USING COLLECTIVE DISCOURSE TO GENERATE SURVEYS OF SCIENTIFIC PARADIGMS

by

Vahed Qazvinian

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
2012

Doctoral Committee:

Professor Dragomir R. Radev, Chair
Associate Professor Lada A. Adamic
Assistant Professor Michael J. Cafarella
Assistant Professor Qiaozhu Mei

To my parents.

ACKNOWLEDGEMENTS

I am most grateful to my advisor Professor Dragomir R. Radev to guide me through four enjoyable and productive years of Ph.D. studies. His guidance and support have been essential not only in the development of this thesis, but also in my personal development as a scientist.

Dragomir Radev is a fantastic advisor, a very brilliant computer scientist, and a great friend. He is also among the most influential people in my life. Dragomir is a wonderful researcher. He has taught me how take initiative and how to think critically. Watching Dragomir approach difficult problems and solve them is both fascinating and inspiring. Dragomir cares about his students a lot and considers their success in their academic lives as his own success. I am most grateful to Dragomir for his support without which my success would have been impossible.

I would also like to thank my dissertation committee, Michael Cafarella, Qiaozhu Mei and Lada Adamic for careful criticism and insightful feedback that improved the quality of this work. Taking the network theory class with Lada was undoubtedly one of my most exciting experiences at Michigan. Lada is an amazing teacher and a truly inspiring researcher. I strongly believe that any graduate student at Michigan in any field of study would benefit from Lada's class.

I was also very fortunate to be Michael Cafarella's first teaching assistant at Michigan. During this time, I found Michael to be an excellent teacher and a wonderful researcher. Working with Mike helped me shape my academic philosophy. I under-

stood that many students have difficulty in seeing how the theoretical material and algorithms in the lectures can be applied to various real-world problems. Michael showed me how teaching solutions to specific theoretical problems is not as effective as showing the students how to think as problem solvers and how to construct solutions in real-world applications.

Finally, I am very thankful to Qiaozhu Mei. I have been working with Qiaozhu in a very interesting, yet challenging project. Qiaozhu is an excellent advisor, a very smart researcher, and an extremely personable individual. I have learned a lot working with Qiaozhu and am very thankful to have known such a great computer scientist .

During two summer internships at Microsoft Research, Redmond, I had the opportunity to work with two wonderful researchers, Chris Brockett, and Wen-tau Scott Yih. Chris is an amazing mentor, who sets high standards in his research. He taught me how to be critical and precise, and have high standards in my research. Scott is a very smart and knowledgeable researcher. He patiently helped me familiarize myself with new Machine Learning techniques and taught me how large-scale machine learning systems work. I am deeply grateful to both Chris and Scott.

During the four years of my Ph.D. studies, I have collaborated with many excellent researchers and scientists. I thank Bonnie Dorr, Ben Shneiderman, Judith Klavans, and Jimmy Lin from the University of Maryland College Park, Saif Mohammad from National Research Council Canada, and Paul Resnick, Rahul Sami, and Qiaozhu Mei from School of Information at Michigan.

I am also thankful to all the former and current members of the CLAIR lab at Michigan: Joshua Gerrish, Alex Gonopolskiy, Ben Nash, Arzucan Ozgur, Ahmed Hassan, Amjad Abu-Jbara, Pradeep Muthukrishnan, Rahul Jha, Wanchen Lu, and

Ben King.

I am very grateful to the administrative staffs in the CSE department at Michigan, specially Dawn Freysinger, Karen Liska, Steve Crang, and Cynthia Watts who made many things seamless. Also many thanks to Professor Smaranda Muresan from Rutgers, Professor Kathleen McKeown from Columbia, and Professor Yejin Choi from Stony Brook University for their hospitality while I was visiting their institutions.

I take this opportunity to thank my very special friends who made life in Ann Arbor enjoyable and memorable. Thank you Jennifer Klein, Siamak Nejad Davarani, Mohammadreza Imani, Farhad Bayatpour, Danial Ehyaie, and Eaman Jahani. I also wish to thank all other friends who shared their joy with me.

I thankfully acknowledge the fellowships and grants that supported my research at Michigan. I was supported by the Rackham Centennial Fellowship Award, National Science Foundation grants "SoCS: Assessing Information Credibility Without Authoritative Sources" as IIS-0968489, and "iOPENER: A Flexible Framework to Support Rapid Learning in Unfamiliar Research Domains" as IIS-0705832.

Above all, I express my heartfelt gratitude to my beloved parents Mehrnoush Toufan Tabrizi and Habibollah Ghazvinian, my brother Vaqef, and my lovely grandmother Roghaiyeh Khalkhali, whose love and prayers have always been with me. Without their love, support, and encouragement none of my achievements would have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

USING COLLECTIVE DISCOURSE TO GENERATE SURVEYS OF SCIENTIFIC PARADIGMS

by
Vahed Qazvinian

Chair: Professor Dragomir R. Radev

This thesis is focused on understanding collective discourse and employing its properties to build better decision support systems. We first define collective discourse as a collective human behavior in content generation. In social media, collective discourse is often a collective reaction to an event. A collective reaction to a well-defined subject emerges in response to an event (a movie release, a breaking story, a newly published paper) in the form of independent writings (movie reviews, news headlines, citation sentences) by many individuals.

In order to understand collective discourse, we perform our analysis on a wide range of real-world datasets from citations to movie reviews. We show that all these datasets exhibit diversity of perspective, a property seen in other collective systems and a criterion in wise crowds. Our experiments also confirm that the network of different perspective co-occurrences exhibits the small-world property with high clustering of different perspectives. Finally, we show that non-expert contributions in collective discourse can be used to answer simple questions that are otherwise hard

to answer.

As a concrete example of collective discourse, we discuss citations to scholarly work. We show how they contain important information that convey the key features and basic underpinnings of a particular field, early and late developments, important contributions, and basic definitions and examples that enable rapid understanding of a field by non-experts. We then present C-LexRank, a system that exploits scientific collective discourse to produce automatically generated, readily consumable technical surveys. Finally, we further extend our experiments to summarize an entire scientific topic. We generate extractive surveys of a set of Question Answering (QA) and Dependency Parsing (DP) papers, their abstracts, and their citation sentences and show that citations have unique survey-worthy information.

CHAPTER I

Introduction

In sociology, the term *collective behavior* is used to denote mass activities that are not centrally coordinated [21]. Collective behavior is different from group behavior in the following ways: (a) it involves limited social interaction, (b) membership is fluid, and (c) it generates weak and unconventional norms [181].

Gordon [69] explained how harvester ants achieve task allocation without any central control and only by means of continual adjustment. Moreover Gordon [69] argued that the cooperative behavior in the ant colony merely results from local interactions between individual ants and not a central controller (emergent behavior). For instance, in ant colonies individual members react to stimuli (in the form of chemical scent) depending only on their local environment. In the absence of a centralized decision maker, ant colonies exhibit complex behavior to solve geometric problems like shortest paths to food or maximum distance from all colony entrances to dispose of dead bodies.

Self organized behavior is not specific to ants. Schools of fish, flocks of birds, herd of ungulate mammals are other examples of complex systems among animal groups [60]. Similarly, pedestrians on a crowded sidewalk exhibit self-organization that leads to forming lanes along which walkers move in the same directions [22]. It

is argued that all examples of complex systems exhibit common characteristics:

1. They are composed of a large number of *inter-connected parts* (i.e., agents)
2. The system is *self-organized* in that there is not central controller.
3. They exhibit *emergent behavior*: properties seen in the group but not observable from the actions of individuals.

Nonlinear behavior has been widely observed in nature in the past. Gordon [69] explains how harvester ants achieve task allocation without any central control and only by means of continual adjustment. Moreover he argues that the cooperative behavior in the ant colony merely results from local interactions between individual ants and not a central controller. For instance, in ant colonies individual members react to stimuli (in the form of chemical scent) depending only on their local environment. In the absence of a centralized decision maker, ant colonies exhibit complex behavior to solve geometric problems like shortest paths to food or maximum distance from all colony entrances to dispose of dead bodies.

Self-organized behavior is not specific to ants. Schools of fish, flocks of birds, herd of ungulate mammals are other examples of complex systems among animal groups [60]. Similarly pedestrians on a crowded sidewalk exhibit self-organization that leads to forming lanes along which walkers move in the same directions [22]. It is argued that all examples of complex systems exhibit common characteristics:

1. They are composed of a large number of inter-connected parts (i.e., agents)
2. The system is self-organized in that there is not central controller.
3. They exhibit emergent behavior: properties seen in the group but not observable from the actions of individuals.

1.1 Collective Discourse

In social sciences, a lot of work has been done on collective systems and their properties [83]. However, there is only little work that studies a collective system in which individual members describe an event or an object. In our work, we focus on the computational analysis of *collective discourse*, a collective behavior seen in interactive content contribution in online social media [162].

In social media, collective discourse [72] is often a *collective reaction* to an event. One scenario leading to collective reaction to a well-defined subject is when an event occurs (a movie is released, a story occurs, a paper is published) and people independently write about it (movie reviews, news headlines, citation sentences). This process of content generation happens over time, and each person chooses the aspects to cover. Each event has an onset and a time of death after which nothing is written about it. Tracing the generation of content over many instances will reveal temporal patterns that will allow us to make sense of the text generated around a particular event.

- **Movie Reviews**

  The first collective discourse appears in the set of reviews that non-expert users write about a movie. The set of online reviews about an object is a perfect case of collective human behavior. Upon its release, each movie, book, or product receives hundreds and thousands of online reviews from non-expert Web users. These reviews, while discussing the same object, focus on different aspects of the object. For instance, in movie reviews, some reviewers solely focus on a few famous actors, while some discuss other aspects like music or screenplay.

  For example, the following excerpts are extracted from user reviews for the movie

Pulp Fiction, and show how non-expert reviewers focus on different aspects of the movie.

"*... starred by many well-known Actors, such as: John Travolta, Samuel L. Jackson, Uma Thurman, Bruce Willis and many. Directed by Quentin Tarantino, the eccentric Director ...*"

"*Pulp Fiction was nominated for seven academy awards and won only one for screen writing ...*"

"*Shocking, intelligent, exciting, hilarious and oddly though-provoking. Best bit: Jackson's Bible quote ...*"

- **Microblogs**

  The second type of collective discourse that we observe in our work is the set of tweets written about a news story. Using Twitter as a corpus of collective discourse does present unusual challenges. In Twitter, posts are limited to 140 characters and often contain information in an unusually compressed form.

  An example of this type of collective discourse is the set of people who spread rumors on social media such as Twitter. In [165], we study several examples of such discourse. One such example is about Sarah Palin's divorce rumor that was popular during the 2008 presidential election campaigns. This dataset contains tweets that are about this story and yet discuss it from different angles. For example, the following tweets are extracted from this dataset and reveal various facts about the story. One aspect is that a blogger has started the spread, and is threatened with libel suit. Another aspect is that the rumor has been debunked on Facebook.

  "*... Palins lawyer threatens divorce blogger with libel suit, gives her the option*

of receiving the summons at her residence http://ow.ly/15JDO6"

"*@jose3030 Palin divorce is supposedly debunked on Facebook, but I think they are just spinning it, until they can announce it.*"

"*RT @mediaite: Sarah Palin uses Facebook to deny unsourced divorce rumors - http://bit.ly/14Xy6hCH.*"

As another Microblog dataset, we collected the tweets that talk about the cancellation rumors of 14 TV shows in August of 2011. For instance, one of our collected datasets is about the rumor that Charlie Sheen might go back to the TV show Two and a Half Men.

"*Charlie Sheen Claims 'Discussions' About Returning to 'Two and a Half Men': In Boston for his national tour, C... http://bit.ly/hIbOWf.*"

"*Charlie Sheen Two And A Half Men' Return Not Happening*"

- **News Headlines**

  Another collective discourse is seen when a story breaks and various news agencies write headlines about it. All such headlines discuss the same story, but view it from different perspectives.

  We collected 25 news clusters from Google News. Each cluster consists of a set of unique headlines about the same story, written by different sources. The following example shows 3 headlines in our datasets that are about hurricane Bill and its damage in Maine.

  "*Hurricane Bill sweeps several people into ocean.*"

  "*7-year-old girl swept away by Bill wave dies after rescue.*"

  "*Maine ranger: wave viewers didn't heed warnings.*"

- **Citation Sentences**

  The final collective discourse example that we study is the set of citation sentences that different scholars write about a specific paper. A citation sentence to an article is a sentence that appears in the literature and cites that paper. Each citation to a paper may or may not discuss one of that paper's contributions.

  A citation sentence is a sentence in an article containing a citation and can contain zero or more nuggets (i.e., non-overlapping contributions) about the cited article. For example the following sentences are a few citation sentences that appeared in the NLP literature in past that talk about Resnik's work.

  "*The STRAND system (Resnik, 1999), for example, uses structural markup information from the pages, without looking at their content, to attempt to align them.*"

  "*Resnik (1999) addressed the issue of language identification for finding Web pages in the languages of interest.*"

  "*Mining the Web for bilingual text (Resnik, 1999) is not likely to provide sufficient quantities of high quality data.*"

- **Other Collective Discourse Examples**

  The study of collective discourse helps us understand new aspects of an object that are hard to identify with a single authoritative view. Collective discourse examples are not limited to the datasets that we have collected. For instance, studying a complete set of introductions about PageRank enables us to learn about its important aspects such as the algorithm, the damping factor, and the Power method, as well as aspects that are less known such as its use in 1940s [61]. Similar examples exist in different TV show synopses, book descriptions, story

narrations and many more.

To understand collective discourse, we are interested in behavior that happens over a short period of time. We focus on topics that are relatively well-defined in scope such as a particular event or a single news event that does not evolve over time. This can eventually be extended to events and issues that are evolving either in time or scope such as elections, wars, or the economy.

In social sciences and the study of complex systems a lot of work has been done to study such collective systems, and their properties such as self-organization [151] and diversity [83, 60]. However, there is little work that studies a collective system in which members individually write summaries.

In this thesis, we will discuss various examples of collective discourse in online social media and in particular citations to scholarly work. We will show how they contain important information that convey the key features and basic underpinnings of a particular field, early and late developments, important contributions, and basic definitions and examples that enable rapid understanding of a field by non-experts. We will then present C-LexRank, a system that exploits scientific collective discourse to produce automatically generated, readily consumable technical surveys.

Collecting collective discourse datasets may not be a straight-forward task. For instance, in scientific literature a citation sentence is normally accompanied with a set of context (background information) sentences that implicitly cite the target paper. In [160], we propose a general framework based on probabilistic inference to extract such context information from scientific papers.

Finally, we further extend our experiments to summarize an entire scientific topic. We generate extractive surveys of a set of Question Answering (QA) and Dependency Parsing (DP) papers, their abstracts, and their citation sentences and show that

citations have unique survey-worthy information.

1.2 Guide to Chapters

In chapter III, we analyze collective discourse and show that it exhibits diversity, a property of general collective systems. We analyze 50 sets of human-written summaries about the same story or artifact and investigate the diversity of perspectives across these summaries. We show how different summaries use various phrasal information units (i.e., *nuggets*) to express the same atomic semantic units, called *factoids*.

In chapter IV, we view collective discourse as a complex system modeled with a network that is bound to a parameter. Such a network can be considered as an ensemble of unweighted graphs, each consisting of edges with weights greater than the cutoff value. We look at this network ensemble as a complex system with a temperature parameter, and refer to it as a *Latent Network*. Our experiments on a number of datasets from different domains show that certain properties of latent networks like clustering coefficient, average shortest path, and connected components exhibit patterns that are significantly divergent from random networks. We explain that these patterns reflect the network phase transition as well as the existence of a strong community structure in document collections. These properties of latent networks can be exploited to predict the network at which the community structure in the network, and thus the clustering quality is best captured.

Our understanding of the emergent behavior in text enables us to design better techniques for common text analysis tasks such as clustering, re-ranking, summarization, salience detection, and more. Chapter V is focused on designing summarization systems that exploits the properties of collective discourse such as diversity

and community structure, and produce diverse summaries of the represented content in a collective discourse.

In chapter VI, we first design a method to automatically extract different perspectives (factoids) from citations. Moreover, we discuss C-LexRank when it is run on words and not documents. We represent the set of words in a corpus as a network, where edges show the similarity of words using the *distributional hypothesis*. By applying C-LexRank on this network, we find communities of words that are more similar to each other whereby each community represents the set of words that relate to one factoid.

Finally in chapters VII, VIII we extend our work from summarizing contributions of single articles to entire scientific topics. More particularly, we present our experiment on using the tools explained in previous chapter for automatic survey generation in chapter VII. We evaluate 2 automatically generated surveys on Question Answering (QA) and Dependency parsing (DP). Finally, in chapter VIII we perform experiments on gold standard datasets that are beyond expensive human annotations. We propose the use of naturally written surveys as gold standards. These gold standard resources include end-of-chapter historical notes from the leading NLP text book of Jurafsky and Matin [89], student summaries from a seminar class, and survey papers written by other scholars.

In Appendix, we list some of the summaries that are generated using our proposed system, C-LexRank, as well as 4 human written summaries on Question Ansewring.

CHAPTER II

Related Work

In this chapter we review the related work in 4 different sections. First, we summarize previous work on collective systems and collective human behavior in general. Then, we look at prior work on modeling natural language as a complex system. Since this thesis is focused on modeling collective discourse as a complex system composed of nodes and edges, it is essential to review previous attempts that model linguistic phenomena as complex systems. Third, we provide a literature review of graph based summarization systems. Systems that represent a set of documents as a network and produce a summary by applying graph based methods such as salience detection and ranking. Finally, we review work on citation analysis. Particularly, we review prior work that has looked at the structure and importance of citations in scholarly work.

2.1 Collective Behavior

Previous work has studied the complex system of natural collective behaviors such as the ant colonies [195] or bird flocks and fish schools [173]. Many of the properties seen in these naturally occurring complex systems are also integral to human systems. Properties such as *self-organization*, *coordination*, and *emergence* are common in real world tasks that involve interaction between humans. For example, teammates in

a sports team constantly adjust their strategies in response to others' actions [174]. Similar phenomenon exists in groups like poker players and traders on eBay.

Previously, it has been argued that diversity is essential in intelligent collective decision-making. Page [151] argued that the diversity of people and groups, which enable new perspectives, leads to better decision making. He found that the diversity of perspectives in a collective system is associated with higher rates of innovation and can enhance the capacity for finding solutions to complex problems. Similarly, Hong and Page [82] showed that a random group of intelligent problem solvers can benefit from diversity and outperform a group of the best problem solvers.

2.2 Language as a Complex System

A *complex system* is a system composed of interconnected parts (agents, processes, etc.) that as a whole exhibit one or more properties called *emergent behavior*. The emergent behavior, which is not obvious from the properties of the individuals, is called to be nonlinear (not derivable from the summations of the activity of individual components). Complex systems have been widely used to explain nonlinear behavior in nature [69, 60].

Naturally occurring complex systems are often represented with systems of equations. The equations from which complex system models are developed generally derive from statistical physics, information theory and non-linear dynamics. They represent organized but unpredictable behaviors of systems of nature that are considered fundamentally complex. Complex Systems are also used to model processes in economics, computer science, physics, and biology [30]. All complex systems have many inter-connected components that can be represented as a network of nodes with edges. Thus, network theory is an important aspect of the study of complex

systems.

The theory of complex systems is helpful in explaining the origin, evolution, and death of human languages and can explain the process of language acquisition as well as the emergent properties of natural languages (e.g., Zipf's law). Moreover, it provides powerful tools such as network theory and agent based modeling, which enable us to provide a unified explanation of various properties seen in different languages.

2.1 Language Evolution and Acquisition

Most of the research on language acquisition has been dominated by the views of Noam Chomsky and his classic questions [32]: (1) what constitutes knowledge of a language? (2) how is this knowledge acquired? (3) how is it put to use? One theory is that what one knows is a grammar, a complex system of rules and constraints, that allows people to distinguish grammatical sequences [179]. Moreover, it is argued that there must be strong innate constraints on the possible forms of grammars [68], and that a child possesses considerable grammatical knowledge at birth [179]. The logical argument behind this claim is due to poverty of stimulus (e.g., the lacking evidence of ungrammatical sequences that a child observes while learning a language) [112] or the fact that no learning process has so far been successful in explaining language acquisition [33]. This theory implies that language is preserved due to genetic transmission.

Various aspects of the theory of language through biological evolution have been criticized and rejected in the literature including the poverty of stimulus [39, 158] and grammar-specific genes [200]. The alternative theory is based on a *complex system* viewpoint. This theory rejects any genetic dimension to language evolution and acquisition, and claims that language is a result of self-organized structures[186].

Structures similar to those result in straight lines in ant colonies and certain patterns in bird flocks.

In the complex systems view of language, individuals lack any "language organ," but preserve information about it in their memories [187]. Language, as we explain later, is transmitted through learning in a cultural fashion. Mutations in the language occur since individual speakers may lack a comprehensive understanding of the language, make errors, or try to be as different from other individuals as possible (giving rise to *diversity*). In fact, Nowak et al. [149] proposed 4 reasons for a language change as part of a cultural evolution: (1) random variation, (2) contact with other languages, (3) hitchhiking on other cultural inventions, and (4) selection for better learnability and communication. This self-organization in the community of individual language users make the language more coherent (i.e., frequent use of a word will give it success, which will in turn preferentially boost the use of that word).

Other attempts have tried to link these two distinct theories and use both genetic evolution and individual adaptation to explain language [81]. In [182] language is modeled as the interaction of 3 separate complex systems: biological evolution, learning, and culture. In their model learning results in cultural evolution which leads to biological evolution. They also argue that innate endowment guides language acquisition (learning).

The Complex System of Language

Previous work argues that the process of language acquisition is a transition of a set of pre-fabricated sequences between two agents (e.g., parent and child) rather than an open choice among all other words in a language [148]. This is also confirmed by usage-based theories, which argue that we learn a language while engaging in a

communicative process that shapes the language [13].

The complex system perspective provides evidence that the process of language acquisition, use, and evolution are not independent from each other but correspond to a single complex system [18]. More formally, this perspective proposes a co-evolution process for language evolution and acquisition. A process that has 4 main components: (1) language usage leads to change, (2) change affects how cues are perceived, (3) perception affects learning, (4) learning affects usage [18].

2.2 Characteristics of Language Complex System

The complex system of language has 7 major characteristics [18]:

1. **Two levels of existence**: This characteristic indicates the existence of language both in individuals (as idiolect) and the community of users (as communal language). Communal language is emergent from the interaction of individual idiolects and idiolects emerge from social interactions of individuals through communal language use. This distinction and connection is a common feature of complex systems. The collective patterns seen in communal level (similar to bird flocks or fish schools) are emergent from long-term local interactions.

2. **Diversity**: Similar to other collective systems (such as bird flocks, ant colonies, or people on crosswalks), there is no ideal or authoritative speaker of a language. Each individual's exposure and experience results in a different idiolect [26]. This variety in idiolects results in high degrees of diversity in language use [208]. More recently, Qazvinian and Radev [162, 163] performed extensive statistical analysis and showed that diversity exists in large degrees in using various phrasal information units (nuggets) that represent the same information. They use a complex system approach and show that collective discourse, a collective system

in which each agent independently contributes content about an artifact or event, exhibits the diversity similar to those seen in other collective systems.

3. **Perpetual dynamics**: Language is very similar to many other complex systems in its natural change. Unlike closed systems, which reduce to an equilibrium, language is in constant change both in its idiolect level and its communal level.
4. **Adaptation through competing factors**: Like many other complex systems, language evolves through various competing factors. Factors like production, brevity, perceptual salience, explicitness, and clarity.
5. **Phase Transition**: Phase transition is often referred to as significant qualitative differences after small quantitative changes in certain parameters. Previous work argues that human language emerges when a set of language-enabling traits (such as sociality, vocal tract control, shared attention, imitation, memory, etc) take a specific form. In this model language is seen as a domain-specific outcome that emerges through the interaction of multiple features, none of which is specific to language. Small change in such features results in a communication means of totally different nature in humans [52].

   In a different work, language use among a group of people is modeled as a complex system, where agents engage in a *collective discourse* [163]. This complex system is bound to a simple temperature parameter. Small change of this parameter makes the network of agents undergo phase transition and exhibit high degrees of community structure, where each community corresponds to a topic of the collective discourse.
6. **Dependence on social structure**: Like many real-world networks that represent various complex systems, the network of linguistic interactions is not formed via random interactions. Mirloy [130] argues that language change is

affected by the social network of language use. The understanding of the role of social networks in language evolution and acquisition remains an important research problem.

Social networks have also been employed in the past to explain certain emergent behaviors in language [207, 57, 59]. For instance, Steyvers and Tenenbaum [189] examined free association networks, WordNet, and the Roget Thesaurus, and noted five different properties in semantic networks. They were all sparse, with a giant component, a small shortest-path length, a high clustering-coefficient, and a scale-free degree distribution. We will review related work on network analysis later in the paper.

7. **Adaptation to human brain** The last characteristics of language complex system listed in [18], is its adaptation to the human brain. This characteristic focuses on the mutual role that language and human brain play in a co-evolution process. This research problem, which is often studied in Neurolinguistics, is beyond the scope of this survey. To learn more on this topic, we encourage the reader to refer to other related work [76, 46, 178, 36, 35].

Computational Models

The theory of language evolution as a complex system is supported by some computational work mainly through simulating models. Prior work in [17, 157] explored a simple communication model in which the world consists of agents. Each agent contains a "meaning vector" and a recurrent neural network for communicating sequences of characters chosen from $\{a, b, c, d\}$. In one communication episode each element in the speaker's meaning vector obtains a value in $[0, 1]$ depending on the meaning that the speaker would like to convey. The listener receives the vector and

process it using its own network. After the sequence has been processed, the output of the listener's network represents its interpretation of the sequence. The neural network is trained using the back-propagation algorithm [177]. The authors showed that after 15,000 iterations, over 92% of the meanings are interpreted accurately by the listeners, suggesting the emergence of some sort of systematic regularities between meanings and their expression as sequences of characters. A central assumption to the model presented by [17] is that agents use their own responses to characters as a means to predict other agents' interpretation of characters. This assumption is somewhat present in primates and more in humans [210].

Similarly, Steels [188] took a complex systems viewpoint of language and models a system in which agents have means of communication but lack specific semantics and language conventions. He further explained that these conventions arise from the interaction of agents in a simulated process. Others have tried to model language as a Nash equilibrium in an evolutionary process [197]. In their work, Trapa and Nowak [197] modeled the complex system of language users in an evolutionary game theory setting, and showed that any random language evolves into a strict Nash equilibrium, in which the total amount of information exchanged between two individuals is more than any other language.

Language Death

The number of living languages today exceeds 6,000 [100]. 52% of the 6,000 languages are spoken by fewer than 10,000 people, and 28% are spoken by less than 1,000 people [71]. According to some estimates 50% to 90% of these languages will vanish by the end of the 21st century [205].

Abrams and Strogatz [1] proposed a simple dynamical complex system for modeling language endangerment. In their model the community of language speakers

has only 2 languages, X and Y, and individuals are monolingual. Intuitively, each individual adopts a language with a frequency proportional to its status, s, a popularity measure of a language expressed as an increasing function of the number of people who speak that language.

As depicted in Figure 2.1 an individual converts from Y to X with a probability of $p_{Y\to X}(x,s)$, where x is the fraction of people speaking X, and $s \in [0,1]$ is the status of X. The minimal model for this language change can be formulated as

$$\frac{dx}{dt} = yP_{Y\to X}(x,s) - xP_{X\to Y}(x,s) \tag{2.1}$$

Equation 2.1 states that the change in the fraction of people speaking X at a point of time equal the fraction of people converting to X minus the fraction abandoning X. By assuming that no one will ever adopt a language that has no speakers (i.e., $P_{x\to y}(0,s) = 0$) or no status (i.e., $P_{x\to y}(x,0) = 0$), [1] conclude that Equation 2.1 has only 2 stable fixed points at $x = 0$ and $x = 1$ suggesting that the two languages cannot coexist stably.

They fit the proposed model on data from 4 languages using Equation 2.2, and show that the exponent a is unexpectedly constant across different languages ($a \approx 1.31$).

$$P_{Y\to X}(x,s) = cx^a s \tag{2.2}$$

The proposed model by [1] could be more realistic if it took a social structure of individuals into account. For instance, this model can not explain the cases of languages that have limited diffusion but are not at risk of extinction. Therefore, a possible extension of this model should consider cultural effect has been proposed before in [154].

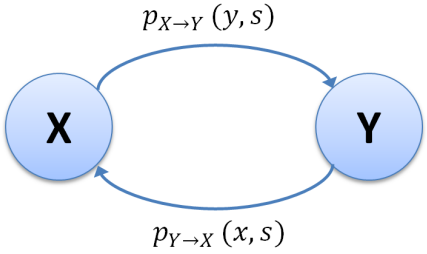


Figure 2.1: In the model proposed by Abrams and Strogatz, each individual is monolingual and with a probability changes her language to the other spoken language in the community.

Patriarca and Leppänen [154] argued that the influence of a language can depend on political or geographical factors. They extended the previous model by assuming the area has 2 zones, and individuals can only interact (speak) with other individual who are in the same zone. They showed that despite the difference in languages' status ($s_A \neq s_B$), the system has a stationary state in which both languages survive. However, they pointed out that the 2 languages are mainly concentrated in different zones, but interact with one another in a narrow area between the two zone. They finally concluded that multiple languages can coexist by acquiring speakers in distinct geographical areas.

Wang and Minett [205] argued that a model for language competition should incorporate bilingual speakers. In their model the sociolinguistic interaction between the speakers of one language (X) and another (Y) is possible via bilingual speakers. As illustrated in Figure 2.2, in this model monolingual speakers of either X or Y can choose to become proficient in the other language, without losing their original language. Moreover, bilingual speakers can opt to lose proficiency in one of the languages they speak.

Previous research that extends Abrams and Strogatz's original model by taking bilingual speakers into account is presented in [131, 132]. They analyzed the evolution

of two coexisting languages (Castilian Spanish and Galician) under Abrams and Strogatz's model, and showed that the two languages can evolve to coexist if they are similar enough [149]. The similarities between Spanish Castilian and Galician allow for limited conversation between monolingual speakers of either languages. Mira and Paredes [131] assumed 3 groups of people: speakers of X, Y, and bilinguals , B and denote by x, y, b the fraction of people in each group where $x + y + b = 1$. Their extension to the model in Equation 2.1 is as follows.

$$\frac{dx}{dt} = yP_{Y\to X} + bP_{B\to X} - x(P_{X\to Y} + P_{X\to B}) \tag{2.3}$$

where

$$P_{X\to B} = cks_Y(1-x)^a \tag{2.4}$$

$$P_{X\to Y} = c(1-k)s_Y(1-x)^a \tag{2.5}$$

Here, a speaker of X will change to Y or become bilingual with probabilities based on a parameter $0 \leq k \leq 1$, which reflects the ease of bilingualism and language similarity. This model reduces to the one by [1] if $k = b = 0$.

Mira and Paredes [131] showed that the extended model successfully fits the data and yields high similarity between the two languages. They argued that for every value of s_X, denoting the status of the language X, there exists $k_{min}(s_X, a)$ such that for all $k < k_{min}$ the language with smaller status dies, and for other values of k both groups survive. Furthermore, Mira et al. [132] simulate the sociolinguistic situation in Galicia using their model up to the year 2100 ($k = 0.80$) and showed that all groups can survive if Galician status is equal to 0.36. However, if Galician language fails to to reach a status of 0.335, Spanish monolingual population will have a sustained growth and the Galician monolingual population will die. Mira et al. [132] emphasized that the disappearance is of the monolingual group does not

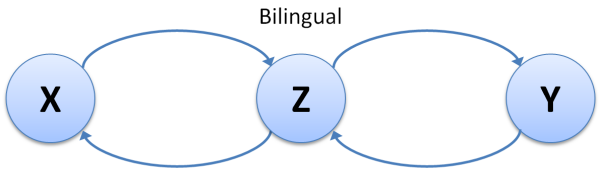


Figure 2.2: In the model proposed by Wang and Minett, each monolingual individual can opt to be bilingual or vice versa.

imply the death of Galician language itself, since it would survive in the bilingual group.

Research on language competition can potentially benefit from recent advancements in modeling social structure, diffusion on networks, and influence in a community [111, 93, 109, 92, 38, 67, 94], and propose models of language evolution that takes social relations into account.

2.3 Complex Networks of Language

Recent achievements in network theory [12] has proved its ability to model complex systems such as the World Wide Web [5, 84, 85, 106], social network of people [207], scientific collaborations [144], biological networks [170], and complex food webs [211]. Similarly, prior work has tried to model language complex system as a network of linguistic entities such as words or phrases [57, 48, 59].

Dorogovtsev and Mendes [48] modeled language as a self-organizing and evolving network of words. The network grows both when new words are born and when new edges emerge between disconnected nodes (words) obeying the following simple rules. At each time step t, the tth node joins the network, and connects to several other nodes following the preferential attachment model of [12]. Preferential attachment states that the new node connects to an older node i with probability proportional to i's degree k_i. Moreover, in this model, at each time step ct new edges emerge in the

network connecting two nodes i and j with probability proportional to the product of their degrees $k_i k_j$ [47].

In this model, at each time t, $1 + 2ct$ new edge tails are distributed preferentially among old nodes. The evolution of the degree of the sth word (the node born at time s) observed at time t (i.e., $k(s,t)$) can be formalized as

$$\frac{\partial k(s,t)}{\partial t} = (1 + 2ct)\frac{k(s,t)}{\int_0^t k(u,t)\mathrm{d}u} \tag{2.6}$$

By approximately solving this stochastic model in Equation 2.6, Dorogovtsev and Mendes [48] showed that this network has a non-stationary power-law degree distribution that is also observed in real-world data [57].

An interesting conclusion from this work is that the number of core words in a lexicon does not depend on the total number of distinct words and is determined by the value of the average rate interconnection c. However, despite confirming the degree distribution seen in human language networks, this model fails to account for important factors such as word extinction or a word's evolution through language use.

2.4 Small-world effect

Statistical relationships between words can be analyzed with the rate of their co-occurrences in sentences. Such co-occurrences could be due to variety of reasons such as syntactical relations or collocations (e.g., United States). The network of words as nodes and co-occurrence relationships as edges is called the co-occurrence network. Ferrer i Cancho and Solé [57] used three quarters of the 10^7 words in the British National Corpus (BNC[1]) and formed a link between any two nodes if the corresponding words had a distance of 2 or smaller (co-occur within 2 words of

[1] http://info.ox.ac.uk/bnc/

each other) in the corpus. With extensive statistical analysis, they showed that this network exhibits small-world properties [207].

The small-word effect is also seen in networks of human interactions [207], metabolite processing [206], food webs [135], electronic circuits [56], and brain neurons [16]. Networks with this property obtain short average shortest paths but large clustering coefficients. Ferrer i Cancho and Solé [57] showed that the average shortest path in co-occurrence networks is below 3.00 and the clustering coefficient is greater than 0.68. This study showed that despite the huge lexicon stored in human brains ($10^4 - 10^5$ words [129]), any word can be reached with fewer than 3 intermediate words, on average. In other words, during the course of a communication, reaching one word from another requires very few steps.

In a similar work, Ferrer i Cancho et al. [59] looked at dependency networks. These networks contain a directed edge from a node s_i to a node s_j if the corresponding word to s_i can be a modifier of the head s_j in a sentence in the corpus. Therefore, the global dependency network consists of all the dependency relations seen in the corpus, and thus the syntactic dependency structure of a sentence corresponds to a subgraph of this large network. Figure 2.3 is an example of a dependency network built using the Stanford parser and dependencies representation [98, 45] extracted from the following sentence:

Television announced that Ashton Kutcher will join the cast of the hit comedy and replace Charlie Sheen.

A *global* dependency network is the network of the dependency graphs of all the sentences in a corpus. Ferrer i Cancho et al. [59] studied global dependency networks of 3 different languages: Czech, German, and Romanian. They showed that these networks exhibit similar properties: they all have small-world structure (i.e., average

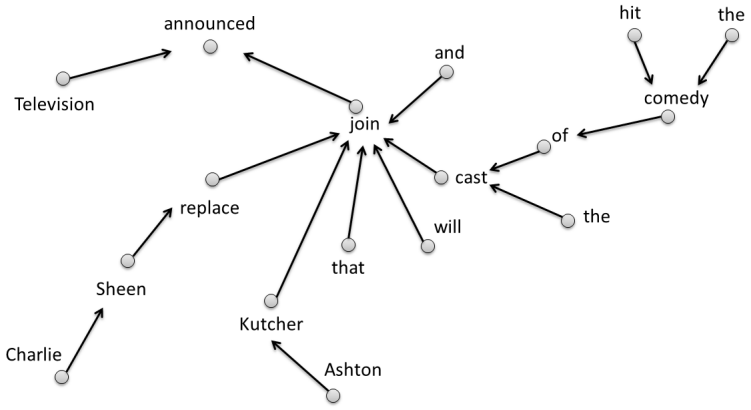


Figure 2.3: Sample dependency network.

shortest path ≈ 3.5) and highly heterogeneous degree distributions, in which the probability that a vertex has degree k obeys a power-law of the form

$$p(k) \propto k^{-\gamma} \tag{2.7}$$

where $\gamma \approx 2.2$. Moreover, and these networks exhibit high degrees of community structure (i.e., hierarchical organization), which implies that syntactic dependency networks define a top-down hierarchical organization that is the basis of phrase-structure formalisms [59]. Similarly, Liu and Xu [118] analyzed the dependency networks of 15 languages and showed that the dependency syntactic networks can reflect morphological variation degrees and morphological complexity in a language. Liu and Hu [117] showed that syntax influences the properties of a complex network, but argued that the scale-free property is a necessary condition but not sufficient to judge whether a network is syntactic or not.

The small-world effect is not limited to BNC or the dependency networks. These properties have also been observed in various semantic networks. Motter et al. [136] analyzed the structure of conceptual networks, and Steyvers and Tenenbaum [189] performed statistical analysis of the large-scale structure of 3 different semantic networks: word associations, WordNet, and Roget's thesaurus.

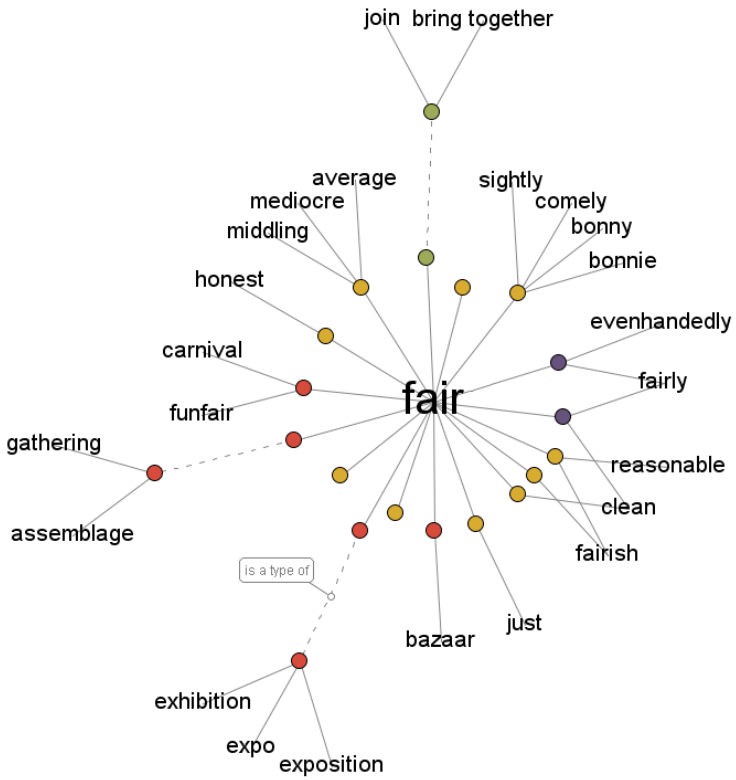


Figure 2.4: Part of the English conceptual network built from a thesaurus.

Motter et al. [136] constructed a conceptual network from the entries in a thesaurus and formed a link between two nodes if they express similar concepts. They showed that this network is sparse and exhibits small-world property with a clustering coefficient equal to 0.53 and average shortest path of 3.2. They argued that this is a result of natural optimization and the network's scale-free property is due to its dynamic character. Moreover, the very small average shortest path is because of the existence of words that correspond to two or more very different concepts. For instance, Figure 2.4 shows part of the English thesaurus network around the word "fair"[2]. This network shows that "fair" is connected to both "carnival" and "reasonable", making them connected through a very short path (length of 2).

The study of conceptual networks was extended in [189] to study word association

[2]Network is built using http://www.visualthesaurus.com/

networks. Word association networks are constructed by asking people the first word that comes to their mind after seeing a particular given word (e.g., dog) [142]. In word (directed) association networks, two word nodes are connected if the cue one word evoked the other as an associative response for at least two of the participants in the database. Roget's thesaurus [175] includes over 29,000 words classified into 1,000 semantic categories. Steyvers and Tenenbaum [189] represented this thesaurus as a bipartite graph where a set of words are connected to a set of semantic classes, and then collapsed this graph to a regular network of words in which two words are connected if they have a common neighbor in the bipartite graph. The WordNet [128] network used in [189], contains more than 120,000 word forms (i.e.m single words and collocations) and 99,000 word meanings. They studied 5 different properties of these 3 semantic networks: sparsity, connectedness, short path lengths, high neighborhood clustering, and power-law degree distributions.

Steyvers and Tenenbaum [189] argued that all 3 semantic networks are very sparse. For instance, in the association networks, a word is only connected to less than 1% of other nodes. However, these networks exhibit high connectivity where a single giant component accounts for the majority of the nodes in the graph (e.g., the largest connected component consists of 96% of all words in association networks and 99% in Roget's thesaurus and WordNet network). Despite sparsity and having giant components, these networks exhibit very short average shortest paths (3 in association networks). This means that on average any two words are 3 association steps away from each other. Moreover, similar to BNC and dependency networks, all these networks obtain singnificantly positive clustering coefficients (significantly larger than a random network). Finally, these networks are scale-free and obtain node degrees from a power-law distribution expressed in Equation 2.7 with $\gamma \approx 3.00$.

These various studies show that there are statistical universals in different language networks, which are similar to those seen in other scale-free networks. This finding suggests that the commonalities are not artifacts of our analysis, but rather some abstract features of self-organization dynamics [189, 183]. An interesting conclusion of the small-world effect in human languages is made in [136]. They emphasized that human memory is associative (i.e., information is retrieved by connecting similar concepts) in which the small-world property of the network maximizes the retrieval efficiency. More precisely, high clustering of the network causes similar pieces of information to be stored together, and low shortest paths make very different pieces of information to be separated only by a few links, guaranteeing a fast search.

2.5 Language Regularities

Different human languages exhibit common markable regularities. Most of prior works have argued that these regularities and similarities are due to common vowels and consonants [105, 156, 70]. [34] model the occurrences of consonants across different languages using a complex network. They build a bipartite graph $G = \langle V_L, V_C, E \rangle$ in which a consonant from one part of the graph V_C is linked with a language in the second group V_L if the language contains that consonant. They analyze the degree distribution in each vertex group and show that the number of consonants in a language follows a β-distribution. That is, for $\alpha > 0$ and $\beta > 0$, the probability that a language has a fraction of x consonants out of all possible consonants is drawn from Equation 2.8, where Γ is the Euler's gamma function.

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{2.8}$$

This distribution is normally characterized by an asymmetric right-skewed distribution. Choudhury et al. [34] fit this model on the data and shows that the best fit

results in $\alpha = 7.06$ and $\beta = 47.64$.

The more interesting observation is about the degree distribution of V_C. Choudhury et al. [34] showed that the degree distribution in V_C (i.e., the distribution of the number of languages a consonant is used in) follows a power-law with two regimes. Two-regime power-law distributions are characterized by two separate power-laws that join at a break point, b. The two parts of the distribution (before and after the break-point) are obtain different exponents in the following equations.

$$p(k) \propto \begin{cases} k^{-\gamma_1} & \text{if } k \leq b \\ k^{-\gamma_2} & \text{if } k \geq b \end{cases}$$

Using empirical analysis, the authors in the mentioned reference suggest that the two exponents for their network are $\gamma_1 = 0.71$ and $\gamma_1 = 2.36$, where the two regimes join at $k = 21$. This simply means that there are two different power-laws. One power-law describes the number of consonants that appear in less than 21 languages (γ_1) and the other describes the number of consonants that appear in more than 21 (γ_2).

Similar work has proposed that vowel systems are the result of self-organization in a population of language users. A self-organization that emerges from language users trying to imitate each other [44, 43].

2.6 Emergent Behavior

The complex systems view of language use has opened new dimensions in explaining some of the properties seen in language. These properties that are not observable from individual language users are called emergent behavior. The emergent behavior in language use is often captured in terms of statistical regularities, regularities that help us develop a complete theory of language.

Zipf's Law

One of the features of natural languages that is widely studied is the statistical distribution of word abundances. Zipf was among the first who studied this property. Zipf's law [218, 217] states that in a corpus of natural language text, the frequency of a word is determined by a power-law function of its rank, r.

$$p(r) \propto r^{-\alpha} \tag{2.9}$$

Equation 2.9 is the simplest form of this representation where $\alpha = 1$. Similarly, we can present the Zipf's law as a function of the frequency f of a word.

$$p(f) \propto f^{-\beta} \tag{2.10}$$

Here $p(r)$ is the probability that a randomly selected word from the corpus is of rank r and $p(f)$ is the probability that a randomly selected word from the corpus has the frequency f. Prior work [57, 141] stated that β can be represented as a function of α of the form

$$\beta = \frac{1}{\alpha} + 1 \tag{2.11}$$

Many prior works have tried to understand the mechanics that results in Zipf's law [19, 58, 110]. Li [110] argued that the same distribution is seen when the corpus is not built using meaningful natural languages, but rather randomly generated text. His model of text generation is based on drawing random characters from a uniform distribution of English characters and the space character, thus each having the probability of 1/27 to be chosen. For instance, the probability of seeing ␣a␣ is $(1/27)^3$ and ␣abc␣ is $(1/27)^5$.

Li [110] argued that seeing a power-law function with an exponent close to 1 is mainly according to the transformation from the word's length to its rank, which

stretches an exponential function to a power-law function [110]. Such a view implies that the structure of the natural language has no effect in the emergence of Zipf's law. However, Ferrer i Cancho and Solé [58] discussed that the distributions seen from such random text are very different from those seen in natural languages. They repeated Li's experiment but replaced the uniform distribution of letters with the frequency of the letters in Herman Melville's Moby Dick. They drew random letters and the space character with the probability proportional to the letter's frequency in Moby Dick. They showed that the new word frequency distributions are more realistic and similar to the ones seen in human written text.

Zipf's law is not limited to English word frequencies in text corpora. Ellis and Ferreira [51] observed the Zipfian distribution in English verb-argument constructions (VACs): verb locative (VL), verb object locative (VOL), and ditransitive (VOO). It has also been observed in non-English languages such as Chinese [176], Greek [75], and Turkish [41].

Gelbukh and Sidorov [66] argued that the exponent of the Zipf's law depends on language. In their work, they processed 39 literature texts for English, Russian, and Spanish chosen randomly from different genres. They showed that the Zipf's exponent is 0.97 for English, and .89 for Russian, with Spanish between the two. They suggested that this is due high "inflectivity" of Russian compared to English, while English is more analytical than Russian. They generalized their finding and conclude that Zipf's law's exponent depends on language.

Zipf's law is observed as an emergent behavior in many other complex systems. It has been used to explain the distribution cities [64], firm sizes [9], scientific citations [171], gene expressions [63], family names [10] and many more.

Heaps' Law

In linguistics, Heaps' law explains the growth of vocabulary represented in a set of documents. This law is normally formulated as shown in Equation 2.12, where V is the size of the vocabulary built after seeing n words in a corpus [79].

$$V \propto n^{\lambda} \tag{2.12}$$

Practically, λ takes a value close to 0.5 in English text corpora. Similar to their work on Zipf's law, Gelbukh and Sidorov [66] also performed experiments to show that the value of the exponent in Heaps' law depends on language.

Some prior work has attempted to show that Heaps' law and Zipf's law are related. Baeza-Yates and Navarro [11] showed that when Zipf's exponent is greater than 1 ($\alpha > 1$) then Heap's exponents (λ) is

$$\lambda = \frac{1}{\alpha}. \tag{2.13}$$

Lü et al. [119] argued that the relation in Equation 2.13 is only an asymptotic solution holds for very-large-size systems with $\alpha > 1$. They refine this result for finite-size systems with $\alpha > 1$ and complement it with $\alpha < 1$.

$$\lambda \propto \begin{cases} \frac{1}{\alpha} & \alpha > 1 \\ 1 & \alpha < 1 \end{cases}$$

2.3 Diversity of Perspectives

In prior work on evaluating independent contributions in content generation, [201] studied IR systems and showed that relevance judgments differ significantly between humans but relative rankings show high degrees of stability across annotators. However, perhaps the closest work to this thesis is [199] in which 40 Dutch students and

10 NLP researchers were asked to summarize a BBC news report, resulting in 50 different summaries. Teufel and van Halteren also used 6 DUC[3]-provided summaries, and annotations from 10 student participants and 4 additional researchers, to create 20 summaries for another news article in the DUC datasets. They calculated the Kappa statistic [29, 101] and observed high agreement, indicating that the task of atomic semantic unit (factoid) extraction can be robustly performed in naturally occurring text, without any copy-editing.

The diversity of perspectives and the unprecedented growth of the factoid inventory also affects evaluation in text summarization. Evaluation methods are either extrinsic, in which the summaries are evaluated based on their quality in performing a specific task [184] or intrinsic where the quality of the summary itself is evaluated, regardless of any applied task [198, 143]. These evaluation methods assess the information content in the summaries that are generated automatically.

Leveraging the diverse range of perspectives has also played a critical role in developing new paraphrase generation systems by providing massive amounts of data that is easily collectable. For instance, Chen and Dolan [31] performed a study and collected highly parallel data, used for training paraphrase generation systems from descriptions that participants wrote for video segments from YouTube. Such parallel corpora of document pairs that represent the same semantic information in different languages have also been extracted from user contributions in Wikipedia and been used for learning translations of words and phrases [212].

In automatic text summarization, a number of previous methods have focused on diversity. Mei et al. [123] introduced a diversity-focused ranking methodology based on reinforced random walks in information networks. Their random walk

[3]Document Understanding Conference

model introduces the rich-gets-richer mechanism to PageRank with reinforcements on transition probabilities between vertices. Mei et al. employ this ranking algorithm on TF-IDF based similarities of sentences, rank documents in the DUC 2004 datasets for summarization, and show improvements in the task.

A similar ranking model is the *Grasshopper* ranking model [216], which leverages an absorbing random walk. This model starts with a regular time-homogeneous random walk, and in each step the node with the highest weight is set as an absorbing state. The multi-view point summarization of opinionated text is discussed in [155]. Paul et al. introduced *Comparative LexRank*, based on the LexRank ranking model [55]. Their random walk formulation is to score sentences and pairs of sentences from opposite viewpoints (clusters) based on both their representativeness of the collection as well as their contrastiveness with each other. Once a lexical similarity graph is built, they modify the graph based on cluster information and perform LexRank on the modified cosine similarity graph.

The most well-known paper that address diversity in summarization is [28], which introduces Maximal Marginal Relevance (MMR). This method is based on a greedy algorithm that picks sentences in each step that are the least similar to the summary so far. There are a few other diversity-focused summarization systems like C-LexRank [159], which employs document clustering. These works have tried to increase diversity in summarizing documents, but do not explain the type of the diversity in their inputs. In this chapter, we give an insightful discussion on the nature of the diversity seen in collective discourse, and will explain why some of the mentioned methods may not work under such environments.

Lin and Hovy [114] discussed the evaluations of summaries on the Document Understanding Conference 2001 data. Although they showed that more than one

gold standard summary is needed to evaluate a system summary, they did not give an estimate on the number of such summaries to achieve a reliable evaluation.

Finally, recent research on analyzing online social media shown a growing interest in mining news stories and headlines because of its broad applications ranging from "meme" tracking and spike detection [108] to text summarization [15]. In similar work on blogs, it is shown that detecting topics [103, 2] and sentiment [152] in the blogosphere can help identify influential bloggers [3, 86] and mine opinions about products [133]. These methods have demonstrated the utility of harnessing the wisdom of disparate crowds in identifying current events and the sentiment they generate.

However, previous research has not addressed the coverage of such distributed contents, whether the marginal information gain with each additional contribution asymptotes and how such contents can best be summarized.

2.4 Lexical Networks

Several properties of lexical networks have been analyzed before [57, 59]. Steyvers and Tenenbaum [189] examined free association networks, WordNet, and the Roget Thesaurus, and noted five different properties in semantic networks. They were all sparse, with a giant component, a small shortest-path length, a high clustering-coefficient, and a scale-free degree distribution.

The evolution of lexical networks over time has also been studied in [48, 27]. These studies found that the resulting network for a text corpus exhibited small-world properties in addition to a power-law degree distribution.

It has been noted that although the standard growth models based on preferential attachment fit the degree distribution of the world wide web and citation networks,

they fail to accurately model the cosine distribution of the linked documents. A mixture model for cosine distribution of linked documents is proposed in [125], which combines preferential attachment with cosine similarity. This model makes use of the idea that authors don't just link to the common pages on the web, but also take into account the content of these pages. Authors tend to link to and cite articles that are related to their own content. Menczer's model generates networks that reproduce the same degree distribution and content distribution of real-world information networks. More formally he showed that at each step t one new page t is added, and m links are created from t to m existing pages. Each of these m pages are selected from $\{i, i < t\}$ with a probability,

$$Pr(i,t) = \begin{cases} \frac{k(i)}{mt} & \text{if } \sigma_c(i,t) > \kappa^* \\ c\sigma_c^{\gamma}(i,t) & \text{otherwise} \end{cases}$$

where $\sigma_c(i,t)$ is the content similarity of two pages i, t, and m, κ^* are constants derived from the data and c is a normalization factor.

They also generate networks by simulating the Open Directory Project (DMOZ) network and a collection of articles published in the Proceedings of the National Academy of Sciences (PNAS). Their results show that their model not only fits the degree distribution, but it fits the similarity distribution, where the probability of a node to be linked is

$$Pr(i) = \alpha\frac{k(i)}{mt} + (1-\alpha)\bar{Pr}(i)$$

where $i < t$ and $\alpha \in [0,1]$ is a preferential attachment parameter.

In the degree-uniform mixture model we have, $\bar{Pr}(i) = \frac{1}{t}$ but a degree-similarity mixture model uses content similarity and results in

$$\bar{Pr}(i) \propto \left(\frac{1}{\sigma_c(i,t)} - 1\right)^{-\gamma}$$

where γ is a constant.

Finally, graph based techniques have been used for other applications in NLP such as summarization [55], and summary evaluation [153]. Finally, Lexrank [55] uses a network representation for multi-document summarization. To do so, it builds a lexical network, in which nodes are sentences and a weighted edge between two nodes shows the lexical similarity.

2.5 Graph-based Summarization Methods

As a representative of graph-based methods applied to summarization, LexRank [55] constructs a graph whose vertices are sentences from all the documents in a cluster. The graph is characterized by a sentence connectivity matrix representing the Markov transition probabilities among vertices. Sentences of high centralities are then selected to form the summary. C-LexRank [159] extended the framework by incorporating community clustering to address the need of covering different aspects of contributions in a scientific work.

Motivated by the similar idea of applying PageRank and HITS [99] on graphs of sentences, Mihalcea and Tarau [127] presented TextRank, a system for keyword extraction and sentence extraction, and successfully apply it to producing extractive summaries. The system is proved scalable to multi-document summarization tasks, and also language-independent [126].

More recent work has integrated link analysis and other techniques as re-ranking to improve the effectiveness of summarization based on graph-based ranking. Wan and Yang [203] incorporated *information richness* and *information novelty* into the criteria of selecting important sentences. These two parameters are determined by a sentence affinity graph reflecting the semantic relationships between sentences.

They also distinguished between intra-document and inter-document links, biasing the latter for information richness computation.

Another optimization ClusterCMRW (and ClusterHITS) proposed in [204] assumes that a given document set covers a few topic themes or subtopics that are of different degrees of importance. The idea of clustering sentences according to subtopics is comparable to C-LexRank. Designed for summarizing scientific contributions, C-LexRank looks for a comprehensive coverage of each subtopic or contribution aspect, while ClusterCMRW (and ClusterHITS) focus at ranking on the cluster level, so that sentence centralities are scaled by the centralities of the clusters in which they belong.

2.6 Citation Analysis

Previous work has analyzed citation and collaboration networks [194, 144] and scientific article summarization [193]. Bradshaw [23, 24] benefited from citations to determine the content of articles and introduce "Reference Directed Indexing" to improve the results of a search engine. Nanba et al. [139, 138] analyzed citation sentences and automatically categorized citations into three groups using 160 pre-defined phrase-based rules. This categorization was then used to build a tool to help researchers analyze citations and write scientific summaries. Nanba and Okumura [140] also discussed the same citation categorization to support a system for writing a survey. They [140, 139] reported that co-citation implies similarity by showing that the textual similarity of co-cited papers is proportional to the proximity of their citations in the citing article.

Previous work has shown the importance of the citation sentences in understanding scientific contributions. Elkiss et al. [50] performed a large-scale study on cita-

tions and their importance. They conducted several experiments on a set of 2,497 articles from the free PubMed Central (PMC) repository[4] and 66 from ACM digital library. Results from this experiment confirmed that the average cosine between sentences in the set of citations to an article is consistently higher than that of its abstract. They also showed that this number is reported to be much greater than the average cosine of the citation sentences with that of a randomly chosen document, and so is for the abstract. Finally, they concluded that the content of citing sentences has much greater uniformity than the content of the corresponding abstract, implying that citations are more focused and contain additional information that does not appear in abstracts.

Nakov and Hearst [137] performed a detailed manual study of citations in the area of molecular interactions and found that the set of citations to a given target paper cover most information found in the abstract of that article, as well as 20% more concepts, mainly related to experimental procedures.

Kupiec et al. [104] used the abstracts of scientific articles as a target summary. They used 188 Engineering Information summaries that are mostly indicative in nature. Kan et al. [90] used annotated bibliographies to cover certain aspects of summarization and suggest guidelines that summaries should also include metadata and critical document features as well as the prominent content-based features.

Siddharthan and Teufel [180] described a new reference task and show high human agreement as well as an improvement in the performance of *argumentative zoning* [192]. In argumentative zoning—a rhetorical classification task—seven classes (Own, Other, Background, Textual, Aim, Basis, and Contrast) are used to label sentences according to their role in the author's argument.

[4]http://www.pubmedcentral.gov

Athar [7] addressed the problem of identifying positive and negative sentiment polarity in citations to scientic papers. Similarly, Athar and Teufel [8] used context-enriched citations to classify scientific sentiment towards a target paper.

Little work has been done on automatic citation extraction from research papers. Kaplan et al. [91] introduced "citation-site" as a block of text in which the cited text is discussed. The mentioned work used a machine learning method for extracting citations from research papers and evaluates the result using an annotated corpus of 38 papers citing 4 articles.

CHAPTER III

Diversity of Perspectives in Collective Discourse

In this chapter, we analyze collective discourse, a collective human behavior in content generation, and show that it exhibits diversity, a property of general collective systems. Using extensive analysis, we propose a novel paradigm for designing summary generation systems that reflect the diversity of perspectives seen in real-life collective summarization. We analyze 50 sets of summaries written by human about the same story or artifact and investigate the diversity of perspectives across these summaries. We show how different summaries use various phrasal information units (i.e., *nuggets*) to express the same atomic semantic units, called *factoids*. We believe that out experiments will give insight into new models of text generation, which is aimed at modeling the process of producing natural language texts, and is best characterized as the process of making choices between alternate linguistic realizations, also known as lexical choice [49, 14, 185].

3.1 Data Annotation

The datasets used in our experiments represent two completely different categories: news headlines, and scientific citation sentences. The *headlines* datasets consist of 25 clusters of news headlines collected from Google News[1], and the *cita-*

[1] news.google.com

tions datasets have 25 clusters of citations to specific scientific papers from the ACL Anthology Network (AAN)[2]. Each cluster consists of a number of unique summaries (headlines or citations) about the same artifact (non-evolving news story or scientific paper) written by different people. Table 3.1 lists some of the clusters with the number of summaries in them.

This table indicates that the cluster size ranges from 10 to more than 100 headlines per cluster. For instance the largest cluster (size=125) is about "Miss Venezuela wins miss universe 2009", and the smallest cluster represents headlines about "Yale lab tech in court" with only 10 headlines on Google.

ID	type	Name	Story/Title	#
1	hdl	miss	Miss Venezuela wins miss universe'09	125
2	hdl	typhoon	Second typhoon hit philippines	100
3	hdl	russian	Accident at Russian hydro-plant	101
4	hdl	redsox	Boston Red Sox win world series	99
5	hdl	gervais	"Invention of Lying" movie reviewed	97
	...	...	...	
25	hdl	yale	Yale lab tech in court	10
26	cit	N03-1017	Statistical Phrase-Based Translation	172
27	cit	P02-1006	Learning Surface Text Patterns ...	72
28	cit	P05-1012	On-line Large-Margin Training ...	71
29	cit	C96-1058	Three New Probabilistic Models ...	66
30	cit	P05-1033	A Hierarchical Phrase-Based Model ...	65
	...	...	...	
50	cit	H05-1047	A Semantic Approach to Recognizing ...	7

Table 3.1: Some of the annotated datasets and the number of summaries in each of them (hdl = headlines; cit = citations)

3.1.1 Nuggets vs. Factoids

We define an annotation task that requires explicit definitions that distinguish between phrases that represent the same or different information units. Unfortunately, there is little consensus in the literature on such definitions. Therefore, we follow [198] and make the following distinction. We define a *nugget* to be a phrasal information unit. Different nuggets may all represent the same atomic semantic unit,

[2]http://clair.si.umich.edu/clair/anthology/

which we call as a *factoid*. In the following headlines, which are randomly extracted from the `redsox` dataset, nuggets are manually underlined.

red sox win 2007 world series

boston red sox blank rockies to clinch world series

boston fans celebrate world series win; 37 arrests reported

These 3 headlines contain 9 nuggets, which represent 5 factoids or classes of equivalent nuggets.

f_1 : {red sox, boston, boston red sox}

f_2 : {2007 world series, world series win, world series}

f_3 : {rockies}

f_4 : {37 arrests}

f_5 : {fans celebrate}

The first headline indicates that the "red sox" are the winning team and have become the champion of the "2007 world series". The second headline mentions that the red sox beat colorado "rockies" to win the world series. The last headline covers other aspects of the story that are absent in the other two: "celebrations" and "arrests". This example suggests that different headlines on the *same story* written independently of one another use different phrases (nuggets) to refer to the same semantic unit (e.g., "red sox" vs. "boston" vs. "boston red sox") or to semantic units corresponding to different aspects of the story (e.g., "37 arrests" vs. "rockies"). In the former case different nuggets are used to represent the same factoid, while in the latter case different nuggets are used to express different factoids. This analogy is similar to the definition of factoids in [199].

The following citation sentences to Koehn's work suggest that a similar phenomenon also happens in citations.

We also compared our model with pharaoh (Koehn et al, 2003).

Koehn et al (2003) find that phrases longer than three words improve performance little.

Koehn et al (2003) suggest limiting phrase length to three words or less.

For further information on these parameter settings, confer (koehn et al, 2003).

where the first author mentions "pharaoh" as a contribution of Koehn et al, but the second and third use different nuggets to represent the same contribution: use of trigrams. However, as the last citation shows, a citation sentence, unlike news headlines, may cover no information about the target paper.

The use of phrasal information as nuggets is an essential element to our experiments, since some headline writers often try to use uncommon terms to refer to a factoid. For instance, two headlines from the `redsox` cluster are:

Short wait for bossox this time

Soxcess started upstairs

The use of different nuggets to represent the same factoid is also seen in smaller datasets. Four randomly selected headlines from the *babies* dataset are as follows.

most babies born this century will live to 100

today's babies are tomorrow's centenarians

100 candles on birthday cake to become common event

most babies in rich nations to see 100

These headlines result in 3 factoids as below.

f_1: {live to 100, tomorrow's centenarians, 100 candles on birthday cake, to see 100}

f_2: {babies born this century, today's babies, most babies}

f_3: {in rich nations}

Following these examples, we asked two annotators to annotate all 1,390 headlines, and 926 citations. The annotators were asked to follow precise guidelines in nugget extraction. Our guidelines instructed annotators to extract non-overlapping phrases from each headline as nuggets. Therefore, each nugget should be a substring of the headline that represents a semantic unit[3].

Previously Lin and Hovy [114] have shown that information overlap judgment is a difficult task for human annotators. To avoid such a difficulty, we enforced our annotators to extract non-overlapping nuggets from a summary to make sure that they are mutually independent and that information overlap between them is minimized.

Finding agreement between annotated well-defined nuggets is straightforward and can be calculated in terms of Kappa. However, when nuggets themselves are to be extracted by annotators, the task becomes less obvious. To calculate the agreement, we annotated 10 randomly selected headline clusters twice and designed a simple evaluation scheme based on Kappa[4]. For each n-gram, w, in a given headline, we look if w is part of any nugget in either human annotations. If w occurs in both or neither, then the two annotators agree on it, and otherwise they do not. Based on this agreement setup, we can formalize the κ statistic as $\kappa = \frac{\Pr(a)-\Pr(e)}{1-\Pr(e)}$ where $Pr(a)$ is the relative observed agreement among annotators, and $Pr(e)$ is the probability that annotators agree by chance if each annotator is randomly assigning categories.

[3]Before the annotations, we lower-cased all summaries and removed duplicates

[4]Previously Qazvinian and Radev [160] have shown high agreement in human judgments in a similar task on citation annotation

	Average κ		
	unigram	bigram	trigram
Human1 vs. Human2	0.76 ± 0.4	0.80 ± 0.4	0.89 ± 0.3

Table 3.2: Agreement between different annotators in terms of average Kappa in 25 headline clusters.

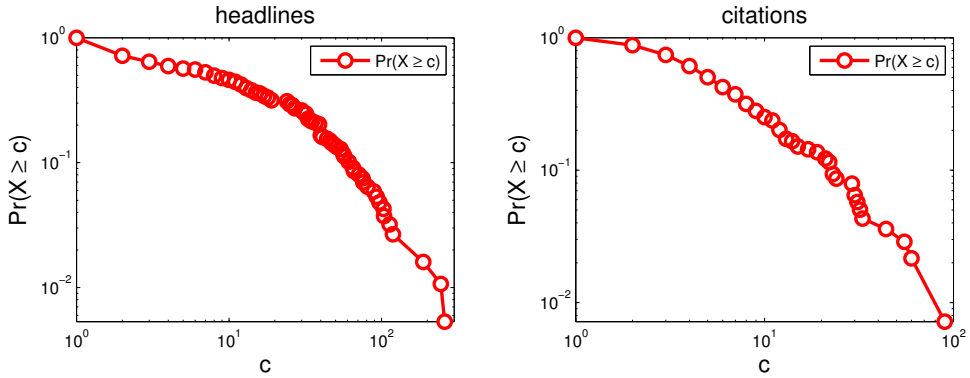


Figure 3.1: The cumulative probability distribution for the frequency of factoids (i.e., the probability that a factoid will be mentioned in c different summaries) across in each category.

Table 3.2 shows the unigram, bigram, and trigram-based average κ between the two human annotators (**Human1**, **Human2**). These results suggest that human annotators can reach substantial agreement when bigram and trigram nuggets are examined, and has reasonable agreement for unigram nuggets.

3.2 Diversity

We study the diversity of ways with which human summarizers talk about the same story or event and explain why such a diversity exists.

3.2.1 Skewed Distributions

Our first experiment is to analyze the popularity of different factoids. For each factoid in the annotated clusters, we extract its count, X, which is equal to the number of summaries it has been mentioned in, and then we look at the distribution of X. Figure 3.1 shows the cumulative probability distribution for these counts (i.e., the probability that a factoid will be mentioned in at least c different summaries) in

both categories.

These highly skewed distributions indicate that a large number of factoids (more than 28%) are only mentioned once across different clusters (e.g., "poor pitching of colorado" in the `redsox` cluster), and that a few factoids are mentioned in a large number of headlines (likely using different nuggets). The large number of factoids that are only mentioned in one headline indicates that different summarizers increase diversity by focusing on different aspects of a story or a paper. The set of nuggets also exhibit similar skewed distributions. If we look at individual nuggets, the `redsox` set shows that about 63 (or 80%) of the nuggets get mentioned in only one headline, resulting in a right-skewed distribution.

The factoid analysis of the datasets reveals two main causes for the content diversity seen in headlines: (1) writers focus on different aspects of the story and therefore write about different factoids (e.g., "celebrations" vs. "poor pitching of colorado"). (2) writer use different nuggets to represent the same factoid (e.g., "redsox" vs. "bosox"). In the following sections we analyze the extent at which each scenario happens.

3.2.2 Factoid Inventory

The emergence of diversity in covering different factoids suggests that looking at more summaries will capture a larger number of factoids. In order to analyze the growth of the factoid inventory, we perform a simple experiment. We shuffle the set of summaries from all 25 clusters in each category, and then look at the number of unique factoids and nuggets seen after reading n^{th} summary. This number shows the amount of information that a randomly selected subset of n writers represent. This is important to study in order to find out whether we need a large number of summaries to capture all aspects of a story and build a complete factoid inventory.

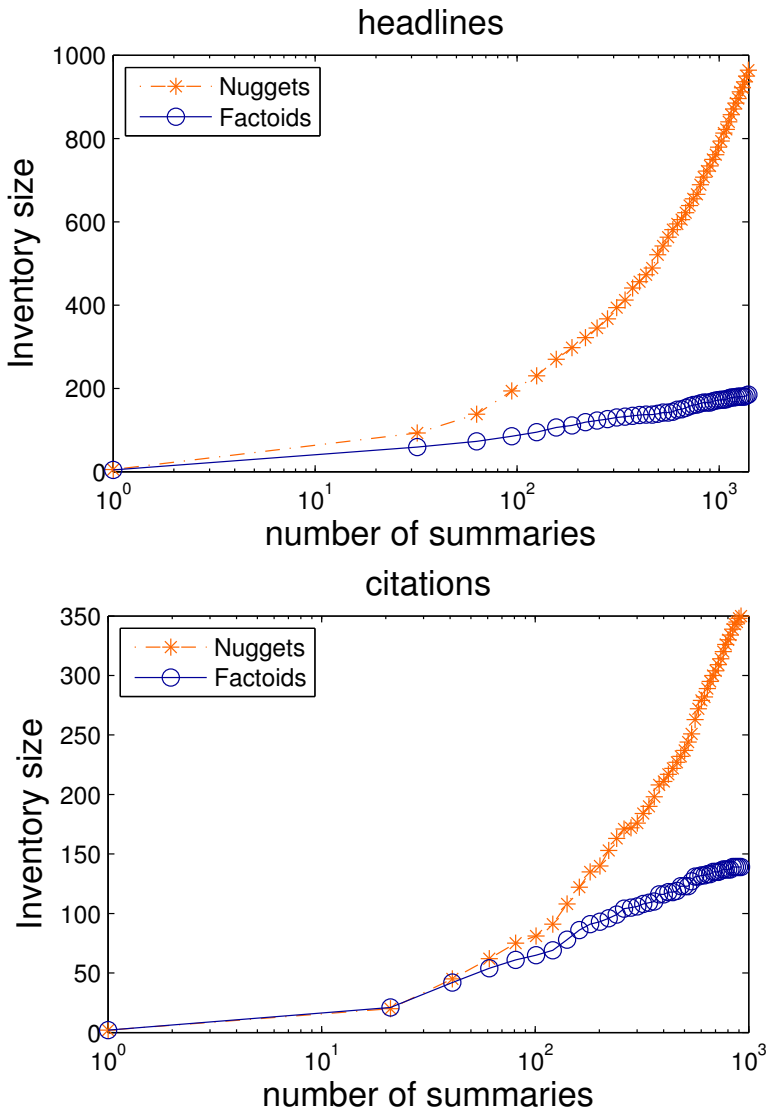


Figure 3.2: The number of unique factoids and nuggets observed by reading n random summaries in all the clusters of each category

The plot in Figure 3.2.1 shows, at each n, the number of unique factoids and nuggets observed by reading n random summaries from the 25 clusters in each category. These curves are plotted on a semi-log scale to emphasize the difference between the growth patterns of the nugget inventories and the factoid inventories[5].

This finding numerically confirms a similar observation on human summary annotations discussed in [198, 199]. In their work, van Halteren and Teufel indicated

[5]Similar experiment using individual clusters exhibit similar behavior

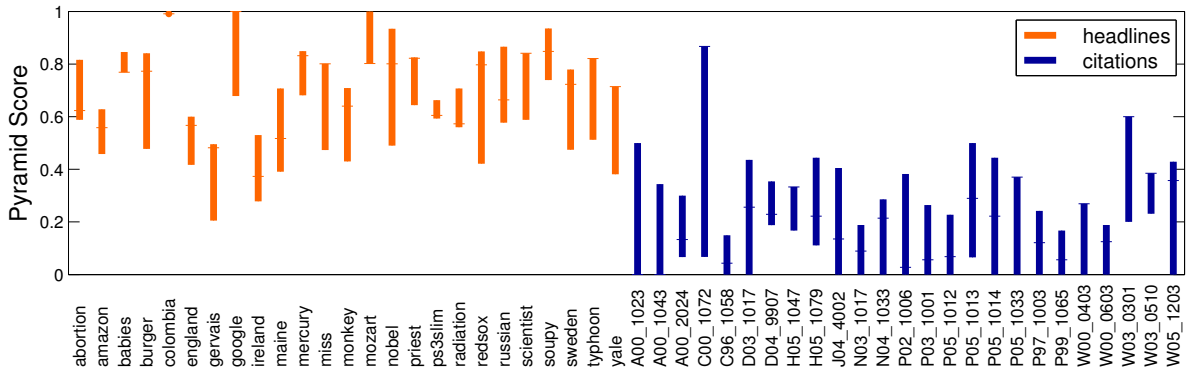


Figure 3.3: The 25th to 75th percentile pyramid score range in individual clusters

that more than 10-20 human summaries are needed for a full factoid inventory. However, our experiments with nuggets of nearly 2,400 independent human summaries suggest that neither the nugget inventory nor the number of factoids will be likely to show asymptotic behavior. However, these plots show that the nugget inventory grows at a much faster rate than factoids. This means that a lot of the diversity seen in human summarization is a result of the so called different *lexical choices* that represent the same semantic units or factoids.

3.2.3 Summary Quality

In previous sections we gave evidence for the diversity seen in human summaries. However, a more important question to answer is whether these summaries all cover important aspects of the story. Here, we examine the quality of these summaries, study the distribution of information coverage in them, and investigate the number of summaries required to build a complete factoid inventory.

The information covered in each summary can be determined by the set of factoids (and not nuggets) and their frequencies across the datasets. For example, in the `redsox` dataset, "red sox", "boston", and "boston red sox" are nuggets that all represent the same piece of information: the red sox team. Therefore, different

summaries that use these nuggets to refer to the red sox team should not be seen as very different.

We use the Pyramid model [143] to value different summary factoids. Intuitively, factoids that are mentioned more frequently are more salient aspects of the story. Therefore, our pyramid model uses the normalized frequency at which a factoid is mentioned across a dataset as its weight. In the pyramid model, the individual factoids fall in tiers. If a factoid appears in more summaries, it falls in a higher tier. In principle, if the term w_i appears $|w_i|$ times in the set of headlines it is assigned to the tier $T_{|w_i|}$. The pyramid score that we use is computed as follows. Suppose the pyramid has n tiers, T_i, where tier T_n is the top tier and T_1 is the bottom. The weight of the factoids in tier T_i will be i (i.e. they appeared in i summaries). If $|T_i|$ denotes the number of factoids in tier T_i, and D_i is the number of factoids in the summary that appear in T_i, then the total factoid weight for the summary is $D = \sum_{i=1}^{n} i \times D_i$. Additionally, the optimal pyramid score for a summary is $Max = \sum_{i=1}^{n} i \times |T_i|$. Finally, the pyramid score for a summary can be calculated as

$$P = \frac{D}{Max}$$

Based on this scoring scheme, we can use the annotated datasets to determine the quality of individual headlines. First, for each set we look at the variation in pyramid scores that individual summaries obtain in their set. Figure 3.3 shows, for each cluster, the variation in the pyramid scores (25th to 75th percentile range) of individual summaries evaluated against the factoids of that cluster. This figure indicates that the pyramid score of almost all summaries obtain values with high variations in most of the clusters For instance, individual headlines from `redsox` obtain pyramid scores as low as 0.00 and as high as 0.93. This high variation confirms the previous observations on diversity of information coverage in different summaries.

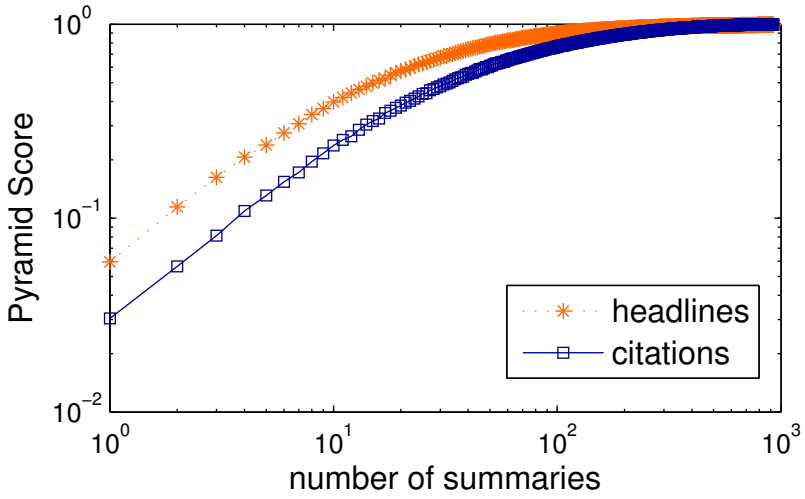


Figure 3.4: Average pyramid score obtained by reading n random summaries shows rapid asymptotic behavior.

Additionally, this figure shows that headlines generally obtain higher values than citations when considered as summaries. One reason, as explained before, is that a citation may not cover any important contribution of the paper it is citing, when headlines generally tend to cover some aspects of the story.

High variation in quality means that in order to capture a larger information content we need to read a greater number of summaries. But how many headlines should one read to capture a desired level of information content? To answer this question, we perform an experiment based on drawing random summaries from the pool of all the clusters in each category. We perform a Monte Carlo simulation, in which for each n, we draw n random summaries, and look at the pyramid score achieved by reading these headlines. The pyramid score is calculated using the factoids from all 25 clusters in each category[6]. Each experiment is repeated $1,000$ times to find the statistical significance of the experiment and the variation from the average pyramid scores.

Figure 3.4 shows the average pyramid scores over different n values in each cate-

[6]Similar experiment using individual clusters exhibit similar results

gory on a log-log scale. This figure shows how pyramid score grows and approaches 1.00 rapidly as more randomly selected summaries are seen.

3.3 Other Collective Discourse Datasets

In this section, we extend our analysis from citations and news headlines to a larger set of collective discourse examples.

3.3.1 Annotations

In addition to citations and headlines, we collect and annotate the following datasets: movie reviews to the same movies from imdb.com, and tweets about the same stories from Twitter.com. For each collective discourse dataset described in Section 1.1, we construct the set of factoids that represent various aspects of a story or a movie or different contributions of a paper.

For the microblogs dataset, we asked two annotators to go over all the tweets and identify a set of factoids that represent different aspects of each rumor. We then manually marked each tweet with the factoid that is relevant to the tweet. Each factoid is usually covered by a number of tweets, and each tweet covers one or more factoids. However, we did not observe any tweets that cover more than 2 factoids in our datasets. The small number of factoids covered by each tweet is most likely due to the length limit enforced by Twitter on each post.

Table 3.3 lists the factoids extracted from the Sarah Palin.s divorce rumor dataset. This table shows that the 414 tweets discuss how "Facebook is used to debunk the rumor," while the "libel suit against the blogger who started the rumor" is only mentioned in 24 tweets of the total 789 tweets.

To calculate the inter-judge agreement, we annotated 100 microblog instances on

Factoid	#tweets	Perspective description
FB	414	debunked on Facebook
FAMILY	106	family values
ALASKA	87	Alaska report's evidence
QUIT	72	resignation and divorce
AFFAIRS	58	affairs
GAY	36	gay marriage ban
CAMP	36	her camp denies the rumor
MONTANA	33	moving to Montana
LIBEL	24	libel suit against the rumor
BLOG	19	blogger who started the rumor

Table 3.3: Different factoids extracted from the Palin dataset with the number of tweets that mention them, and short descriptions.

Sarah Palin twice, and calculated the statistic as

$$\kappa = \frac{Pr(a) - Pr(\epsilon)}{1 - Pr(\epsilon)}$$

where $Pr(a)$ is the relative observed agreement among the two annotators on the 10 factoids from Table 3.3, and $Pr(\epsilon)$ is the probability that annotators agree by chance if each annotator is randomly assigning categories. Based on this formulation, we reach a value of 0.913 in κ, and 93% agreement between the two annotators. Previously we showed high agreement in human judgments for extracting factoids from other datasets such as news headlines and citations ($\kappa \approx 0.8$) [162].

For the movie review clusters, we downloaded the list of cast names as well as the list of plot keywords provided for each movie by IMDB, as the set of factoids about the movie[7].

Table 3.4 lists the average number of factoids for each collective discourse corpus. For the Movie reviews, there is an average of 131 factoids per movie, and for citations, headlines and microblogs, our annotators identify an average of 5, 7, and 3 factoids respectively.

[7]We admit that the set of cast names and plot keywords provided by IMDB does not include all the factoids about the movie. However, since creating gold standard data from complete user reviews is fairly arduous, and we did not pursue manual annotations for movies.

Dataset	Number of factoids
Movie reviews	131.31 ± 52.67
Microblogs	2.93 ± 2.05
News headlines	7.48 ± 4.02
Citations	5.48 ± 1.96

Table 3.4: Average number of factoids in various collective discourse corpora.

3.3.2 Diversity

Surow [190] defines 4 criteria for a crowd to be wise: (1) people in the crowd should have diverse knowledge of facts (diversity); (2) people should act independently and their opinion should not be affected by that of others (independence); (3) people should have access to local knowledge (decentralization); and (4) a mechanism should exist to turn individual judgments into collective intelligence (aggregation).

Previously, we showed that citation and headline datasets exhibit skewed factoid distributions and diversity. Here, we present evidence that the individuals who engage in collective discourse in other examples also have diverse perspectives and interpret things differently.

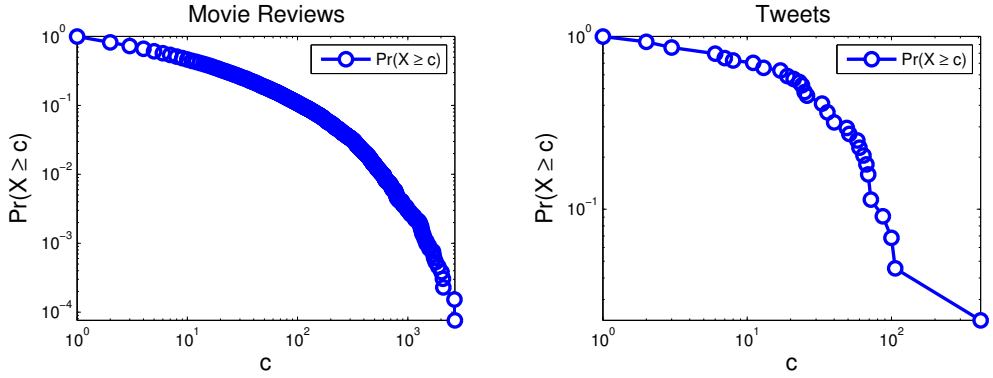


Figure 3.5: The cumulative probability distribution for the frequency of factoids (i.e., the probability that a factoid will be mentioned in c different summaries) across in each corpus.

Novelty and Redundancy

To investigate the diversity of perspectives, we look at the frequency distribution of various factoids in different corpora by extracting the number of individuals that mention each factoid, f, in the annotated clusters. Figure 3.5 shows the log-log scale cumulative probability distribution for these counts (i.e., the probability that a factoid will be mentioned by at least c different people) in the movie review and microblog corpora. This figure suggests that factoid mention frequencies exhibit a highly skewed distribution with many factoids mentioned only once and a very few factoids mentioned by a large number of people. For instance, in the Pulp Fiction example, "Bruce Willis" and "Quentin Tarantino" are very popular factoids and most reviewers mention them, while "Rene Beard", "Frank Whaley" (two other actors), or "Jackson's bible quote" are among many factoids that are not as frequently mentioned.

3.4 Small-world of Factoids

Recent research has shown that a wide range of natural graphs such as the biological networks [170], food webs [135], electronic circuits [56], brain neurons [16], and human languages [57] exhibit the small-world property. This common characteristic can be detected from two basic statistical properties: the clustering coefficient $\mathcal{C}$, and the average shortest path length ℓ.

The clustering coefficient of a graph measures the number of closed triangles in the graph. The clustering coefficient describes how likely it is that two neighbors of a vertex are connected [146]. Watts and Strogatz [207] define the clustering coefficient as the average of the local clustering values for each vertex.

$$\mathcal{C} = \frac{\sum_{i=1}^{n} c_i}{n}$$

The local clustering coefficient, c_i for the ith vertex is the number of triangles connected to vertex i divided by the total possible number of triangles connected to vertex. Watts and Strogatz [207] show that small-world networks are highly clustered and obtain relatively short paths (i.e., ℓ is small). These networks are usually studied in contrast with random networks in which both ℓ and $\mathcal{C}$ obtain small values.

To understand the relationship between various aspects of a story or subject and to study the relationship between different individuals' contributions, we analyze the network of factoids.

For each dataset, we build a network in which nodes represent different factoids and there is an edge between two nodes if the corresponding factoids have been mentioned together in at least 10 documents. Using these networks, we would like to investigate whether there are many factoid pairs that co-occur in individual user contributions, and whether there are communities of factoids that co-occur more frequently than others. For each network, we use the same number of nodes and edges and generate a random network using the Erdös–Rényi model, which sets an edge between each pair of nodes with equal probability, independently of the other edges [53].

Table 3.5 lists the average clustering coefficient ($\mathcal{C}$) and the average shortest path length (ℓ) in the networks built using factoid co-occurrences. This table confirms that the clustering coefficient in the factoid networks is generally significantly greater than random networks of the same size. Moreover, this table confirms that the average shortest paths in the random networks are small.

Ferrer i Cancho and Solé [57] and Motter et al., [136] perform similar experiments

$\mathcal{C}$	$\mathcal{C}_{random}$	ℓ	ℓ_{random}
0.814	0.072	1.613	2.627

Table 3.5: Average clustering coefficient ($\mathcal{C}$) and the average shortest path length (ℓ) in the networks of the collective discourse corpora and the corresponding random networks.

and show that the word co-occurrence and word synonymy networks have small-world properties. However, we believe that this is the first work that shows the small-world effect in human language at the factoid level (network of concepts). This finding further justifies the conclusion made by [136], who emphasize that human memory is associative (i.e., information is retrieved by connecting similar concepts) in which the small-world property of the network maximizes the retrieval efficiency. More precisely, high clustering of the network causes similar pieces of information to be stored together, and low shortest paths make very different pieces of information to be separated only by a few links, guaranteeing a fast search.

3.5 Wise Crowds

Previous work has studied crowd wisdom in online content contributions. Wikipedia for instance, has been named as an example of a successful collective effort. Kittur et al., [96] study user contributions in Wikipedia and suggest that the main workload is Wikipedia is driven by "common" users and that the admin influence has dramatically decreased over years. Furthermore, Kittur [97] show that adding more editors to an article results in higher article quality when appropriate coordination techniques are used. In this section, we present some evidence of wisdom in collective discourse that is not achievable from individuals or from smaller groups. In our experiments, we try to answer a simple question about a movie just by using its set of reviews.

The question we try to answer is to find each movie's genre. As the gold standard,

Rank	Genre	S_g	relevance
1	**action**	0.241	1
2	**sci-fi**	0.124	1
3	war	0.105	0
4	**fantasy**	0.087	1
5	history	0.086	0
6	animation	0.062	0
7	**adventure**	0.051	1
8	romance	0.039	0
9	drama	0.025	0
10	family	0.023	0

Table 3.6: Top 10 genres extracted for the movie "Avatar" from user reviews.

we collected the genres for each of the 100 movies for which we had user reviews. Each movie is associated with a few (3-4) genres out of a total of 19 genre names.

To extract the list of possible genres for a movie, we match all the genre names against the reviews and rank them based on their relative frequency. More particularly, the score of each genre, g for a movie with N reviews ($D_1 ... D_N$) is calculated as

$$S_g = \frac{\sum_{i=1}^{N} \mathbf{1}_{D_i \text{ mentions } g}}{N}$$

Table 3.6 lists the top 10 genres retrieved for the movie "Avatar" form user reviews together with the score of each genre and the relevance according to the gold standard that we obtained from IMDB. This table shows an example in which all the 4 genre names for Avatar are among the 7 most frequently genres mentioned by non-expert users.

To evaluate the ranked list of retrieved genre names, we use Mean Average Precision and F-score. The Mean Average Precision (MAP) for a set of queries (movie names in our experiments) is calculated as the mean of the average precision scores for each query. The average precision for each query, q is calculated as

Method	MAP	95% C.I.	$F_{\beta=3}$	95% C.I.
Reviews	0.698	[0.657 , 0.740]	0.550	[0.499 , 0.600]
Random	0.260	[0.229 , 0.290]	0.140	[0.101 , 0.179]

Table 3.7: Mean Average Precision and F-score for genre extraction from a set of reviews (C.I.: Confidence Interval).

$$AP_q = \frac{\sum_{k=1}^{N} Precision@k \times rel(k)}{\text{number of relevant genres}} \tag{3.1}$$

where $rel(k)$ obtains a value of 1 if the kth retrieved genre is correct and 0 otherwise. We also calculate $F_{\beta=3}$ when top 3 genres from the top of the ranked list are retrieved as relevant. Table 3.7 lists the results of this experiment.

To see how useful the set of reviews is for this particular task, we compare it with ranking genre names randomly and repeating the experiment. As Table 6 shows, using simple mention frequency measures provides significant improvements over guessing the genre randomly.

The numbers in Table 3.7 are calculated using all the user reviews collected for each movie (ranging from a few hundreds to a few thousands per movie). Here, we would like to investigate if having more reviews will give us a more accurate estimate of the genres associated with each movie.

Figure 3.6 plots the 95% confidence interval of MAP versus the number of randomly selected user reviews used to rank the genres for each movie. This figure, which is plotted on a semi-log scale, shows that the quality of ranking grows rapidly by the 100th randomly selected review and exhibits asymptotic behavior when more reviews are visited.

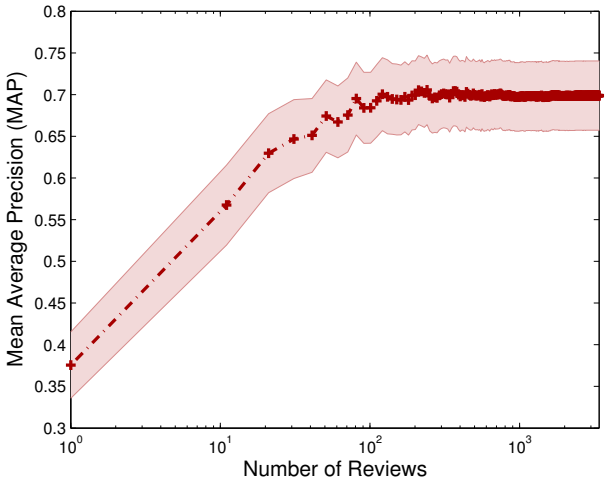


Figure 3.6: Mean Average Precision (MAP) versus the number of reviews used to extract each movie genre. (The shaded area shows 95% confidence interval for each MAP result)

3.6 Conclusion

We studied collective discourse and investigated diverse perspectives when a number of non-expert Web users engage in collective behavior and generate content on the Web. We show that the set of people who discuss the same story or subject have diverse perspectives, introducing new aspects that have not been previously discussed by others.

We analyzed a wide range of collective discourse examples, from movie reviews and news stories to scientific citations and microblogs. To the best of our knowledge this is the first work that studies the diversity in perspectives, and the small world-effect in factoid co-occurrences. We also perform an experiment that provides some evidence of collective intelligence in the collectively written set of reviews by non-expert users.

The ultimate goal of this work is to develop models of collective discourse. The models would be informed by empirical analysis of varied and large-scale datasets and would address various aspects of collective discourse: motivation behind continuous

contributions, heterogeneity and diversity in perspectives, and collective intelligence from collaboration. By formulating simple stochastic models of individual and group behavior, we may be able to predict phenomena on the macro level of discourse. We will be trying to address these questions by developing state of the art technologies in computational linguistics, network science and social theories of mass communications.

CHAPTER IV

Community Structure

In this chapter, we model the pair-wise similarities of a set of documents as a weighted network with a single *cutoff* parameter. Such a network can be thought of an ensemble of unweighted graphs, each consisting of edges with weights greater than the cutoff value. We look at this network ensemble as a complex system with a temperature parameter, and refer to it as a *Latent Network*. Our experiments on a number of datasets from two different domains show that certain properties of latent networks like *clustering coefficient*, *average shortest path*, and *connected components* exhibit patterns that are significantly divergent from randomized networks. We explain that these patterns reflect the network *phase transition* as well as the existence of a community structure in document collections. Using numerical analysis, we show that we can use the aforementioned network properties to predict the clustering Normalized Mutual Information (NMI) with high correlation ($\overline{\rho} > 0.9$). Finally we show that our clustering method significantly outperforms other baseline methods ($\overline{\text{NMI}} > 0.5$)

4.1 Introduction

Lexical networks are graphs that show relationship (e.g., semantic, similarity, dependency, etc.) between linguistic entities (e.g., words, sentences, or documents) [57].

One specific type of lexical networks include those in which edges represent a similarity relation between documents. These networks are fully connected, weighted, and symmetric (if the similarity measure is symmetric).

If we apply a cutoff value $c \in [0, 1]$, and prune the edges with values smaller than c, we will have an ordinary binary lexical network (i.e., an unweighted network in which edges denote a binary relationship). Therefore, at each value c, we have a different network. In other words, binding a network with a cutoff parameter c on edge weights as the single parameter of the network, will result in an ensemble of networks with different properties. We refer to this ensemble of networks as a *latent network*. More accurately, a latent network, $\mathcal{L}$, is an ensemble of lexical networks that are originated from the same document collection and differ by the value of a single parameter.

In our work, we analyze different properties of latent networks when the cutoff value changes, and will discuss how the network undergoes different phases and exhibits high degrees of community structure. Finally, we propose a predictive model to estimate the best cutoff value for which the network community structure is maximum and use this estimation for clustering the document collection.

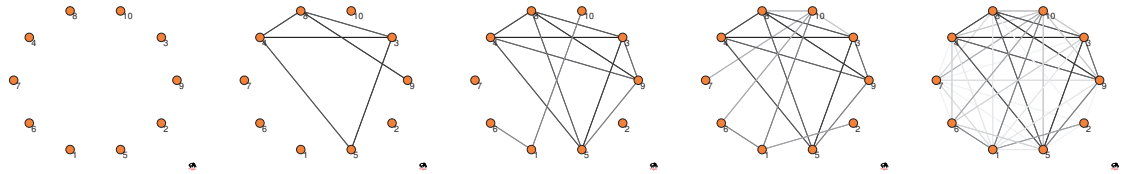
4.1.1 Data

For our experiments, we use the data from [162] on collective discourse, a collective human behavior in content generation. This data contains 50 different datasets of collective discourse from two completely different domains: news *headlines*, and scientific *citation sentences*. Each set consists of a number of unique headlines or citations about the same non-evolving news story or scientific paper.

Table 4.1 lists some of these datasets with the number of documents in them.

ID	type	Name	Story/Title	#
1	hdl	miss	Venezuela wins miss universe 2009	125
2	hdl	typhoon	Second typhoon hit philippines	100
3	hdl	russian	Accident at Russian hydro-plant	101
	...	...	...	
25	hdl	yale	Yale lab tech in court	10
26	cit	N03-1017	Statistical Phrase-Based Translation	172
27	cit	P02-1006	Learning Surface Text Patterns ...	72
28	cit	P05-1012	On-line Large-Margin Training ...	71
	...	...	...	
50	cit	H05-1047	A Semantic Approach To Recognizing TE	7

Table 4.1: The datasets and the number of documents in each of them (hdl = headlines; cit = citations)



$(a)\tau = 0.00$ $(b)\tau = 0.25$ $(c)\tau = 0.50$ $(d)\tau = 0.75$ $(e)\tau = 1.00$

Figure 4.1: Lexical network for the `Yale` dataset at 5 different τ values

4.1.2 Annotation

Following [159], we asked a number of annotators to read each set and extract different *facts* that are covered in each sentence. Each fact is an aspect of the news story or a contribution of the cited paper.

For example, one of the annotated datasets, `Yale`, is the set of the headlines about a murder incident at Yale. The manual annotation of the `Yale` dataset has resulted in 4 facts or classes:

f_1 : {annie le, yale student}

f_2 : {former yale lab tech, raymond clark}

f_3 : {plea, court}

f_4 : {murder, killing}

Table 4.2 shows the headlines with sentence-to-fact assignments in the `Yale` dataset.

ID	sentence	f_1 f_2 f_3 f_4
1	annie le slay suspect raymond clark due in court	1 1 1 0
2	attorneys to spar today over sealed annie le file	1 0 0 0
3	former yale lab tech due in court	0 1 1 0
4	former yale lab tech due in court for murder charge	0 1 1 1
5	photos: accused yale lab tech due in court today	0 1 1 0
6	raymond clark due in court	0 1 1 0
7	suspect in yale student killing to enter plea	1 0 1 1
8	yale lab tech murder suspect expected	0 1 0 1
9	yale lab tech murder suspect expected to plead not guilty	0 1 1 1
10	yale slaying suspect due in court	0 0 1 0

Table 4.2: Full Annotation of the `Yale` dataset results in a fact distribution matrix of sentences.

The full annotation of each dataset results in a number of facts (representing classes) and a fact distribution matrix.

4.2 Network Properties

One way to look at a latent network is to use a physical point of view. The network is a complex system, and the temperature of this system will determine the interaction of the nodes. Here, nodes with smaller similarities will join each other at higher temperatures. In fact, the temperature of this system can be interpreted as

$$\tau = 1 - \text{cutoff} \tag{4.1}$$

increasing which will cause more nodes to connect to each other.

Figure 4.1 shows the cosine similarity-based latent network for the 10 documents in the `Yale` dataset at 5 different τ values. At $\tau = 0$ (cutoff = 1.00) all the edges are pruned and the network is empty, while on the other end of the spectrum, where $\tau = 1$ all edges with positive weights are present.

A simple 2-D visualization of a latent network does not reveal much information about it. Describing different aspects of the network structure is easier when looking at quantitative network properties. We observe some of the latent networks' properties over different network temperatures. Starting at $\tau = 0$ and gradually increasing

it till it reaches $\tau = 1$ will cause more edges to emerge and network properties to change.

4.2.1 Number of Edges

Increasing the temperature τ (and thus decreasing the cutoff) will cause different edges to appear in the network according to the distribution of edge weights. To compare the number of edges in different networks we use the normalized number of edges at each τ based on Equation 4.2, in which $e(\tau)$ is the number of edges at temperature τ, and n is the total number of documents (nodes).

$$ne(\tau) = \frac{2e(\tau)}{n(n-1)} \tag{4.2}$$

4.2.2 Number of Connected Nodes

Another property that we are interested in is the number of nodes that have positive degrees at each τ. The number of connected nodes quantifies the distribution of $e(\tau)$ edges between n nodes. Here, we normalize this number by the total number of nodes in the graph based on Equation 4.3.

$$nn(\tau) = \frac{|\{i|k_i(\tau) > 0\}|}{n} \tag{4.3}$$

where $k_i(\tau)$ is the degree of node i at temperature τ.

4.2.3 Connected Components

A *connected component* (*cc*) of a graph is a subgraph in which there is a path between any two node pairs. The pattern in which smaller components merge into larger components or join the *largest connected component* (*lcc*) can quantify community structure in a network. Here, we observe the number of different connected components and the size of the largest connected component at each network tem-

perature τ.

$$ncc(\tau) = \frac{\#\ cc(\tau)}{n}; \qquad nlcc(\tau) = \frac{|lcc(\tau)|}{n} \tag{4.4}$$

In a network, where community structure is weak, new nodes join the largest connected component one-by-one, and the giant component includes most of the nodes in the graph. However, in a network with an inherent community structure, we expect to see the formation of smaller separate connected components that will only merge in high temperatures.

4.2.4 Average Shortest Path and Diameter

In graph theory, the shortest path between two vertices is path with the smallest number of edges. In network analysis, the average shortest path (asp) of a network is the mean of all shortest path lengths between reachable vertices. Moreover, the *diameter* (d) of a network is defined as the length of the longest shortest path. We observe the normalized average shortest path ($nasp$) and the normalized diameter (nd) of each network at different values of τ.

$$nasp(\tau) = \frac{asp(\tau)}{n}; \qquad nd(\tau) = \frac{d(\tau)}{n} \tag{4.5}$$

4.2.5 Clustering Coefficient

The clustering coefficient of a graph measures the number of closed triangles in the graph. The clustering coefficient describes how likely it is that two neighbors of a vertex are connected. In social networks, it can represent the idea that "the friends of my friends are my friends." [146]. A high clustering coefficient indicates a network with many triangles, where most of my friends are connected, a low clustering coefficient indicates a network with few triangles, where most of my friends are not friends with each other.

Watts and Strogatz's definition of clustering coefficient [207] is based on a local clustering value for each vertex that is averaged over the entire network. The clustering coefficient for a given vertex i is the number of triangles connected to vertex i divided by the total possible number of triangles connected to vertex i. More formally, in a undirected graph, if $m_i(\tau)$ is the number of i's neighbors that are connected at temperature τ, and $k_i(\tau)$ is the degree of node i at τ, then the clustering coefficient of i can be defined by Equation 4.6.

$$c_i(\tau) = \frac{2m_i(\tau)}{k_i(\tau)(k_i(\tau) - 1)} \tag{4.6}$$

The global clustering coefficient of the network is defined by Equation 4.7. Higher global clustering coefficient values of a network would imply the existence of groups of nodes in the network that are densely connected.

$$cc(\tau) = \frac{1}{n} \sum_i c_i(\tau) \tag{4.7}$$

4.3 Phase Transition

Increasing τ in a latent network will cause new edges to emerge and network properties to change. We observe these changes in non-overlapping intervals of $\tau \in [0, 1]$.

The solid black lines in Figure 4.2 show 4 network properties for the `Yale` dataset: clustering coefficient, average shortest path, number of connected components, and the size of the largest connected component. This figure also plots the same properties for a network of the same size and edge weights, but in which edges are randomly assigned to node pairs. We can think of this randomization as a random permutation of edges that preserves the number and the weights of edges.

Figure 4.2 reveals a lot of information about the structure of the `Yale` latent

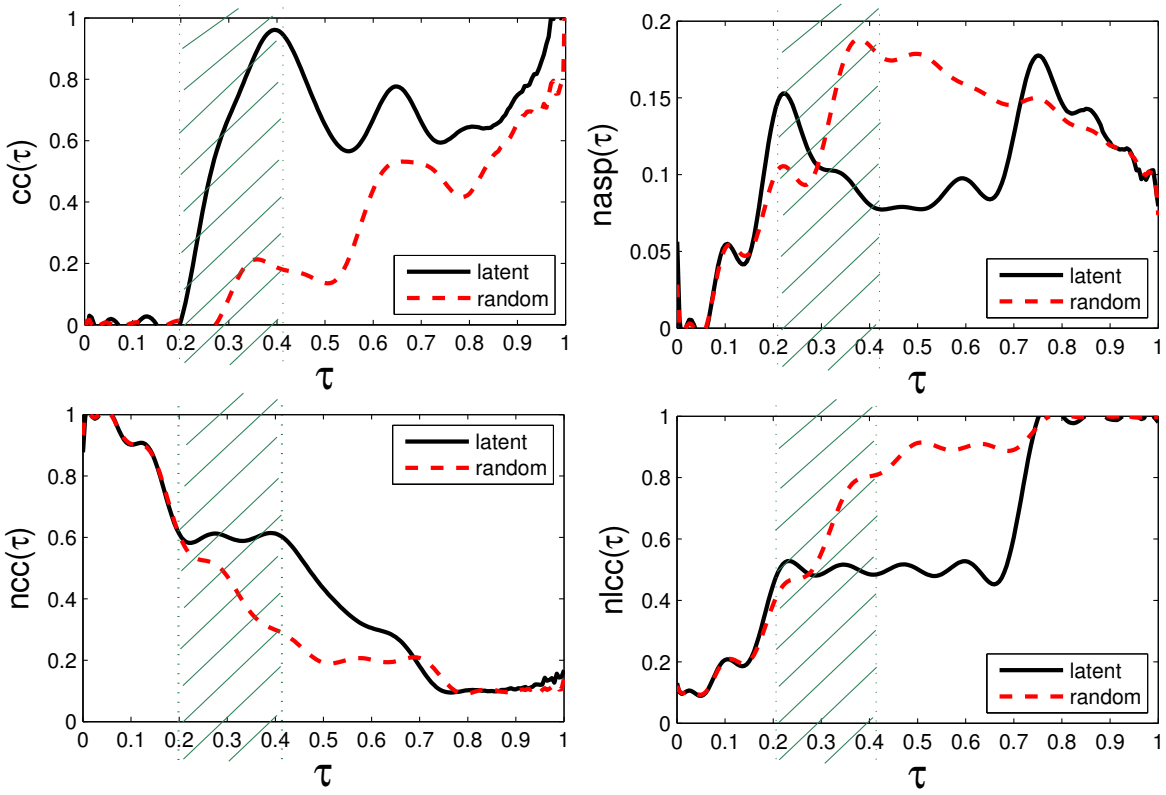


Figure 4.2: Clustering coefficient (*cc*), average shortest path (*nasp*), connected components (*ncc*), and largest connected component (*nlcc*) in the `Yale` latent network over τ, compared with a randomized network of the same size.

network. When $\tau = 0$ where the network is empty the latent network and the randomized version are identical. For values of $\tau \in [0, 0.2]$, the two networks exhibit similar behavior: clustering coefficient is very small, shortest path lengths increase, the number of connected components decrease and the largest connected component get bigger. However, for values of $\tau > 0.2$ the two networks show different behavior until τ is approximately greater than 0.8, where both networks become very dense and exhibit similar patterns again.

We refer to each of these intervals, in which the network has a different behavior, as a *phase*. One such phase is when the network's different connected components exhibit high degrees of community structure. The shaded area in Figure 4.2 ($\tau \in [0.2, 0.4]$) shows a phase in which the clustering coefficient spikes; shortest paths, unlike the randomized network, get smaller; the number of connected components is non-decreasing; and the largest connected component does not get larger. These patterns suggest the formation of dense communities in this interval because of two reasons: (1) Nodes connect to smaller components rather than the giant component. (2) Current components in the graph get denser rather than joining each other. Our goal in the rest of this chapter is to predict a value $\hat{\tau}$ that best characterizes this phase, and for which the network has the best clustering of nodes represented by different connected components.

To cluster the network at each τ, we simply assign all the nodes in a connected component to the same cluster, and assign isolated (degree = 0) nodes to separate individual clusters. To evaluate this clustering we use the fact distribution matrices from the annotations and calculate the *normalized mutual information (NMI)* proposed by [121]. Let's assume $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ is the set of clusters and

$\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$ is the set of classes. Then,

$$\text{NMI}(\Omega, \mathbb{C}) = \frac{I(\Omega; \mathbb{C})}{[H(\Omega) + H(\mathbb{C})]/2} \tag{4.8}$$

where $I(\Omega; \mathbb{C})$ is the mutual information:

$$I(\Omega, \mathbb{C}) = \sum_k \sum_j P(\omega_k \cap c_j) \log \frac{P(\omega_k \cap c_j)}{P(\omega_k)P(c_j)} \tag{4.9}$$

$$= \sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log \frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|} \tag{4.10}$$

where $P(\omega_k)$, $P(c_j)$, and $P(\omega_k \cap c_j)$ are the probabilities of a document being in cluster ω_k, class c_j, and in the intersection of ω_k and c_j, respectively. Here, H is entropy:

$$H(\Omega) = -\sum_k P(\omega_k) \log P(\omega_k) \tag{4.11}$$

$$= -\sum_k \frac{|\omega_k|}{N} \log \frac{|\omega_k|}{N} \tag{4.12}$$

$I(\Omega; \mathbb{C})$ in Equation 4.9 measures the amount of information that we would lose about the classes without the cluster assignments. The normalization factor ($[H(\Omega)+ H(\mathbb{C})]/2$) in Equation 4.8 enables us to trade off the quality of the clustering against the number of clusters, since entropy tends to increase with the number of clusters. For example, $H(\Omega)$ reaches its maximum when each document is assigned to a separate cluster. Because NMI is normalized, we can use it to compare cluster assignments with different numbers of clusters. Moreover, $[H(\Omega) + H(\mathbb{C})]/2$ is a tight upper bound for $I(\Omega; \mathbb{C})$, making NMI obtain values between 0 and 1 [121].

The evolution of a latent network over τ can be illustrated using a dendrogram, and characterized by the quality of the clustering that the connected components produce. Figure 4.3 shows $\text{NMI}(\Omega, \mathbb{C})$ versus τ in the `Yale` dataset aligned with a clustering dendrogram. The shaded area in the plot ($\tau \in [0.2, 0.4]$) shows the area

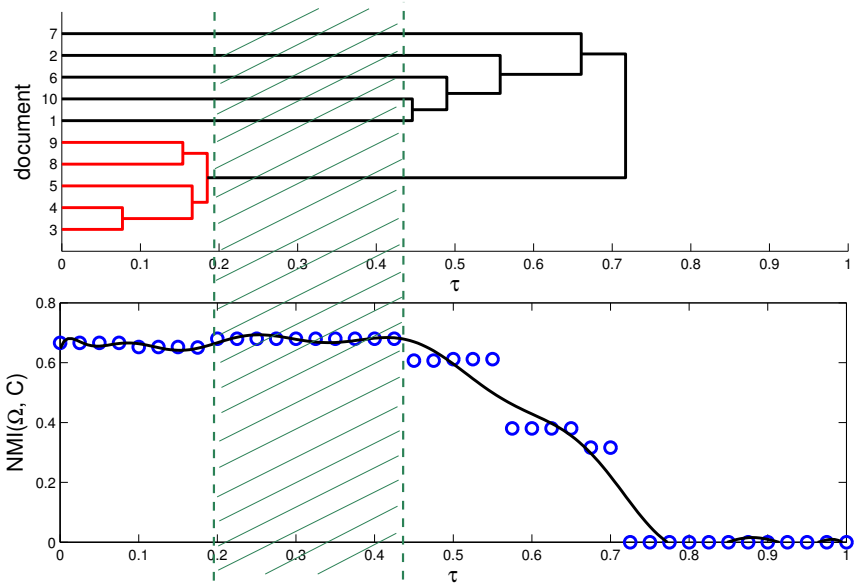


Figure 4.3: The dendrogram for the `Yale` dataset's latent network. Different sentences join into connected components at different temperatures.

in which any cut on the dendrogram will result in a maximum community structure characterized by NMI.

4.3.1 Optimization

To find the best cut on the dendrogram, we propose a model that is similar to the Information Bottleneck method [40] in optimizing clustering mutual information.

We build an L_1-regularized log-linear model [6] on τ and 7 network-based features discussed before to predict $\text{NMI}(\Omega, \mathbb{C})$ at each τ. Let's suppose $\Phi : X \times Y \to \mathbb{R}^D$ is a function that maps each (x, y) to a vector of feature values. Here, the feature vector is the vector of coefficients corresponding to τ and 7 different network properties, and the parameter vector $\theta \in \mathbb{R}^D$ ($D = 8$ in our experiments) assigns a real-valued weight to each feature. This estimator chooses θ to minimize the sum of least squares and a regularization term R.

$$\hat{\theta} = arg\min_{\theta}\{\frac{1}{2}\sum_{i}||\langle\theta, x_i\rangle - y_i||_2^2 + R(\theta)\} \tag{4.13}$$

where the regularizer term $R(\theta)$ is the weighted L_1 norm of the parameters.

$$R(\theta) = \alpha \sum_j |\theta_j| \tag{4.14}$$

Here, α is a parameter that controls the amount of regularization (set to 0.1 in our experiments).

To optimize the L_1-regularized objective function, we use the *orthant-wise limited-memory quasi-Newton* algorithm (OWL-QN), which is a modification of L-BFGS that allows it to effectively handle the discontinuity of the gradient [6]. This algorithm works quite well in practice, and typically reaches convergence in even fewer iterations than standard L-BFGS [65].

4.3.2 NMI Prediction

Using 5-fold cross validation scheme in each category, we predict the value of $\text{NMI}(\Omega, \mathbb{C})$ ($\widehat{\text{NMI}}$) at each τ for each network. In each fold the training data consists of 20 networks. For each network $\mathcal{L}_i$, we observe 201 values of τ ($\tau \in [0, 1]$ with increments of 0.005), and calculate NMI and 7 network-based features in $\mathcal{L}_i$. We use all the observations from 20 networks to predict NMI at each τ in the test networks.

The result of this experiment is a list of $\widehat{\text{NMI}}$s at each τ for each network. Table 4.3 shows the average correlation between NMIs and $\widehat{\text{NMI}}$s in each category, using different features. The highest correlation is when we use all the features. However, clustering coefficient seems to play as an important indicator of the clustering quality.

4.3.3 Domain Adaptation

To generalize the effectiveness of network-level features in predicting cluster quality, we design the following experiment. We first use τ and 7 network features to train a model on all the 25 networks from the citations category (at 201 equally spaced values of $\tau \in [0, 1]$) and use this model to predict NMI at each τ for each

Features	headlines		citations		Mean
τ+ ne + nn +	$\overline{\rho}$	95% C.I.	$\overline{\rho}$	95% C.I.	
ncc + $nlcc$	0.904	[0.844, 0.964]	0.923	[0.857, 0.989]	0.913
$nasp$ + nd	0.861	[0.790, 0.932]	0.886	[0.796, 0.976]	0.873
cc	0.907	[0.856, 0.958]	0.805	[0.716, 0.894]	0.856
all	0.906	[0.845, 0.967]	0.923	[0.864, 0.982]	**0.914**
C.I. = Confidence Interval					

Table 4.3: Average Pearson Correlation coefficient between clustering NMI and predicted NMI at different τ values for each network, using various features

headlines		citations		Mean
$\overline{\rho}$	95% C.I.	$\overline{\rho}$	95% C.I.	
0.865	[0.786, 0.945]	0.929	[0.867, 0.991]	0.897
C.I. = Confidence Interval; Features: all.				

Table 4.4: Average prediction correlation when the model is trained on the other category.

headline network. We also do the same experiment when the model is trained on headline networks and tested on citation networks. Table 4.4 reports the average correlation between predicted NMIs and actual NMIs at various τ values.

4.3.4 Clustering

We have shown the effectiveness of network-based features in predicting the clustering quality. Here, we employ our model to find a good clustering of a document collection. Our clustering works by simply applying the best clustering τ_c, the temperature that results in the highest predicted NMI:

$$\tau_c = \arg\max_{\tau \in [0,1]} \widehat{\text{NMI}}(\tau) \tag{4.15}$$

Applying τ_c to a latent network means pruning all the edges whose weight is below the cutoff from Equation 4.1. We then simply, assign all the nodes in each connected component to a single cluster. Here, to build a predictive model of NMI, we follow our first experiment, and perform a 5-fold cross validation for each category. We compare the results of this experiment with 3 clustering systems: Random, Modularity-based, and K-means.

Random

The Random clustering, randomly assigns each document to one of k clusters. Here we assigned k to be the number of classes in each dataset ($|f|$). Although Random is basically a weak baseline, using $|f|$ as the number of classes makes is stronger.

Modularity-based

Modularity is a measure of network community division quality and is based on the measure of assortative mixing [145]. Here we explain Newman's definition of modularity as defined in [147, 38]. Consider a division in the network with k communities. Let's define e as the community matrix. e is a $k \times k$ symmetric matrix in which e_{ij} is the fraction of all edges in the network that link a vertex in community i to a vertex in community j. The trace of this matrix is the fraction of edges that link vertices within the same community.

$$\text{Tr } e = \sum_i e_{ii} \tag{4.16}$$

A good division should result in a high value of the trace matrix. Let's also define the row sums as $a_i = \sum_j e_{ij}$, which represents the fraction of edges that connect to vertices in community i. In a random network in which edges fall between nodes regardless of any community structure, we would have $e_{ij} = a_i a_j$. In such a network, a_i^2 shows the fraction of edges within the community i. Given this setting, modularity is defined as Equation 4.17

$$Q = \sum_i (e_{ii} - a_i^2) \tag{4.17}$$

If the number of within-community edges is no better than random, we will have $Q = 0$. Higher values of Q indicates strong community structure, while $Q = 1$ is the maximum value Q can obtain.

Method	headlines		citations		Mean
	NMI	95% C.I.	NMI	95% C.I.	
Random	0.183	[0.124, 0.243]	0.272	[0.201, 0.343]	0.227
K-means(4)	0.310	[0.244, 0.377]	0.333	[0.253, 0.413]	0.321
K-means(f)	0.364	[0.289, 0.439]	0.378	[0.298, 0.458]	0.371
Modularity	0.254	[0.193, 0.315]	0.298	[0.234, 0.362]	0.276
Latent network	**0.489**	[0.425, 0.553]	**0.575**	[0.515, 0.635]	**0.532**
C.I. = Confidence Interval					

Table 4.5: Average clustering Normalized Mutual Information (NMI) for each method, in each category.

The modularity-based algorithm [147] uses *edge betweenness* to do the clustering. Edge betweenness in a network is an extension of the *node betweenness* definition [62], and measures the number of shortest paths in the graph that fall on the given edge. Intuitively, removing edges will high betweenness values will cause node pair to become more separated and form communities. Thus this algorithm iteratively removes edges with highest betweenness values, and stops when modularity is maximal.

K-means

We finally used two variants of the K-means algorithm as baselines. In the first one, we run K-means on each collection with a constant number of clusters (k= 4 in our experiments), and in the second one we assign k to be the number of classes from the annotations in each dataset ($k = |f|$).

Table 4.5 lists the average NMI achieved by each method in each category. As this table shows the latent network model can achieve high values of NMI in clustering while outperforming other state of the art algorithms.

CHAPTER V

Citation Summarization using C-LexRank

Researchers and scientists increasingly find themselves in the position of having to quickly understand large amounts of technical material. Our goal is to effectively serve this need by using bibliometric lexical link mining and summarization techniques to generate summaries of scientific literature. In this chapter, we show how we can use citations to produce an automatically generated, readily consumable, technical survey. We first propose C-LexRank, a model of summarizing single scientific articles based on citations, which employs community detection and extracts salient information-rich sentences.

5.1 Introduction

In today's rapidly expanding disciplines, scientists and scholars are constantly faced with the daunting task of keeping up with knowledge in their field. In addition, the increasingly interconnected nature of real-world tasks often requires experts in one discipline to rapidly learn about other areas in a short amount of time. Cross-disciplinary research requires scientists in areas such as linguistics, biology, and sociology to learn about computational approaches and applications, e.g., computational linguistics, biological modeling, social networks. Authors of journal articles and books must write accurate surveys of previous work, ranging from short sum-

maries of related research to in-depth historical notes. Interdisciplinary review panels are often called upon to review proposals in a wide range of areas, some of which may be unfamiliar to panelists. Thus, they must learn about a new discipline "on the fly" in order to relate their own expertise to the proposal.

Our goal is to effectively serve these needs by combining two currently available technologies: (1) bibliometric lexical link mining that exploits the structure of citations and (2) summarization techniques that exploit the content of the material in both the citing and cited papers.

It is generally agreed upon that manually written abstracts are good summaries of individual papers. More recently, we argued in [159] that *citation sentences* (i.e., set of sentences that appear in other papers and cite a given article) are useful in creating a summary of important contributions of a research paper. Moreover, Qazvinian and Radev [160] showed the usefulness of using implicit citations (*context sentences*) in summary generation. Teufel [192] argued that citations could contain subjective content, and that this content can be exploited for summary generation. Additional work [134] demonstrated the usefulness of citations for producing multi-document surveys of scientific articles. Follow-on work indicated that further improvements to citation handling enables the production of more fluent summaries [209].

In our work, we compare and contrast the usefulness of abstracts and of citations in automatically generating a technical survey on a given topic from multiple research papers. Moreover, we develop a summarization model that exploits citations to produce a multi-faceted summary of scientific contributions.

5.2 Background

Automatically creating technical surveys is significantly distinct from traditional multi-document summarization. Below we describe the primary characteristics of a technical survey and we present different types of input texts that we used for the production of surveys.

Technical Survey

In the case of multi-document summarization, the goal is to produce a readable presentation of multiple documents, whereas in the case of technical survey creation, the goal is to convey the key features and basic underpinnings of a particular field, early and late developments, important contributions and findings, contradicting positions that may reverse trends or start new sub-fields, and basic definitions and examples that enable rapid understanding of a field by non-experts.

A prototypical example of a technical survey is that of "chapter notes," i.e., short (50–500 word) descriptions of sub-areas found at the end of chapters of textbook, such as [89]. One might imagine producing such descriptions automatically, then hand-editing them and refining them for use in an actual textbook.

Previously Mohammad et al. [134] conducted a human analysis of these chapter notes and revealed a set of conventions, an outline of which is provided here (with example sentences in italics):

1. Introductory/opening statement: *The earliest computational use of X was in Y, considered by many to be the foundational work in this area.*
2. Definitional follow up: *X is defined as Y.*
3. Elaboration of definition (e.g., with an example): *Most early algorithms were based on Z.*

4. Deeper elaboration, e.g., pointing out issues with initial approaches: *Unfortunately, this model seems to be wrong.*
5. Contrasting definition: *Algorithms since then...*
6. Introduction of additional specific instances / historical background with citations: *Two classic approaches are described in Q.*
7. References to other summaries: *R provides a comprehensive guide to the details behind X.*

The notion of *text level categories* or *zoning* of technical papers—related to the survey components enumerated above—has been investigated previously in [139] and [193]. These earlier works focused on the *analysis* of scientific papers based on their rhetorical structure and on determining the portions of papers that contain new results, comparisons to earlier work, etc. The work described here focuses on the *synthesis* of technical surveys based on knowledge gleaned from rhetorical structure not unlike that of the work of these earlier researchers, but guided by structural patterns along the lines of the conventions listed above.

Although our current approach to survey creation does not yet incorporate a fully pattern-based component, our ultimate objective is to apply these patterns to guide the creation and refinement of the final output. As a first step toward this goal, we use citation sentences (closest in structure to the patterns identified by convention 7 above) to pick out the most important content for survey creation.

Scholarly Texts

Published research on a particular topic can be summarized from two different kinds of sources: (1) where an author describes her own work and (2) where others describe an author's work (usually in relation to their own work). The author's

description of her own work can be found in her paper. How others perceive her work is spread across other papers that cite her work.

Traditionally, technical survey generation has been tackled by summarizing a set of research papers pertaining to the topic. However, individual research papers usually come with manually-created "summaries"—their abstracts. The abstract of a paper may have sentences that set the context, state the problem statement, mention how the problem is approached, and the bottom-line results—all in 200 to 500 words. Thus, using only the abstracts (instead of full papers) as input to a summarization system is worth exploring.

Whereas the abstract of a paper presents what the authors think to be the important aspects of a paper, the citations to a paper captures what others in the field perceive as the contributions of the paper. The two perspectives are expected to have some overlap in their content, but the citations also contain additional information not found in abstracts [50, 137]. For example, authors may describe how a particular methodology from one paper was combined with another from another paper to overcome some of the drawbacks of each. A citation is also an indicator of what contributions described in a paper were influential over time.

Another feature that distinguishes citations texts from abstracts is that citations tend to have a certain amount of redundant information. This is because multiple papers may describe the same contributions of a target paper. This redundancy can be exploited to determine the important contributions of the target paper.

Our goal is to test the hypothesis that an effective technical survey will reflect information on research not only from the perspective of its authors but also from the perspective of others who use, commend, discredit, or add to it. Before describing our experiments with technical papers, abstracts, and citations, we first summarize

relevant prior work that used these sources of information as input.

5.3 Citation Summarization

The ACL Anthology Network[1] (AAN) is a manually curated anthology built on top of the ACL Anthology[2] [20]. AAN includes all the papers published by ACL and related organizations as well as the Computational Linguistics journal over a period of four decades. AAN consists of more than $18,000$ papers from more than $14,000$ authors, each distinguished with a unique ACL ID, together with their full-texts, abstracts, and citation information. It also includes other valuable metadata such as author affiliations, citation and collaboration networks, and various centrality measures [167, 88].

To study citations across different areas within Computational Linguistics, we first extracted six different clusters of papers from AAN corresponding to 6 different NLP topics: Dependency Parsing (DP), Phrase-based Machine Translation (PBMT), Text Summarization (Summ), Question Answering (QA), Textual Entailment (TE), and Conditional Random Fields (CRF). To build each cluster, we matched the topic phrase against the title and the content of AAN papers, and picked 5 highest cited papers. Table 5.1 shows the number of articles and citation sentences in each cluster.

Next, we designed an annotation task that requires explicit definitions that distinguish between phrases that represent the same or different information units. Unfortunately, there is little consensus in the literature on such definitions. Therefore, following [198, 162] we made the following distinction. We define a *nugget* to be a phrasal information unit (i.e., any phrase that would contain some information about the contributions of the cited paper). Different nuggets may all represent the same

[1] http://clair.si.umich.edu/anthology/
[2] http://www.aclweb.org/anthology-new/

atomic semantic unit, which we refer to as a *factoid*. In the context of citations, a factoid refers to a unique contribution of a target paper mentioned in a citation sentence. For example, the following set of citations to Eisner's famous parsing paper[3] illustrate the set of factoids about this paper and suggest that different authors who cite a particular paper may discuss different contributions (factoids) of that paper.

In the context of DPs, this edge based factorization method was proposed by **(Eisner, 1996)**.

Eisner (1996) *gave a generative model with a cubic parsing algorithm based on an edge factorization of trees.*

Eisner (1996) *proposed an $O(n^3)$ parsing algorithm for PDG.*

If the parse has to be projective, Eisner's bottom-up-span algorithm **(Eisner, 1996)** *can be used for the search.*

This example also suggests that different authors use different wordings (nuggets) to represent the same factoids. For instance, *cubic parsing* and *$O(n^3)$ parsing algorithm* are two nuggets that represent the same factoid about (Eisner, 1996).

Another example, which we will use throughout the paper, is the paper by Cohn and Blunsom (2005)[4] (identified with the ACL ID W05-0622 in Table 5.1). This paper is cited in 9 different sentences within AAN. All of these sentences are listed in Table 5.2. In each sentence, the nuggets extracted by the annotators are underlined. As this table suggests, a citation sentence may not discuss any of the contributions of the cited paper. For instance, the last sentence does not contain any factoids about (Cohn & Blunsom, 2005). The nuggets that are identified using the citation to (Cohn

[3]Eisner, J. (1996). Three new probabilistic models for dependency parsing: An exploration. In Proceedings of the 34th Annual Conference of the Association for Computational Linguistics (ACL-96), pp. 340–345.

[4]Cohn, T., & Blunsom, P. (2005). Semantic role labelling with tree conditional random Fields. In Proceedings of the Ninth Conference on Computational Natural Language Learning, pp. 169-172.

	ACL ID	Title	Year	# of citations
DP	C96-1058	Three New Probabilistic Models For Dependency Parsing ...	1996	66
	P97-1003	Three Generative, Lexicalized Models For Statistical Parsing	1997	50
	P99-1065	A Statistical Parser For Czech	1999	54
	P05-1013	Pseudo-Projective Dependency Parsing	2005	40
	P05-1012	On-line Large-Margin Training Of Dependency Parsers	2005	71
PBMT	N03-1017	Statistical Phrase-Based Translation	2003	172
	W03-0301	An Evaluation Exercise For Word Alignment	2003	11
	J04-4002	The Alignment Template Approach To Statistical Machine Translation	2004	49
	N04-1033	Improvements In Phrase-Based Statistical Machine Translation	2004	23
	P05-1033	A Hierarchical Phrase-Based Model For Statistical Machine Translation	2005	65
Summ	A00-1043	Sentence Reduction For Automatic Text Summarization	2000	19
	A00-2024	Cut And Paste Based Text Summarization	2000	20
	C00-1072	The Automated Acquisition Of Topic Signatures ...	2000	19
	W00-0403	Centroid-Based Summarization Of Multiple Documents ...	2000	28
	W03-0510	The Potential And Limitations Of Automatic Sentence Extraction ...	2003	14
QA	A00-1023	A Question Answering System Supported By Information Extraction	2000	13
	W00-0603	A Rule-Based Question Answering System For Reading ...	2002	19
	P02-1006	Learning Surface Text Patterns For A Question Answering System	2002	72
	D03-1017	Towards Answering Opinion Questions: Separating Facts From Opinions ...	2003	39
	P03-1001	Offline Strategies For Online Question Answering ...	2003	27
TE	D04-9907	Scaling Web-Based Acquisition Of Entailment Relations	2004	12
	H05-1047	A Semantic Approach To Recognizing Textual Entailment	2005	7
	H05-1079	Recognising Textual Entailment With Logical Inference	2005	9
	W05-1203	Measuring The Semantic Similarity Of Texts	2005	17
	P05-1014	The Distributional Inclusion Hypotheses And Lexical Entailment	2005	10
CRF	N03-1023	Weekly Supervised Natural Language Learning ...	2003	29
	N04-1042	Accurate Information Extraction from Research Papers ...	2004	24
	W05-0622	Semantic Role Labelling with Tree CRFs	2005	9
	P06-1009	Discriminative Word Alignment with Conditional Random Fields	2006	33
	W06-1655	A Hybrid Markov/Semi-Markov CRF for Sentence Segmentation	2006	20

DP: Dependency Parsing, PBMT: Phrase-based Machine Translation, Summ: Text Summarization, QA: Question Answering, TE: Textual Entailment, CRF: Conditional Random Fields

Table 5.1: Papers chosen from clusters for single document summarization, with their publication year, and the number of citing sentences in AAN's 2008 release.

& Blunsom, 2005) account for a total number of 3 factoids (contributions) identified for this paper: semantic role labeling, tree structures, and a pipelined approach.

5.3.1 C-LexRank

In this section we describe C-LexRank as a method to extract citing sentences that cover a diverse set of factoids. Our method works by modeling the set of citations as a network of sentences and identifying communities of sentences that cover similar factoids. Once a good division of sentences is made, we extract salient sentences from different communities. Figure 5.1 illustrates a toy example that depicts C-LexRank's process.

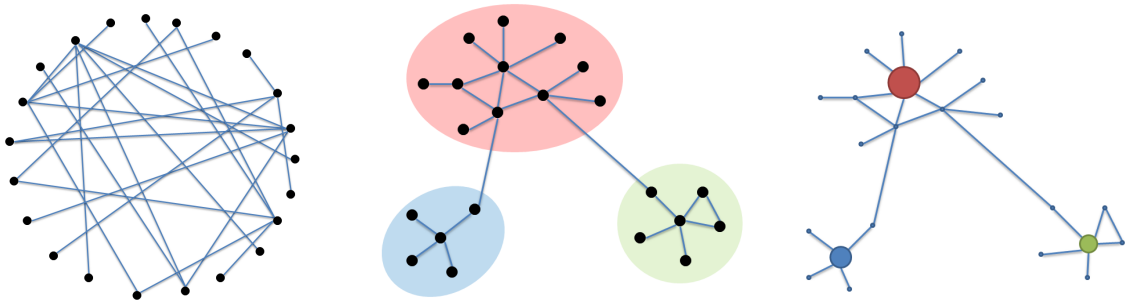
In the first step (as shown in Figure 5.1 a), we model the set of sentences that cite a specific paper with a network in which vertices represent citing sentences and undirected weighted edges show the degree of semantic relatedness between node

1	Our parsing model is based on a conditional random field model, however, unlike previous TreeCRF work, e.g., (**Cohn and Blunsom**, 2005; Jousse et al., 2006), we do not assume a particular tree structure, and instead find the most likely structure and labeling.
2	Some researchers (Xue and Palmer, 2004; Koomen et al., 2005; **Cohn and Blunsom, 2005**; Punyakanok et al., 2008; Toutanova et al., 2005; Toutanova et al., 2008) used a pipelined approach to attack the task.
3	They have been used for tree labelling, in XML tree labelling (Jousse et al., 2006) and semantic role labelling tasks **(Cohn and Blunsom, 2005)**.
4	Finally, probabilistic models have also been applied to produce the structured output, for example, generative models (Thompson, Levy, and Manning 2003), sequence tagging with classifiers (M'arquez et al. 2005; Pradhan et al.2005b), and Conditional Random Fields on tree structures **(Cohn and Blunsom 2005)**.
5	As for SRL on news, most researchers used the pipelined approach, i.e., dividing the task into several phases such as argument identification, argument classification, global inference, etc., and conquering them individually (Xue and Palmer, 2004; Koomen et al., 2005; **Cohn and Blunsom**, 2005; Punyakanok et al., 2008; Toutanova et al., 2005; Toutanova et al., 2008).
6	Although T-CRFs are relatively new models, they have already been applied to several NLP tasks, such as semantic role labeling, semantic annotation, word sense disambiguation, image modeling (**Cohn and Blunsom**, 2005; Tang et al., 2006; Jun et al., 2009; Awasthi et al., 2007).
7	The model can be used for tasks like syntactic parsing (Finkel et al., 2008) and semantic role labeling **(Cohn and Blunsom, 2005)**.
8	Regarding novel learning paradigms not applied in previous shared tasks, we find Relevant Vector Machine (RVM), which is a kernelbased linear discriminant inside the framework of Sparse Bayesian Learning (Johansson and Nugues, 2005) and Tree Conditional Random Fields (T-CRF) **(Cohn and Blunsom, 2005)**, that extend the sequential CRF model to tree structures.
9	We use CRFs as our models for both tasks **(Cohn and Blunsom, 2005)**.

Table 5.2: The set of citing sentences to the AAN paper W05-0622 (Cohn & Blunsom, 2005). Each nugget extracted by the annotators is underlined.

pairs, normally quantified by a similarity measure. We refer to this network as the *Citation Summary Network* of an article. The similarity function should ideally assign high scores to sentence pairs that have the same factoids, and should assign low scores to sentences that talk about different contributions of the target paper.

Previously, Qazvinian and Radev [159] examined 7 different similarity measures including TF-IDF with various IDF databases, longest common sub-sequence, generation probability [54], and the Levenstein distance on a training set of citations. They showed that the cosine similarity measure that employs TF-IDF vectors assigns higher similarities to pairs that contain the same factoids. Following [159], we use the cosine similarity between TF-IDF vector models that employ a general IDF corpus to construct the citation summary network of each article.



(a) Citation summary network (b) Community structure (c) C-LexRank output

Figure 5.1: An illustration of C-LexRank algorithm in a toy citation summary network

Community Structure

We generate summaries by extracting representative sentences from the citation summary network. Intuitively, a good summary should include sentences that represent different contributions of a paper. Therefore, a good sentence selection from the citation summary network will include vertices that are similar to many other vertices and which are not very similar to each other. On the other hand, a bad selection would include sentences that are only representing a small set of vertices in

the graph. This is very similar to the concept of maximizing social influence in social networks [93]. Figure 5.2 shows a toy example in which the selected two nodes in the citation summary networks represent a small subset of vertices (left) and a larger subset of vertices (right). In our work we try to select vertices that maximize the size of the set of vertices that they represent. We achieve this by detecting different vertex communities in the citation summary network.

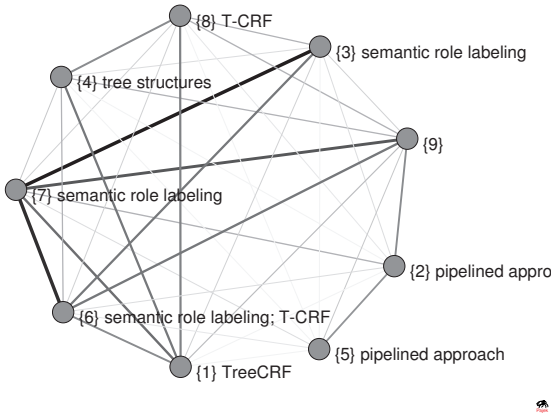
(a) Bad sentence selection

(b) Good sentence selection

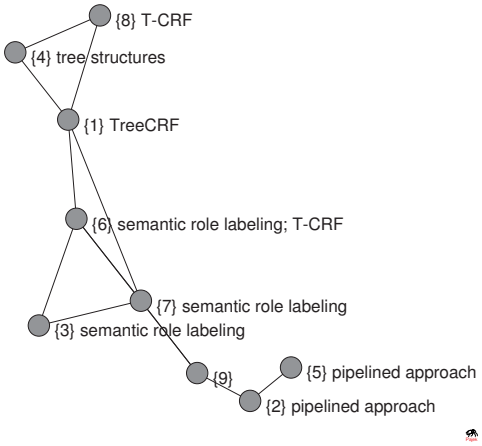
Figure 5.2: An illustration of vertex coverage by selecting representative nodes as a summary. Selecting two similar vertices will cause the summary to cover fewer contributions of the target paper in (a), while selecting less similar vertices as the summary will increase the coverage of the summary (b).

In order to find vertex communities and thus a good sentence selection, we exploit the small-world property of citation summary networks. A network is called *small-world*, if most of its vertices are not neighbors of each other, but can be reached from one another by a small number of steps [207]. Recent research has shown that a wide range of natural graphs such as biological networks [170], food webs [135], brain neurons [16] and human languages [57] exhibit the small-world property.

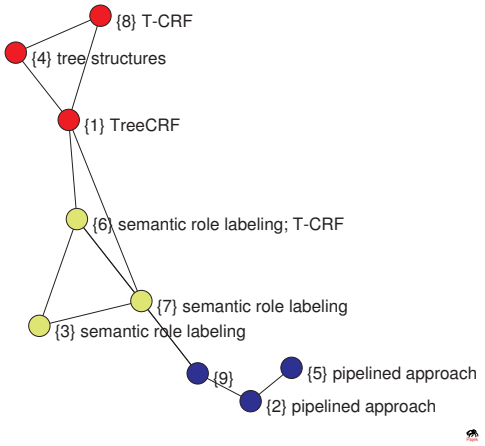
This common characteristic can be detected using two basic statistical properties: the clustering coefficient C, and the average shortest path length ℓ. The clustering coefficient of a graph measures the number of closed triangles in the graph. It describes how likely it is that two neighbors of a vertex are connected [146]. Watts and



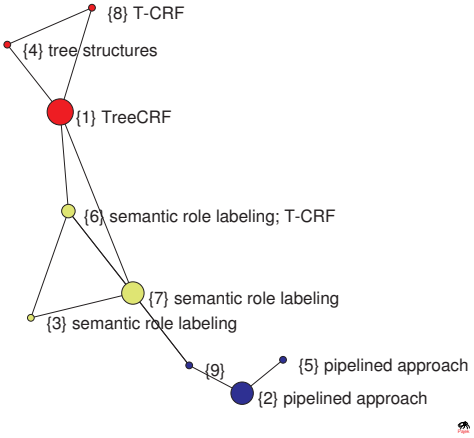
(a) Citation summary network



(b) Latent community structure



(c) Clustering output



(d) C-LexRank ranking

Figure 5.3: Illustration of the C-LexRank algorithm on the citation summary network of (Cohn & Blunsom, 2005). In the network (a), the nodes are citation sentences (annotated with their nuggets from Table 5.2), and each edge is the cosine similarity between the corresponding node pairs. (b) shows that the network has an underlying structure which is captured by C-LexRank in (c). Finally, (d) shows the C-LexRank output where node diameter is proportional to its LexRank value within the cluster.

Strogatz [207] define the clustering coefficient as the average of the local clustering values for each vertex.

$$C = \frac{\sum_{i=1}^{n} c_i}{n} \tag{5.1}$$

The local clustering coefficient c_i for the ith vertex is the number of triangles connected to vertex i divided by the total possible number of triangles connected

to vertex i. Watts and Strogatz [207] show that small-world networks are highly clustered and obtain relatively short paths (i.e., ℓ is small). We showed in previous work that citation summary networks are highly clustered in which C obtains values that are significantly larger than random networks [161]. This suggests that citation summary networks have an inherent community structure whereby each community consists of the citing sentences that discuss the same factoids.

Figure 5.3 (a) illustrates a real citation summary network built using the citation sentences in Table 5.2 in which each node is labeled with its corresponding nugget. With some re-arrangement of the nodes in Figure 5.3 (b), it becomes clear that the citation summary network of this paper has an underlying community structure in which sentences that cover similar factoids are closer to each other and form communities. For instance, in this network there are at least 3 observable communities: one that is about "tree structure," one about "semantic role labeling" and the last one about the "pipelined approach" as proposed by (Cohn & Blunsom, 2005).

In order to detect these communities automatically we use modularity. *Modularity*, [147], is a measure to evaluate the divisions that a community detection algorithm generates. For a division with g groups, they define matrix $\mathbf{e}_{g\times g}$ whose component e_{ij} is the fraction of edges in the original network that connect vertices in components i, j. Then the modularity Q can be defined as

$$Q = \sum_{i} e_{ii} - \sum_{ijk} e_{ij}e_{ki} \tag{5.2}$$

Intuitively, Q is the fraction of all the edges that are embedded within communities minus the expected value of the same quantity in a network with the same degrees but in which edges are placed at random regardless of the community structure.

Graph clustering methods aim at finding a division that for which the average number of intra-cluster edges is significantly greater and that of inter-cluster edges.

Here, we employ the network clustering algorithm described in [38]. In their work, Clauset et al. propose a hierarchical agglomeration algorithm which works by greedily optimizing the modularity in a linear running time for sparse graphs. More particularly, their method continuously merges vertex or cluster pairs with the highest similarity and stops when modularity reaches the maximum value. Figure 5.3 (c) shows how the clustering algorithm detects factoid communities in the citation summary network of (Cohn & Blunsom, 2005). In this figure, we have color-coded nodes based on their community. The clustering algorithm assigns sentences 1, 4 and 8 (which are all about the tree structures) to one cluster; sentences 3, 6 and 7 (which are all about semantic role labeling) to another cluster; and finally assigns sentences 2, 5 and 9 (sentences 2 and 5 are both about pipelined approach) to the last cluster.

To evaluate how well the clustering method works in all of our datasets, we calculated both the *purity* and the *normalized mutual information* (NMI) for the divisions in each citation set, extracted using the community detection algorithm. Purity [215] is a method in which each cluster is assigned to the class with the majority vote in the cluster, and the accuracy of this assignment is then measured by dividing the number of correctly assigned documents by N. More formally:

$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_{k} \max_{j} |\omega_k \cap c_j| \tag{5.3}$$

where $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ is the set of clusters and $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$ is the set of classes. ω_k is interpreted as the set of documents in the cluster ω_k and c_j as the set of documents in the class c_j.

We also calculate the *normalized mutual information* (NMI) [121]. Let's assume $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ is the set of clusters and $\mathbb{C} = \{c_1, c_2, \ldots, c_J\}$ is the set of

classes. Then,

$$\text{NMI}(\Omega, \mathbb{C}) = \frac{I(\Omega; \mathbb{C})}{[H(\Omega) + H(\mathbb{C})]/2} \tag{5.4}$$

where $I(\Omega; \mathbb{C})$ is the mutual information:

$$I(\Omega, \mathbb{C}) = \sum_k \sum_j P(\omega_k \cap c_j) \log \frac{P(\omega_k \cap c_j)}{P(\omega_k)P(c_j)} \tag{5.5}$$

$$= \sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log \frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|} \tag{5.6}$$

where $P(\omega_k)$, $P(c_j)$, and $P(\omega_k \cap c_j)$ are the probabilities of a document being in cluster ω_k, class c_j, and in the intersection of ω_k and c_j, respectively.

Here, H is entropy

$$H(\Omega) = -\sum_k P(\omega_k) \log P(\omega_k) \tag{5.7}$$

$$= -\sum_k \frac{|\omega_k|}{N} \log \frac{|\omega_k|}{N} \tag{5.8}$$

$I(\Omega; \mathbb{C})$ in Equation 5.5 measures the amount of information that we would lose about the classes without the cluster assignments. The normalization factor ($[H(\Omega) + H(\mathbb{C})]/2$) in Equation 5.4 enables us to trade off the quality of the clustering against the number of clusters, since entropy tends to increase with the number of clusters. For example, $H(\Omega)$ reaches its maximum when each document is assigned to a separate cluster. Because NMI is normalized, we can use it to compare cluster assignments with different numbers of clusters. Moreover, $[H(\Omega)+H(\mathbb{C})]/2$ is a tight upper bound for $I(\Omega; \mathbb{C})$, making NMI obtain values between 0 and 1. Table 5.3 lists the average Purity and NMI for each paper in our collected dataset, as well as the same numbers for a division of the same size, but in which vertices are randomly assigned to clusters.

	average	95% C.I.
purity$(\Omega, \mathbb{C})$	**0.461**	[0.398, 0.524]
purity$(\Omega_{\text{random}}, \mathbb{C})$	0.389	[0.334, 0.445]
NMI$(\Omega, \mathbb{C})$	**0.312**	[0.251, 0.373]
NMI$(\Omega_{\text{random}}, \mathbb{C})$	0.182	[0.143, 0.221]
C.I.: Confidence Interval.		

Table 5.3: Average purity and normalized mutual information (NMI) in the evaluated datasets

Ranking

Once the graph is clustered and communities are formed, we extract sentences from different clusters to build a summary. We start with the largest cluster and extract sentences using LexRank [55] within each cluster. In other words, for each cluster Ω_i we made a lexical network of *the sentences in that cluster* (N_i). Using LexRank we can find the most central sentences in N_i as salient sentences of Ω_i to include in the main summary. We choose, for each cluster Ω_i, the most salient sentence of Ω_i, and if we have not reached the summary length limit, we do that for the second most salient sentences of each cluster, and so on. The cluster selection is in order of decreasing size. Figure 5.3 (d) shows the citation summary network of (Cohn & Blunsom, 2005) in which each node is plotted with a size proportional to its LexRank value within its cluster. This figure shows how C-LexRank emphasizes on picking a diverse set of sentences covering a diverse set of factoids.

Previously, we mentioned that factoids with higher weights appear in a greater number of sentences, and clustering aims to cluster such fact-sharing sentences in the same communities. Thus, starting with the largest community is important to ensure that the system summary first covers the factoids that are more frequently mentioned in other citation sentences and thus are more important.

The last sentence in the example in Table 5.2 is as follows. "We use CRFs as our models for both tasks (Cohn and Blunsom, 2005)." This sentence shows that

Our parsing model is based on a conditional random field model, however, unlike previous **TreeCRF** work, e.g., (Cohn and Blunsom, 2005; Jousse et al., 2006), we do not assume a particular **tree structure**, and instead find the most likely structure and labeling. Some researchers (Xue and Palmer, 2004; Koomen et al., 2005; Cohn and Blunsom, 2005; Punyakanok et al., 2008; Toutanova et al., 2005; Toutanova et al., 2008) used a **pipelined approach** to attack the task. The model can be used for tasks like syntactic parsing (Finkel et al., 2008) and **Semantic Role Labeling** (Cohn and Blunsom, 2005).

Table 5.4: The 100 word summary constructed using C-LexRank for (Cohn & Blunsom, 2005) together with the factoids shown in bold face.

a citation may not cover any contributions of the target paper. Such sentences are assigned by the community detection algorithm in C-LexRank to clusters to which they are semantically most similar. The intuition behind employing LexRank within each cluster is to try to avoid extracting such sentences for the summary, since LexRank within a cluster enforces picking the most central sentence in that cluster. In order to verify this, we also try a variant of C-LexRank in which we do not pick sentences from clusters based on their salience in the cluster, but rather in a round-robin fashion, in which all the sentences within a cluster are equally likely to be picked. We call this variant *C-RR*.

Table 5.4 shows the 100 word summary constructed using C-LexRank for (Cohn & Blunsom, 2005), in which different nuggets are illustrated in bold. This summary is a perfect summary in terms of covering the different factoids about the paper. It includes citing sentences that talk about **tree CRF**, **pipelined approach**, and **Semantic Role Labeling**, which are the three main contributions of (Cohn & Blunsom, 2005).

5.4 Other Methods

In our experiments in Section 5.5 we compare C-LexRank to a number of other summarization systems described below.

5.4.1 Random

For each set, this method simply chooses citations in random order without replacement. Since a citing sentence may cover no information about the cited paper (as in the last sentence in Table 5.2), randomization has the drawback of selecting citations that have no valuable information in them. Moreover, the random selection procedure is more prone to produce redundant summaries, if it selects citing sentences that discuss the same factoid.

5.4.2 LexRank

LexRank [55] works by first building a graph of all the documents (D_i) in a cluster. The edges between corresponding nodes (d_i) represent the cosine similarity between them if the cosine value is above a threshold (0.10 following [55]). Once the network is built, the system finds the most central sentences by performing a random walk on the graph.

$$p(d_j) = (1 - \lambda)\frac{1}{|D|} + \lambda \sum_{d_i} p(d_i)P(d_i \rightarrow d_j) \tag{5.9}$$

5.4.3 MMR

Maximal Marginal Relevance (MMR) is proposed in [28] and is widely used algorithm in generating summaries that reflect the diversity of perspectives in the source documents [42]. It uses the pairwise cosine similarity matrix and greedily chooses sentences that are the least similar to those already in the summary. In particular,

$$MMR = \arg \min_{D_i \in D-A} \left[\max_{D_j \in A} Sim(D_i, D_j) \right] \tag{5.10}$$

where A is the set of documents in the summary, initialized to $A = \emptyset$. In equation 5.10, a sentence D_i that is not in the summary A is chosen such that its highest similarity to summary sentences $\max_{D_j \in A} Sim(D_i, D_j)$ is minimum among all unselected sentences.

5.4.4 DivRank

Unlike other time-homogeneous random walks (e.g., PageRank), DivRank does not assume that the transition probabilities remain constant over time. DivRank uses a *vertex-reinforced random walk* model to rank graph nodes based on a diversity based centrality. The basic assumption in DivRank is that the transition probability from a node to other is reinforced by the number of previous visits to the target node [123]. Particularly, let's assume $p_T(u, v)$ is the transition probability from any node u to node v at time T. Then,

$$p_T(d_i, d_j) = (1 - \lambda).p^*(d_j) + \lambda.\frac{p_0(d_i, d_j).N_T(d_j)}{D_T(d_i)} \tag{5.11}$$

where $N_T(d_j)$ is the number of times the walk has visited d_j up to time T and

$$D_T(d_i) = \sum_{d_j \in V} p_0(d_i, d_j) N_T(d_j) \tag{5.12}$$

Here, $p^*(d_j)$ is the prior distribution that determines the preference of visiting vertex d_j, and $p_0(u, v)$ is the transition probability from u to v prior to any reinforcement. Mei et al. argue that the random walk could stay at the current state at each time, and therefore assumes a hidden link from each node to itself, thus defining $p_0(u, v)$ as

$$p_0(u, v) = \begin{cases} \alpha.\frac{w(u,v)}{deg(u)} & \text{if } u \neq v \\ 1 - \alpha & \text{if } u = v \end{cases} \tag{5.13}$$

Here, we try two variants of this algorithm: *DivRank*, in which $p^*(d_j)$ is uniform, and *DivRank with priors* in which $p^*(d_j) \propto l(D_j)^{-\beta}$, where $l(D_j)$ is the number of the words in the document D_j and β is a parameter (0.1 in our experiments). This prior distribution assigns larger probabilities to shorter sentences which will increase their likelihood of being salient. This will enforce more sentences to be included in the summary, and might increase the factoid coverage. In our experiments, we follow [123] and set $\lambda = 0.90$ and $\alpha = 0.25$.

5.4.5 Trimmer

Trimmer is a sentence-compression tool that extends the scope of an extractive summarization system by generating multiple alternative sentence compressions of the most important sentences in target documents [214]. Trimmer compressions are generated by applying linguistically-motivated rules to mask syntactic components of a parse of a source sentence. The rules can be applied iteratively to compress sentences below a configurable length threshold, or can be applied in all combinations to generate the full space of compressions.

Trimmer can leverage the output of any constituency parser that uses the Penn Treebank conventions. At present, the Stanford Parser [98] is used. The set of compressions is ranked according to a set of features that may include metadata about the source sentences, details of the compression process that generated the compression, and externally calculated features of the compression.

Summaries are constructed from the highest scoring compressions, using the metadata and maximal marginal relevance [28] to avoid redundancy and over-representation of a single source. The summarizer contains a redundancy score using an index index

	Length: 100 words		Length: 200 words	
Method	pyramid	95% C.I.	pyramid	95% C.I.
Random	0.560	[0.465,0.655]	0.763	[0.692,0.834]
MMR	0.600	[0.501,0.699]	0.761	[0.685,0.838]
LexRank	0.604	[0.511,0.696]	0.784	[0.725,0.844]
DivRank	0.644	[0.580,0.709]	0.769	[0.704,0.834]
DivRank (with priors)	0.632	[0.545,0.719]	0.778	[0.716,0.841]
Trimmer	0.571	[0.477,0.665]	0.772	[0.705,0.840]
C-RR	0.513	[0.436,0.591]	0.755	[0.678,0.832]
C-LexRank	**0.647**	[**0.565,0.730**]	**0.799**	[**0.732,0.866**]
C.I.=Confidence Interval				

Table 5.5: Comparison of different ranking systems.

of the words (w) in the document set.

$$\sum_{w} \log(\lambda.P(w|\text{summary}) + (1-\lambda).P(w|\text{corpus})) \tag{5.14}$$

where λ is a weighting factor (set to 0.3 in our experiments).

5.5 Experiments

We used the 30 sets of citations listed in Table 5.1 and employ C-LexRank to produce 2 extractive summaries with different lengths (100 and 200 words) for each set. In addition to C-LexRank and C-RR, we also performed the same experiments with the baseline methods described in Section 5.4, most of which are aimed at leveraging diversity in summarization.

5.5.1 Evaluation

To evaluate our system, we use the pyramid evaluation method [143]. Each factoid in the citations to a paper corresponds to a *summary content unit (SCU)* in [143].

The score given by the pyramid method for a summary is the ratio of the sum of the weights of its factoids to the sum of the weights of an optimal summary. This score ranges from 0 to 1, and high scores show the summary content contain more heavily weighted factoids. We believe that if a factoid appears in more sentences of the citation summary than another factoid, it is more important, and thus should

be assigned a higher weight. To weight the factoids we build a pyramid, and each factoid falls in a tier. Each tier shows the number of sentences a factoid appears in. Thus, the number of tiers in the pyramid is equal to the citation summary size. If a factoid appears in more sentences, it falls in a higher tier. So, if the factoid f_i appears $|f_i|$ times in the citation summary it is assigned to the tier $T_{|f_i|}$.

The pyramid score formula that we use is computed as follows. Suppose the pyramid has n tiers, T_i, where tier T_n on top and T_1 on the bottom. The weight of the factoids in tier T_i will be i (i.e. they appeared in i sentences). If $|T_i|$ denotes the number of factoids in tier T_i, and D_i is the number of factoids in the *summary* that appear in T_i, then the total factoid weight for the summary is

$$D = \sum_{i=1}^{n} i \times D_i \tag{5.15}$$

Additionally, the optimal pyramid score for a summary with X factoids, is

$$Max = \sum_{i=j+1}^{n} i \times |T_i| + j \times (X - \sum_{i=j+1}^{n} |T_i|) \tag{5.16}$$

where $j = \max_i(\sum_{t=i}^{n} |T_t| \geq X)$. Subsequently, the pyramid score for a summary is calculated as

$$P = \frac{D}{Max} \tag{5.17}$$

Table 5.5 shows the average pyramid score of the summaries generated using different methods with different lengths. Longer summaries result in higher pyramid scores since the amount of information they cover is greater than shorter summaries. C-LexRank outperforms all other methods that leverage diversity as well as random summaries and LexRank. The results in this table also suggest that employing LexRank within each cluster is essential for the selection of salient citing sentences, as the average pyramid scores from C-RR, where sentences are picked in a round-robin fashion, are lower.

5.5.2 Relative Utility

Since inter-judge agreement measured by Precision and percent agreement are low for extractive summaries, it is practically impossible to write summarizers which are optimized for these measures. Relative Utility is an multi-document summarization evaluation system, which provides an intuitive mechanism and takes into account human judge disagreement on sentences that belong in a summary [168, 191].

In this section, we perform more evaluations on the proposed systems and different summary lengths using *Relative Utility* [191]. In Relative Utility (RU), a number of judges, $N(N > 1)$, are asked to assign utility scores to all n sentences in a set of documents. The sentence utility score of judge i over n sentences is defined as follows.

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

The summary based utility vector for judge i is then defined as

$$\vec{U_i'} = \{\delta_{i,1} \cdot u_{i,1}, \delta_{i,2} \cdot u_{i,2}, \cdots, \delta_{i,n} \cdot u_{i,n}\}$$

where $\delta_{i,j}$ is the summary characteristic function for judge i and sentence j. Accordingly, the total self-utility and total extractive self-utility for judge i are respectively defined as follows.

$$U_i = \sum_{j=1}^{n} u_{i,j} \tag{5.18}$$

$$U_i' = \sum_{j=1}^{n} \delta_{i,j} \cdot u_{i,j} \tag{5.19}$$

For a summarization system, RU will be computed as its performance against the human judges divided by the maximum possible performance. In other words, the ratio of the sum of its cross-utility with the totality of human judges and the maximum utility U' achievable at a given summary length e:

	Length: 100 words	
Method	RU	95% C.I.
Random	0.234	[0.176,0.292]
MMR	0.228	[0.170,0.286]
LexRank	0.238	[0.186,0.290]
DivRank	0.240	[0.204,0.277]
DivRank (with priors)	0.204	[0.152,0.256]
Trimmer	0.074	[0.032,0.116]
C-RR	0.178	[0.135,0.221]
C-LexRank	**0.245**	**[0.187,0.302]**
C.I.=Confidence Interval		

Table 5.6: Comparison of different ranking systems using Relative Utility (RU)

$$S = \frac{\sum_{j=1}^{n} \delta_{s,j} \cdot \sum_{i=1}^{N} u_{i,j}}{U'} \tag{5.20}$$

In this formula, $\sum_{i=1}^{N} u_{i,j}$ is the utility assigned by the totality of judges to a given sentence j extracted by the summarizer. In the setting of our annotations for citations to scientific papers, we assume that each factoid represents a judge. And each judge assigns the same utility to any sentence containing that factoid. Intuitively, each factoid judges a sentence to determine whether that sentence has any nuggets representing the factoid. The utility that a judge assigns to such sentences in our work is proportional to the count of such nuggets (therefore, more important factoids will be assigned higher utility values).

Table 5.6 shows the average RU value for each of the summarization systems, when summaries of 100 words are generated. This Table suggests that the experimental results from Relative Utility are consistent with those of pyramid scores. For instance, C-LexRank is the best system using both measures and DivRank and LexRank follow. Moreover, this Table suggests that RU is not a suitable evaluation measure for abstractive summarization systems such as Trimmer. Intuitively, the judges assign utility scores to sentences in the original documents. Therefore, systems such as Trimmer, which alter original sentences could utility assignments intractable.

CHAPTER VI

Factoid Extraction

6.1 Contribution Extraction

In the first part of this chapter, we present an approach to summarize a single scientific paper, by extracting its contributions from the set of citation sentences written in other papers. Our methodology, as explained in [164] is based on extracting significant keyphrases from the set of citation sentences and using these keyphrases to build the summary. Comparisons show how this methodology excels at the task of single paper summarization, and how it out-performs other multi-document summarization methods.

6.1.1 Introduction

In recent years statistical physicists and computer scientists have shown great interest in analyzing complex adaptive systems. The study of such systems can provide valuable insight on the behavioral aspects of the involved agents with potential applications in economics and science. One such aspect is to understand what motivates people to provide the $n + 1^{st}$ review of an artifact given that they are unlikely to add something significant that has not already been said or emphasized. Citations are part of such complex systems where articles use citations as a way to mention different contributions of other papers, resulting in a collective system.

The focus of this work is on the corpora created based on citation sentences. A citation sentence is a sentence in an article containing a citation and can contain zero or more *nuggets* (i.e., non-overlapping contributions) about the cited article. For example the following sentences are a few citation sentences that appeared in the NLP literature in past that talk about Resnik's work.

The STRAND system **(Resnik, 1999)***, for example, uses structural markup information from the pages, without looking at their content, to attempt to align them.*

Resnik (1999) *addressed the issue of language identification for finding Web pages in the languages of interest.*

Mining the Web for bilingual text **(Resnik, 1999)** *is not likely to provide sufficient quantities of high quality data..*

The set of citations is important to analyze because human summarizers have put their effort collectively but independently to read the target article and cite its important contributions. This has been shown in other work too [50, 139, 159, 124, 134]. In this work, we introduce a technique to summarize the set of citation sentences and cover the major contributions of the target paper. Our methodology first finds the set of keyphrases that represent important information units (i.e., nuggets), and then finds the best set of k sentences to cover more, and more important nuggets.

Our results confirm the effectiveness of the method and show that it outperforms other state of the art summarization techniques. Moreover, as shown in the chapter, this method does not need to calculate the full cosine similarity matrix for a document cluster, which is the most time consuming part of the mentioned baseline methods.

ACL-ID	Title
N03-1017	Statistical Phrase-Based Translation
P02-1006	Learning Surface Text Patterns For A Question Answering System
P05-1012	On-line Large-Margin Training Of Dependency Parsers
C96-1058	Three New Probabilistic Models For Dependency Parsing: An Exploration
P05-1033	A Hierarchical Phrase-Based Model For Statistical Machine Translation
P97-1003	Three Generative, Lexicalized Models For Statistical Parsing
P99-1065	A Statistical Parser For Czech
J04-4002	The Alignment Template Approach To Statistical Machine Translation
D03-1017	Towards Answering Opinion Questions: Separating Facts From Opinions ...
P05-1013	Pseudo-Projective Dependency Parsing
W00-0403	Centroid-Based Summarization Of Multiple Documents: Sentence Extraction, ...
P03-1001	Offline Strategies For Online Question Answering: Answering Questions Before ...
N04-1033	Improvements In Phrase-Based Statistical Machine Translation
A00-2024	Cut And Paste Based Text Summarization
W00-0603	A Rule-Based Question Answering System For Reading Comprehension Tests
A00-1043	Sentence Reduction For Automatic Text Summarization
C00-1072	The Automated Acquisition Of Topic Signatures For Text Summarization
W05-1203	Measuring The Semantic Similarity Of Texts
W03-0510	The Potential And Limitations Of Automatic Sentence Extraction For Summarization
W03-0301	An Evaluation Exercise For Word Alignment
A00-1023	A Question Answering System Supported By Information Extraction
D04-9907	Scaling Web-Based Acquisition Of Entailment Relations
P05-1014	The Distributional Inclusion Hypotheses And Lexical Entailment
H05-1047	A Semantic Approach To Recognizing Textual Entailment
H05-1079	Recognising Textual Entailment With Logical Inference

Table 6.1: List of papers chosen from AAN for evaluation together with the number of sentences citing each.

6.1.2 Data

In order to evaluate our method, we use the ACL Anthology Network (AAN), which is a collection of papers from the Computational Linguistics journal and proceedings from ACL conferences and workshops and includes more than $13,000$ papers [167]. We use 25 manually annotated papers from [159], which are highly cited articles in AAN. Table 6.1 shows the ACL ID and the title of these papers.

The annotation guidelines asked a number of annotators to read the citation summary of each paper and extract a list of the main contributions of that paper. Each item on the list is a non-overlapping contribution (nugget) perceived by reading the citation summary. The annotation strictly instructed the annotators to focus on the citing sentences to do the task and not their own background on the topic. Then, extracted nuggets are reviewed and those nuggets that have only been mentioned by 1 annotator are removed. Finally, the union of the rest is used as a set of nuggets representing each paper.

Fact	Occurrences
f_1: " Supervised Learning"	5
f_2: " instance/concept relations"	3
f_3: "Part-of-Speech tagging"	3
f_4: "filtering QA results"	2
f_5: "lexico-semantic information"	2
f_6: "hyponym relations"	2

Table 6.2: Nuggets of P03-1001 extracted by annotators.

	unique	all	max freq
unigrams	229,631	7,746,792	437,308
bigrams	2,256,385	7,746,791	73,957
3-grams	5,125,249	7,746,790	3,600
4-grams	6,713,568	7,746,789	2,408

Table 6.3: Statistics on the abstract corpus in AAN used as the background data

Table 6.2 lists the nuggets extracted by annotators for P03-1001.

6.1.3 Methodology

Our methodology assumes that each citation sentence covers 0 or more nuggets about the cited papers, and tries to pick sentences that maximize nugget coverage with respect to summary length.

These nuggets are essentially represented using keyphrases. Therefore, we try to extract significant keyphrases in order to represent nuggets each sentence contains. Here, the keyphrases are expressed using N-grams, and thus these building units are the key elements to our summarization. For each citation sentence d_i, our method first extracts a set of important keyphrases, D_i, and then tries to find sentences that have a larger number of important and non-redundant keyphrases. In order to take the first step, we extract statistically significantly frequent N-grams (up to $N = 4$) from each citing sentence and use them as the set of representative keyphrases for that citing sentence.

6.1.4 Automatic Keyphrase Extraction

A list of keyphrases for each citation sentence can be generated by extracting N-grams that occur significantly frequently in that sentence compared to a large corpus of such N-grams. Our method for such an extraction is inspired by the previous work by Tomokiyo and Hurst [196].

A language model, $\mathcal{M}$, is a statistical model that assigns probabilities to a sequence of N-grams. Every language model is a probability distribution over all N-grams and thus the probabilities of all N-grams of the same length sum up to 1. In order to extract keyphrases from a text using statistical significance we need two language models. The first model is referred to as the *Background Model ($\mathcal{BM}$)* and is built using a large text corpus. Here we build the BM using the text of all the paper abstracts provided in AAN[1]. The second language model is called the *Foreground Model ($\mathcal{FM}$)* and is the model built on the text from which keyphrases are being extracted. In this work, the set of all citation sentences that cite a particular target paper are used to build a foreground language model.

Let g^i be an N-gram of size i and $C_{\mathcal{M}}(g^i)$ denote the count of g^i in the model $\mathcal{M}$. First, we extract the counts of each N-grams in both the background ($\mathcal{BM}$) and the foreground corpora ($\mathcal{FM}$).

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

$$
\begin{aligned}
M_{\mathcal{BM}} &= \sum_{g^i \in \{\mathcal{BM} \cup \mathcal{FM}\}} 1 \\
N_{\mathcal{BM}} &= \sum_{g^i \in \{\mathcal{BM} \cup \mathcal{FM}\}} C_{\mathcal{BM}}(g^i) \\
N_{\mathcal{FM}} &= \sum_{g^i \in \mathcal{FM}} C_{\mathcal{FM}}(g^i) \\
\hat{p}_{\mathcal{FM}}(g^i) &= C_{\mathcal{FM}}(g^i)/N_{\mathcal{FM}} \\
\hat{p}_{\mathcal{BM}}(g^i) &= (C_{\mathcal{BM}}(g^i) + 1)/(M_{\mathcal{BM}} + N_{\mathcal{BM}})
\end{aligned}
$$

The last equation is also known as Laplace smoothing [122] and handles the N-grams in the foreground corpus that have a 0 occurrence frequency in the background corpus. Next, we extract N-grams from the foreground corpus that have significant frequencies compared to the frequency of the same N-grams in the background model and its individual terms in the foreground model.

To measure how randomly a set of consecutive terms are forming an N-gram, Tomokiyo and Hurst [196] use point-wise divergence. In particular, for an N-gram of size i, $g^i = (w_1 w_2 \cdots w_i)$,

$$
\delta_{g^i}(\mathcal{FM}^i \| \mathcal{FM}^1) = \hat{p}_{\mathcal{FM}}(g^i) \log\left(\frac{\hat{p}_{\mathcal{FM}}(g^i)}{\prod_{j=1}^{i} \hat{p}_{\mathcal{FM}}(w_j)}\right)
$$

This equation shows the extent to which the terms forming g^i have occurred together randomly. In other words, it indicates the extent of information that we lose by assuming independence of each word by applying the unigram model, instead of the N-gram model.

In addition, to measure how randomly a sequence of words appear in the foreground model with respect to the background model, we use point-wise divergence as

well. Here, point-wise divergence defines how much information we lose by assuming that g^i is drawn from the background model instead of the foreground model:

$$\delta_{g^i}(\mathcal{FM}^i \| \mathcal{BM}^i) \quad = \quad \hat{p}_{\mathcal{FM}}(g^i) \log(\frac{\hat{p}_{\mathcal{FM}}(g^i)}{\hat{p}_{\mathcal{BM}}(g^i)})$$

We set the criteria of choosing a sequence of words as significant to be whether it has positive point-wise divergence with respect to both the background model, and individual terms of the foreground model. In other words we extract all g^i from $\mathcal{FM}$ for which the both properties are positive:

$$\delta_{g^i}(\mathcal{FM}^i \| \mathcal{BM}^i) > 0$$

$$\delta_{g^i}(\mathcal{FM}^i \| \mathcal{FM}^1) \geq 0$$

The equality condition in the second equation is specifically set to handle unigrams, in which $\hat{p}_{\mathcal{FM}}(g^i) = \prod_{j=1}^{i} \hat{p}_{\mathcal{FM}}(w_j)$.

In order to handle the text corpora and building the language models, we have used the CMU-Cambridge Language Model toolkit [37]. We use the set of citation sentences for each paper to build foreground language models. Furthermore, we employ this tool and make the background model using nearly 11,000 abstracts from AAN. Table 6.1.3 summarizes some of the statistics about the background data.

Once keyphrases (significant N-grams) of each sentence are extracted, we remove all N-grams in which more than half of the terms are stopwords. For instance, we remove all stopword unigrams, if any, and all bigrams with at least one stopword in them. For 3-grams and 4-grams we use a threshold of 2 and 3 stopwords respectively. After that, the set of remaining N-grams is used to represent each sentence and to

(Corley and Mihalcea, 2005) applied or utilized lexical based word overlap measures.
{overlap measures, word overlap, lexical based, utilized lexical}

Table 6.4: Example: citation sentence for W05-1203 written by D06-1621, and its extracted bigrams.

build summaries. Table 6.4 shows an example of a citation sentence from D06-1621 citing W05-1203 (Corley and Mihalcea, 2005), and its extracted bigrams.

6.1.5 Sentence Selection

After extracting the set of keyphrases for each sentence, d_i, the sentence is represented using its set of N-grams, denoted by D_i. Then, the goal is to pick sentences (sets) for each paper that cover more important and non-redundant keyphrases. Essentially, keyphrases that have been repeated in more sentences are more important and could represent more important nuggets. Therefore, sentences that contain more frequent keyphrases are more important. Based on this intuition we define the reward of building a summary comprising a set of keyphrases S as

$$f(S) = |S \cap A|$$

where A is the set of all keyphrases from sentences not in the summary.

The set function f has three main properties. First, it is non-negative. Second, it is monotone (i.e., for every set v we have $f(S+v) \geq f(S)$). Third, f is sub-modular. The submodularity means that for a set v and two sets $S \subseteq T$ we have

$$f(S + v) - f(S) \geq f(T + v) - f(T)$$

Intuitively, this property implies that adding a set v to S will increase the reward at least as much as it would to a larger set T. In the summarization setting, this means that adding a sentence to a smaller summary will increase the reward of the

summary at least as much as adding it to a larger summary that subsumes it. The following theorem formalizes this and is followed by a proof.

Theorem 1. *The reward function f is submodular.*

Proof

We start by defining a gain function $\mathcal{G}$ of adding sentence (set) D_i to $\mathcal{S}_{k-1}$ where $\mathcal{S}_{k-1}$ is the set of keyphrases in a summary built using $k-1$ sentences, and D_i is a candidate sentence to be added:

$$\mathcal{G}(D_i, \mathcal{S}_{k-1}) = f(\mathcal{S}_{k-1} \cup D_i) - f(\mathcal{S}_{k-1})$$

Simple investigation through a Venn diagram proof shows that $\mathcal{G}$ can be re-written as

$$\mathcal{G}(D_i, \mathcal{S}_{k-1}) = |D_i \cap (\cup_{j \neq i} D_j) - \mathcal{S}_{k-1}|$$

Let's denote $D_i \cap (\cup_{j \neq i} D_j)$ by $\cap_i$. The following equations prove the theorem.

$$\mathcal{S}_{k-1} \subseteq \mathcal{S}_k$$

$$\mathcal{S}'_{k-1} \supseteq \mathcal{S}'_k$$

$$\cap_i \cap \mathcal{S}'_{k-1} \supseteq \cap_i \cap \mathcal{S}'_k$$

$$\cap_i - \mathcal{S}_{k-1} \supseteq \cap_i - \mathcal{S}_k$$

$$|\cap_i - \mathcal{S}_{k-1}| \geq |\cap_i - \mathcal{S}_k|$$

$$\mathcal{G}(D_i, \mathcal{S}_{k-1}) \geq \mathcal{G}(D_i, \mathcal{S}_k)$$

$$f(\mathcal{S}_{k-1} \cup D_i) - f(\mathcal{S}_{k-1}) \geq f(\mathcal{S}_k \cup D_i) - f(\mathcal{S}_k)$$

Here, $\mathcal{S}'_k$ is the set of all N-grams in the vocabulary that are not present in $\mathcal{S}_k$. The gain of adding a sentence, D_i, to an empty summary is a non-negative value.

$$\mathcal{G}(D_i, \mathcal{S}_0) = C \geq 0$$

Summary generated using bigram-based keyphrases	
ID	Sentence
P06-1048:1	Ziff-Davis Corpus Most previous work (Jing 2000; Knight and Marcu 2002; Riezler et al 2003; Nguyen et al 2004a; Turner and Charniak 2005; McDonald 2006) has relied on automatically constructed parallel corpora for training and evaluation purposes.
J05-4004:18	Between these two extremes, there has been a relatively modest amount of work in sentence simplification (Chandrasekar, Doran, and Bangalore 1996; Mahesh 1997; Carroll et al 1998; Grefenstette 1998; Jing 2000; Knight and Marcu 2002) and document compression (Daume III and Marcu 2002; Daume III and Marcu 2004; Zajic, Dorr, and Schwartz 2004) in which words, phrases, and sentences are selected in an extraction process.
A00-2024:9	The evaluation of sentence reduction (see (Jing, 2000) for details) used a corpus of 500 sentences and their reduced forms in human-written abstracts.
N03-1026:17	To overcome this problem, linguistic parsing and generation systems are used in the sentence condensation approaches of Knight and Marcu (2000) and Jing (2000).
P06-2019:5	Jing (2000) was perhaps the first to tackle the sentence compression problem.

Table 6.5: Bigram-based summary generated for A00-1043.

By induction, we will get

$$\mathcal{G}(D_i, \mathcal{S}_0) \geq \mathcal{G}(D_i, \mathcal{S}_1) \geq \cdots \geq \mathcal{G}(D_i, \mathcal{S}_k) \geq 0$$

□

Theorem 1 implies the general case of submodularity:

$$\forall m, n, 0 \leq m \leq n \leq |D| \Rightarrow \mathcal{G}(D_i, \mathcal{S}_m) \geq \mathcal{G}(D_i, \mathcal{S}_n)$$

Maximizing this submodular function is an NP-hard problem [95]. A common way to solve this maximization problem is to start with an empty set, and in each iteration pick a set that maximizes the gain. It has been shown before in [102] that if f is a submodular, nondecreasing set function and $f(\emptyset) = 0$, then such a greedy algorithm finds a set $\mathcal{S}$, whose gain is at least as high as $(1 - 1/e)$ of the best possible solution. Therefore, we can optimize the keyphrase coverage as described in Algorithm of Table 6.6.

6.1.6 Experimental Setup

We use the annotated data described in Section 6.1.2. In summary, the annotation consisted of two parts: nugget extraction and nugget distribution analysis. Five annotators were employed to annotate the sentences in each of the 25 citation

Algorithm 1 The greedy algorithm for summary generation

$k \leftarrow$ the number of sentences in the summary
$D_i \leftarrow$ keyphrases in d_i
$\mathcal{S} \leftarrow \emptyset$
for $l = 1$ to k **do**
    $s_l \leftarrow \arg\max_{D_i \in D} |D_i \cap (\cup_{j \neq i} D_j)|$
    $\mathcal{S} \leftarrow \mathcal{S} \cup s_l$
    for $j = 1$ to $|D|$ **do**
        $D_j \leftarrow D_j - s_l$
    end for
end for
return S

Table 6.6: The greedy algorithm for summary generation

summaries and write down the nuggets (non-overlapping contributions) of the target paper. Then using these nugget sets, each sentence was annotated with the nuggets it contains. This results in a sentence-fact matrix that helps with the evaluation of the summary. The summarization goal and the intuition behind the summarizing system is to select a few (5 in our experiments) sentences and cover as many nuggets as possible. Each sentence in a citation summary may contain 0 or more nuggets and not all nuggets are mentioned an equal number of times. Covering some nuggets (contributions) is therefore more important than others and should be weighted highly.

To capture this property, the pyramid score seems the best evaluation metric to use. We use the pyramid evaluation method [143] at the sentence level to evaluate the summary created for each set. We benefit from the list of annotated nuggets provided by the annotators as the ground truth of the summarization evaluation. These annotations give the list of nuggets covered by each sentence in each citation summary, which are equivalent to the *summarization content unit (SCU)* as described in [143].

The pyramid score for a summary is calculated as follows. Assume a pyramid that has n tiers, T_i, where tier $T_i > T_j$ if $i > j$ (i.e., T_i is not below T_j, and that if a

nugget appears in more sentences, it falls in a higher tier.). Tier T_i contains nuggets that appeared in i sentences, and thus has weight i. Suppose $|T_i|$ shows the number of nuggets in tier T_i, and Q_i is the size of a subset of T_i whose members appear in the summary. Further suppose Q shows the sum of the weights of the facts that are covered by the summary. $Q = \sum_{i=1}^{n} i \times Q_i$.

In addition, the optimal pyramid score for a summary with X facts, is

$$Max = \sum_{i=j+1}^{n} i \times |T_i| + j \times (X - \sum_{i=j+1}^{n} |T_i|)$$

where $j = \max_i(\sum_{t=i}^{n} |T_t| \geq X)$. The pyramid score for a summary is then calculated as follows.

$$P = \frac{Q}{Max}$$

This score ranges from 0 to 1, and a high score shows the summary contains more heavily weighted facts.

6.1.7 Baselines and Gold Standards

To evaluate the quality of the summaries generated by the greedy algorithm, we compare its pyramid score in each of the 25 citation summaries with those of a gold standard, a random summary, and four other methods. The gold standards are summaries created manually using 5 sentences. The 5 sentences are manually selected in a way to cover as many nuggets as possible with higher priority for the nuggets with higher frequencies. We also created random summaries using Mead [166]. These summaries are basically a random selection of 5 sentences from the pool of sentences in the citation summary. Generally we expect the summaries created by the greedy method to be significantly better than random ones.

In addition to the gold and random summaries, we also used 4 baseline state of the art summarizers: LexRank, the clustering C-RR and C-LexRank, and Maximal Marginal Relevance (MMR). LexRank [55] works based on a random walk on the cosine similarity of sentences and prints out the most frequently visited sentences. Said differently, LexRank first builds a network in which nodes are sentences and edges are cosine similarity values. It then uses the eigenvalue centralities to find the most central sentences. For each set, the top 5 sentences on the list are chosen for the summary.

The clustering methods, C-RR and C-LexRank, work by clustering the cosine similarity network of sentences. In such a network, nodes are sentences and edges are cosine similarity of node pairs. Clustering would intuitively put nodes with similar nuggets in the same clusters as they are more similar to each other. The C-RR method as described in [159] uses a round-robin fashion to pick sentences from each cluster, assuming that the clustering will put the sentences with similar facts into the same clusters. Unlike C-RR, C-LexRank uses LexRank to find the most salient sentences in each cluster, and prints out the most central nodes of each cluster as summary sentences.

Finally, MMR uses the full cosine similarity matrix and greedily chooses sentences that are the least similar to those already selected for the summary [28]. In particular,

$$MMR = \arg \min_{d_i \in D-A} \left[\max_{d_j \in A} Sim(d_i, d_j) \right]$$

where A is the set of sentences in the summary, initially set to $A = \emptyset$. This method is different from ours in that it chooses the least similar sentence to the summary in each iteration.

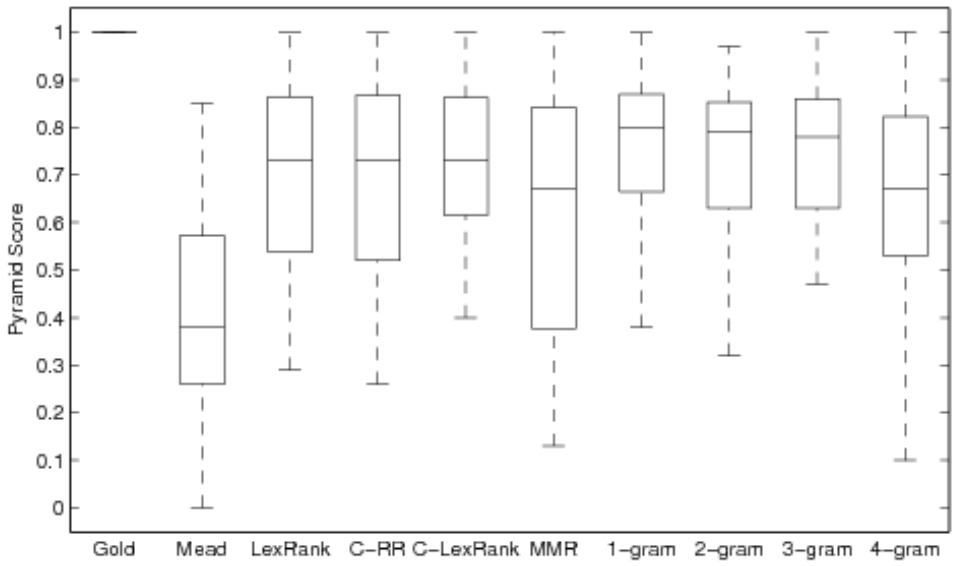


Figure 6.1: Evaluation Results (summaries with 5 sentences): The median pyramid score over 25 datasets using different methods.

6.1.8 Results and Discussion

As mentioned before, we use the text of the abstracts of all the papers in AAN as the background, and each citation set as a separate foreground corpus. For each citation set, we use the method described in Section 6.1.4 to extract significant N-grams of each sentence. We then use the keyphrase set representation of each sentence to build the summaries using Algorithm 6.6. For each of the 25 citation summaries, we build 4 different summaries using unigrams, bigrams, 3-grams, and 4-grams respectively. Table 6.5 shows a 5-sentence summary created using algorithm 6.6 for the paper A00-1043 [87].

The pyramid scores for different methods are reported in Figure 6.1 together with the scores of gold standards, manually created to cover as many nuggets as possible in 5 sentences, as well as summary evaluations of the 4 baseline methods

described above. This Figure shows how the keyphrase based summarization method when employing N-grams of size 3 or smaller, outperforms other baseline systems significantly. More importantly, Figure 6.1 also indicates that this method shows more stable results and low variation in summary quality when keyphrases of size 3 or smaller are employed. In contrast, MMR shows high variation in summary qualities making summaries that obtain pyramid scores as low as 0.15.

Another important advantage of this method is that we do not need to calculate the cosine similarity of the pairs of sentences, which would add a running time of $O(|D|^2|V|)$ in the number of documents, $|D|$, and the size of the vocabulary $|V|$ to the algorithm.

6.1.9 Conclusion

This chapter presents a summarization methodology that employs keyphrase extraction to find important contributions of scientific articles. The summarization is based on citation sentences and picks sentences to cover nuggets (represented by keyphrases) or contributions of the target papers. In this setting the best summary would have as few sentences and at the same time as many nuggets as possible. In this work, we use point-wise KL-divergence to extract statistically significant N-grams and use them to represent nuggets. We then apply a new set function for the task of summarizing scientific articles. We have proved that this function is submodular and concluded that a greedy algorithm will result in a near-optimum set of covered nuggets using only 5 sentences. Our experiments in this paper confirm that the summaries created based on the presented algorithm are better than randomly generated summary, and also outperform other state of the art summarization methods in most cases. Moreover, we show how this method generates more stable summaries with lower variation in summary quality when N-grams of size 3 or smaller are employed.

6.2 Communities of Contributions

In previous chapters we showed that the diversity seen in human summaries could be according to different nuggets or phrases that represent the same factoid. Ideally, a summarizer that seeks to increase diversity should capture this phenomenon and avoid covering redundant nuggets. C-LexRank achieves this by finding communities of sentences that represent the same factoid. However in datasets such as citations a sentence may contain more than one factoid. Assigning each sentence to one cluster will ignore this phenomenon.

In this chapter, we discuss C-LexRank when it is run on words and not documents. We represent the set of words in a corpus as a network, where edges show the similarity of words using the *distributional hypothesis*. By applying C-LexRank on this network, we find communities of words that are more similar to each other whereby each community represent the set of words that relate to one factoid.

6.2.1 Distributional Similarity

Measuring the semantic relatedness of words is a fundamental problem in natural language process and and has many useful applications, including textual entailment, word sense disambiguation, information retrieval and automatic thesaurus discovery. Existing approaches can be roughly categorized into two kinds: knowledge-based and corpus-based, where the former includes graph-based algorithms and similarity measures operating on a lexical database such as WordNet [25, 4] and the latter consists of various kinds of vector space models (VSMs) constructed with the help of a large collection of text [172, 169].

Corpus-based vector space models follow the standard *distributional hypothesis*, which states that words appearing in the same *contexts* tend to have similar mean-

ing [73, 107]. Each target word is thus represented by a high-dimensional sparse term-vectors that consists of words occurring in its context.

The definition of a context varies from the neighboring words [120] to words that are linked in a syntactic dependency structure [115, 150]. Nevertheless, previous study shows that the performance differences of different context definitions are limited [4]. For simplicity and scalability, we use the following bag-of-words approach to construct term vectors.

Given a corpus, we first collect terms within a window of $[-3, +3]$ centered at each occurrences of a target word. This bag-of-words representation is then mapped to the TF-IDF term vector: each term is weighted by $\log(freq) \times \log(N/df)$, where $freq$ is the number of times the term appears in the collection, df the document frequency of the term in the whole corpus and N the number of total documents.

Particularly, each word w_i is represented by a term vector ℓ_i, using the words that have a surface distance of 3 or smaller to w_i anywhere in the cluster. In other words, ℓ_i contains any word that co-occurs with w_i in a 4-gram in the cluster. This *bag of words* representation of words enables us to find the word-pair similarities.

$$sim(w_i, w_j) = \frac{\vec{\ell_i} \cdot \vec{\ell_j}}{\sqrt{|\vec{\ell_i}||\vec{\ell_j}|}} \tag{6.1}$$

We use the pair-wise similarities of words in each cluster, and build a network of words and their similarities. Intuitively, words that appear in similar contexts are more similar to each other and will have a stronger edge between them in the network. Therefore, similar words, or words that appear in similar contexts, will form communities in this graph. Ideally, each community in the word similarity network would represent a factoid. To find the communities in the word network we

use [38], a hierarchical agglomeration algorithm which works by greedily optimizing the modularity in a linear running time for sparse graphs.

The community detection algorithm will assign to each word w_i, a community label C_i. For each community, we use LexRank to rank the words using the similarities in Equation 6.1, and assign a score to each word w_i as $S(w_i) = \frac{R_i}{|C_i|}$, where R_i is the rank of w_i in its community, and $|C_i|$ is the number of words that belong to C_i. Figure 6.2.1 shows part of the word similarity graph in the `redsox` cluster, in which each node is color-coded with its community. This figure illustrates how words that are semantically related to the same aspects of the story fall in the same communities (e.g., "police" and "arrest"). Finally, to rank sentences, we define the score of each document D_j as the sum of the scores of its words.

$$p_{ds}(D_j) = \sum_{w_i \in D_j} S(w_i)$$

Intuitively, sentences that contain higher ranked words in highly populated communities will have a smaller score. To rank the sentences, we sort them in an ascending order, and cut the list when its size is greater than the length limit.

6.2.2 Other Methods

Random

For each cluster in each category (citations and headlines), this method simply gets a random permutation of the summaries. In the headlines datasets, where most of the headlines cover some factoids about the story, we expect this method to perform reasonably well since randomization will increase the chances of covering headlines that focus on different factoids. However, in the citations dataset, where a citing sentence may cover no information about the cited paper, randomization has

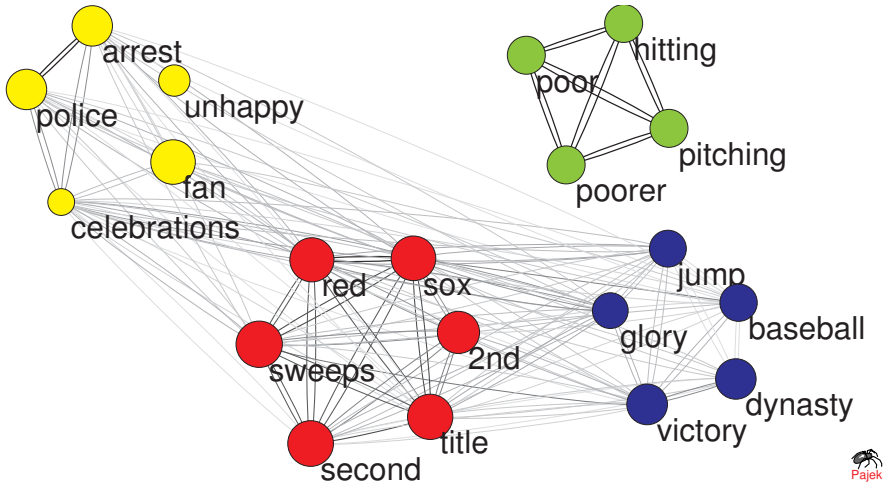


Figure 6.2: Part of the word similarity graph in the redsox cluster

the drawback of selecting citations that have no valuable information in them.

LexRank

LexRank [55] works by first building a graph of all the documents (D_i) in a cluster. The edges between corresponding nodes (d_i) represent the cosine similarity between them is greater than a threshold (0.10 following [55]). Once the network is built, the system finds the most central sentences by performing a random walk on the graph.

$$p(d_j) = (1 - \lambda)\frac{1}{|D|} + \lambda \sum_{d_i} p(d_i)P(d_i \rightarrow d_j) \tag{6.2}$$

MMR

Maximal Marginal Relevance (MMR) [28] uses the pairwise cosine similarity matrix and greedily chooses sentences that are the least similar to those already in the summary. In particular,

$MMR = \arg\min_{D_i \in D-A}\left[\max_{D_j \in A} Sim(D_i, D_j)\right]$

where A is the set of documents in the summary, initialized to $A = \emptyset$.

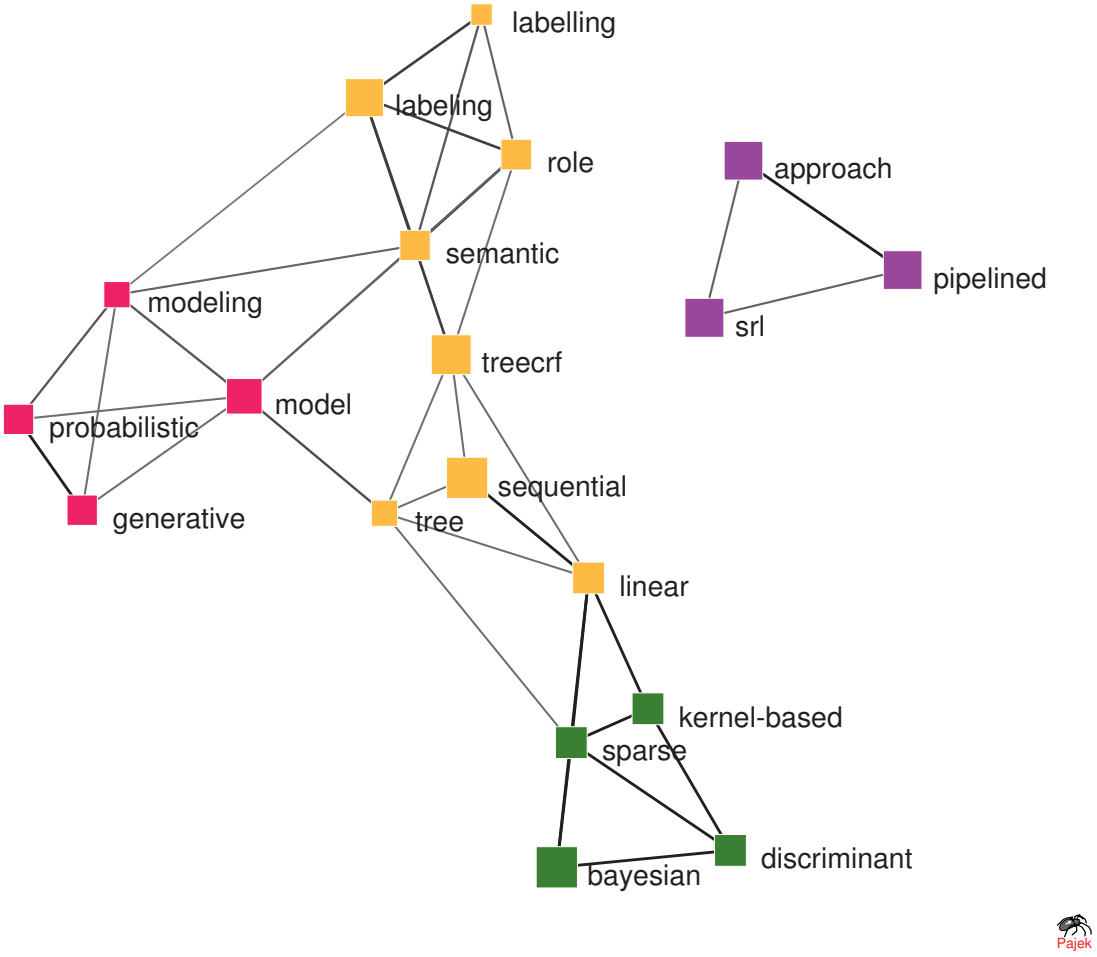


Figure 6.3: Part of the word similarity graph in the citation cluster

DivRank

Unlike other time-homogeneous random walks (e.g., PageRank), DivRank does not assume that the transition probabilities remain constant over time. DivRank uses a *vertex-reinforced random walk* model to rank graph nodes based on a diversity based centrality. The basic assumption in DivRank is that the transition probability from a node to other is reinforced by the number of previous visits to the target node [123]. Particularly, let's assume $p_T(u, v)$ is the transition probability from any node u to node v at time T. Then,

$$p_T(d_i, d_j) = (1 - \lambda).p^*(d_j) + \lambda.\frac{p_0(d_i, d_j).N_T(d_j)}{D_T(d_i)} \tag{6.3}$$

where $N_T(d_j)$ is the number of times the walk has visited d_j up to time T and

$$D_T(d_i) = \sum_{d_j \in V} p_0(d_i, d_j) N_T(d_j) \tag{6.4}$$

Here, $p^*(d_j)$ is the prior distribution that determines the preference of visiting vertex d_j. We try two variants of this algorithm: **DivRank**, in which $p^*(d_j)$ is uniform, and **DivRank with priors** in which $p^*(d_j) \propto l(D_j)^{-\beta}$, where $l(D_j)$ is the number of the words in the document D_j and β is a parameter ($\beta = 0.8$).

C-LexRank

C-LexRank is a clustering-based model in which the cosine similarities of document pairs are used to build a network of documents. Then the the network is split into communities, and the most salient documents in each community are selected [159]. C-LexRank focuses on finding communities of documents using their cosine similarity. The intuition is that documents that are more similar to each

Method	headlines		citations		Mean
	pyramid	95% C.I.	pyramid	95% C.I.	
R	0.928	[0.896, 0.959]	0.716	[0.625, 0.807]	0.822
MMR	0.930	[0.902, 0.960]	0.766	[0.684, 0.847]	0.848
LR	0.918	[0.891, 0.945]	0.728	[0.635, 0.822]	0.823
DR	0.927	[0.900, 0.955]	0.736	[0.667, 0.804]	0.832
DR(p)	0.916	[0.884, 0.949]	0.764	[0.697, 0.831]	0.840
C-LR	0.942	[0.919, 0.965]	0.781	[0.710, 0.852]	0.862
WDS	0.931	[0.905, 0.958]	0.813	[0.738, 0.887]	**0.872**

R=Random; LR=LexRank; DR=DivRank; DR(p)=DivRank with Priors; C-LR=C-LexRank; WDS=Word Distributional Similarity; C.I.=Confidence Interval

Table 6.7: Comparison of different ranking systems

other contain similar factoids. We expect C-LexRank to be a strong ranker, but incapable of capturing the diversity caused by using different phrases to express the same meaning. The reason is that different nuggets that represent the same factoid often have no words in common (e.g., "victory" and "glory") and won't be captured by a lexical measure like cosine similarity.

6.2.3 Experiments

We use each of the systems explained above to rank the summaries in each cluster. Each ranked list is then cut at a certain length (50 words for headlines, and 150 for citations) and the information content in the remaining text is examined using the pyramid score.

Table 6.7 shows the average pyramid score achieved by different methods in each category. The method based on the distributional similarities of words outperforms other methods in the citations category. All methods show similar results in the headlines category, where most headlines cover at least 1 factoid about the story and a random ranker performs reasonably well. Table 6.8 shows top 3 headlines from 3 rankers: word distributional similarity (WDS), C-LexRank, and MMR. In this example, the first 3 headlines produced by WDS cover two important factoids:

Method	Top 3 headlines
WDS	1: how sweep it is 2: fans celebrate red sox win 3: red sox take title
C-LR	1: world series: red sox sweep rockies 2: red sox take world series 3: red sox win world series
MMR	1:red sox scale the rockies 2: boston sweep colorado to win world series 3: rookies respond in first crack at the big time
C-LR=C-LexRank; WDS=Word Distributional Similarity	

Table 6.8: Top 3 ranked summaries of the redsox cluster using different methods

"red sox's winning the title" and "fans celebrating". However, the second factoid is absent in the other two.

6.2.4 Conclusion

We proposed a ranking system that employs word distributional similarities to identify semantically equivalent words, and compared it with a wide range of summarization systems that leverage diversity.

In the future, we plan to move to content from other collective systems on Web. In order to generalize our findings, we plan to examine blog comments, online reviews, and tweets (that discuss the same URL). We also plan to build a generation system that employs the Yule model [213] to determine the importance of each aspect (e.g. who, when, where, etc.) in order to produce summaries that include diverse aspects of a story.

Our work has resulted in a publicly available dataset [2] of 25 annotated news clusters with nearly $1,400$ headlines, and 25 clusters of citation sentences with more than 900 citations. We believe that this dataset can open new dimensions in studying diversity and other aspects of automatic text generation.

[2] http://www-personal.umich.edu/ vahed/data.html

CHAPTER VII

Survey Generation

7.1 Survey Generation

In this chapter, we present our experiment on using the tools explained in previous sections for automatic survey generation. Our evaluation experiments for survey generation are on a set of papers in the research area of Question Answering (QA) and another set of papers on Dependency parsing (DP). The two sets of papers were compiled by selecting all the papers in AAN that had the words *Question Answering* and *D*ependency Parsing, respectively, in the title and the content. There were 10 papers in the QA set and 16 papers in the DP set. We also compiled the citation texts for the 10 QA papers and the citation texts for the 16 DP papers.

7.1.1 Data Preparation

Our goal is to determine if citation texts do indeed have useful information that one will want to put in a survey and if so, how much of this information is *not* available in the original papers and their abstracts. For this we evaluate each of the automatically generated surveys using two separate approaches: nugget-based pyramid evaluation and ROUGE (described in the two subsections below).

Two sets of gold standard data were manually created from the QA and DP

citation texts and abstracts, respectively:[1] (1) We asked three impartial judges to identify important nuggets of information worth including in a survey. (2) We asked four fluent speakers of English to create 250-word surveys of the datasets. Then we determined how well the different automatically generated surveys perform against these gold standards. If the citation texts have only redundant information with respect to the abstracts and original papers, then the surveys of citation texts will not perform better than others.

Nugget-Based Annotations

For our first approach we used a nugget-based evaluation methodology [116, 143, 80, 202]. We asked three impartial annotators (knowledgeable in NLP but not affiliated with the project) to review the citation texts and/or abstract sets for each of the papers in the QA and DP sets and manually extract prioritized lists of 2–8 "nuggets," or main contributions, supplied by each paper. Each nugget was assigned a weight based on the frequency with which it was listed by annotators as well as the priority it was assigned in each case. Our automatically generated surveys were then scored based on the number and weight of the nuggets that they covered. This evaluation approach is similar to the one adopted by [159], but adapted here for use in the multi-document case.

The annotators had two distinct tasks for the QA set, and one for the DP set: (1) extract nuggets for each of the 10 QA papers, based only on the citation texts for those papers; (2) extract nuggets for each of the 10 QA papers, based only on the abstracts of those papers; and (3) extract nuggets for each of the 16 DP papers, based only on the citation texts for those papers.[2]

[1] Creating gold standard data from complete papers is fairly arduous, and was not pursued.

[2] We first experimented using only the QA set. Then to show that the results apply to other datasets, we asked human annotators for gold standard data on the DP citation texts. Additional experiments on DP abstracts were not pursued because this would have required additional human annotation effort to establish a point we had already

We obtained a weight for each nugget by reversing its priority out of 8 (e.g., a nugget listed with priority 1 was assigned a weight of 8) and summing the weights over each listing of that nugget.[3]

To evaluate a given survey, we counted the number and weight of nuggets that it covered. Nuggets were detected via the combined use of annotator-provided regular expressions and careful human review. Recall was calculated by dividing the combined weight of covered nuggets by the combined weight of all nuggets in the nugget set. Precision was calculated by dividing the number of distinct nuggets covered in a survey by the number of sentences constituting that survey, with a cap of 1. F-measure, the weighted harmonic mean of precision and recall, was calculated with a beta value of 3 in order to assign the greatest weight to recall. Recall is favored because it rewards surveys that include highly weighted (important) factoids, rather than just a great number of factoids.

Table 7.1 gives the F-measure values of the 250-word surveys manually generated by humans. The surveys were evaluated using the nuggets drawn from the QA citation texts, QA abstracts, and DP citation texts. The average of their scores (listed in the rightmost column) may be considered a good score to aim for by the automatic summarization methods.

7.1.2 Experiments

We used four summarization systems for our survey-creation approach: *Trimmer, LexRank, C-LexRank,* and *C-RR.*

Trimmer is a sentence-compression tool that extends the scope of an extractive summarization system by generating multiple alternative sentence compressions of

made with the QA set, i.e., that abstracts are useful for survey creation.

[3]Results obtained with other weighting schemes that ignored priority ratings and multiple mentions of a nugget by a single annotator showed the same trends as the ones shown by the selected weighting scheme, but the latter was a stronger distinguisher among the four systems.

	Human Performance: Pyramid F-measure				
	Human1	Human2	Human3	Human4	Average
Input: QA citation surveys					
QA–CT nuggets	0.524	0.711	0.468	0.695	0.599
QA–AB nuggets	0.495	0.606	0.423	0.608	0.533
Input: QA abstract surveys					
QA–CT nuggets	0.542	0.675	0.581	0.669	0.617
QA–AB nuggets	0.646	0.841	0.673	0.790	0.738
Input: DP citation surveys					
DP–CT nuggets	0.245	0.475	0.378	0.555	0.413

Table 7.1: Pyramid F-measure scores of human-created surveys of QA and DP data. The surveys are evaluated using nuggets drawn from QA citation texts (QA–CT), QA abstracts (QA–AB), and DP citation texts (DP–CT).

the most important sentences in target documents [214]. Trimmer compressions are generated by applying linguistically-motivated rules to mask syntactic components of a parse of a source sentence. The rules can be applied iteratively to compress sentences below a configurable length threshold, or can be applied in all combinations to generate the full space of compressions.

Trimmer can leverage the output of any constituency parser that uses the Penn Treebank conventions. At present, the Stanford Parser [98] is used. The set of compressions is ranked according to a set of features that may include metadata about the source sentences, details of the compression process that generated the compression, and externally calculated features of the compression.

Summaries are constructed from the highest scoring compressions, using the metadata and maximal marginal relevance [28] to avoid redundancy and over-representation of a single source. The summarizer contains a cheap-to-calculate redundancy score using an index index of the words (w) in the document set.

$$\sum_{w} \log(\lambda.P(w|\text{summary}) + (1-\lambda).P(w|\text{corpus})) \tag{7.1}$$

where λ is a weighting factor (set to 0.3 in our experiments)

We automatically generated surveys for both QA and DP from three different

	System Performance: Pyramid F-measure				
	Random	C-LexRank	C-RR	LexRank	Trimmer
Input: QA citation surveys					
QA–CT nuggets	0.321	0.434	0.268	0.295	**0.616**
QA–AB nuggets	0.305	0.388	0.349	0.320	**0.543**
Input: QA abstract surveys					
QA–CT nuggets	0.452	0.383	**0.480**	0.441	0.404
QA–AB nuggets	**0.623**	0.484	0.574	0.606	0.622
Input: QA full paper surveys					
QA–CT nuggets	0.239	**0.446**	0.299	0.190	0.199
QA–AB nuggets	0.294	**0.520**	0.387	0.301	0.290
Input: DP citation surveys					
DP–CT nuggets	0.219	0.231	0.170	**0.372**	0.136
Input: DP abstract surveys					
DP–CT nuggets	**0.321**	0.301	0.263	0.311	0.312
Input: DP full paper surveys					
DP–CT nuggets	0.032	0.000	0.144	*	**0.280**

Table 7.2: Pyramid F-measure scores of automatic surveys of QA and DP data. The surveys are evaluated using nuggets drawn from QA citation texts (QA–CT), QA abstracts (QA–AB), and DP citation texts (DP–CT). * LexRank is computationally intensive and so was not run on the DP-PA dataset (about 4000 sentences). (Highest scores for each input source are shown in bold.)

types of documents: (1) full papers from the QA and DP sets—*QA and DP full papers (PA)*, (2) only the abstracts of the QA and DP papers—*QA and DP abstracts (AB)*, and (3) the citation texts corresponding to the QA and DP papers—*QA and DP citations texts (CT)*.

We generated twenty four (4x3x2) surveys, each of length 250 words, by applying Trimmer, LexRank, C-LexRank on the three data types (citation texts, abstracts, and full papers) for both QA and DP. (Table 7.5 shows a fragment of one of the surveys automatically generated from QA citation texts.) We created six (3x2) additional 250-word surveys by randomly choosing sentences from the citation texts, abstracts, and full papers of QA and DP. We will refer to them as *random surveys*.

Table 7.2 gives the F-measure values of the surveys generated by the four automatic summarizers, evaluated using nuggets drawn from the QA citation texts, QA abstracts, and DP citation texts. The table also includes results for the baseline random summaries.

	Human Performance: ROUGE-2				
	human1	human2	human3	human4	average
Input: QA citation surveys					
QA–CT refs.	0.1807	0.1956	0.0756	0.2019	0.1635
QA–AB refs.	0.1116	0.1399	0.0711	0.1576	0.1201
Input: QA abstract surveys					
QA–CT refs.	0.1315	0.1104	0.1216	0.1151	0.1197
QA-AB refs.	0.2648	0.1977	0.1802	0.2544	0.2243
Input: DP citation surveys					
DP–CT refs.	0.1550	0.1259	0.1200	0.1654	0.1416

Table 7.3: ROUGE-2 scores obtained for each of the manually created surveys by using the other three as reference. ROUGE-1 and ROUGE-L followed similar patterns.

When we used the nuggets from the abstracts set for evaluation, the surveys created from abstracts scored higher than the corresponding surveys created from citation texts and papers. Further, the best surveys generated from citation texts outscored the best surveys generated from papers. *When we used the nuggets from citation sets for evaluation,* the best automatic surveys generated from citation texts outperform those generated from abstracts and full papers. All these pyramid results demonstrate that citation texts can contain useful information that is not available in the abstracts or the original papers, and that abstracts can contain useful information that is not available in the citation texts or full papers.

Among the various automatic summarizers, Trimmer performed best at this task, in two cases exceeding the average human performance. Note also that the random summarizer outscored the automatic summarizers in cases where the nuggets were taken from a source different from that used to generate the survey. However, one or two summarizers still tended to do well. This indicates a difficulty in extracting the overlapping survey-worthy information across the two sources.

ROUGE evaluation

Table 7.3 presents ROUGE scores [113] of each of human-generated 250-word surveys against each other. The average (last column) is what the automatic surveys

	System Performance: ROUGE-2				
	Random	C-LexRank	C-RR	LexRank	Trimmer
Input: QA citation surveys					
QA–CT refs.	0.11561	**0.17013**	0.09522	0.13501	0.16984
QA–AB refs.	0.08264	**0.11653**	0.07600	0.07013	0.10336
Input: QA abstract surveys					
QA–CT refs.	0.04516	0.05892	**0.06149**	0.05369	0.04114
QA–AB refs.	0.12085	0.13634	0.12190	**0.20311**	0.13357
Input: QA full paper surveys					
QA–CT refs.	0.03042	0.03606	0.03599	**0.28244**	0.03986
QA–AB refs.	0.04621	0.05901	0.04976	**0.10540**	0.07505
Input: DP citation surveys					
DP–CT refs.	0.10690	**0.13164**	0.08748	0.04901	0.10052
Input: DP abstract surveys					
DP–CT refs.	0.07027	0.07321	0.05318	**0.20311**	0.07176
Input: DP full paper surveys					
DP–CT refs.	0.03770	0.02511	0.03433	*	**0.04554**

Table 7.4: ROUGE-2 scores of automatic surveys of QA and DP data. The surveys are evaluated by using human references created from QA citation texts (QA–CT), QA abstracts (QA–AB), and DP citation texts (DP–CT). These results are obtained after Jack-knifing the human references so that the values can be compared to those in Table 4. * LexRank is computationally intensive and so was not run on the DP full papers set (about 4000 sentences). (Highest scores for each input source are shown in bold.)

Most of work in QA and paraphrasing focused on folding paraphrasing knowledge into question analyzer or answer locator Rinaldi et al, 2003; Tomuro, 2003.
In addition, number of researchers have built systems to take reading comprehension examinations designed to evaluate children's reading levels.
Charniak et al, 2000; Hirschman et al, 1999; Ng et al, 2000; Riloff and Thelen, 2000; Wang et al, 2000. so-called " definition " or " other " questions at recent TREC evaluations Voorhees, 2005 serve as good examples.
To better facilitate user information needs, recent trends in QA research have shifted towards complex, context-based, and interactive question answering Voorhees, 2001; Small et al, 2003; Harabagiu et al, 2005. [And so on.]

Table 7.5: First few sentences of the QA citation texts survey generated by Trimmer.

can aim for. We then evaluated each of the random surveys and those generated by the four summarization systems against the references. Table 7.4 lists ROUGE scores of surveys when the manually created 250-word survey of the QA citation texts, survey of the QA abstracts, and the survey of the DP citation texts, were used as gold standard.

When we use manually created citation text surveys as reference, then the surveys generated from citation texts obtained significantly better ROUGE scores than the surveys generated from abstracts and full papers ($p < 0.05$) [Result 1]. This shows that crucial survey-worthy information present in citation texts is not available, or hard to extract, from abstracts and papers alone. Further, the surveys generated from abstracts performed significantly better than those generated from the full papers ($p < 0.05$) [Result 2]. This shows that abstracts and citation texts are generally denser in survey-worthy information than full papers.

When we use manually created abstract surveys as reference, then the surveys generated from abstracts obtained significantly better ROUGE scores than the surveys generated from citation texts and full papers ($p < 0.05$) [Result 3]. Further, and more importantly, the surveys generated from citation texts performed significantly better than those generated from the full papers ($p < 0.05$) [Result 4]. Again, this shows that abstracts and citation texts are richer in survey-worthy information. These results also show that abstracts of papers and citation texts have some overlapping information (Result 2 and Result 4), but they also have a significant amount of unique survey-worthy information (Result 1 and Result 3).

Among the automatic summarizers, C-LexRank and LexRank perform best. This is unlike the results found through the nugget-evaluation method, where Trimmer performed best. This suggests that Trimmer is better at identifying more useful

nuggets of information, but C-LexRank and LexRank are better at producing unigrams and bigrams expected in a survey. To some extent this may be due to the fact that Trimmer uses smaller (trimmed) fragments of source sentences in its summaries.

CHAPTER VIII

Expert-written Historical Notes

8.1 Historical Notes

Researchers and scholars often face the problem of keeping up with the ever increasing number of publications in their fields of research. In addition, research is increasingly becoming inter-disciplinary, bridging different areas and forcing researchers to familiarize themselves with new areas. For instance, cancer researchers often have to quickly move into a new area in their research; a pathologist may want to learn about new medical devices or drug developments; business researchers are interested in understanding user behavior in an online community; and social scientists may be interested in learning new computational models that explain certain social phenomena. Inter-disciplinary review panels and funding agencies often need to make decisions on proposals from a wide range of newly emerging areas. Thus they have to learn about the development of ideas in a new discipline and be able to relate their expertise to the proposals.

In this chapter, we present Surveyor, a summary generation system that addresses such needs by generating surveys of key developments on a research topic. The solution that we propose uses different sources of information (i.e., source texts and citations) and exploits the citation network to produce summaries that compete

expert-written surveys.

Previous work has noted the difference between conventional multi-document summarization and summarizing scientific literature [134]. In the case of multi-document summarization, the goal is to produce a readable presentation of multiple documents, whereas in the case of technical survey creation, the goal is to convey the key features and basic underpinnings of a particular field, temporal developments, important contributions, emergence of sub-fields, and basic definitions and examples that enable rapid understanding of a field by non-experts.

One example of expert-written surveys is the set of end-of-chapter summaries and "historical notes" that appear at the end of chapters in the *Speech and Language Processing* textbook [89]. Each summary or historical note is about a sub-field in Natural Language Processing (NLP), and includes information about the background, early and recent developments, state-of-the-art results, etc. Table 8.1 shows parts of the historical notes that Jurafsky and Martin wrote for "Machine Translation." The example shows that this survey includes various information for non-expert readers including some history, early developments, toolkits, evaluations and additional references and tutorials for further reading.

The goal of this chapter is to present a framework that generates summaries similar to the one in Table 8.1. We first present our data preparation, including scanning and parsing citations in [89] and the ACL Anthology Network [167], which is used as the source for summary generation in Section 8.2. We propose a new approach that repurposes both citations and source text of papers and exploits the citation graph to build a survey in Section 8.3. Finally, Section 8.4 present our experiments and results on the Jurafsky and Martin textbook.

Historical Notes: Machine Translation	
history	*Work on models of the process and goals of translation goes back at least to Saint Jerome in the fourth century (Kelley, 1979)...*
early work	*... At the same time, the IBM group, drawing directly on algorithms for speech recognition (many of which had themselves been developed originally at IBM!) proposed the Candide system, based on the IBM statistical models we have described (Brown et al., 1990, 1993) ...*
tools	*... Progress was made hugely easier by the development of publicly-available toolkits, particularly tools extended from the EGYPT toolkit developed by the Statistical Machine Translation team in during the summer 1999 research workshop at the Center for Language and Speech Processing at the Johns Hopkins University. These include the GIZA++ aligner, developed by Franz Joseph Och by extending the GIZA toolkit (Och and Ney, 2003), which implements IBM models 1-5 as well as the HMM alignment model ...*
evaluations	*... These included the use of doze and Shannon tasks to measure intelligibility as well as a metric of edit distance from a human translation, the intuition that underlies all modern automatic evaluation metrics like BLEU ...*
other resources	*... Nirenburg et al. (2002) is a comprehensive collection of classic readings in MT. Knight (1999b) is an excellent tutorial introduction to statistical MT ...*

Table 8.1: Part of the historical note in [89] signifying the history, early and late developments and evaluation in "machine translation"

8.2 Datasets

In this section, we first describe the ACL Anthology Network, which is used as the source dataset for generating system surveys. We then explain our gold standard preparation from the Jurafsky and Martin textbook.

8.2.1 The ACL Anthology Network

The ACL Anthology[1] includes all papers published by ACL and related organizations as well as the Computational Linguistics journal over a period of four decades. [167] have further processed this Anthology to produce the the ACL Anthology Network (AAN)[2]. The AAN includes more than $16,000$ papers, each distinguished with a unique ACL ID, together with their full-texts, abstracts, and citation information. It also includes other valuable meta-data such as author affiliations, citation and col-

[1] http://www.aclweb.org/anthology-new/
[2] http://clair.si.umich.edu/clair/anthology/

laboration networks, and various centrality measures [167, 88]. In our experiments, we generate a set of automatic summaries using the papers in AAN.

8.2.2 Gold Standard Preparation

We use 2 sets of gold standards both extracted from the Jurafsky and Martin textbook[3] *Speech and Language Processing* [89]: end-of-chapter summaries and the historical notes.

End-of-chapter Summaries

We were fortunate to obtain the end of chapter summaries in the JM book in text format. Each summary is generally a few paragraphs long and explains the main points discussed in the chapter. We will refer to these gold standards as **chapter summaries**.

Historical Notes

We also use the **historical notes** at the end of each chapter in the JM book as the second set of gold standards. Each historical note, corresponding to one chapter, is generally 1-2 pages long and summarizes the history, early developments and the state-of-art methods in each NLP topic.

In order to prepare this gold standard, we first scanned the historical notes of the chapters as well as the references in the JM book. Next, we used a commercial OCR tool to convert the scanned files to plain text. We further processed the OCR output by removing end-of-line hyphens and fixing sentence fragments and line breaks.[4] Cleaning-up references included identifying entry boundaries and combining multiple lines corresponding to one entry.

[3] we use the shorthand "JM book" in the rest of this paper

[4] Parsing the bibliographies from the OCR output is more challenging than historical notes because of the smaller fonts and frequent out-of-vocabulary words such as author names. However, OCR errors are tolerated in bibliographies since we use minimum edit distance to find the corresponding papers in AAN.

Chapter	src	cit	$\lvert\mathcal{B}_L\rvert$	$\lvert\mathcal{B}_R\rvert$	$E_{\mathcal{B}}$
Words and Transducers	14	255	489	3,484	202,940
N-grams	5	73	97	2,083	32,690
Part-of-Speech Tagging	16	657	1,261	3,385	344,886
Hidden Markov and Maximum Entropy Models	2	432	659	525	187,905
Phonetics	1	12	81	216	17,496
Speech Synthesis	4	54	126	920	29,357
Automatic Speech Recognition	2	27	103	401	21,566
Speech Recognition: Advanced Topics	7	189	445	2,007	96,467
Syntactic Parsing	4	131	246	763	63,673
Dialog and Conversational Agents	11	170	368	3,745	114,281

Table 8.2: List of chapter historical notes used in our experiments together with the number of source papers extracted from historical notes (src), the number of citing papers extracted from AAN (cit), size of the left ($\mathcal{B}_L$) and right ($\mathcal{B}_R$) components in the bi-partite graph, and number of edges in the graph ($E_{\mathcal{B}}$).

We used the extracted references and citations in each historical note to extract the set of papers that are cited by [89] and are part of AAN. We use these papers as the seed source papers to generate automatic summaries. To extract the list of AAN papers that are cited in each historical note, we first map each reference in the JM book to an AAN paper. First, for each reference we represent it by a vector of metadata that consists of the author names, title (stop words removed), canonical name of the venue, and publication year. We then compare these vectors with AAN metadata and find the closest match by computing the minimum edit distance of corresponding metadata vectors when the publication dates agree. Finally, we manually verify the output of the above procedure and correct mismatches.

8.3 Approach

Previous work on scientific survey generation have compared surveys that are generated from different sources such as citations and source paper texts [159, 124, 134]. However, none of these approaches combine these heterogeneous information sources to produce automatic surveys.

In our approach, we investigate the usefulness of combining different information

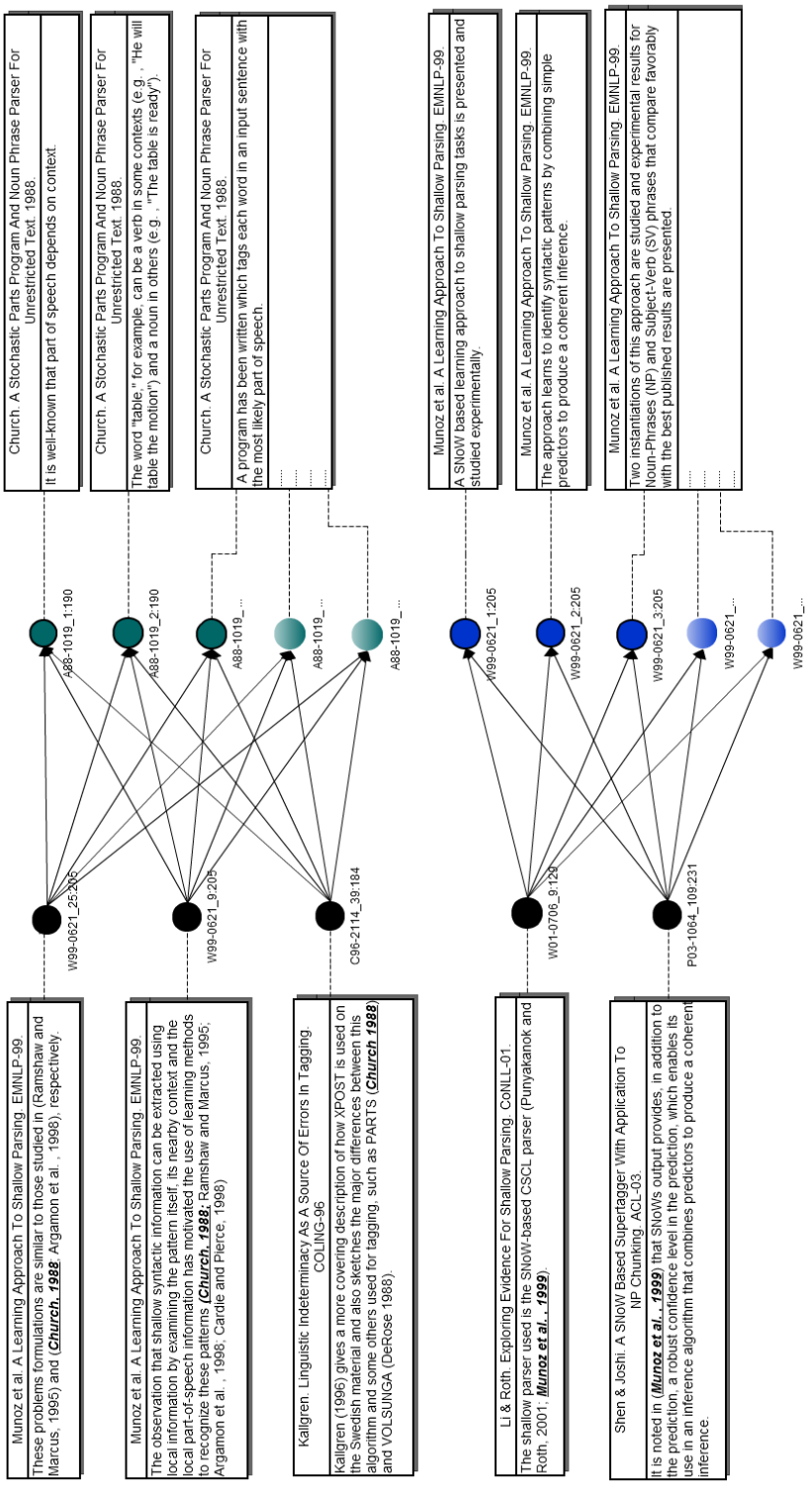


Figure 8.1: A mini-model of the bi-partite graph for Chapter 5 (Part-of-Speech Tagging)

sources and producing summaries that are both affected by source paper text and citation information. For a set of papers in the same scientific topic, we extract survey-worthy sentences from the source texts that cover contributions recognized by other scholars in citations, and extract citations that cover contributions that are recognized by the authors in the source text.

In our algorithm, we model the set of papers in a scientific topic t as a bi-partite graph, $\mathcal{B}$ with a left and a right component ($\mathcal{B}_L$, $\mathcal{B}_R$). Each node in $\mathcal{B}_L$ is a citation sentence to one or more papers in t extracted from AAN, and each node in $\mathcal{B}_R$, represents a sentence extracted from the source text of a paper in t. We construct the edges in $\mathcal{B}$ by connecting each citing sentence to all the source sentences in the papers it cites. Each edge in $\mathcal{B}$ is assigned a weight equal to the cosine similarity of the TF-IDF term vectors of the two sentences it connects. Figure 8.1 illustrates part of the bi-partite graph built for the "Part-of-Speech Tagging" chapter in the JM book.

To build the summaries we are interested in citations and source sentences that cover important contributions in the given scientific topics. Intuitively, contributions that both the paper authors and other scholars recognize as significant are important and should be extracted. Surveyor extracts citations that cover important contributions mentioned in the source papers as well as source sentences that discuss important factoids recognized by others in citations.

8.3.1 Ranking

The inherent duality in the source papers and citations suggests that the problem could be addressed by applying the HITS algorithm [99] to iteratively assign hub and authority scores to citations and source sentences respectively. The induction process is as follows. Each citation sentence $c \in \mathcal{B}_L$ is associated with a hub score

h_c, and each source sentence $s \in \mathcal{B}_R$ is associated with an authority score a_s. These scores are initialized with a value of 1.0. Hub and authority scores are iteratively updated using the following equations.

$$a_s^{(i+1)} = \sum_{c \in nei(s)} \frac{h_c^{(i)}}{H^{(i)}} \tag{8.1}$$

$$h_c^{(i+1)} = \sum_{s \in nei(c)} \frac{a_s^{(i)}}{A^{(i)}} \tag{8.2}$$

where a source sentence s is in a citation sentence, c's neighborhood ($s \in nei(c)$) if there is an edge between s and c in $\mathcal{B}$ (c cites the paper that contains s), and their cosine similarity is greater than a threshold (i.e., $cos(s, c) > \theta$). Here, $H^{(i)}$ and $A^{(i)}$ are normalization factors:

$$H^{(i)} = (\sum_{c \in \mathcal{B}_L} {h_c^{(i)}}^2)^{1/2} \tag{8.3}$$

$$A^{(i)} = (\sum_{s \in \mathcal{B}_R} {a_s^{(i)}}^2)^{1/2} \tag{8.4}$$

In our experiments, we set $\theta = 0.1$. This ranking gives us top authorities (source sentences) and top hubs (citations) with which we build two different summaries: $\mathbf{HITS_{src}}$ and $\mathbf{HITS_{cit}}$. Although these summaries are built from different sources (i.e., source papers and citations) they are affected by each other. In other words, the scores and thus extraction of top citations affects the extraction of top source sentences and vice versa.

8.3.2 Adding Weights

In previous section, we described the basic version of our system in which the edges are considered as binary connections (if the cosine similarity is above a threshold).

We would like to investigate the effect of similarity on sentence extraction. In other words, instead of applying a threshold we use the actual edge weights and modify Equations 8.1, 8.2 as follows.

$$a_s^{(i+1)} = \sum_{c \in nei(s)} \frac{w_{cs} \cdot h_c^{(i)}}{H^{(i)}} \tag{8.5}$$

$$h_c^{(i+1)} = \sum_{s \in nei(c)} \frac{w_{sc} \cdot a_s^{(i)}}{A^{(i)}} \tag{8.6}$$

where w_{sc} is the is the edge weight between vertices s and c, calculated as the TF-IDF based cosine similarity between their corresponding sentences.

Intuitively, this modification will take into account the similarity of a sentence with its neighbors rather than the number of connections, and would result in summaries that contain more *lexically* salient sentences. The weighted ranking gives us top authorities (source sentences) and top hubs (citations) with which we build two different summaries: **HITS$_{\text{src}}$ with weights** and **HITS$_{\text{cit}}$with weights**.

8.3.3 Citation Bias

The downside of the current HITS-based sentence extraction is that it assumes equal importance for the papers in a given topic. However, contributions from highly cited papers are intuitively more important. To address this issue, we propose an improvement inspired by [123] and modify equations 8.1, 8.2 to include a prior distribution of prestige.

$$a_s^{(i+1)} = (1-\lambda) \cdot p^*(s) + \lambda \cdot \sum_{c \in nei(s)} \frac{h_c^{(i)}}{H^{(i)}} \tag{8.7}$$

$$h_c^{(i+1)} = (1-\lambda) \cdot p^*(c) + \lambda \cdot \sum_{s \in nei(c)} \frac{a_s^{(i)}}{A^{(i)}} \tag{8.8}$$

Here, $p^*(v)$ is a distribution which represents the prior preference of vertex v. When $p^*(v)$ is uniform, the left component is similar to the random jumping probabilities in PageRank. Other possible choices for $p^*(v)$ include a topic sensitive distribution, inspired by personalized jumping in personalized PageRank [77, 78]. In Equations 8.7, 8.8 λ obtains a value between 0 and 1. When $\lambda = 1$, Equations 8.7, 8.8 lead to the standard HITS algorithm. In our experiments, we set $\lambda = 0.75$.

The prior distribution allows us to favor citation sentences that are from more impactful papers. Therefore we define the prior distributions as the normalized citation frequency of the paper

$$p^*(v) = \frac{C_v + 1}{\sum_{v \in \mathcal{B}} C_v + |\mathcal{B}|} \tag{8.9}$$

where C_v is the number of citations to the paper that contains sentence v. Equations 8.7, 8.8 give us top authorities (source sentences) and top hubs (citations) with which we build two different summaries: **HITS$_{\text{src}}$ with priors** and **HITS$_{\text{cit}}$with priors**.

	System Performance: Rouge-1 Gold Standard: Historical Notes				
Method	**src**	95% C.I.	**cit**	95% C.I.	Mean
LexRank	0.150	[0.110, 0.190]	0.212	[0.189, 0.235]	0.181
C-LexRank	0.183	[0.147, 0.220]	0.187	[0.158, 0.217]	0.185
HITS	0.202	[0.162, 0.243]	0.152	[0.120, 0.185]	0.177
HITS with weights	**0.216**	**[0.195, 0.237]**	0.200	[0.178, 0.222]	0.208
HITS with priors	0.207	[0.182, 0.233]	0.138	[0.100, 0.177]	0.173
HITS with weights/priors	0.204	[0.187, 0.221]	**0.215**	**[0.181, 0.249]**	**0.209**

Table 8.3: Average Rouge-1 scores of automatic surveys of the 10 chapters listed in Table 8.2 evaluated using historical notes as reference (C.I.: Confidence Interval).

8.4 Experiments

Using the procedure described in section 8.2.2, we extract the list of source papers from 10 chapters' historical notes in the JM book. For each chapter, the papers cited

in its historical note are used as the source papers (**src**) and the set of AAN papers that cite them are used as citing papers (**cit**). Table 8.2 summarizes the list of chapter historical notes used in our experiments together with the number of source papers, citing papers extracted from AAN, the size of the left ($\mathcal{B}_L$) and right ($\mathcal{B}_R$) components in the bi-partite graph, and number of edges in the graph ($E_{\mathcal{B}}$).

For each chapter we generate 2×2 summaries using the **cit** and **src** papers with a length equal to the length of chapter's **chapter summaries** and that of chapter's **historical notes**. We evaluate these summaries using Rouge [113], and compare them with two state-of-the-art methods in scientific survey generation: LexRank and C-LexRank.

8.4.1 Baseline Methods

LexRank

LexRank [55] works by first building a graph of all the documents (D_i) in a cluster. The edges between corresponding nodes (d_i) represent the cosine similarity between them if the cosine value is above a threshold (0.10 following [55]). Once the network is built, the system finds the most central sentences by performing a random walk on the graph.

$$p(d_j) = (1 - \lambda)\frac{1}{|D|} + \lambda \sum_{d_i} p(d_i) P(d_i \rightarrow d_j) \tag{8.10}$$

C-LexRank

C-LexRank, as discussed in Chapter V, is a clustering-based summarization system that is proposed by [159] to summarize different scientific perspectives. In C-LexRank, we first create a full connected network in which nodes are sentences and edges are cosine similarities. To create summaries, C-LexRank constructs a fully

connected network in which vertices are sentences and edges are cosine similarities calculated using the TF-IDF vectors of citation sentences. It then employs a hierarchical agglomeration clustering algorithm proposed by [38] to find communities of sentences that discuss the same scientific contributions.

Once the graph is clustered and communities are formed, we extract sentences from different clusters to build a summary. We start with the largest cluster and extract sentences using LexRank within each cluster. In other words, for each cluster Ω_i they make a lexical network of *the sentences in that cluster* (N_i). LexRank extracts the most central sentences in N_i as salient sentences of Ω_i to include in the main summary. For each cluster Ω_i, the most salient sentence of Ω_i is extracted until the summary length limit is reached. The cluster selection is in order of decreasing size.

8.4.2 Results and Discussion

Table 8.3 lists the average Rouge-1 scores of different automatic summaries with each chapter's **historical notes** chosen as the gold standard and its length as the automatic summary length. Similarly, Table 8.4 summarizes the average Rouge-1 scores of different system summaries when **chapter summaries** are used as reference.

	System Performance: Rouge-1				
	Gold Standard: Chapter Summaries				
Method	**src**	95% C.I.	**cit**	95% C.I.	Mean
LexRank	0.205	[0.141, 0.269]	0.232	[0.203, 0.260]	0.218
C-LexRank	0.188	[0.129, 0.246]	0.198	[0.140, 0.256]	0.193
HITS	0.233	[0.191, 0.274]	0.161	[0.122, 0.200]	0.197
HITS with weights	**0.242**	**[0.215, 0.268]**	0.222	[0.183, 0.260]	0.232
HITS with priors	0.205	[0.170, 0.239]	0.129	[0.094, 0.165]	0.167
HITS with weights/priors	0.235	[0.198, 0.271]	**0.241**	**[0.198, 0.284]**	**0.238**

Table 8.4: Average Rouge-1 scores of automatic surveys of the 10 chapters listed in Table 8.2 evaluated using chapter summaries as reference (C.I.: Confidence Interval).

	Part of the automatic summary
early developments	*During the early stages of the Penn Treebank project, the initial automatic POS assignment was provided by PARTS (Church 1988), a stochastic algorithm developed at AT&T Bell Labs.*
methods	*As shown by Klein and Manning (2002, 2004), the extension to inducing trees for words instead of P-O-S tags is rather straight-forward since there exist several unsupervised part-of-speech taggers with high accuracy, which can be combined with unsupervised parsing (see e.g. Schütze 1996; Clark 2000).*
ambiguity problem	*Jardino and Adda (1994), Schütze (1997) and Clark (2000) have attempted to address the ambiguity problem to a certain extent.*

Table 8.5: Part of the automatic survey generated using **HITS with weights** for "part-of-speech" tagging signifying early work, state-of-the-art, etc.

Both of these tables show that the HITS method that employs weights on graph edges leads to significantly better results than other methods both when the summaries are generated from citations (cit) or source texts (src). Moreover, these tables suggest that our proposed method (HITS with weights/priors) outperforms the state-of-the-art methods and baselines whether summaries are generated using source texts (src) or citations (cit) and whether evaluated against historical notes or chapter summaries. Table 8.5 shows part of the automatic survey generated using **HITS with weights** for "part-of-speech tagging" signifying some early work, state-of-the-art, etc.

We repeat the same experiments using Rouge-L. Figure 8.2 summarizes the average Rouge-L score for automatic summaries generated using source texts (src) and citations (cit). This figure confirms that Rouge-L results follow a similar pattern as Rouge-1. The results in Figure 8.2 and Tables 8.4, 8.3 also suggest that surveys generated using citations are consistently better that those generated from source texts in LexRank and C-LexRank. However, when the two summaries are generated using both sources affecting each other in a bi-partite graph, summaries from source and citations obtain similar qualities on average.

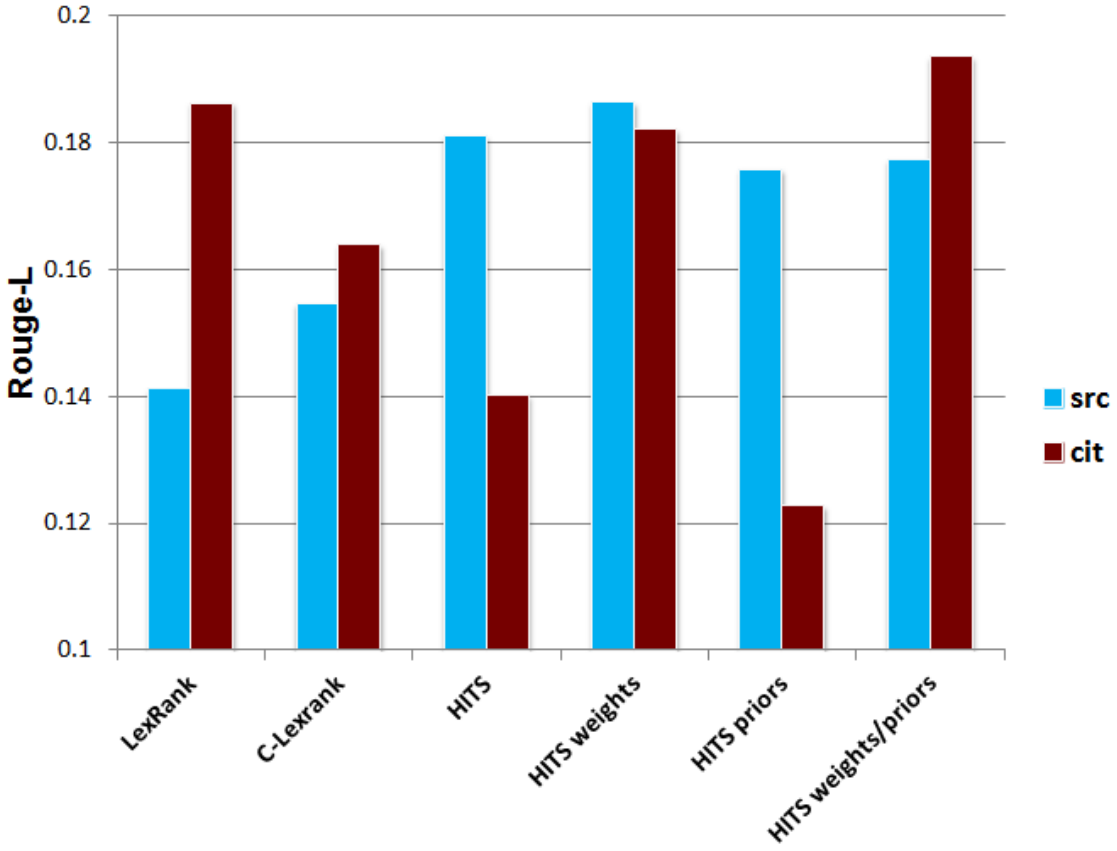


Figure 8.2: Average Rouge-L scores of automatic surveys of the 10 chapters listed in Table 8.2 using chapter summaries and historical notes as reference

One of the collaborators in this work organized an NLP seminar previously. As part of the seminar, the students in the class took turns to present surveys of specific topics in NLP and Information Retrieval (IR) and wrote chapter-length surveys of their topics. In future work, we plan to make use of the surveys written by NLP students as gold standard in evaluations. Compared to the chapters from JM book, these topics are more specific and close to the latest development in NLP and IR. Examples include Sentiment and Polarity Extraction, Science Maps, Spectral graph-based methods for NLP, Information Diffusion In Graphs, Financial Networks and Query Expansion.

Here, we are using the papers cited in each chapter of the JM textbook as seed source papers (i.e. we assume that the set of seminal papers on each topic are known). However in the science community, there are thousands more papers that are related to a given topic. In the future, we will work on a method of automatically identifying the most influential papers that represent a specific topic from the vast range of publications.

CHAPTER IX

Conclusion and Future Direction

9.1 Conclusion

In linguistics, discourse structure [72] is composed of three separate but interrelated components: the structure of the sequence of utterances, a structure of purposes and focus of attention. In social media, *collective discourse* is composed of a set of (independent) utterances about an object or an artifact focused on different aspects of that object [162].

The study of collective discourse is especially more important on the Web. With the growth of Web 2.0, millions of individuals engage in collective discourse. They participate in online discussions, share their opinions, and generate content about the same artifacts, objects, and news events in Web portals like amazon.com, epinions.com, imdb.com and so forth. This massive amount of text is mainly written on the Web by non-expert individuals with different perspectives, and yet exhibits accurate knowledge as a whole.

In social media, collective discourse is often a collective reaction to an event. A collective reaction to a well-defined subject emerges in response to an event (a movie release, a breaking story, a newly published paper) in the form of independent writings (movie reviews, news headlines, citation sentences) by many individuals.

9.1.1 Summary of Contributions

The main question we tried to answer in this thesis is: "Can we exploit properties of collective discourse to have better understanding of the world without an authoritative view on information?" In particular, we explore properties of "Scholarly collective discourse in citations to reach a better understanding of emergent scientific contributions in a research field?" We had 2 main sub-goals related to these questions: (i) investigate properties of collective discourse such as diversity of perspectives, skewed distribution of factoids, and community structure, (ii) develop methodology to extract communities of perspectives and generate extractive summaries that leverage the observed diversity. Finally, we applied our methodologies to a broader application: generating surveys of scientific topics. Here we summarize our contributions in this thesis.

In Chapter II we review prior relevant work in four sections. First, we summarize previous work on collective systems and collective human behavior in general. Then, we look at work on modeling natural language as a complex system. Next, we provide a literature review of graph based summarization systems that represent a set of documents as a network and produce a summary by applying graph based methods such as salience detection and ranking. Last, we review work on citation analysis. Particularly, we review prior work that has looked at the structure and importance of citations in scholarly work.

Chapter III is focused on studying collective discourse and investigating diverse perspectives when a number of non-expert Web users engage in collective behavior and generate content on the Web. We show that the set of people who discuss the same story or subject have diverse perspectives, introducing new aspects that have not been previously discussed by others.

Our experiments on two different categories of human-written summaries (headlines and citations) showed that a lot of the diversity seen in human summarization comes from different nuggets that may actually represent the same semantic information (i.e., factoids). We showed that the factoids exhibit a skewed distribution model, and that the size of the nugget inventory exhibits asymptotic behavior even with a large number of summaries. We also showed high variation in summary quality across different summaries in terms of pyramid score, and that the information covered by reading n summaries has a rapidly growing asymptotic behavior as n increases.

In Chapter IV, we define latent network, an ensemble of similarity networks between documents, and show how we can exploit its properties to predict the best cutoff at which the community structure in the network, and thus the clustering quality is maximum. We will pursue 3 ideas in future. (1) Apply the clustering technique to other tasks like text summarization and perform an extensive extrinsic evaluation of the clustering technique. (2) Extend our datasets to an even wider range of document types. (3) Examine the relation between phase transition in document collections and the underlying Zipfian distribution. Such a model would enable us to explain why some certain patterns are seen in document networks but not other social networks.

Chapter V investigated the usefulness of directly summarizing citation sentences (set of sentences that cite a paper) in the automatic creation of technical surveys. We proposed C-LexRank, a graph-based summarization model and generated summaries of 30 single scientific articles selected from 6 different topics in the ACL Anthology Network (AAN): Dependency Parsing (DP), Phrase-based Machine Translation (PBMT), Text Summarization (Summ), Question Answering (QA), Textual Entail-

ment (TE), and Conditional Random Fields (CRF). We compared C-LexRank with a number of state-of-the-art summarization systems (LexRank, DivRank, and Trimmer).

Chapter VI is focused on extracting factoids from the set of documents in collective discourse. Particularly, we first presented a summarization methodology that employs keyphrase extraction to find important contributions of scientific articles. The summarization is based on citation sentences and picks sentences to cover nuggets (represented by keyphrases) or contributions of the target papers. We used point-wise KL-divergence to extract statistically significant N-grams and use them to represent nuggets. We then applied a new set function for the task of summarizing scientific articles. We have proved that this function is submodular and concluded that a greedy algorithm will result in a near-optimum set of covered nuggets using only 5 sentences.

Our experiments in this paper confirm that the summaries created based on the presented algorithm are better than randomly generated summaries, and also outperform other state of the art summarization methods in most cases. Moreover, we showed how this method generates more stable summaries with lower variation in summary quality when N-grams of size 3 or smaller are employed.

In addition to extracting keyphrases as nuggets, we also discussed finding communities of words and phrases that represent a factoid in Chapter VI. we discussed C-LexRank when it is applied on words and not documents. We represented the set of words in a corpus as a network, where edges show the similarity of words using the *distributional hypothesis*. By applying C-LexRank on this network, we found communities of words that are more similar to each other whereby each community represents the set of words that relate to one factoid.

In Chapter VII, we extend our experiments on single paper summarization to generating entire surveys of scientific topics. We generated surveys of a set of Question Answering (QA) and Dependency Parsing (DP) papers, their abstracts, and their citation sentences using four state-of-the-art summarization systems (C-LexRank, C-RR, LexRank, and Trimmer). We then used two different approaches, nugget-based pyramid and ROUGE, to evaluate the surveys. The results from both approaches and all four summarization systems show that both citation sentences and abstracts have unique survey-worthy information. These results also demonstrate that multi-document summarization—especially technical survey creation—benefits considerably from citations.

Finally, in Chapter VIII, we performed some additional experiments on gold standard datasets that are beyond expensive human annotations. We believe that resources such as end-of-chapter historical notes in the leading NLP text book of Jurafsky and Matin [89], student summaries from a seminar class, and survey papers written by other scholars provide valuable gold standard data that are naturally generated. Moreover, we presented a framework based on the HITS algorithm that employs heterogeneous information (i.e., citations and source texts) to generate surveys of scientific paradigms. Using Rouge evaluations, we showed that our proposed system, Surveyor, generates summaries that have higher quality than the state-of-the-art methods that use only one source of information (either citations or source papers) when compared with end of chapter summaries and historical notes in the Jurafsky and Martin NLP textbook.

9.2 Future Directions

One promising future direction is to extend our exploration on collective intelligence, particularly in understanding how intelligence emerges by means of language use in various domains on the Web. Besides work on collective systems on the Web, we envision several future opportunities to go beyond my current research.

9.2.1 Decision Support Systems

We would like to go beyond developing models of online collective intelligence to further build information extraction tools that benefit from broader Web users contributions in contrast to authoritative information sources. Such tools will help us build better decision support systems that use the crowd wisdom for specific tasks. For instance, cancer researchers often have to quickly move into a new area in their research – a pathologist may want to learn about new medical devices or drug developments. Business researchers are interested in understanding user behavior in an online community; the National Science Foundation (NSF) is interested in understanding the development and evolution of new ideas; Internet users may want to know the general public opinion about a movie or a restaurant; and so on. Using collective discourse and crowd opinion to build solutions for these problems is both novel and challenging. In our work, we plan to build a general framework for a decision support system that can be easily adapted to specific tasks and new domains with as little effort as possible.

9.2.2 Identifying Misinformation

Beyond Natural Language Processing (NLP) another future direction is the study of complex systems and in particular social network analysis. We would like to benefit from theories in complex systems and graph mining to explore the social network of

people as it intersects with language use. Increasingly, language use on the Internet is largely influenced by the social networks of people [74]. We believe that there is an opportunity both in understanding the development of language on the Web as well as the structure of the social networks of interacting people. We can address problems such as rumor in microblogs, where some people collectively spread rumors on social media, by detecting and highlighting misinformation.

9.2.3 Paraphrase Acquisition

Previously, we showed that different citations to the same paper they discuss various contributions of the cited paper. Moreover we discussed that the number of factoids (contributions) show asymptotic behavior when the number of citations grow (i.e., the number of contributions of a paper is limited). Therefore, intuitively multiple citations to the same paper may refer to the same contributions of that paper. Since these sentences are written by different authors, they often use different wording to describe the cited factoid. This enables us to use the set of citing sentence pairs that cover the same factoids to create data sets for paraphrase extraction. For example, the sentences below both cite (Turney, 2002) and highlight the same aspect of Turneys work using slightly different wordings. Therefore, this sentence pair can be considered paraphrases of each other.

> "In (Turney, 2002), an unsupervised learning algorithm was proposed to classify reviews as recommended or not recommended by averaging sentiment annotation of phrases in reviews that contain adjectives or adverbs."
>
> "For example, Turney (2002) proposes a method to classify reviews as recommended/not recommended, based on the average semantic orientation of the review."

Similarly, "Eisner (1996) gave a cubic parsing algorithm" and "Eisner (1996) proposed an O(n3)" could be considered paraphrases of each other. Paraphrase annotation of citing sentences consists of manually labeling which sentence consists of what factoids. Then, if two citing sentences consist of the same set of factoids, they are labeled as paraphrases of each other. As a proof of concept, we annotated 25 papers from AAN using the annotation method described above. This data set consisted of 33,683 sentence pairs of which 8,704 are paraphrases (i.e., discuss the same factoids or contributions).

9.2.4 Datasets

Our work on analyzing citations and understanding their importance has resulted in automatically extracted sets of citation sentences as part of the ACL Anthology Network[1]. Moreover, we have made datasets of 25 annotated news clusters with nearly 1,400 headlines, and 30 clusters of citation sentences with more than 900 citations publicly available[2]. We believe that these datasets can open new dimensions in studying diversity and other aspects of automatic text generation.

[1]http://clair.eecs.umich.edu/aan/index.php
[2]http://www-personal.umich.edu/ vahed/data.html

APPENDICES

APPENDIX A

Sample Automatic Summaries

In this appendix, we list some of the summaries generated using C-LexRank on the datasets in Table 5.1.

C96-1058	**Title:** Three New Probabilistic Models For Dependency Parsing ...
	Summary: At both training and run time, edges are scored independently, and Eisner's $O(n^3)$ decoder (Eisner, 1996) is used to find the optimal parse. In dependency reparsing we focus on unlabeled dependencies, as described by Eisner (1996). Eisner (1996a, 1996b) describes several dependency-based models that are also closely related to the models in this article. Eisner (Eisner, 1996) proposed an $O(n^3)$ parsing algorithm for PDG. In many dependency parsing models such as (Eisner, 1996) and (Macdonald et al, 2005), the score of a dependency tree is the sum of the scores of the dependency links, which are computed independently of other links.
N03-1017	**Title:** Statistical Phrase-Based Translation
	Summary: The phrase-based decoder extracts phrases from the word alignments produced by GIZA++, and computes translation probabilities based on the frequency of one phrase being aligned with another (Koehn et al, 2003). We use the model of Koehn et al (2003) as a baseline for our experiments. Currently, the most successful such systems employ so-called phrase-based methods that translate input text by translating sequences of words at a time [Och, 2002; Zens et al, 2002; Koehn et al, 2003; Vogel et al, 2003; Tillmann, 2003] phrase-based machine translation systems make use of a language model trained for the target language.
A00-1043	**Title:** Sentence Reduction For Automatic Text Summarization
	Summary: Many algorithms exploit parallel corpora (Jing 2000; Knight and Marcu 2002; Riezler et al 2003; Nguyen et al 2004a; Turner and Charniak 2005; Mcdonald 2006) to learn the correspondences between long and short sentences in a supervised manner, typically using a rich feature space induced from parse trees. Jing and Mckeown (2000) and Jing (2000) propose a cut-and-paste strategy as a computational process of automatic abstracting and a sentence reduction strategy to produce concise sentences. Sentence reduction the task of the sentence reduction module, described in detail in (Jing, 2000), is to remove extraneous phrases from extracted sentences.

Table A.1: The output of C-LexRank summarization system for 3 papers from Table 5.1 in 3 topics: DP, MT, and Summ.

A00-1023	**Title:** A Question Answering System Supported By Information Extraction
	Summary: Examples of using NLP and IE in question answering include shallow parsing [Kupiec 1993] [Srihari & Li 2000], deep parsing [Li et al 2002] [Litkowski 1999] [Voorhees 1999], and IE [Abney et al 2000] [Srihari & li 2000]. If the expected answer types are typical named entities, information extraction engines (Bikel et al 1999, Srihari and li 2000) are used to extract candidate answers. We use a qa system supported by increasingly sophisticated levels of ie [Srihari & Li 2000] [Li et al 2002].
D04-9907	**Title:** Scaling Web-Based Acquisition Of Entailment Relations
	Summary: Many recent efforts have also focused on extracting binary semantic relations between entities, such as entailments (Szpektor et al 2004), is-a (Ravichandran and Hovy 2002), part-of (Girju et al 2003), and other relations. The tease algorithm (Szpektor et al, 2004) is an unsupervised method for acquiring entailment relations from the web for a given input template. Szpektor et al (2004) automatically identify anchors in web corpus data. Many recent efforts have also focused on extracting semantic relations between entities, such as entailments (Szpektor et al 2004), is-a (Ravichandran and Hovy 2002), part-of (Girju et al 2006), and other relations.
W05-0622	**Title:** Semantic Role Labelling with Tree CRFs
	Summary: Our parsing model is based on a conditional random field model, however, unlike previous TreeCRF work, e.g., (Cohn and Blunsom, 2005; Jousse et al., 2006), we do not assume a particular tree structure, and instead find the most likely structure and labeling. Some researchers (Xue and Palmer, 2004; Koomen et al., 2005; Cohn and Blunsom, 2005; Punyakanok et al., 2008; Toutanova et al., 2005; Toutanova et al., 2008) used a pipelined approach to attack the task. The model can be used for tasks like syntactic parsing (Finkel et al., 2008) and Semantic Role Labeling (Cohn and Blunsom, 2005).

Table A.2: The output of C-LexRank summarization system for 3 papers from Table 5.1 in 3 topics: QA, TE, and CRF.

APPENDIX B

Expert Summaries

In this appendix, we list the 4 manual summaries that were written for Question Answering for experiments in Chapter VII.

Question Answering Summaries; Reference: Abstracts
Expert 1
Automatic Question Answering (QA) is intended to model aspects of dialog processing in an evaluative task (Voorhees, 2005). Approaches to QA include: (1) the use of surface text patterns for a Maximum Entropy approach (Ravichandran et al., 2003; Ravichandran and Hovy, 2002); (2) multi-layered answer construction from factual information scattered in different documents (Saquete et al. 2004); (3) rule-based extraction of sentences that best answer a set of questions (Riloff and Thelen, 2000); (4) a combination of syntactic and semantic features and machine learning techniques (Wang et al., 2000); (5) interactive QA wherein users of information systems to pose questions in natural language and obtain relevant answers (Small et al., 2003); (6) intelligent use of paraphrases to increase the likelihood of finding the answer to the users question (Rinaldi et al., 2003). In the layered approach, complex temporal questions are first decomposed into simpler ones and then answers of each simple question are re-composed, fulfilling the temporal restrictions of the original complex question. This approach achieved 85% precision and 71% recall. By contrast, the rule-based approach achieved 40% accuracy. In the interactive approach, clarification dialogue was often needed to negotiate with the user the exact scope and intent of the question. In evaluating components of QA systems, Srihari and Li (2000) demonstrate that: (1) Named Entity tagging is an important component for QA; (2) an NL shallow parser provides a structural basis for questions, and (iii) high-level domain independent IE can result in a QA breakthrough.
Expert 2
These abstracts focused on system descriptions. Srihari and Li (2000) presents Textract, which uses named entity tagging to improve QA performance. Ravichandran and Hovy (2002) and Ravichandran et al. (2003) describe a QA system based on automatically generated text patterns. They reported results on the TREC-10 question set. Riloff and Thelen (2000) presents Quarc, which finds the sentence in a document that best answers a query using heuristic rules. Wang et al. (2000) presents a system that uses machine learning on syntactic and semantic features to perform QA. These abstracts described domain applications. Rinaldi et al. (2003) presents a paraphrase-based system tailored to technical domains. Squete et al. (2004) presents a system for handling complex questions, i.e. where the answer requires information from several documents, with focus on a module dealing with temporal complexity. Small et al. (2003) presents HITIQA, an interactive system that provides answers to questions for which the type of the answer is dependent on the content of the document. Users participate in a clarification dialogue to negotiate the scope and intent of questions. These abstracts dealt with evaluation. Voorhes (2005) describes the TREC 2004 QA track. The task was to give answers to a series of questions in a dialogue. This task was a good granularity for evaluation because it was small enough to represent a single user interation, but large enough to avoid scores skewed by single-question answers. Damerau (1981) provides data about the actual use of Transformational Question Answering during 1978, including problematic inputs.

Table B.1: Sample expert surveys of Question Answering using abstracts.

Question Answering Summaries; Reference: Citations
Expert 1
Question answering (QA) requires selection of question keywords, query generation, and answer generation (Abney et al, 2000) (Moldovan et al, 2000) (Srihari and Li, 2000). QA differs from IR: (i) instead of keyword terms, the query is a natural language question (ii) instead of documents or URLs, a list of candidate answers are returned. Factoid QA systems extract query words from the question, perform IR, identify likely answers, rank and select the best (Harabagiu et al, 2001; Hovy et al, 2001; Srihari and Li, 2000; Abney et al, 2000). Definition questions involve nuggets about a particular person, entity, event. If answers are named entities, information extraction is used (Bikel et al 1999, Srihari and Li 2000). Manually generated rules may be used to select answer sentences (Riloff and Thelen, 2000). Many approaches learn lexical semantic relations through co-occurrence patterns (Hearst, 1992; Ravichandran and Hovy, 2002; Moldovan et al, 2004). Saquete et al (2004) decompose complex temporal questions into simpler ones. Semantics-poor techniques (Soubbotin, 2002; Ittycheriah et al, 2002), yield answers to factoid questions. More complex tasks, e.g., disambiguation, require consideration of text meaning (Harabagiu et al, 2000) (Radev et al, 2000) (Srihari and Li, 2000). Relations may be selected through manual selection of entity pairs (Brin, 1998). The web has been employed for pattern acquisition, query expansion (Yang et al, 2003), and answer validation (Magnini et al, 2002). Ravichandran and Hovy (2002) use an ontology to extract answers from surface text patterns. Many researchers have used paraphrases for QA (Agichtein et al, 2001; Florence et al, 2003; Rinaldi et al, 2003; Tomuro, 2003; Lin and Pantel, 2001).
Expert 2
In Question Answering the query is a natural language question, and candidate answers are returned in response to a query. QA is supported by Natural Language Processing, Information Extraction and Named Entity Tagging (Srihari and Li 2000). Good answers to factoid and list questions include interesting nuggets about a particular person, organization, entity or event (Voorhees 2005). The web has been employed to learn surface patterns automatically from trivia question and answer pairs. Improvements may come from using a sentence splitter, anaphora resolution, and clustering of similar snippets (Ravichandran and Hovy 2002). (Saquete et al. 2004) decompose complex questions into simpler ones. Using lexical patterns to identify answers was competitive (Soubbotin and Soubbotin 2002). Rote extractors look for textual contexts that convey a relationship between concepts, such as m characters to the left or the right (Brin 1998), or the longest common substring of several contexts (Agichtein and Gravano 2000). More complex tasks require consideration of text meaning, which has motivated work on QA systems that incorporate semantic representation, ontologies, reasoning and inference engines (Moldovan et al. 2003). (Rinaldi et al. 2003) used paraphrasing knowledge in the question analyzer and answer locator, using minimal logical forms to represent texts and questions. Systems have been designed to take reading comprehension examinations (Charniak et al. 2000, Wang et al. 2000). The answer to a why question often precedes/follows the sentence with the highest number of matching words (Riloff and Thelen 2000). Trends have shifted towards interactive query answering (Harabagiu et al. 2005).

Table B.2: Sample expert surveys of Question Answering using citations.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Daniel M. Abrams and Steven H. Strogatz. Modelling the dynamics of language death. *Nature*, 424(6951):900, 2003.

[2] Eytan Adar, Daniel S. Weld, Brian N. Bershad, and Steven S. Gribble. Why we search: visualizing and predicting user behavior. In *Proceedings of the 16th International Conference on World Wide Web (WWW-07)*, pages 161–170, New York, NY, USA, 2007.

[3] Eytan Adar, Li Zhang, Lada A. Adamic, and Rajan M. Lukose. Implicit structure and the dynamics of Blogspace. In *WWW'04, Workshop on the Weblogging Ecosystem*, 2004.

[4] Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL '09)*, pages 19–27, 2009.

[5] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Diameter of the World Wide Web. *Nature*, 401(6749):130–131, September 1999.

[6] Galen Andrew and Jianfeng Gao. Scalable training of l1-regularized log-linear models. In *Proceedings of the 24th Annual International Conference on Machine Learning (ICML-07)*, pages 33–40, 2007.

[7] Awais Athar. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, HLT-SS '11, pages 81–87, 2011.

[8] Awais Athar and Simone Teufel. Context-enhanced citation sentiment detection. pages 597–601, Montréal, Canada, June 2012. Association for Computational Linguistics.

[9] Robert L. Axtell. Zipf distribution of US firm sizes. *Science*, 293(5536):1818, 2001.

[10] Seung Ki Baek, Hoang Anh, Tuan Kiet, and Beom Jun Kim. Family name distributions: Master equation approach. *Physical Review E*, 76, 2007.

[11] Ricardo Baeza-Yates and Gonzalo Navarro. Block addressing indices for approximate text retrieval. In *Proceedings of the Sixth International Conference on Information and Knowledge Management (CIKM-97)*, pages 1–8, 1997.

[12] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[13] Michael Barlow and Suzanne Kemmer, editors. *Usage-Based Models of Language*. CSLI Publications, Stanford, CA, 2000.

[14] Regina Barzilay and Lillian Lee. Bootstrapping lexical choice via multiple-sequence alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-02)*, pages 164–171, 2002.

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

[16] Danielle Smith Bassett and Ed Bullmore. Small-world brain networks. *The neuroscientist*, 12(6):512–523, 2006.

[17] John Batali. Computational simulations of the emergence of grammar. In J. R. Hurford, M. Studdert-Kennedy, and Knight C., editors, *Approaches to the Evolution of Language: Social and Cognitive Bases*, pages 405–426. Cambridge, 1998.

[18] Clay Beckner, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, and Tom Schoenemann. Language is a complex adaptive system: Position paper. *Language Learning*, 59(s1):1–26, December 2009.

[19] Christian Biemann. *Unsupervised and Knowledge-free Natural Language Processing in the Structure Discovery Paradigm*. PhD thesis, University of Leipzig, Germany, 2007.

[20] Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, 2008.

[21] Herbert Blumer. Collective behavior. *In Lee, Alfred McClung, Ed., Principles of Sociology.*, 1951.

[22] Nino Boccara. *Modeling complex systems*. Springer Verlag, 2010.

[23] Shannon Bradshaw. *Reference Directed Indexing: Indexing Scientific Literature in the Context of Its Use*. PhD thesis, Northwestern University, 2002.

[24] Shannon Bradshaw. 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*, 2003.

[25] Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32:13–47, March 2006.

[26] Joan Bybee. *Frequency of Use and the Organization of Language*. Oxford University Press, 2007.

[27] Silvia M. G. Caldeira, Thierry C. Petit Lobão, R. F. S. Andrade, Alexis Neme, and J. G. V. Miranda. The network of concepts in written texts. *The European Physical Journal B-Condensed Matter*, 49(4):523–529, 2006.

[28] Jaime G. Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98)*, pages 335–336, 1998.

[29] Jean Carletta. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254, 1996.

[30] Georges Chapouthier. Mosaic structures – a working hypothesis for the complexity of living organisms. *E-Logos (Electronic Journal for Philosophy)*, (17), 2009.

[31] David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. pages 190–200, Portland, Oregon, USA, June 2011.

[32] Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA, 1965.

[33] Noam Chomsky. *Reflections on Language*. Pantheon Books, New York, 1975.

[34] Monojit Choudhury, Animesh Mukherjee, Anupam Basu, and Niloy Ganguly. Analysis and synthesis of the distribution of consonants over languages: A complex network approach. In *Proceedings of the COLING/ACL on Main conference (poster sessions)*, pages 128–135, 2006.

[35] Morten H. Christiansen. *Infinite Languages, Finite Minds: Connectionism, Learning and Linguistic Structure*. PhD thesis, University of Edinburgh, Scotland, 1994.

[36] Morten H. Christiansen and Nick Chater. Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489–509, 2008.

[37] Phillip Clarkson and Roni Rosenfeld. Statistical language modeling using the CMU-cambridge toolkit. *Proceedings ESCA Eurospeech*, 47:45–148, 1997.

[38] Aaron Clauset, Mark E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70(6):066111, Dec 2004.

[39] Walter Daelemans, Steven Gillis, and Gert Durieux. The acquisition of stress: a data-oriented approach. *Computational Linguistics*, 20(3):421–451, 1994.

[40] Bing Tian Dai, Nick Koudas, Beng Chin Ooi, Divesh Srivastava, and Suresh Venkatasubramanian. Rapid identification of column heterogenity. In *Proceedings of the IEEE International Conference on Data Mining (ICDM-06)*, 2006.

[41] Gökhan Dalkilic and Yalcin Cebi. Zipf's law and Mandelbrot's constants for Turkish language using Turkish corpus (TurCo). *Lecture notes in computer science*, pages 273–282, 2004.

[42] D. Das and A.F.T. Martins. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4:192–195, 2007.

[43] Bart de Boer. Emergence of vowel systems through self-organisation. *AI Communications*, 13(1):27–40, 2000.

[44] Bart de Boer. Self-organization in vowel systems. *Journal of Phonetics*, 28(4):441–465, 2000.

[45] Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *The International Conference on Language Resources and Evaluation (LREC)*, volume 6, pages 449–454. Citeseer, 2006.

[46] Terrence W. Deacon. *The Symbolic Species: The Co-evolution of Language and the Brain*. Norton, New York, 1997.

[47] Sergey N. Dorogovtsev and José Fernando F. Mendes. Scaling behaviour of developing and decaying networks. *Europhysics Letters*, 52(1):33–39, October 2000.

[48] Sergey N. Dorogovtsev and José Fernando F. Mendes. Language as an evolving word Web. *Proceedings of the Royal Society of London B*, 268(1485):2603–2606, 2001.

[49] Michael Elhadad. Using argumentation in text generation. *Journal of Pragmatics*, 24:189–220, 1995.

[50] Aaron Elkiss, Siwei Shen, Anthony Fader, Güneş Erkan, David States, and Dragomir R. Radev. 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, 2008.

[51] Nick C. Ellis and Fernando Ferreira-junior. Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal*, 93(3):370–385, 2009.

[52] Jeffrey L. Elman. Connectionist views of cognitive development, where next? *TRENDS in Cognitive Sciences*, 9(3), 2005.

[53] Paul Erdös and Alfréd Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–60, 1960.

[54] Güneş Erkan. Language model-based document clustering using random walks. In *Proceedings of the HLT-NAACL conference*, pages 479–486, New York City, USA, June 2006. Association for Computational Linguistics.

[55] Güneş Erkan and Dragomir R. Radev. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research (JAIR)*, 2004.

[56] Ramon Ferrer i Cancho, Christiaan Janssen, and Ricard V. Solé. The topology of technology graphs: Small world patterns in electronic circuits. *Physical Review E*, 64(4):046119–1–046119–5, October 2001.

[57] Ramon Ferrer i Cancho and Ricard V. Solé. The small-world of human language. *Proceedings of the Royal Society of London B*, 268(1482):2261–2265, November 2001.

[58] Ramon Ferrer i Cancho and Ricard V. Solé. Zipf's law and random texts. *Advances in Complex Systems*, 5(1):1–6, 2002.

[59] Ramon Ferrer i Cancho, Ricard V. Solé, and Reinhard Köhler. Patterns in syntactic dependency networks. *PRE*, 69(5), May 26 2004.

[60] Len Fisher. *The Perfect Swarm: The Science of Complexity in Everyday Life*. Basic Books, 2009.

[61] Massimo Franceschet. PageRank: Stand on the shoulders of giants. Report, Department of Mathematics and Computer Science, University of Udine, 2010.

[62] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[63] Chikara Furusawa and Kunihiko Kaneko. Zipf's law in gene expression. *Physical review letters*, 90(8):88102, 2003.

[64] Xavier Gabaix. Zipf's law for cities: An explanation. *Quarterly journal of Economics*, 114(3):739–767, 1999.

[65] Jianfeng Gao, Galen Andrew, Mark Johnson, and Kristina Toutanova. A comparative study of parameter estimation methods for statistical natural language processing. In *Proceedings of the 45th Annual Conference of the Association for Computational Linguistics (ACL-07)*, 2007.

[66] Alexander F. Gelbukh and Grigori Sidorov. Zipf and Heaps laws' coefficients depend on language. In *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '01, pages 332–335, 2001.

[67] Michelle Girvan and Marm E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences (PNAS)*, 99(12):7821–7826, June 2002.

[68] E. Mark Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967.

[69] Deborah M. Gordon. *Ants at work: how an insect society is organized*. Free Press., 1999.

[70] Joseph Harold Greenberg. *Language universals: With special reference to feature hierarchies*. Walter De Gruyter Inc., 2005.

[71] Barbara F. Grimes. Ethnologue: Languages of the world. *Dallas, TX: Summer Institute of Linguistics*, 1996.

[72] Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Comput. Linguist.*, 12:175–204, July 1986.

[73] Zelig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

[74] Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev. What's with the attitude? identifying sentences with attitude in online discussions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-10)*, pages 1245–1255, October 2010.

[75] Nick Hatzigeorgiu, George Mikros, and George Carayannis. Word length, word frequencies and Zipf's law in the Greek language. *Journal of Quantitative Linguistics*, 8(3):175–185, 2001.

[76] Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. The faculty of language: what is it, who has it, and how did it evolve? *Science*, 298:1569–1579, 2002.

[77] Taher H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th International Conference on World Wide Web (WWW-02)*, pages 517–526, 2002.

[78] Taher H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):784–796, 2003.

[79] Harold Stanley Heaps. *Information retrieval: Computational and theoretical aspects*. Academic Press, Inc. Orlando, FL, USA, 1978.

[80] Wesley Hildebrandt, Boris Katz, and Jimmy Lin. Overview of the TREC 2003 question-answering track. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL '04)*, 2004.

[81] Geoffrey E. Hinton and Stephen J. Nowlan. How learning can guide evolution. *Complex Systems*, 1(1):495–502, June 1987.

[82] Lu Hong and Scott E. Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385 – 16389, 2004.

[83] Lu Hong and Scott E. Page. Interpreted and generated signals. *Journal of Economic Theory*, 144(5):2174–2196, 2009.

[84] Bernardo A. Huberman and Lada A. Adamic. Growth dynamics of the World Wide Web. *Nature*, 401(6749), September 9, 1999.

[85] Bernardo A. Huberman, Peter L. Pirolli, James E. Pitkow, and Rajan M. Lukose. Strong Regularities in World Wide Web Surfing. *Science*, 280(5360):95–97, 1998.

[86] Akshay Java, Pranam Kolari, Tim Finin, and Tim Oates. Modeling the spread of influence on the blogosphere. In *Proceedings of the 15th International Conference on World Wide Web (WWW-06)*, 2006.

[87] Hongyan Jing. Sentence reduction for automatic text summarization. In *Proceedings of the sixth conference on Applied Natural Language Processing*, pages 310–315, Morristown, NJ, USA, 2000.

[88] Mark T. Joseph and Dragomir R. Radev. Citation analysis, centrality, and the ACL Anthology. Technical Report CSE-TR-535-07, University of Michigan. Department of Electrical Engineering and Computer Science, 2007.

[89] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics (2nd edition)*. Prentice-Hall, 2008.

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

[91] Dain Kaplan, Ryu Iida, and Takenobu Tokunaga. Automatic extraction of citation contexts for research paper summarization: A coreference-chain based approach. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 88–95, Suntec City, Singapore, August 2009.

[92] Brian Karrer, Elizaveta Levina, and Mark E. J. Newman. Robustness of community structure in networks. *Physical Review E*, 77:046119, 2008.

[93] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD-09)*, pages 137–146. ACM, 2003.

[94] David Kempe, Jon Kleinberg, and Éva Tardos. Influential nodes in a diffusion model for social networks. In *Proceedings of the 32nd International Colloquium on Automata, Languages and Programming (ICALP-05)*, Lisboa, Portugal, 2005.

[95] Samir Khuller, Anna Moss, and Joseph (Seffi) Naor. The budgeted maximum coverage problem. *Inf. Process. Lett.*, 70(1):39–45, 1999.

[96] Aniket Kittur, Ed Chi, Bryan A. Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *alt.CHI at 25th Annual ACM Conference on Human Factors in Computing Systems (CHI-2007)*, 1(2), 2007.

[97] Aniket Kittur and Robert E. Kraut. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work (CSCW-2008)*, pages 37–46, 2008.

[98] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Conference of the Association for Computational Linguistics (ACL-03)*, pages 423–430, 2003.

[99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:604–632, September 1999.

[100] Michael E. Krauss. The world's languages in crisis. *Language*, 68(1):4–10, 1992.

[101] Klaus Krippendorff. *Content Analysis: An Introduction to its Methodology*. Beverly Hills: Sage Publications, 1980.

[102] Ariel Kulik, Hadas Shachnai, and Tami Tamir. Maximizing submodular set functions subject to multiple linear constraints. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA-09)*, pages 545–554, Philadelphia, PA, USA, 2009. Society for Industrial and Applied Mathematics.

[103] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the bursty evolution of blogspace. In *Proceedings of the 12th International Conference on World Wide Web (WWW-03)*, pages 568–576, New York, NY, USA, 2003.

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

[105] Peter Ladefoged and Ian Maddieson. *The sounds of the world's languages*. Wiley-Blackwell, 1996.

[106] Steve Lawrence and C. Lee Giles. Accessibility of information on the Web. *Nature*, 400(6740):107–109, July 8 1999.

[107] Lillian Lee. Measures of distributional similarity. In *Proceedings of the 37th Annual Conference of the Association for Computational Linguistics (ACL-99)*, pages 25–32, 1999.

[108] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD-09)*, pages 497–506, 2009.

[109] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD-07)*, pages 420–429, 2007.

[110] Wentian Li. Random texts exhibit Zipf's law-like word frequency distribution. *IEEE-TOIT*, 38(6):1842–1845, November 1992.

[111] Erez Lieberman, Christoph Hauert, and Martin A. Nowak. Evolutionary dynamics on graphs. *Nature*, 433(7023):312–316, 2005.

[112] David Lightfoot. *The Language Lottery: Toward a Biology of Grammar*. MIT press, Cambridge, MA, 1982.

[113] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL workshop on Text Summarization Branches Out*, 2004.

[114] Chin-Yew Lin and Eduard Hovy. Manual and automatic evaluation of summaries. In *ACL-Workshop on Automatic Summarization*, 2002.

[115] Dekang Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING/ACL-98)*, pages 768–774, Stroudsburg, PA, USA, 1998.

[116] Jimmy J. Lin and Dina Demner-Fushman. Methods for automatically evaluating answers to complex questions. *Information Retrieval*, 9(5):565–587, 2006.

[117] Haitao Liu and Fengguo Hu. What role does syntax play in a language network? *Europhysics Letters*, 83(18002), July 2008.

[118] Haitao Liu and Chunshan Xu. Can syntactic networks indicate morphological complexity of a language? *Europhysics Letters*, 93:28005, 2011.

[119] Linyuan Lü, Zi-Ke Zhang, and Tao Zhou. Zipf's law leads to Heaps' law: analyzing their relation in finite-size systems. *PloS One*, 5(12):e14139, 2010.

[120] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, 28(2):203–208, 1996.

[121] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[122] Christopher D. Manning and Hirich Schutze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, London, England, 2002.

[123] Qiaozhu Mei, Jian Guo, and Dragomir Radev. Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD-10)*, pages 1009–1018, 2010.

[124] Qiaozhu Mei and ChengXiang Zhai. Generating impact-based summaries for scientific literature. In *Proceedings of the 46th Annual Conference of the Association for Computational Linguistics (ACL-08)*, pages 816–824, 2008.

[125] Filippo Menczer. Evolution of document networks. *Proceedings of the National Academy of Sciences (PNAS)*, 101(1):5261–5265, April 6, 2004.

[126] Rada Mihalcea. Language independent extractive summarization. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 49–52, 2005.

[127] Rada Mihalcea and Paul Tarau. TextRank: Bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, July 2004.

[128] George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[129] George A. Miller and Patricia M Gildea. How children learn words. *Scientific American*, pages 94–99, 1987.

[130] Lesley Milroy. *Language change and social networks*. Blackwell, Oxford, 1980.

[131] Jorge Mira and Ángel Paredes. Inter-linguistic similarity and language death dynamics. *EPL (Europhysics Letters)*, 69:1031, 2005.

[132] Jorge Mira, Luis F Seoane, and Ángel Paredes. Can two languages coexist within the same community of speakers? *Contributions to science*, 6(1):21–26, 2011.

[133] Gilad Mishne and Natalie Glance. Predicting movie sales from blogger sentiment. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, 2006.

[134] Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. Using citations to generate surveys of scientific paradigms. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL '09)*, pages 584–592, Boulder, Colorado, June 2009.

[135] Jose M. Montoya and Ricard V. Solé. Small world patterns in food webs. *Journal of theoretical biology*, 214(3):405–412, 2002.

[136] Adilson E. Motter, Alessandro P. S. de Moura, Ying-Cheng Lai, and Partha Dasgupta. Topology of the conceptual network of language. *Physical Revire E*, 65(065102), June 25, 2002.

[137] Divoli A. Nakov, P. and M. Hearst. Do peers see more in a paper than its authors? *Advances in Bioinformatics, special issue on Literature Mining Solutions for Life Science Research*, 2012.

[138] Hidetsugu Nanba, Takeshi Abekawa, Manabu Okumura, and Suguru Saito. Bilingual PRESRI: Integration of multiple research paper databases. In *Proceedings of RIAO 2004*, pages 195–211, Avignon, France, 2004.

[139] Hidetsugu Nanba, Noriko Kando, and Manabu Okumura. Classification of research papers using citation links and citation types: Towards automatic review article generation. In *Proceedings of the 11th SIG Classification Research Workshop*, pages 117–134, Chicago, USA, 2004.

[140] Hidetsugu Nanba and Manabu Okumura. Towards multi-paper summarization using reference information. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI-99)*, pages 926–931, 1999.

[141] S Naranan. Statistical laws in information-science, language and system of natural-numbers – some striking similarities. *Journal of Scientific and Industrial Research*, 51(8-9):736–755, 1992.

[142] Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402, 2004.

[143] Ani Nenkova and Rebecca Passonneau. 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)*, 2004.

[144] Mark E. J. Newman. The structure of scientific collaboration networks. *PNAS*, 98(2):404–409, 2001.

[145] Mark E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67:026126, 2003.

[146] Mark E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[147] Mark E. J. Newman. Analysis of weighted networks. *Physical Review E*, 70–056131, 2004.

[148] Partha Niyogi. *The Computational Nature of Language Learning and Evolution*. MIT Press, Cambridge, MA, April 2006.

[149] Matrin A. Nowak, Natalia L. Komarova, and Partha Niyogi. Computational and evolutionary aspects of language. *Nature*, 417(6889):611–617, 2002.

[150] Sebastian Padò and Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, June 2007.

[151] Scott E. Page. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press, 2007.

[152] Bo Pang and Lillian Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42st Annual Conference of the Association for Computational Linguistics (ACL-04)*, 2004.

[153] Thiago Alexandre Salgueiro Pardo, Lucas Antiqueira, Maria das Graças Volpe Nunes, Osvaldo N. Oliveira Jr., and Luciano da Fontoura Costa. Modeling and evaluating summaries using complex networks. In *Proceedings of Computational Processing of the Portuguese Language, the Seventh International Workshop (PROPOR-06)*, 2006.

[154] Marco Patriarca and Teemu Leppänen. Modeling language competition. *Physica A: Statistical Mechanics and its Applications*, 338(1-2):296–299, 2004.

[155] Michael Paul, ChengXiang Zhai, and Roxana Girju. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-10)*, pages 66–76, 2010.

[156] Steven Pinker. *The language instinct*. Harper Perennial, 1995.

[157] Andrei Popescu-Belis and John Batali. Incremental simulations of the emergence of grammar: Towards complex sentence-meaning mappings. In *Third International Conference on the Evolution of Language*, pages 187–190, 2000.

[158] Geoffrey K. Pullum. Learnability, hyperlearning, and the poverty of the stimulus. In *Proceedings of the 22nd Annual Meeting of the Berkeley Linguistics Society*, Berkeley CA, 1996. Berkeley Linguistics Society.

[159] Vahed Qazvinian and Dragomir R. Radev. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, Manchester, UK, 2008.

[160] Vahed Qazvinian and Dragomir R. Radev. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual Conference of the Association for Computational Linguistics (ACL-10)*, pages 555–564, Uppsala, Sweden, July 2010.

[161] Vahed Qazvinian and Dragomir R. Radev. Exploiting phase transition in latent networks for clustering. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI-11)*, 2011.

[162] Vahed Qazvinian and Dragomir R. Radev. Learning from collective human behavior to introduce diversity in lexical choice. In *Proceedings of the 49th Annual Conference of the Association for Computational Linguistics (ACL-11)*, pages 1098–1108, 2011.

[163] Vahed Qazvinian and Dragomir R. Radev. Computational analysis of collective discourse. In *Proceedings of Collective Intelligence (CI-2012)*, 2012.

[164] Vahed Qazvinian, Dragomir R. Radev, and Arzucan Özgür. Citation summarization through keyphrase extraction. In *Proceedings of the 23nd International Conference on Computational Linguistics (COLING-10)*, pages 895–903, 2010.

[165] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-11)*, pages 1589–1599, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[166] Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. MEAD - a platform for multidocument multilingual text summarization. In *LREC 2004*, Lisbon, Portugal, May 2004.

[167] Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. The ACL anthology network corpus. In *ACL workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, Singapore, 2009.

[168] Dragomir R. Radev and Daniel Tam. Summarization evaluation using relative utility. In *Proceedings of the twelfth international conference on Information and knowledge management (CIKM)*, pages 508–511. ACM, 2003.

[169] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web (WWW-11)*, pages 337–346, 2011.

[170] Erzsébet Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551, 2002.

[171] Sidney Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 4(2):131–134, 1998.

[172] Joseph Reisinger and Raymond J. Mooney. Multi-prototype vector-space models of word meaning. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL '10)*, June 2010.

[173] Craig W. Reynolds. Flocks, herds and schools: A distributed behavioral model. *ACM SIGGRAPH Computer Graphics*, 21(4):25–34, 1987.

[174] Michael E. Roberts. *Human Collective Behavior*. PhD thesis, Indiana University, 2008.

[175] Peter Mark Roget. *Roget's Thesaurus of English words and phrases.* TY Crowell co., 1911.

[176] Ronald Rousseau and Qiaoqiao Zhang. Zipf's data on the frequency of chinese words revisited. *Scientometrics*, 24:201–220, 1992.

[177] David E. Rumelhart and Ronald J. Williams Geoffrey E. Hinton. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[178] P. Thomas Schoenemann. Evolution of the size and functional areas of the human brain. *Annual Review of Anthropology*, 35:379–406, 2006.

[179] Mark S. Seidenberg. Language acquisition and use: Learning and applying probabilistic constraints. *Science*, 275:1559–1603, 1997.

[180] Advaith Siddharthan and Simone Teufel. Whose idea was this, and why does it matter? attributing scientific work to citations. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL '07)*, 2007.

[181] Neil J. Smelser. *Theory of Collective Behavior*. Free Press, 1963.

[182] Kenny Smith, Henry Brighton, and Simon Kirby. Complex systems in language evolution: the cultural emergence of compositional structure. *Advances in Complex Systems*, 6(4):537–558, 12 2003.

[183] Ricard V. Solé, Bernat Corominas Murtra, Sergi Valverde, and Luc Steels. Language networks: their structure, function and evolution. *Trends in Cognitive Sciences*, 12(42):343–352, 2005.

[184] Karen Spärck-Jones. Automatic summarizing: factors and directions. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in automatic text summarization*, chapter 1, pages 1 – 12. The MIT Press, 1999.

[185] Manfred Stede. Lexicalization in natural language generation: a survey. *Artificial Intelligence Review*, 8:309–336, 1995.

[186] Luc Steels. Self-organizing vocabularies. In Christopher G. Langton and Katsunori Shimohara, editors, *Artificial Life V*, pages 179–184, Nara, Japan, 1996.

[187] Luc Steels. The synthetic modeling of language origins. *Evolution of Communication*, 1(1):1–34, 1997.

[188] Luc Steels. The emergence of grammar in communicating autonomous robotic agents. In *Proceedings of the 14th European Conference on Artificial Life (ECAI-00)*, pages 764–769, 2000.

[189] Mark Steyvers and Joshua B. Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78, 2005.

[190] James Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* Doubleday, May 2004.

[191] Daniel Tam, Dragomir R. Radev, and Gunes Erkan. Single-document and multi-document summary evaluation using relative utility. Technical Report CSE-TR-538-07, University of Michigan. Department of Electrical Engineering and Computer Science, 2007.

[192] Simone Teufel. Argumentative Zoning for Improved Citation Indexing. *Computing Attitude and Affect in Text: Theory and Applications*, pages 159–170, 2005.

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

[194] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-06)*, pages 103–110, Sydney, Australia, July 2006.

[195] Guy Theraulaz and Eric Bonabeau. Modelling the collective building of complex architectures in social insects with lattice swarms. *Journal of Theoretical Biology*, 177(4):381–400, 1995.

[196] Takashi Tomokiyo and Matthew Hurst. A language model approach to keyphrase extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 33–40, 2003.

[197] Peter E. Trapa and Martin A. Nowak. Nash equilibria for an evolutionary language game. *Journal of Mathematical Biology*, 41(2):172–188, 2000.

[198] Hans van Halteren and Simone Teufel. Examining the consensus between human summaries: initial experiments with factoid analysis. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop*, pages 57–64, Morristown, NJ, USA, 2003.

[199] Hans van Halteren and Simone Teufel. Evaluating information content by factoid analysis: human annotation and stability. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, Barcelona, 2004.

[200] Faraneh Vargha-Khadem, Kate Watkins, Kattie Alcock, Paul Fletcher, and Richard Passingham. Praxic and nonverbal cognitive deficits in a large family with a genetically transmitted speech and language disorder. *Proceedings of the National Academy of Sciences (PNAS)*, 92:930–933, 1995.

[201] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-98)*, pages 315–323, 1998.

[202] Ellen M. Voorhees. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-2003)*, 2003.

[203] Xiaojun Wan and Jianwu Yang. Improved affinity graph based multi-document summarization. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL '06)*, pages 181–184, 2006.

[204] Xiaojun Wan and Jianwu Yang. Multi-document summarization using cluster-based link analysis. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-08)*, pages 299–306, 2008.

[205] William S-Y. Wang and James W. Minett. The invasion of language: emergence, change and death. *Trends in Ecology & Evolution*, 20(5):263–269, 2005.

[206] Xiao Fan Wang and Guanrong Chen. Complex networks: small-world, scale-free and beyond. *Circuits and Systems Magazine, IEEE*, 3(1):6–20, 2003.

[207] Duncan J. Watts and Steven Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, June 1998.

[208] Uriel Weinreich, Willian Labov, and Marvin I. Herzog. Empirical foundations for a theory of language change. In Winfred P. Lehmann and Yakov Malkiel, editors, *Directions for historical linguistics. A symposium*, pages 95–195. University of Texas Press, Austin, TX, 1968.

[209] Michael Alan Whidby. Citation handling: Processing citation texts in scientific documents. Master's thesis, University of Maryland, Department of Computer Science, College Park, MD, 2012.

[210] Andrew Whiten. *Natural theories of mind: Evolution, development and simulation of everyday mindreading.* Basil Blackwell, 1991.

[211] Richard J. Williams and Neo D. Martinez. Simple rules yield complex food webs. *Nature*, 404(6774):180–182, March 9, 2000.

[212] Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. Learning discriminative projections for text similarity measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 247–256, Portland, Oregon, USA, June 2011.

[213] George Udny Yule. A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:21–87, 1925.

[214] David M. Zajic, Bonnie J. Dorr, Jimmy Lin, and Richard Schwartz. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing and Management (Special Issue on Summarization)*, 2007.

[215] Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis. Technical Report TR #01–40, Department of Computer Science, University of Minnesota, Minneapolis, MN, 2001.

[216] Xiaojin Zhu, Andrew Goldberg, Jurgen Van Gael, and David Andrzejewski. Improving diversity in ranking using absorbing random walks. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (HLT-NAACL '07)*, pages 97–104, 2007.

[217] George K. Zipf. *The psycho-biology of language: an introduction to dynamic philology.* Houghton Mifflin, 1935.

[218] George K. Zipf. *Human Behaviour and the Principle of Least Effort.* Addison-Wesley, Cambridge, MA, 1949.