is completely human written, but should score low on informativeness because it comes from a single paper and thus contains sentences about only a specific sub-topic. Our assessors preferred this output 80% of the time compared to C-Lexrank due to its high coherence. However, this system obtains a ROUGE-1 score of 0.35, the lowest among all the systems by a large margin. This shows that this summary has high coherence but very low informativeness. This is confirmed by the scores on the quality questions, which show that these summaries have high score for all questions, but contain a lot of paper specific text irrelevant to the topic in general, as indicated by the low scores for avoiding useless text (2.93).

6.4 Publications

A large part of the work presented in this chapter has been published in the following AAAI article: Jha *et al.* (2015b).

CHAPTER VII

Conclusions and Future Work

This chapter provides the conclusions of this work, lists some limitations and provides pointers for future work. This thesis has presented work towards building a system that can automatically generate informative and readable surveys of scientific topics. This is an important problem given the exponential increase in the number of scientific publications in all scientific fields which is making it increasingly difficult for researchers to stay on top of the current research.

Our goal was to explore the different components needed for an end-to-end system that can take a query representing a scientific topic as an input (e.g *question answering*) and automatically generate a survey article summarizing the past research on the topic. A system to do this successfully should be able to search and find relevant papers in a given corpus, aggregate text describing those papers, and then use content and coherence models to extract relevant text segments and arrange them in a readable fashion to build a survey article for the topic. We set up automatic and human evaluations for each of these components, evaluated existing methods for building these components, and created new algorithms when previous methods fell short. We now summarize the main contributions of this work.

7.1 Main Contributions

7.1.1 Query Handling and Document Retrieval

There wasn't a lot of prior work on doing query based retrieval of scientific documents, and so we had to start with standard information retrieval algorithms for this task. We first built an evaluation dataset using surveys for seven topics in NLP and evaluated these existing algorithms. We found that these systems faced some problems due to the changing scientific terminology. Based on this, we proposed a simple model called *Restricted Expansion* that combined pure lexical search and citation information to retrieve relevant papers for a query. This algorithm led to a four fold improvement in both precision and recall compared to simple tf-idf search, and a two fold improvement compared to tf-idf search sorted with citation Pagerank.

Restricted Expansion gets acceptable results, but one of the areas that need more work is handling queries with higher granularity and detecting invalid queries. Valid queries are those queries for which it is theoretically possible to build a survey using the data available in our corpus. These can be coarse high level topics that are established topics in an area such as *word sense disambiguation*, *question answering*, etc. Our system can handle these queries successfully. A user can also submit finer topics such as *supervised methods for word sense disambiguation*, *question classification for question answering*, etc. These are valid queries, but our current system cannot handle such queries at this time, we propose to do this in our future work. In addition, a user can submit invalid queries for which it is unlikely to find information in our corpus. These include topics at a lower level of granularity than our corpus itself, for example, *computer science*, *artificial intelligence* etc. and topics that are peripheral to the corpus such as *machine learning*, *speech recognition* etc. Detecting such topics and graceful degradation is also part of the proposed research.

7.1.2 Content Models

Once the relevant documents are retrieved, we need to use content models to assign importance scores to sentences which will be part of the survey. Again, we first evaluated existing methods for this task to understand the limitations of future work. We created a factoid dataset with factoids extracted from surveys and tutorials for seven topics in NLP. We then annotated 2,625 citing sentences to evaluate existing citation based content models.

We used pyramid evaluation to compare three methods: Centroid, Lexrank and C-Lexrank. Evaluation showed that these models did a good job retrieving informative sentences, with the best performing model, Lexrank, achieving an average pyramid score of 0.81. However, error analysis showed that certain factoids related to background (e.g. definitions) were consistently missing from citing sentences.

Based on this, we decided to change the problem formulation for this task where instead of using only citing sentences, we used entire sections of papers that have a high concentration of citing sentences but also have background sentences. For our experiments, we used introduction sections of citing papers for the 7 NLP topics and annotated these 3,425 new sentences with factoids. We used this data set to evaluate Lexrank and C-Lexrank. This new formulation also allowed us to adapt a Bayesian content model called TopicSum for the task of survey generation. Based on the new structure of the data, we also proposed a new HITS based algorithm called HitSum that achieved better pyramid scores compared to all the existing algorithms.

The three content modelling algorithms (C-Lexrank, TopicSum, and HitSum) tend to capture different pieces of information from the same data. Thus, we hypothesize that methods to combine these content models might lead to improvements. We have presented some early experiments in this direction by combining Lexrank and TopicSum using a simple linear model, but more exploration in this direction is likely to yield more improvements.

7.1.3 Building Readable Surveys

One of the main shortcoming of the existing work on survey generation has been the lack of readability in generated summaries. Our evaluations on 7 topics made it clear that even though the surveys generated by existing methods are quite informative, their usability in a real use-case scenario is quite limited given how difficult it is to read the surveys. Therefore, in this thesis, we focused on making the surveys readable and coherent.

A coherent and readable piece of text has two important linguistic qualities: cohesion and coherence. Cohesion can be loosely defined as the property of the text "holding together". The sentences should be linked to each other with lexical cues so that the text reads as a combined unit without any breaks. Coherence has more to do with the structure of the text, the information should be presented in a hierarchical order that makes sense. We developed two components that would ensure that our generated surveys would be both coherent and cohesive. For coherence, we developed an HMM based model that tracks the subtopics that should be presented in the survey and the natural order of presenting these subtopics. A discourse model was developed to track the linguistic context of sentences in the generated summary so that the sentences connect to each other in a more natural way. Both these components were combined together in our summarization algorithm that produced much more readable and coherent surveys.

We ran manual evaluation of our new summarization algorithm on 15 topics in NLP where human assessors were asked to compare summaries generated by our system with those generated by prior methods. Summaries generated by our system were preferred 5 times as much as those generated by C-Lexrank, an existing state-of-the-art survey generation system. Additionally, pyramid evaluation results showed that our system does this without sacrificing the informativeness of the resulting summaries too much.

7.2 Limitations and Future Work

There are two main limitations of this work which we now describe along with pointers to future work that might alleviate these problems.

7.2.1 Evaluation Corpus

For the experiments in this thesis, we use various topics in NLP for evaluation. This limits the scope of the work somewhat, and a more comprehensive evaluation across different scientific disciplines would certainly be valuable. However, there were two strong reasons for working with the topic of NLP. First, the field of NLP has a wonderful resource, the ACL Anthology Network (AAN), that contains manually cleaned citation data, full text as well as metadata for all the major venues in the field. This obviated the overhead of collecting large amounts of publications and cleaning them up and helped us focus more on the experimental work. Secondly, given that we are researchers in NLP, it was easier for us to do initial assessment of the summaries generated by our system for the various topics. We suspect this would be difficult for other topics. However, now that enough progress has been made in terms of methods for building all the components of the pipeline, the next step would be to create evaluation data for other scientific fields, test the methods on them, and make any modifications needed to make the system more general. We suspect that the methods will need only minor modifications to adapt them to a wider array of scientific disciplines. We have collected data from some other fields that we hope to experiment with in the future:

Pubmed Central We have a citation network of close to 500,000 articles from Pubmed Central that are in our repository. In addition, a complete citation network of Pubmed is available from the Open Citations Corpus [1].

[1] http://opencitations.wordpress.com/2011/07/01/the-citation-processing-pipeline-and-the-open-citations-corpus/

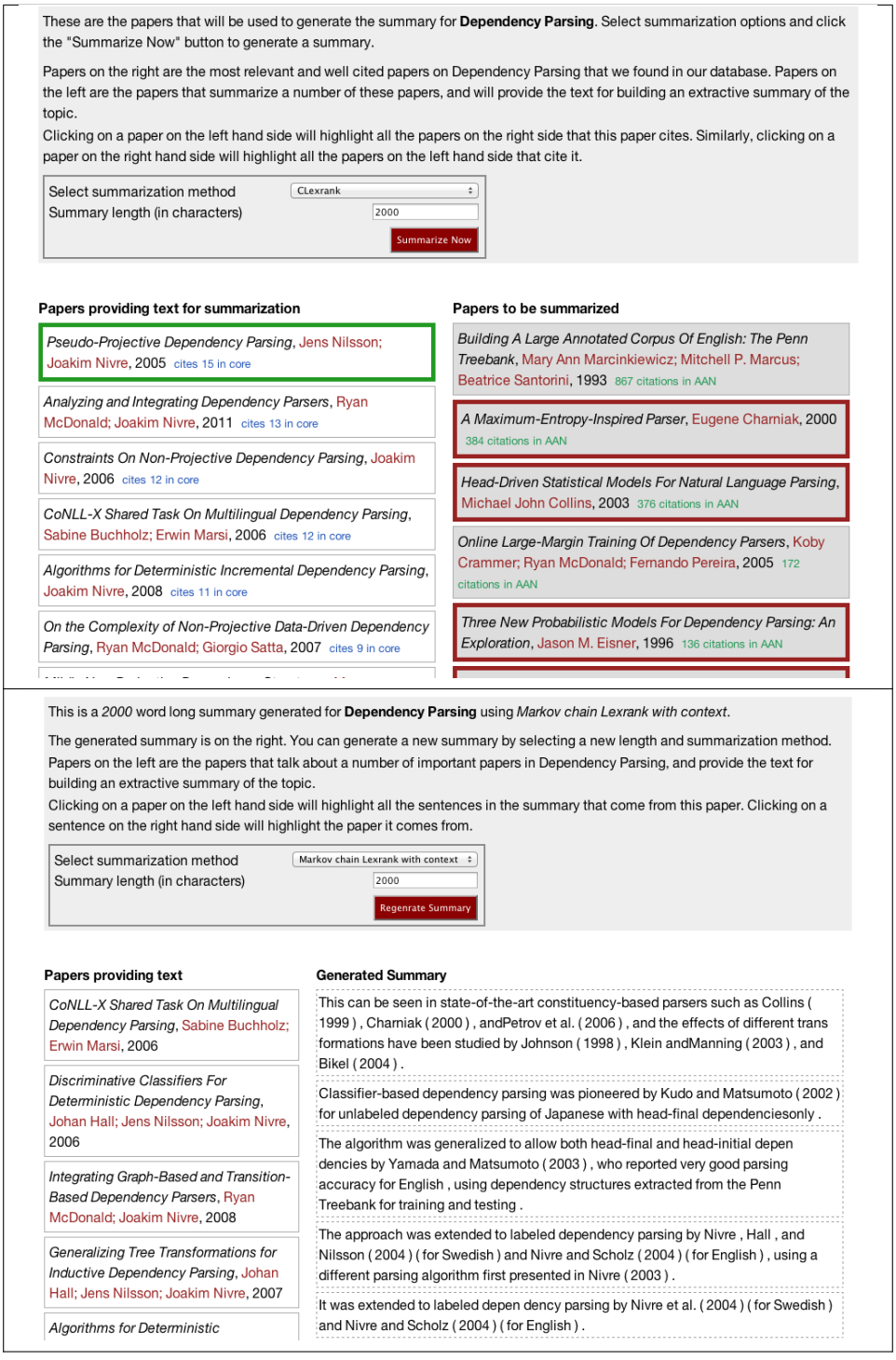

Figure 7.1: Screenshots of a prototype survey generation system that can be deployed on the web.

Arxiv Arxiv is a publicly available repository that is the main source of scholarly exchange for researchers working high energy physics. Two Arxiv datasets are available for experimentation. Data from KDD 2003 shared task contains metadata and citation network for papers in high energy physics till 2003. The citation network for all papers in Arxiv is also available from related-work.net[2].

DBLP This corpus contains bibliographic information for publications in computer science. Complete metadata (but without abstracts) and citation network for papers in DBLP is available through ArnetMiner[3].

7.2.2 User Testing

We ran several experiments to evaluate the effectiveness of our system. Arguably though, laboratory experiments done in a controlled setting can only go so far and a true test of a survey generation system like this is to be deployed in the real world where actual users can play with it. This was beyond the scope of this thesis, but we have made some progress in this direction. Figure 7.1 shows screenshots of a prototype survey generation system that we have built. The system allows users to find papers related to a NLP topic and builds a survey for the topic using the methods presented in this thesis.

With some engineering effort, the current algorithms could be used to build a complete system that can be deployed "in the wild". This will allow us to gather user data to understand how well the system is doing. Additionally, we can try to explore ways to record user interactions with the survey generation system and use them to build more evaluation data. For example, the system can provide users with the ability to edit the automatically generated summary and store it for future use. The edits done by users can be used as evaluation data for understanding what is

[2]http://blog.related-work.net/data/
[3]http://arnetminer.org/citation

missing in the current approach and to build evaluation data.

Despite the limitations, our experimental results are encouraging and indicate that an end-to-end system for automatically generating surveys of scientific articles is possible given the data available in scientific papers on these topics and the current state of NLP and machine learning research.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abu-Jbara, Amjad, & Radev, Dragomir. 2011. Coherent citation-based summarization of scientific papers. *In: Proceedings of the 49th Annual Conference of the Association for Computational Linguistics (ACL-11)*.

Abu-Jbara, Amjad, & Radev, Dragomir. 2012. Reference Scope Identification in Citing Sentences. *Pages 80–90 of: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL HLT '12. Stroudsburg, PA, USA: Association for Computational Linguistics.

Acuna, Daniel E., Allesina, Stefano, & Kording, Konrad P. 2012. Future impact: Predicting scientific success. *Nature*, **489**(7415), 201–202.

Afantenos, StergosD., Doura, Irene, Kapellou, Eleni, & Karkaletsis, Vangelis. 2004. Exploiting Cross-Document Relations for Multi-document Evolving Summarization. *Pages 410–419 of:* Vouros, George A., & Panayiotopoulos, Themistoklis (eds), *Methods and Applications of Artificial Intelligence*. Lecture Notes in Computer Science, vol. 3025. Springer Berlin Heidelberg.

Airoldi, Edoardo M., Erosheva, Elena A., Fienberg, Stephen E., Joutard, Cyrille, Love, Tanzy, & Shringarpure, Suyash. 2010. Reconceptualizing the classification of PNAS articles. *Proceedings of the National Academy of Sciences*, **107**(49), 20899–20904.

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. 2009. h-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, May.

Alonso i Alemany, Laura, & Fuentes Fort, Maria. 2003. Integrating Cohesion and Coherence for Automatic Summarization.

Amancio, D. R., Nunes, M. G. V., Oliveira, Jr., O. N., & Costa, L. d. F. 2010. Good practices for a literature survey are not followed by authors while preparing scientific manuscripts. *ArXiv e-prints*, **1005.3063**(May).

Anderson, Ashton, McFarland, Dan, & Jurafsky, Dan. 2012. Towards a Computational History of the ACL: 1980-2008. *Pages 13–21 of: Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics.

Angrosh, M. A., Cranefield, Stephen, & Stanger, Nigel. 2013. Conditional Random Field Based Sentence Context Identification: Enhancing Citation Services for the Research Community. *Pages 59–68 of: Proceedings of the First Australasian Web Conference - Volume 144*. AWC '13. Darlinghurst, Australia, Australia: Australian Computer Society, Inc.

Arbesman, Samuel, & Christakis, Nicholas A. 2011. Eurekometrics: Analyzing the Nature of Discovery. *PLoS Comput Biol*, **7**(6), e1002072.

Athar, Awais. 2011. Sentiment Analysis of Citations using Sentence Structure-Based Features. *Pages 81–87 of: Proceedings of the ACL 2011 Student Session*. Portland, OR, USA: Association for Computational Linguistics.

Athar, Awais, & Teufel, Simone. 2012a. Context-enhanced Citation Sentiment Detection. *Pages 597–601 of: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL HLT '12. Stroudsburg, PA, USA: Association for Computational Linguistics.

Athar, Awais, & Teufel, Simone. 2012b. Detection of Implicit Citations for Sentiment Detection. *Pages 18–26 of: Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*. Jeju Island, Korea: Association for Computational Linguistics.

Baldwin, Breck, & Morton, Thomas S. 1998. Dynamic Coreference-Based Summarization. *In: In Proceedings of the Third Conference on Empirical Methods in Natural Language Processing (EMNLP-3*.

Banea, Carmen, Hassan, Samer, Mohler, Michael, & Mihalcea, Rada. 2012. Unt: A supervised synergistic approach to semantic text similarity. *Pages 635–642 of: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Bartneck, Christoph, & Hu, Jun. 2009. Scientometric Analysis of the CHI Proceedings. *Pages 699–708 of: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. New York, NY, USA: ACM.

Barzilay, Regina, & Elhadad, Michael. 1997. Using Lexical Chains for Text Summarization. *Pages 10–17 of: In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*.

Barzilay, Regina, & Lapata, Mirella. 2005. Modeling Local Coherence: An Entity-based Approach. *Pages 141–148 of: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics.

Barzilay, Regina, & Lee, Lillian. 2004. Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization. *Pages 113–120 of:* Susan Dumais, Daniel Marcu, & Roukos, Salim (eds), *HLT-NAACL 2004: Main Proceedings*. Boston, Massachusetts, USA: Association for Computational Linguistics.

Barzilay, Regina, & McKeown, Kathleen R. 2005. Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, **31**(3), 297–328.

Barzilay, Regina, McKeown, Kathleen R., & Elhadad, Michael. 1999. Information Fusion in the Context of Multi-Document Summarization. *Pages 550–557 of: ACL'99*.

Barzilay, Regina, Elhadad, Noémie, & McKeown, Kathleen R. 2001a. Sentence Ordering in Multidocument Summarization.

Barzilay, Regina, Elhadad, Noemie, & McKeown, Kathleen R. 2001b. Sentence Ordering in Multidocument Summarization. *Pages 1–7 of: Proceedings of the First International Conference on Human Language Technology Research*. HLT '01. Stroudsburg, PA, USA: Association for Computational Linguistics.

Bateman, John A. 1993. *Using Text Structure and Text Planning to Guide Text Summarization*.

Baxendale, P. B. 1958. Machine-made Index for Technical Literature: An Experiment. *IBM J. Res. Dev.*, **2**(4), 354–361.

Beaugrande, Robert-Alain de, & Dressler, Wolfgang U. 1981. *Introduction to text linguistics*.

Berendt, B., Krause, B., & Kolbe-Nusser, S. 2010. Intelligent scientific authoring tools: Interactive data mining for constructive uses of citation networks. *Information Processing & Management*, **46**(1), 1–10.

Bergler, Sabine, Witte, René, Khalifé, Michelle, Li, Zhuoyan, & Rudzicz, Frank. 2003. Using Knowledge-Poor Coreference Resolution for Text Summarization.

Bergstrom, Carl T., West, Jevin D., & Wiseman, Marc A. 2008. The EigenfactorTM Metrics. *Journal of Neuroscience*, **28**(45), 11433–11434.

Bethard, Steven, & Jurafsky, Dan. 2010. Who Should I Cite: Learning Literature Search Models from Citation Behavior. *Pages 609–618 of: Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. CIKM '10. New York, NY, USA: ACM.

Birnholtz, Jeremy, Guha, Shion, Yuan, Y. Connie, Gay, Geri, & Heller, Caren. 2013. Cross-campus collaboration: A scientometric and network case study of publication activity across two campuses of a single institution. *Journal of the American Society for Information Science and Technology*, **64**(1), 162–172.

Bletsas, Aggelos, & Sahalos, John N. 2009. Hirsch index rankings require scaling and higher moment. *J. Am. Soc. Inf. Sci. Technol.*, **60**(12), 2577–2586.

Boguraev, Branlair, & Kennedy, Christopher. 1997. Salience-based Content Characterisation of Text Documents. *Pages 2–9 of: Advances in Automatic Text Summarization*. The MIT Press.

Bollegala, Danushka, Okazaki, Naoaki, & Ishizuka, Mitsuru. 2010. A bottom-up approach to sentence ordering for multi-document summarization. *Information Processing & Management*, **46**(1), 89 – 109.

Bonzi, Susan. 1982. Characteristics of a Literature as Predictors of Relatedness Between Cited and Citing Works. *Journal of the American Society for Information Science*, **33**(4), 208–216.

Bonzi, Susan, & Snyder, H. W. 1991. Motivations for citation: A comparison of self citation and citation to others. *Scientometrics*, **21**(2), 245–254.

Borner, Katy, Maru, Jeegar T., & Goldstone, Robert L. 2004. The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences*, **101**(suppl 1), 5266–5273.

Bornmann, Lutz, & Mutz, Ruediger. 2014. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *ArXiv e-prints*, **1402.4578**(Feb.).

Bosma, Wauter. 2005. Extending answers using discourse structure. *In: In RANLP Workshop on Crossing Barriers in Text Summarization Research*.

Bosma, Wauter E. 2008 (Mar.). *Discourse oriented summarization*. PhD Thesis, University of Twente. publisher: Centre for Telematics and Information Technology University of Twente, publisherlocation: Enschede, ISSN: 1381-3617, ISBN: 9789036526494, Numberofpages: 205.

Bradshaw, Shannon. 2003. Reference Directed Indexing: Redeeming Relevance for Subject Search in Citation Indexes. *In: Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries*.

Braun, Tibor, Glnzel, Wolfgang, & Schubert, Andrs. 2006. A Hirsch-type index for journals. *Scientometrics*, **69**(1), 169173.

Brembs, Bjorn, & Munafo, Marcus. 2013. Deep Impact: Unintended consequences of journal rank. *ArXiv e-prints*, **1301.3748**(Jan.).

Brody, Tim, Harnad, Stevan, & Carr, Leslie. 2006. Earlier Web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, **57**(8), 1060–1072.

Brunn, Meru, Chali, Yllias, & Pinchak, Christopher J. 2001. Text Summarization Using Lexical Chains.

Carbonell, Jaime G., & Goldstein, Jade. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. *Pages 335–336 of: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98).*

Carlson, L., Conroy, J. M., Marcu, D., O'Leary, D. P., Okurowski, M. E., Taylor, A., & Wong, W. 2001. An empirical study of the relation between abstracts, extracts, and the discourse structure of texts. *In: Proceedings of the DUC-2001 Workshop on Text Summarization.*

Castillo, Carlos, Donato, Debora, & Gionis, Aristides. 2007. Estimating number of citations using author reputation. *Page 107117 of: String processing and information retrieval.*

Chan, Samuel W. K., Lai, Tom B. Y., Gao, W. J., & Tsou, Benjamin K. 2000. Mining Discourse Markers for Chinese Textual Summarization.

Cheang, Brenda, Chu, Samuel Kai Wah, Li, Chongshou, & Lim, Andrew. 2014. A multidimensional approach to evaluating management journals: Refining pagerank via the differentiation of citation types and identifying the roles that management journals play. *Journal of the Association for Information Science and Technology*, **65**(12), 2581–2591.

Chen, Chaomei. 2012. Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology*, **63**(3), 431–449.

Christensen, Janara, Mausam, Soderland, Stephen, & Etzioni, Oren. 2013. Towards Coherent Multi-Document Summarization. *In: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2013).*

Chubin, Daryl E., & Moitra, Soumyo D. 1975. Content Analysis of References: Adjunct or Alternative to Citation Counting? *Social Studies of Science*, **5**(4), pp. 423–441.

Conroy, John, & O'leary, Dianne P. 2001. *Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition.* Tech. rept. In SIGIR.

Conroy, John M., Schlesinger, Judith D., O'leary, Dianne P., & Okurowski, Mary Ellen. 2001. Using HMM and Logistic Regression to Generate Extract Summaries. *Pages 13–14 of: In DUC 01 Conference Proceedings.*

Cordier, Stephane. 2012. A measure of similarity between scientific journals and of diversity of a list of publications. *ArXiv e-prints*, **1210.6510**(Oct.).

Cormode, Graham, Muthukrishnan, S., & Yan, Jinyun. 2012a. Scienceography: the study of how science is written. *ArXiv e-prints*, **1202.2638**(Feb.).

Cormode, Graham, Ma, Qiang, Muthukrishnan, S., & Thompson, Brian. 2012b. Socializing the h-index. *ArXiv e-prints*, **1211.7133**(Nov.).

Costas, Rodrigo, van Leeuwen, Thed N., & van Raan, Anthony F. J. 2009. Is scientific literature subject to a sell-by-date? A general methodology to analyze the durability of scientific documents. *ArXiv e-prints*, **0907.1455**(July).

Crookes, Graham. 1986. Towards a Validated Analysis of Scientific Text Structure. *Applied Linguistics*, **7**(1), 57–70.

Da Cunha, Iria, & Wanner, Leo. 2005. Towards the Automatic Summarization of Medical Articles in Spanish: Integration of textual, lexical, discursive and syntactic criteria. *Pages 46–51 of: Crossing Barriers in Text Summarization Research (RANLP-2005)*. INCOMA Ltd.

Daumé, III, Hal, & Marcu, Daniel. 2002. A noisy-channel model for document compression. *Pages 449–456 of: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics.

Daumé, III, Hal, & Marcu, Daniel. 2006. Bayesian Query-focused Summarization. *Pages 305–312 of: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. ACL-44. Stroudsburg, PA, USA: Association for Computational Linguistics.

Daume III, Hal, & Marcu, Daniel. 2004. Generic Sentence Fusion is an Ill-Defined Summarization Task. *Pages 96–103 of:* Marie-Francine Moens, Stan Szpakowicz (ed), *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Barcelona, Spain: Association for Computational Linguistics.

de Marneffe, Marie-Catherine, MacCartney, Bill, & Manning, Christopher D. 2006. Generating typed dependency parses from phrase structure parses. *Pages 449–454 of: The International Conference on Language Resources and Evaluation (LREC)*, vol. 6. Citeseer.

De Meo, Pasquale, Ferrara, Emilio, Fiumara, Giacomo, & Provetti, Alessandro. 2011. Generalized Louvain Method for Community Detection in Large Networks. *CoRR*, **abs/1108.1502**.

de Solla Price, Derek. 1951. Quantitative Measures of the Development of Science. *Archives Internationales d'Histoire des Sciences*, **4**, 85–93.

de Solla Price, Derek J. 1970. Citation measures of hard science, soft science, technology, and nonscience. *Pages 3–22 of:* Nelson, Carnot E., & Pollock, Donald K.

(eds), *Communication among scientists and engineers*. Lexington: Heath Lexington Book.

de Solla Price, Derek J. 1974. *Little Science, Big Science*.

Dehghani, Leila, Jahromi, Reza Basirian, & Ganjoo, Mazyar. 2011. Citations to highly-cited researchers by their co-authors and their self-citations: How these factors affect highly-cited researchers' h-index in Scopus. *Webology*, **8**(2).

Deutsch, Karl W., Platt, John, & Senghaas, Dieter. 1971. Conditions Favoring Major Advances in Social Science. *Science*, **171**(3970), 450–459.

Dietz, Laura, Bickel, Steffen, & Scheffer, Tobias. 2007. Unsupervised Prediction of Citation Influences. *Pages 233–240 of: Proceedings of the 24th International Conference on Machine Learning*. ICML '07. New York, NY, USA: ACM.

Ding, Ying, Yan, Erjia, Frazho, Arthur, & Caverlee, James. 2009. PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, **60**(11), 2229–2243.

Ding, Ying, Zhang, Guo, Chambers, Tamy, Song, Min, Wang, Xiaolong, & Zhai, Chengxiang. 2014. Content-based citation analysis: The next generation of citation analysis. *Journal of the Association for Information Science and Technology*, **65**(9), 1820–1833.

Doboli, Simona, Zhao, Fanshu, & Doboli, Alex. 2014. New measures for evaluating creativity in scientific publications. *ArXiv e-prints*, **1406.7582**(June).

Dong, Yuxiao, Johnson, Reid A., & Chawla, Nitesh V. 2014. Will This Paper Increase Your h-index? Scientific Impact Prediction. *ArXiv e-prints*, **1412.4754**(Dec.).

Doran, William, Stokes, Nicola, Carthy, Joe, & Dunnion, John. 2004. Assessing the Impact of Lexical Chain Scoring Methods and Sentence Extraction Schemes on Summarization. *Pages 627–635 of:* Gelbukh, Alexander (ed), *Computational Linguistics and Intelligent Text Processing*. Lecture Notes in Computer Science, vol. 2945. Springer Berlin Heidelberg.

Dunne, Cody, Shneiderman, Ben, Gove, Robert, Klavans, Judith, & Dorr, Bonnie. 2012. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, **63**(12), 2351–2369.

Edmundson, H. P. 1969. New Methods in Automatic Extracting. *J. ACM*, **16**(2), 264–285.

Egghe, Leo. 2006. Theory and practise of the g-index. *Scientometrics*, **69**, 131–152.

Egghe, Leo. 2014. A good normalized impact and concentration measure. *Journal of the Association for Information Science and Technology*, **65**(10), 2152–2154.

El-Arini, Khalid, & Guestrin, Carlos. 2011. Beyond Keyword Search: Discovering Relevant Scientific Literature. *Pages 439–447 of: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '11. New York, NY, USA: ACM.

Elhadad, Noemie, & McKeown, Kathleen R. 2001. Towards generating patient specific summaries of medical articles. *In: In Proceedings of NAACL-2001 Automatic Summarization Workshop*.

Elkiss, Aaron, Shen, Siwei, Fader, Anthony, Erkan, Güneş, States, David, & Radev, Dragomir R. 2008. Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, **59**(1), 51–62.

Erdi, Peter, Makovi, Kinga, Somogyvari, Zoltan, Strandburg, Katherine, Tobochnik, Jan, Volf, Peter, & Zalanyi, Lazlo. 2012. Prediction of Emerging Technologies Based on Analysis of the U.S. Patent Citation Network. *ArXiv e-prints*, **1206.3933**(June).

Erkan, Güneş, & Radev, Dragomir R. 2004. LexRank: Graph-based Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research (JAIR)*.

Erosheva, Elena, Fienberg, Stephen, & Lafferty, John. 2004. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, **101**(suppl 1), 5220–5227.

Eysenbach, Gunther. 2011. Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, **13**(4).

Farzindar, Atefeh, & Lapalme, Guy. 2004. Legal text summarization by exploration of the thematic structures and argumentative roles. *Pages 27–38 of: In Text Summarization Branches Out Conference held in conjunction with ACL 2004*.

Ferrara, Emilio, & Romero, Alfonso E. 2013. Scientific impact evaluation and the effect of self-citations: Mitigating the bias by discounting the h-index. *Journal of the American Society for Information Science and Technology*, **64**(11), 2332–2339.

Frantzi, Katerina T., Ananiadou, Sophia, & Tsujii, Jun-ichi. 1998. The C-value/NC-value Method of Automatic Recognition for Multi-Word Terms. *Pages 585–604 of: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*. ECDL '98. London, UK, UK: Springer-Verlag.

Fried, Daniel, & Kobourov, Stephen G. 2013. Maps of Computer Science. *ArXiv e-prints*, **1304.2681**(Apr.).

Frijters, Raoul, van Vugt, Marianne, Smeets, Ruben, van Schaik, Ren, de Vlieg, Jacob, & Alkema, Wynand. 2010. Literature Mining for the Discovery of Hidden Connections between Drugs, Genes and Diseases. *PLoS Comput Biol*, **6**(9), e1000943.

Fu, Lawrence D., & Aliferis, Constantin. 2008. Models for predicting and explaining citation count of biomedical articles. *AMIA Annu Symp Proc*, 222–226.

Fukuda, Satoshi, Nanba, Hidetsugu, & Takezawa, Toshiyuki. 2012. Extraction and Visualization of Technical Trend Information from Research Papers and Patents. *D-Lib Magazine*, **18**(7/8).

Fung, Pascale, & Ngai, Grace. 2006. One Story, One Flow: Hidden Markov Story Models for Multilingual Multidocument Summarization. *ACM Trans. Speech Lang. Process.*, **3**(2), 1–16.

Fung, Pascale, Ngai, Grace, & Cheung, Chi-Shun. 2003. Combining Optimal Clustering and Hidden Markov Models for Extractive Summarization. *Pages 21–28 of: Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*. Sapporo, Japan: Association for Computational Linguistics.

Garfield, Eugene. 1964. Can Citation Indexing Be Automated? *Statistical Assoc. Methods for Mechanized Documentation, Symposium Proceedings*.

Garfield, Eugene. 1972. Citation indexing-its theory and application in science, technology, and humanities. *Ph.D. thesis, University of California at Berkeley.Ph.D. thesis, University of California at Berkeley.*

Garfield, Eugene. 2006a. Citation indexes for science. A new dimension in documentation through association of ideas. *International Journal of Epidemiology*, **35**(5), 1123–1127.

Garfield, Eugene. 2006b. The history and meaning of the journal impact factor. *JAMA: the journal of the American Medical Association*, **295**(1), 9093.

Garfield, Eugene, Sher, Irving H., & Torpie, R. J. 1984. *The Use of Citation Data in Writing the History of Science*. Philadelphia, Pennsylvania, USA: Institute for Scientific Information Inc.

Gaudry, Pierrick. 2006. Secure H-numbers. *J. Craptology 3*.

Gehrke, Johannes, Ginsparg, Paul, & Kleinberg, Jon. 2003. Overview of the 2003 KDD Cup. *SIGKDD Explor. Newsl.*, **5**(2), 149–151.

Gerrish, Sean M., & Blei, David M. 2010. *A Language-based Approach to Measuring Scholarly Impact*.

Giles, C. Lee, & Councill, Isaac G. 2004. Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(51), 17599–17604.

Glänzel, Wolfgang, Schlemmer, Balázs, & Thijs, Bart. 2003. Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, **58**(3), 571–586.

Grosz, Barbara J., & Sidner, Candace L. 1986. Attention, intentions, and the structure of discourse. *Comput. Linguist.*, **12**(July), 175–204.

Grosz, Barbara J., Weinstein, Scott, & Joshi, Aravind K. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Comput. Linguist.*, **21**(2), 203–225.

Grover, Claire, Hachey, Ben, & Korycinski, Chris. 2003. *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*. Chap. Summarising Legal Texts: Sentential Tense and Argumentative Roles.

Guerini, Marco, Pepe, Alberto, & Lepri, Bruno. 2012. Do Linguistic Style and Readability of Scientific Abstracts affect their Virality? *CoRR*, **abs/1203.4238**.

Guha, Shion, Steinhardt, Stephanie, Ahmed, Syed Ishtiaque, & Lagoze, Carl. 2013. Following Bibliometric Footprints: The ACM Digital Library and the Evolution of Computer Science. *Pages 139–142 of: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*. JCDL '13. New York, NY, USA: ACM.

Gupta, Sonal, & Manning, Christopher D. 2011. Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers. *Pages 1–9 of: IJCNLP*.

Haghighi, Aria, & Vanderwende, Lucy. 2009. Exploring Content Models for Multi-document Summarization. *Pages 362–370 of: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. NAACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics.

Hahn, Udo. 1990. Topic parsing: Accounting for text macro structures in full-text analysis. *Information Processing & Management*, **26**(1), 135 – 170. Special Issue: Natural Language Processing and Information Retrieval.

Halevi, Gali, & Moed, Henk F. 2013. The thematic and conceptual flow of disciplinary research: A citation context analysis of the journal of informetrics , 2007. *Journal of the American Society for Information Science and Technology*, **64**(9), 1903–1913.

Hall, David, Jurafsky, Daniel, & Manning, Christopher D. 2008. Studying the history of ideas using topic models. *Page 363371 of: Proceedings of the conference on empirical methods in natural language processing*.

Halliday, Michael A.K., & Hasan, Ruqaiya. 1976. *Cohesion in English*. London: Longman.

Hamblin, Terence J. 1981. Fake. *British Medical Journal (Clinical research ed.)*, **283**, 1671–1674.

Havemann, Frank, & Larsen, Birger. 2014. Bibliometric Indicators of Young Authors in Astrophysics: Can Later Stars be Predicted? *ArXiv e-prints*, **1404.3084**(Apr.).

He, Qi, Pei, Jian, Kifer, Daniel, Mitra, Prasenjit, & Giles, Lee. 2010. Context-aware Citation Recommendation. *Pages 421–430 of: Proceedings of the 19th International Conference on World Wide Web*. WWW '10. New York, NY, USA: ACM.

He, Qi, Kifer, Daniel, Pei, Jian, Mitra, Prasenjit, & Giles, C. Lee. 2011. Citation Recommendation Without Author Supervision. *Pages 755–764 of: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11. New York, NY, USA: ACM.

Hearst, Marti A. 1994. Multi-paragraph Segmentation of Expository Text. *Pages 9–16 of: Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*. ACL '94. Stroudsburg, PA, USA: Association for Computational Linguistics.

Hirsch, Jorge E. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, **102**(46), 16569–16572.

Hirsch, Jorge E. 2007. Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, **104**(49), 1919319198.

Hirsch, Jorge E. 2010. An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, **85**(3), 741–754.

Hoang, Cong Duy Vu, & Kan, Min-Yen. 2010. Towards automated related work summarization. *Pages 427–435 of: Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. COLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics.

Horn, Daniel B., Finholt, Thomas A., Birnholtz, Jeremy P., Motwani, Dheeraj, & Jayaraman, Swapnaa. 2004. Six Degrees of Jonathan Grudin: A Social Network Analysis of the Evolution and Impact of CSCW Research. *Pages 582–591 of: Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*. CSCW '04. New York, NY, USA: ACM.

Hou, Wen-Ru, Li, Ming, & Niu, Deng-Ke. 2011. Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution: citation frequency of individual articles in other papers more fairly measures their scientific contribution than mere presence in refere. *BioEssays : news and reviews in molecular, cellular and developmental biology*, **33**(10), 724–7.

Hovy, Eduard. 1993. Automated Discourse Generation Using Discourse Structure Relations. **63**, 341–385.

Huang, Wenyi, Kataria, Saurabh, Caragea, Cornelia, Mitra, Prasenjit, Giles, C. Lee, & Rokach, Lior. 2012. Recommending Citations: Translating Papers into References. *Pages 1910–1914 of: Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. CIKM '12. New York, NY, USA: ACM.

Huang, Wenyi, Wu, Zhaohui, Liang, Chen, Mitra, Prasenjit, & Giles, C. Lee. 2015. A Neural Probabilistic Model for Context Based Citation Recommendation. *In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.*

Ioannidis, John P. A. 2005a. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, **294**(2), 218–228.

Ioannidis, John P. A. 2005b. Why Most Published Research Findings Are False. *PLoS Med*, **2**(8), e124.

Jha, Rahul, Abu-Jbara, Amjad, & Radev, Dragomir R. 2013. A System for Summarizing Scientific Topics Starting from Keywords. *In: Proceedings of The Association for Computational Linguistics (short paper).*

Jha, Rahul, Finegan-Dollak, Catherine, King, Ben, Coke, Reed, & Radev, Dragomir R. 2015a. Content Models for Survey Generation: A Factoid-Based Evaluation. *Page to appear of: Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Jha, Rahul, Coke, Reed, & Radev, Dragomir R. 2015b. Surveyor: A System for Generating Coherent Survey Articles for Scientific Topics. *In: Proceedings of the Twenty-Ninth AAAI Conference.*

Jiang, Jay J., & Conrath, David W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *Pages 19–33 of: Proc. of the Int'l. Conf. on Research in Computational Linguistics.*

Jiang, Jing, & Zhai, ChengXiang. 2005. Accurately Extracting Coherent Relevant Passages Using Hidden Markov Models.

Jing, Hongyan. 2002. Using Hidden Markov Modeling to Decompose Human-written Summaries. *Comput. Linguist.*, **28**(4), 527–543.

Jing, Hongyan, & McKeown, Kathleen R. 2000. Cut and Paste Based Text Summarization. *Pages 178–185 of: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference.* NAACL 2000. Stroudsburg, PA, USA: Association for Computational Linguistics.

Kan, Min-Yen, Klavans, Judith L., & McKeown, Kathleen R. 2002. Using the Annotated Bibliography as a Resource for Indicative Summarization. *In: The International Conference on Language Resources and Evaluation (LREC).*

Kaplan, Dain, Iida, Ryu, & Tokunaga, Takenobu. 2009 (August). Automatic Extraction of Citation Contexts for Research Paper Summarization: A Coreference-chain based Approach. *Pages 88–95 of: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries.*

Karamuftuoglu, Murat. 2002. An Approach to Summarization Based on Lexical Bonds.

King, Ben, Jha, Rahul, & Radev, Dragomir R. 2014. Heterogeneous Networks and Their Applications: Scientometrics, Name Disambiguation, and Topic Modeling. *Transactions of the Association for Computational Linguistics*, **2**, 1–14.

Kleinberg, Jon M. 1999. Authoritative sources in a hyperlinked environment. *J. ACM*, **46**(September), 604–632.

Klosik, David F., & Bornholdt, Stefan. 2013. The citation wake of publications detects Nobel laureates' papers. *ArXiv e-prints*, **1301.7471**(Jan.).

Knight, Kevin, & Marcu, Daniel. 2000. Statistics-Based Summarization - Step One: Sentence Compression. *Pages 703–710 of: Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. AAAI Press.

Knight, Kevin, & Marcu, Daniel. 2002. Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression. *Artif. Intell.*, **139**(1), 91–107.

Kostoff, Ronald N., del Rio, J. Antonio, Humenik, James A., Garcia, Esther Ofilia, & Ramirez, Ana Maria. 2001. Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology*, **52**(13), 1148–1156.

Kraaij, Wessel, Spitters, Martijn, & van der Heijden, Martine. 2001. Combining a mixture language model and Naive Bayes for multi-document summarisation. *In: Proceedings of the DUC2001 workshop (SIGIR2001), New Orleans*.

Küçüktunç, Onur, Saule, Erik, Kaya, Kamer, & Çatalyürek, Ümit V. 2014. Diversifying Citation Recommendations. *ACM Trans. Intell. Syst. Technol.*, **5**(4), 55:1–55:21.

Kuhn, Michael, Campillos, Mnica, Gonzlez, Paula, Jensen, Lars Juhl, & Bork, Peer. 2008. Large-scale prediction of drugtarget relationships. {*FEBS*} *Letters*, **582**(8), 1283 – 1290. (1) The Digital, Democratic Age of Scientific Abstracts (2) Targeting and Tinkering with Interaction Networks.

Kuhn, Thomas S. 1970. *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Kuhn, Tobias, Perc, Matjaz, & Helbing, Dirk. 2014. Inheritance Patterns in Citation Networks Reveal Scientific Memes. *Physical Review X*, **4**(4), 041036.

Kupiec, Julian, Pedersen, Jan, & Chen, Francine. 1995. A trainable document summarizer. *Pages 68–73 of: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-95)*.

Lapata, Mirella. 2003. Probabilistic Text Structuring: Experiments with Sentence Ordering. *Pages 545–552 of: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Sapporo, Japan: Association for Computational Linguistics.

Lapata, Mirella. 2005. Automatic evaluation of text coherence: models and representations. *Pages 1085–1090 of: In the Intl. Joint Conferences on Artificial Intelligence*.

Leacock, Claudia, & Chodorow, Martin. 1998. Combining local context and WordNet similarity for word sense identification. *Pages 265–283 of:* Fellfaum, Christiane (ed), *MIT Press*.

Leeuwen, Theo V. 2004. *Introducing Social Semiotics: An Introductory Textbook*.

Lehmam, Abderrafih. 1999. Text Structuration Leading to an Automatic Summary System: RAFI. *Inf. Process. Manage.*, **35**(2), 181–191.

Lehman, Harvey C. 1947. The Exponential Increase of Man's Cultural Output. *Social Forces*, **25**(3), pp. 281–290.

Leydesdorff, Loet, & Wouters, Paul. 1999. Between texts and contexts: Advances in theories of citation? (A rejoinder). *Scientometrics*, **44**(2), 169–182.

Li, Rui, Chambers, Tamy, Ding, Ying, Zhang, Guo, & Meng, Liansheng. 2014. Patent citation analysis: Calculating science linkage based on citing motivation. *Journal of the Association for Information Science and Technology*, **65**(5), 1007–1017.

Liddy, Elizabeth DuRoss. 1991. Discourse-level Structure of Empirical Abstracts: An Exploratory Study. *Information Processing and Management*, **27**(1), 550–81.

Liddy, Elizabeth DuRoss, Bonzi, Susan, Katzer, Jeffrey, & Oddy, E. 1987. A Study of Discourse Anaphora in Scientific Abstracts. **38**(4), 255–261.

Lin, Chin-Yew. 2004a. Looking for a Few Good Metrics: Automatic Summarization Evaluation - How Many Samples are Enough? *In: Proceedings of the NTCIR Workshop 4*.

Lin, Chin-Yew. 2004b. ROUGE: A Package for Automatic Evaluation of summaries. *In: Proceedings of the ACL workshop on Text Summarization Branches Out*.

Lin, Chin-Yew, & Hovy, Eduard. 2000. The Automated Acquisition of Topic Signatures for Text Summarization. *Pages 495–501 of: Proceedings of the 18th Conference on Computational Linguistics - Volume 1*. COLING '00. Stroudsburg, PA, USA: Association for Computational Linguistics.

Lin, Dekang. 1998. An Information-Theoretic Definition of Similarity. *Pages 296–304 of: Proceedings of the Fifteenth International Conference on Machine Learning*. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Liu, John S., Chen, Hsiao-Hui, Ho, Mei Hsiu-Ching, & Li, Yu-Chen. 2014a. Citations with different levels of relevancy: Tracing the main paths of legal opinions. *Journal of the Association for Information Science and Technology*, **65**(12), 2479–2488.

Liu, Shengbo, Chen, Chaomei, Ding, Kun, Wang, Bo, Xu, Kan, & Lin, Yuan. 2014b. Literature retrieval based on citation context. *Scientometrics*, **101**(2), 1293–1307.

Liu, Yuxian, & Rousseau, Ronald. 2014. Citation analysis and the development of science: A case study using articles by some Nobel prize winners. *Journal of the Association for Information Science and Technology*, **65**(2), 281–289.

Louis, Annie, Joshi, Aravind, & Nenkova, Ani. 2010. Discourse Indicators for Content Selection in Summarization. *Pages 147–156 of: Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. SIGDIAL '10. Stroudsburg, PA, USA: Association for Computational Linguistics.

Luhn, H. P. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, **2**(2), 159–165.

Luong, Minh-thang, Nguyen, Thuy D., & Kan, Min-yen. 2010. Logical Structure Recovery in Scholarly Articles with Rich Document Features. *IJDLS*.

MacRoberts, Michael H., & MacRoberts, Barbara R. 1984. The Negational Reference: Or the Art of Dissembling. *Social Studies of Science*, **14**(1), pp. 91–94.

Mani, Inderjeet, & Bloedorn, Eric. 1997. Multi-document Summarization by Graph Search and Matching. *Pages 622–628 of: Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*. AAAI'97/IAAI'97. AAAI Press.

Mani, Inderjeet, Bloedorn, Eric, & Gates, Barbara. 1998. Using Cohesion and Coherence Models for Text Summarization.

Mani, Inderjeet, Gates, Barbara, & Bloedorn, Eric. 1999. Improving Summaries by Revising Them. *Pages 558–565 of: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. ACL '99. Stroudsburg, PA, USA: Association for Computational Linguistics.

Manjunatha, J. N., Sivaramakrishnan, K. R., Pandey, Raghavendra Kumar, & Murthy, M Narasimha. 2003. Citation prediction using time series approach KDD Cup 2003 (task 1). *SIGKDD Explor. Newsl.*, **5**(2), 152–153.

Mann, William C., & Thompson, Sandra A. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, **8**(3), 243–281.

Marcu, Daniel. 1995. Discourse Trees Are Good Indicators of Importance in Text. *Pages 123–136 of:* Mani, Inderjeet, & Maybury, Mark T. (eds), *Advances in Automatic Text Summarization*.

Marcu, Daniel. 1997. From Discourse Structures to Text Summaries.

Marcu, Daniel. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA, USA: MIT Press.

Marcu, Daniel. 2001. Discourse-based summarization in DUC-2001. *In: In Workshop on text summarization (DUC.*

Martin, T., Ball, B., Karrer, B., & Newman, Mark E. J. 2013. Coauthorship and citation in scientific publishing. *ArXiv e-prints*, **1304.0473**(Apr.).

McKeown, Kathleen R., Klavans, Judith L., Hatzivassiloglou, Vasileios, Barzilay, Regina, & Eskin, Eleazar. 1999. Towards Multidocument Summarization by Reformulation: Progress and Prospects. *Pages 453–460 of: Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence.* AAAI '99/IAAI '99. Menlo Park, CA, USA: American Association for Artificial Intelligence.

Mei, Qiaozhu, & Zhai, ChengXiang. 2008. Generating Impact-Based Summaries for Scientific Literature. *Pages 816–824 of: Proceedings of the 46th Annual Conference of the Association for Computational Linguistics (ACL-08).*

Midorikawa, N. 1983. Citation analysis of physics journals: Comparison of subfields of physics. *Scientometrics*, **5**, 361–374.

Mihalcea, Rada, & Tarau, Paul. 2004 (July). TextRank: Bringing Order into Texts. *In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04).*

Mihalcea, Rada, & Welch, Charles. 2015. What Women Want: Analyzing Research Publications to Understand Gender Preferences in Computer Science.

Milard, Béatrice. 2014. The social circles behind scientific references: Relationships between citing and cited authors in chemistry publications. *Journal of the Association for Information Science and Technology*, **65**(12), 2459–2468.

Miller, Tristan. 2003. Latent Semantic Analysis and the Construction of Coherent Extracts. *Pages 270–277 of: Proceedings of the International Conference RANLP-2003 (Recent Advances in Natural Language Processing.* John Benjamins, Amsterdam/Philadelphia.

Mimno, David, & McCallum, Andrew. 2007. Mining a digital library for influential authors. *Pages 105–106 of: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries.* JCDL '07. New York, NY, USA: ACM.

Mohammad, Saif, Dorr, Bonnie, Egan, Melissa, Hassan, Ahmed, Muthukrishan, Pradeep, Qazvinian, Vahed, Radev, Dragomir, & Zajic, David. 2009. Using citations to generate surveys of scientific paradigms. *Pages 584–592 of: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* NAACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics.

Moonesinghe, Ramal, Khoury, Muin J, & Janssens, A. Cecile J. W. 2007. Most Published Research Findings Are FalseBut a Little Replication Goes a Long Way. *PLoS Med*, **4**(2), e28.

Moravcsik, Michael J., & Murugesan, Poovanalingam. 1975. Some results on the function and quality of citations. *Social Studies of Science*, **5**, 86–92.

Morris, Jane, & Hirst, Graeme. 1991. Lexical Cohesion Computed by Thesaural Relations As an Indicator of the Structure of Text. *Comput. Linguist.*, **17**(1), 21–48.

Mryglod, O., Kenna, R., Holovatch, Y., & Berche, B. 2012. Absolute and specific measures of research group excellence. *ArXiv e-prints*, **1210.0732**(Oct.).

Nakov, Preslav I., Schwartz, Ariel S., & Hearst, Marti A. 2004. Citances: Citation Sentences for Semantic Analysis of Bioscience Text. *In: In Proceedings of the SIGIR04 workshop on Search and Discovery in Bioinformatics*.

Nanba, Hidetsugu, & Okumura, Manabu. 1999. Towards Multi-paper Summarization Using Reference Information. *Pages 926–931 of: IJCAI '99: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Nanba, Hidetsugu, & Okumura, Manabu. 2000. Producing More Readable Extracts by Revising Them. *Pages 1071–1075 of: Proceedings of the 18th Conference on Computational Linguistics - Volume 2*. COLING '00. Stroudsburg, PA, USA: Association for Computational Linguistics.

Nanba, Hidetsugu, Kando, Noriko, & Okumura, Manabu. 2004a. Classification of Research Papers using Citation Links and Citation Types: Towards Automatic Review Article Generation. *Pages 117–134 of: Proceedings of the 11th SIG Classification Research Workshop*.

Nanba, Hidetsugu, Kando, Noriko, & Okumura, Manabu. 2004b. Classification of Research Papers using Citation Links and Citation Types: Towards Automatic Review Article Generation. *Pages 117–134 of: Proceedings of the 11th SIG Classification Research Workshop*.

Nascimento, Cristiano, Laender, Alberto H.F., da Silva, Altigran S., & Gonçalves, Marcos André. 2011. A Source Independent Framework for Research Paper Recommendation. *Pages 297–306 of: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. JCDL '11. New York, NY, USA: ACM.

Nenkova, Ani. 2005. Discourse Factors in Multi-document Summarization. *Pages 1654–1655 of: Proceedings of the 20th National Conference on Artificial Intelligence - Volume 4*. AAAI'05. AAAI Press.

Nenkova, Ani, & McKeown, Kathleen. 2011. Automatic Summarization. *Foundations and Trends in Information Retrieval*, **5**(2-3), 103–233.

Nenkova, Ani, & Passonneau, Rebecca. 2004. Evaluating content selection in summarization: The pyramid method. *In: Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL '04).*

Newman, Mark. 2010. *Networks: An Introduction.* New York, NY, USA: Oxford University Press, Inc.

Ono, Kenji, Sumita, Kazuo, & Miike, Seiji. 1994. Abstract Generation Based on Rhetorical Structure Extraction. *Pages 344–348 of: Proceedings of the 15th Conference on Computational Linguistics - Volume 1.* COLING '94. Stroudsburg, PA, USA: Association for Computational Linguistics.

Otterbacher, Jahna, Erkan, Gunes, & Radev, Dragomir R. 2009. Biased LexRank: Passage Retrieval Using Random Walks with Question-based Priors. *Inf. Process. Manage.*, **45**(1), 42–54.

Otterbacher, Jahna C., Radev, Dragomir R., & Luo, Airong. 2002. Revisions That Improve Cohesion in Multi-document Summaries: A Preliminary Study. *Pages 27–36 of: Proceedings of the ACL-02 Workshop on Automatic Summarization - Volume 4.* AS '02. Stroudsburg, PA, USA: Association for Computational Linguistics.

Paice, Chris D., & Jones, P. A. 1993. The Identification of Important Concepts in Highly Structured Technical Papers. *Pages 69–78 of:* Korfhage, R., Rasmussen, E., & Willett, P. (eds), *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93).*

Pan, Raj Kumar, Kaski, Kimmo, & Fortunato, Santo. 2012 (Sep). *World citation and collaboration networks: uncovering the role of geography in science.* Tech. rept. arXiv:1209.0781. Comments: 8 pages, 5 figures + Appendix, The world citation and collaboration networks at both city and country level are available at http://becs.aalto.fi/rajkp/datasets.html.

Papineni, Kishore, Roukos, Salim, Ward, Todd, & Zhu, Wei-Jing. 2002. BLEU: a method for automatic evaluation of machine translation. *Pages 311–318 of: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics.

Pennebaker, James W., Francis, Martha E., & Booth, Roger J. 2001. *Linguistic Inquiry and Word Count.* Mahwah, NJ: Lawerence Erlbaum Associates.

Penner, Orion, Pan, Raj K., Petersen, Alexander M., & Fortunato, Santo. 2013. The case for caution in predicting scientists' future impact. *ArXiv e-prints*, **1304.0627**(Apr.).

Pepe, Alberto, & Kurtz, Michael J. 2012. A measure of total research impact independent of time and discipline.

Petersen, Alexander M., Fortunato, Santo, Pan, Raj K., Kaski, Kimmo, Penner, Orion, Riccaboni, Massimo, Stanley, H. Eugene, & Pammolli, Fabio. 2013. Reputation and Impact in Academic Careers. *ArXiv e-prints*, **1303.7274**(Mar.).

Peterson, George J., Press, Steve, & Dill, Ken A. 2010. Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences*, **107**(37), 16023–16027.

Pia, M. G., Basaglia, T., Bell, Z. W., & Dressendorfer, P. V. 2012. Publication patterns in HEP computing. *Journal of Physics Conference Series*, **396**(6), 062015.

Porter, Alan L., & Rafols, Ismael. 2009. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, **81**(3), 719–745.

Prasad, Rashmi, Dinesh, Nikhil, Lee, Alan, Miltsakaki, Eleni, Robaldo, Livio, Joshi, Aravind, & Webber, Bonnie. 2008. The Penn Discourse TreeBank 2.0. *In: In Proceedings of LREC*.

Qazvinian, Vahed. 2012. *Using Collective Discourse to Generate Surveys of Scientific Paradigms*. Ph.D. thesis.

Qazvinian, Vahed, & Radev, Dragomir R. 2008a. Scientific Paper Summarization Using Citation Summary Networks. *In: Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*.

Qazvinian, Vahed, & Radev, Dragomir R. 2008b. Scientific Paper Summarization Using Citation Summary Networks. *Pages 689–696 of: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK: Coling 2008 Organizing Committee.

Qazvinian, Vahed, & Radev, Dragomir R. 2010a. Identifying Non-Explicit Citing Sentences for Citation-Based Summarization. *Pages 555–564 of: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics.

Qazvinian, Vahed, & Radev, Dragomir R. 2010b (July). Identifying Non-Explicit Citing Sentences for Citation-Based Summarization. *Pages 555–564 of: Proceedings of the 48th Annual Conference of the Association for Computational Linguistics (ACL-10)*.

Radev, Dragomir, Allison, Timothy, Blair-Goldensohn, Sasha, Blitzer, John, Çelebi, Arda, Dimitrov, Stanko, Drabek, Elliott, Hakim, Ali, Lam, Wai, Liu, Danyu, Otterbacher, Jahna, Qi, Hong, Saggion, Horacio, Teufel, Simone, Topper, Michael, Winkel, Adam, & Zhang, Zhu. 2004a (May). MEAD - a platform for multidocument multilingual text summarization. *In: LREC 2004*.

Radev, Dragomir R. 2000. A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-document Structure. *Pages 74–83 of: Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue - Volume 10*. SIGDIAL '00. Stroudsburg, PA, USA: Association for Computational Linguistics.

Radev, Dragomir R., & Tam, Daniel. 2003. Summarization Evaluation Using Relative Utility. *Pages 508–511 of: CIKM2003*.

Radev, Dragomir R., Jing, Hongyan, Stys, Malgorzata, & Tam, Daniel. 2004b. Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, **40**(6), 919–938.

Radev, Dragomir R., Muthukrishnan, Pradeep, Qazvinian, Vahed, & Abu-Jbara, Amjad. 2013. The ACL anthology network corpus. *Language Resources and Evaluation*, 1–26.

Radicchi, Filippo. 2012. In science "there is no bad publicity": Papers criticized in technical comments have high scientific impact. *CoRR*, **abs/1209.4997**.

Rafols, Ismael, & Meyer, Martin. 2009. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, **82**(2), 263–287.

Redner, S. 2004. Citation Statistics From More Than a Century of Physical Review. *ArXiv Physics e-prints*, **physics/0407137**(July).

Resnik, Philip. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Pages 448–453 of: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Sarigöl, Emre, Pfitzner, Rene, Scholtes, Ingo, Garas, Antonios, & Schweitzer, Frank. 2014. Predicting Scientific Success Based on Coauthorship Networks. *ArXiv e-prints*, **1402.7268**(Feb.).

Schreiber, Michael. 2013a. How relevant is the predictive power of the h-index? A case study of the time-dependent Hirsch index. *ArXiv e-prints*, **1301.2060**(Jan.).

Schreiber, Michael. 2013b. Inconsistencies of the Highly-Cited-Publications Indicator. *ArXiv e-prints*, **1302.6391**(Feb.).

Serpa, F. G., Graves, Adam M., & Javier, Artjay. 2012. Statistical Common Author Networks (SCAN). *CoRR*, **abs/1208.3101**.

Shahaf, Dafna, Guestrin, Carlos, & Horvitz, Eric. 2012. Metro Maps of Science. *Pages 1122–1130 of: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '12. New York, NY, USA: ACM.

Shi, Xiaolin, Leskovec, Jure, & McFarland, Daniel A. 2010. Citing for high impact. *Pages 49–58 of: Proceedings of the 10th annual joint conference on Digital libraries*. JCDL '10. New York, NY, USA: ACM.

Sidiropoulos, Antonis, Katsaros, Dimitrios, & Manolopoulos, Yannis. 2014. Identification of Influential Scientists vs. Mass Producers by the Perfectionism Index. *ArXiv e-prints*, **1409.6099**(Sept.).

Silber, H. Grogory, & McCoy, Kathleen F. 2002. Efficiently Computed Lexical Chains As an Intermediate Representation for Automatic Text Summarization. *Comput. Linguist.*, **28**(4), 487–496.

Sim, Yanchuan, Smith, Noah A., & Smith, David A. 2012. Discovering Factions in the Computational Linguistics Community. *Pages 22–32 of: Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics.

Spiegel-Rösing, Ina. 1977. Science Studies: Bibliometric and Content Analysis. *Social Studies of Science*, **7**(1), 97–113.

Stede, Manfred. 2012. *Discourse processing*. Synthesis lectures on human language technologies. San Rafael, Calif.: Morgan & Claypool Publishers. Part of: Synthesis digital library of engineering and computer science.

Stirling, Andy. 2007. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, **4**(15), 707–719.

Strohman, Trevor, Croft, W. Bruce, & Jensen, David. 2007. Recommending Citations for Academic Papers. *Pages 705–706 of: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '07. New York, NY, USA: ACM.

Sugiyama, Kazunari, & Kan, Min-Yen. 2013. Exploiting Potential Citation Papers in Scholarly Paper Recommendation. *Pages 153–162 of: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*. JCDL '13. New York, NY, USA: ACM.

Sun, Xiaoling, Kaur, Jasleen, Milojevic, Stasa, Flammini, Alessandro, & Menczer, Filippo. 2012. Social Dynamics of Science. *ArXiv e-prints*, **1209.4950**(Sept.).

Swales, John. 1981. *Aspects of Article Introductions*. Aston ESP research reports. Language Studies Unit, University of Aston in Birmingham.

Swanson, Don R. 1986. Undiscovered Public Knowledge. *The Library Quarterly: Information, Community, Policy*, **56**(2), pp. 103–118.

Swanson, Don R. 1990. Medical literature as a potential source of new knowledge. *Bull Med Libr Assoc*, **78**(1), 29–37.

Swanson, Don R., & Smalheiserf, Neil. 1996. Undiscovered Public Knowledge: a Ten-Year Update. *KDD-96 Proceedings.*

Szanto-Varnagy, Adam, Pollner, Peter, Vicsek, Tamas, & Farkas, Illes J. 2014. Scientometrics: Untangling the topics. *ArXiv e-prints*, **1403.2140**(Mar.).

Tan, Chenhao, & Lee, Lillian. 2014. A Corpus of Sentence-level Revisions in Academic Writing: A Step towards Understanding Statement Strength in Communication. *ArXiv e-prints*, **1405.1439**(May).

Tang, Rong. 2008. "Citation characteristics and intellectual acceptance of scholarly monographs". *College and Research Libraries*, **69**, 356–369.

Tatsioni, Athina, Bonitsis, Nikolaos G., & Ioannidis, John P.A. 2007. Persistence of contradicted claims in the literature. *JAMA*, **298**(21), 2517–2526.

Teufel, Simone. 1999. *Argumentative Zoning: Information Extraction from Scientific Text.* Tech. rept.

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

Teufel, Simone, Siddharthan, Advaith, & Tidhar, Dan. 2006a. Automatic classification of citation function. *In: In Proc. of EMNLP-06.*

Teufel, Simone, Siddharthan, Advaith, & Tidhar, Dan. 2006b (July). Automatic classification of citation function. *Pages 103–110 of: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-06).*

Thelwall, Mike, Haustein, Stefanie, Larivire, Vincent, & Sugimoto, Cassidy R. 2013. Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLoS ONE*, **8**(5), e64841.

Thompson, Geoff, & Yiyun, Ye. 1991. Evaluation in the Reporting Verbs Used in Academic Papers. *Applied Linguistics*, **12**(4), 365–382.

Tiantian, Zhu, & Lan, Man. 2013. ECNUCS: Measuring Short Text Semantic Equivalence Using Multiple Similarity Measurements. *Pages 124–131 of: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity.* Association for Computational Linguistics.

Van Dijk, T.A. 1979. Recalling and summarizing complex discourse. *In: W. Burghardt and K. Holker, (Eds.) Textverarbeitung/ Text Processing.*

Van Raan, AnthonyF. 2004. Sleeping Beauties in science. *Scientometrics*, **59**(3), 467–472.

Van Zyl, J. Martin, & Van Der Merwe, Sean. 2012. An empirical study to order citation statistics between subject fields. *ArXiv e-prints*, **1210.2246**(Oct.).

Velden, Theresa, Haque, Asif-ul, & Lagoze, Carl. 2010. A new approach to analyzing patterns of collaboration in co-authorship networks: mesoscopic analysis and interpretation. *Scientometrics*, **85**(1), 219–242.

Viana, Matheus P., Amancio, Diego R., & Costa, Luciano da F. 2013. On time-varying collaboration networks. *ArXiv e-prints*, **1302.4092**(Feb.).

Waltman, Ludo, van Eck, Nees Jan, & Wouters, Paul. 2013. Counting publications and citations: Is more always better? *ArXiv e-prints*, **1301.4597**(Jan.).

Wan, Xiaojun, & Liu, Fang. 2014a. Are all literature citations equally important? Automatic citation strength estimation and its applications. *Journal of the Association for Information Science and Technology*, **65**(9), 1929–1938.

Wan, Xiaojun, & Liu, Fang. 2014b. WL-index: Leveraging citation mention number to quantify an individual's scientific impact. *Journal of the Association for Information Science and Technology*, **65**(12), 2509–2517.

Wang, Dashun, Song, Chaoming, & Barabási, Albert-László. 2013. Quantifying Long-Term Scientific Impact. *Science*, **342**(6154), 127–132.

Wang, Senzhang, Xie, Sihong, Zhang, Xiaoming, Li, Zhoujun, Yu, Philip S., & Shu, Xinyu. 2014. Future Influence Ranking of Scientific Literature. *ArXiv e-prints*, **1407.1772**(July).

Wei, Tian, Li, Menghui, Wu, Chensheng, Yan, XiaoYong, Fan, Ying, Di, Zengru, & Wu, Jinshan. 2013. Do scientists trace hot topics? *ArXiv e-prints*, **1303.5596**(Mar.).

Weinstock, Melvin. 1971. *Citation Indexes*. Vol. 5. Kent, A. (ed.), Encyclopedia of Library and Information Science.

West, Jevin D., Jacquet, Jennifer, King, Molly M., Correll, Shelley J., & Bergstrom, Carl T. 2012. The role of gender in scholarly authorship. *ArXiv e-prints*, **1211.1759**(Nov.).

White, Howard D. 2004. Citation Analysis and Discourse Analysis Revisited. *Applied Linguistics*, **25**(1), 89–116.

Winiarczyk, Ryszard, Gawron, Piotr, Miszczak, Jaroslaw A., Pawela, Lukasz, & Puchala, Zbigniew. 2012. Analysis of patent activity in the field of quantum information processing. *ArXiv e-prints*, **1212.2439**(Dec.).

Wolf, Florian, & Gibson, Edward. 2005. Representing Discourse Coherence: A Corpus-Based Study. *Comput. Linguist.*, **31**(2), 249–288.

Wu, Zhibiao, & Palmer, Martha. 1994. Verbs Semantics and Lexical Selection. *Pages 133–138 of: Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*. ACL '94. Stroudsburg, PA, USA: Association for Computational Linguistics.

Yan, Rui, Tang, Jie, Liu, Xiaobing, Shan, Dongdong, & Li, Xiaoming. 2011. Citation count prediction: learning to estimate future citations for literature. *Pages 1247–1252 of: Proceedings of the 20th ACM international conference on Information and knowledge management*. CIKM '11. New York, NY, USA: ACM.

Yan, Rui, Huang, Congrui, Tang, Jie, Zhang, Yan, & Li, Xiaoming. 2012. To better stand on the shoulder of giants. *Pages 51–60 of: Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*. JCDL '12. New York, NY, USA: ACM.

Yin, Xiaoshi, Huang, Jimmy Xiangji, & Li, Zhoujun. 2011. Mining and modeling linkage information from citation context for improving biomedical literature retrieval. *Information Processing & Management*, **47**(1), 53–67.

Yogatama, Dani, Heilman, Michael, O'Connor, Brendan, Dyer, Chris, Routledge, Bryan R., & Smith, Noah A. 2011. Predicting a scientific community's response to an article. *Pages 594–604 of: Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics.

Yun, Jinhyuk, Kim, Pan-Jun, & Jeong, Hawoong. 2014. Anatomy of Scientific Evolution. *ArXiv e-prints*, **1405.0917**(May).

Zhang, Chun-Ting. 2009. The e-Index, Complementing the h-Index for Excess Citations. *PLoS ONE*, **4**(5), e5429+.

Zhang, Han, Fiszman, Marcelo, Shin, Dongwook, Wilkowski, Bartlomiej, & Rindflesch, Thomas C. 2013. Clustering cliques for graph-based summarization of the biomedical research literature. *BMC Bioinformatics*, **14**, 182.

Zhao, Dangzhi, & Strotmann, Andreas. 2014. In-text author citation analysis: Feasibility, benefits, and limitations. *Journal of the Association for Information Science and Technology*, **65**(11), 2348–2358.

Zhu, Xiaodan, Turney, Peter, Lemire, Daniel, & Vellino, Andr. 2015. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, **66**(2), 408–427.

Ziman, John M. 1968. *Public knowledge: An essay concerning the social dimension of science*. London: Cambridge U.P.

Zuckerman, Harriet. 1967. Nobel Laureates in Science: Patterns of Productivity, Collaboration, and Authorship. *American Sociological Review*, **32**(3), pp. 391–403.