Text Summarization in Data Mining

Colleen E. Crangle

ConverSpeech LLC, 60 Kirby Place, Palo Alto, California 94301, USA
crangle@converspeech.com
www.converspeech.com

Abstract. Text summarizers automatically construct summaries of a natural-language document. This paper examines the use of text summarization within data mining, identifying the potential summarizers have for uncovering interesting and unexpected information. It describes the current state of the art in commercial summarization and current approaches to the evaluation of summarizers. The paper then proposes a new model for text summarization and suggests a new form of evaluation. It argues that for summaries to be truly useful within data mining, they must include concepts abstracted from the text in addition to sentences extracted from the text. The paper uses two news articles to illustrate its points.

1 Introduction

To summarize a piece of writing is to present the main points in a concise form. Work on automated text summarization began over 40 years ago [1]. The growth of the Internet invigorated this work in recent years [2], and summarization systems are beginning to be applied in areas such as healthcare and digital libraries [3]. Several commercially available text summarizers are now on the market. Examples include Capito from Semiotis, Inxight's summarizer, the Brevity summarizer from LexTek International, the Copernic summarizer, TextAnalyst from Megaputer, and Whiskey™ from Converspeech. These programs work by automatically extracting selected sentences from a piece of writing.

A true summary succinctly expresses the gist of a document, revealing the essence of its content. This paper examines the use of text summarization within data mining for uncovering interesting and unexpected information. It describes the current state of the art in summarization systems and current approaches to the evaluation of summarizers. The paper then proposes a new model for text summarization and suggests a new form of evaluation. It argues that for summaries to be truly useful within data mining, they must include concepts abstracted from the text in addition to sentences extracted. Such summarizers offer a potential not yet exploited in data mining.

2 Summarizers in Data Mining

Much of the information crucial to an organization exists in the form of unstructured text data. That is, the information does not reside in a database with well-defined

D. Bustard, W. Liu, and R. Sterritt (Eds.): Soft-Ware 2002, LNCS 2311, pp. 332–347, 2002.

methods of organization and access, but is expressed in natural language and is contained within various documents such as web pages, e-mail messages, and other electronic documents. The process of identifying and extracting valuable information from such data repositories is known as text data mining. Tools to do the job must go beyond simple keyword indexing and searching. They must determine, at some level, what a document is about.

2.1 Text Data Mining

Keyword indexing and searching can provide a specific answer to a specific question, such as "*What is the deepest lake in the United States*?" with the answer being found in a piece of text such as: "*Crater Lake, at 1,958 feet (597 meters) deep, is the seventh deepest lake in the world and the deepest in the United States*."

Keyword indexing and searching can also provide answers to more complex questions, such as *"What geological processes formed the three deepest lakes in the world*?" Several sources will probably have to be consulted, their information fused, and interpretations made (what counts as a geological process versus a human intervention), and conclusions drawn. But standard keyword indexing and searching will probably suffice to find the pieces of text needed.

Text data mining goes beyond question answering. It seeks to uncover interesting and useful patterns in large repositories of text, answering questions may not yet have been posed. The focus is on discovery, not simply finding what is sought. The focus is on uncovering unexpected content in text and unexpected relationships between pieces of text, not simply the text itself.

2.2 Summaries in Text Data Mining

Summaries aid text data mining in at least the following ways:

- An information analyst—whether a social scientist, a member of the intelligence community, or a market researcher—uses summaries to guide her examination of data repositories that are so large she cannot possibly read everything or even browse the repository adequately. Summaries suggest what documents should be read in their entirety, which should be read together or in sequence, and so on.
- Summaries of the individual documents in a collection can reveal similarities in their content. The summaries then form the basis for clustering the documents or categorizing them into specified groups. Applications include Internet portal management, evaluating free-text responses to survey questions, help-desk automation for responses to customer queries, and so on. The very process of categorizing or clustering two document summaries into the same group can reveal an unexpected relationship between the documents.
- The summary of a collection of related documents taken together can reveal aggregated information that exists only at the collection level. In biomedicine, for example, Swanson has used summaries together with additional information-extraction techniques to form a new and interesting clinical hypothesis [4].

An interesting and significant form of indeterminacy creeps into summarization. It results from the inherent indeterminacy of meaning in natural language. Summaries,

whether produced by a human abstractor or a machine, are generally thought to be good if they capture the author's intent, that is, succinctly present the main points the author intended to make. (There are other kinds of summarization in which sentences are extracted relative to a particular topic, a technique that is a form of information extraction.) So-called neutral summaries, however, those that aim to capture the author's intent, can succeed only to the extent that the author had a clear intent and expressed it adequately. What if the author had no clear intent or was an inadequate writer? Poor writers abound, and most short written communications, such as e-mail messages or postings to electronic bulletin boards, are messy in content and execution. Do automated text summarizers reveal anything useful in these cases? If the summarization technique is itself valid, the answer is that the summary reveals what the piece of text is really about, whether the author intended it or not.

Various studies have explored the indeterminacy of meaning in language, and the extent to which meaning depends on the context in which language is used [5, 6]. Author's intent does not bound meaning nor fully determine the content of a document. When documents are pulled together and their collective content is examined, there generally is no single author anyway whose intent could dominate.

An automated summarizer that reveals what a text is really about, independent of authorial intent, is a powerful tool in data mining. It has the potential to reveal new and interesting information in a document or a collection of documents.

The pressing and practical concern is how to evaluate any given summarizer; that is, how do we know whether or not it produces good summaries? What counts as a good summary, and does that judgment depend on the purpose the summary is to serve? Within data mining, for example, summaries that revealed unexpected content or unexpected relationships between documents would be of the greatest value. The next section looks at current work in summarization evaluation.

3 Evaluating Summarizers

A group representing academic, U.S. government, and commercial interests has been working over the past few years to draw up guidelines for the design and evaluation of summarization systems. This work arose out of the TIDES Program (Translingual Information Detection, Extraction, and Summarization) sponsored by the Information Technology Office (ITO) of the U.S. Defense Advanced Research Project Agency (DARPA). In a related effort, the National Institute of Standards and Technology (NIST) of the U.S. government has initiated a new evaluation series in text summarization. Called the Document Understanding Conference (DUC), this initiative has resulted in the production of reference data—documents and summaries—for summarizer training and testing. (See http://www-nlpir.nist.gov/projects/duc/ for further information.)

A key task accomplished by these initiatives was the compilation of sets of test documents. In these sets the important sentences (or sentence fragments) within each document are annotated by human evaluators, and/or for each document, human-generated abstracts of various lengths are provided.

An early example of a summary data set consisted of eight news articles published by seven news providers, New York Times, CNN, CBS, Fox News, BBC, Reuters,

and Associated Press, on June 3rd, 2000, the eve of the meeting between Presidents Clinton and Putin. One of these articles is used below.

Several approaches to summarizer evaluation have been identified. They include:

- Using the annotated sentences.
- For each document, counting how many of the annotated (i.e., important) sentences are included in the summary. A simple measure of percent agreement can be applied, or the traditional measures of recall and precision.[1]
- Using the abstracts.
- Counting how many of the sentences in the human-generated abstracts are represented by sentences in the summaries. A simple measure of percent agreement can be applied, or the traditional measures of recall and precision.
- Using a question-answering task.
- For a given set of pre-determined questions, counting how many of the questions can be answered using the summary? The more questions can be answered, the better the summary.
- Using the utility method of Radev [7] .
- Using the content-based measures of Donaway [8].

To illustrate a simple evaluation, consider the following test document. The underlined sentences are those considered important by the human evaluators.

CLINTON TAKES STAR WARS PLAN TO RUSSIA

- US president Bill Clinton has arrived in Moscow for his first meeting with Russia's new president Vladmir Putin. The two heads of state will meet on Saturday night for an informal dinner before getting down to business on Sunday.
- High on the agenda will be the United State's plans to build a missile shield in Alaska. Russia opposes the shield as it contravenes a pact signed by the two countries in 1972 which bans any anti-missile devices.
- Clinton—in his last few months of office and keen to make his mark in American history—will be seeking to secure some sort of concession from Putin.
- The Russian leader has said that he will suggest an alternative to the US system.
- Kremlin officials said Putin would propose a system that would shoot down the missiles with interceptors shortly after they were fired rather than high in their trajectory.
- "We'll talk about it in Russia," Clinton told reporters before leaving Berlin for Moscow. "It won't be long now." Accompanying the President is US Secretary of State Madeline Albright. "What's new is that Putin is signalling that he is open to discuss it, that he is ready for talks," she said. "We will discuss it."
- Arms control will not be the only potentially troublesome issue. US National Security Adviser Sandy Berger said last week Clinton would raise human rights and press freedom.

Here is an automatically generated summary of this text:

---

[1] **Recall** refers to the number of annotated sentences correctly included in the summary, divided by the total number of annotated sentences. **Precision** refers to the number of annotated sentences correctly included in the summary, divided by the total number of sentences (correctly or incorrectly) included in the summary.

CLINTON TAKES STAR WARS PLAN TO RUSSIA
- US president Bill Clinton has arrived in Moscow for his first meeting with Russia's new president Vladmir Putin.
- The two heads of state will meet on Saturday night for an informal dinner before getting down to business on Sunday.
- High on the agenda will be the United State's plans to build a missile shield in Alaska.
- Russia opposes the shield as it contravenes a pact signed by the two countries in 1972 which bans any anti-missile devices.
- Clinton—in his last few months of office and keen to make his mark in American history—will be seeking to secure some sort of concession from Putin.

This extraction summary has three of the five important sentences, and of its six sentences (including the heading) three are considered important. Simple recall and precision figures of 60% and 50% result.

All the current evaluation approaches assume that a summary is produced by extracting sentences. Are there other ways to think about summarization? Are there also new ways to think about evaluating summarizers? This author would argue yes, particularly in the context of data mining. In data mining, we are interested in discovering, not merely finding, information. We may need to dig beneath the surface of a text to make such discoveries. §6 returns to these questions, after a brief review of the state of the art in text summarization and presentation of a new model for summarization.

4 Text Summarization: The State of the Art

Current summarizers work by extracting key sentences from a document. As yet, there is no summarizer on the market or even within the research community that truly fuses information to create a set of new sentences to represent the document's content. In general, summarizers simply extract sentences. They differ in the methods they use to select those sentence. There are two main kinds of methods involved, that may be used separately or in combination:

1. *Heuristic methods, based largely on insight into how human, professional abstractors work*. Many of these heuristics exploit document organization. So, for example, sentences in the opening and closing paragraphs are more likely to be in the summary. Some heuristics exploit the occurrence of cue phrases such as "in conclusion" or "important."
2. *Methods based on identifying key words, phrases, and word clusters*. The document is analyzed using statistical and/or linguistic techniques to identify the words, phrases or word clusters that by their frequency and co-occurrence are thought to represent the content of the document. Then sentences containing or related to these words and phrases are selected.

The techniques commercial summarizers use to identify key words and phrases are often proprietary and can only be inferred from the extracted sentences. What is readily seen, however, is whether or not the method identifies concepts in the text. Concepts are expressed using words and phrases that may or may not appear within the text. Concept identification as opposed to key word and phrase identification is a cru-

cial differentiating factor between summarizers. Summaries that contain true abstractions from the text are more likely to reveal unexpected, sometimes hidden, information within documents and surprising relationships between documents. A true abstraction summarizer can be a powerful tool for text data mining.

It is important from a scientific point of view to devise objective measures to evaluate summarizers. However, given that the output of a summarizer is itself natural-language text, some human judgment is inescapable. The DUC initiative relies heavily on human evaluators.

Based on informal testing of several dozen documents of various kinds—business and marketing documents (regulatory filing, product description, business news article), personal communications (fax, e-mail, letter), non-technical pieces (long essay, short information piece, work of fiction), scientific articles, and several documents that pose specific challenges (threaded bulletin board messages, enumerations in text, program code in text)—what follows is an intuitive judgment of the state of commercially available summarizers.

Current summarizers are able to produce adequate sentence-extraction summaries of articles that have the following characteristics:

- The article is well written and well organized.
- It is on one main topic.
- It is relatively short (600-2,000 words).
- It is informational, for example, a newspaper article or a technical article in an academic journal. It is not a work of the imagination, such as fiction, or an opinion piece or general essay.
- It is devoid of layout features such as enumerations, indented quotations, or blocks of program code. (Although some summarizers use heuristics that take headings into account, for example, summarizers typically ignore or strip a document of most of its layout features.)

Some summarizers perform limited post-processing "smoothing" on the sentences they list in an attempt to give coherence and fluency to the summary. This post-processing includes:

- Removing inappropriate connecting words and phrases. If a sentence in the document begins with a connecting phrase—for example, "Furthermore" or "Although"—and that sentence is selected for the summary, the connecting phrase must be removed from the summary because it probably no longer plays the connecting role it was meant to.
- Resolving anaphora and co-reference. When a sentence is selected for inclusion in the summary, the pronouns (and other referring phrases) in it have to be resolved. That is, the summarizer has to make clear what that pronoun (or referring phrase) refers to. For example, suppose a document contains the following sentences: "*Newcompany Inc. has recently reported record losses. If it continues to lose money, it risks strong shareholder reaction. The company yesterday announced new measures to…*"

  If the second sentence is selected for the summary, the word "it" has to be resolved to refer to Newcompany Inc.; otherwise, the reader will have no idea what "it" is, and will naturally relate "it" to whatever entity is named in the preceding sentence of the summary. Ideally, the summary sentence should appear as: "*If it [Newcompany Inc.] continues to loose money, it risks strong shareholder reaction.*"

If the third sentence is selected for the summary, the phrase "the company" should similarly be identified as referring to Newcompany Inc. and not any other company that may be named in a preceding sentence of the summary. Anaphoric and co-reference resolution is very difficult; not surprisingly, current commercial summarizers incorporate very few, if any, of these techniques.

Current research into summarization has a strong emphasis on post-processing techniques.

5 A New Model for Summarization

The standard model of summary production is represented by the sequence shown in Figure 1.

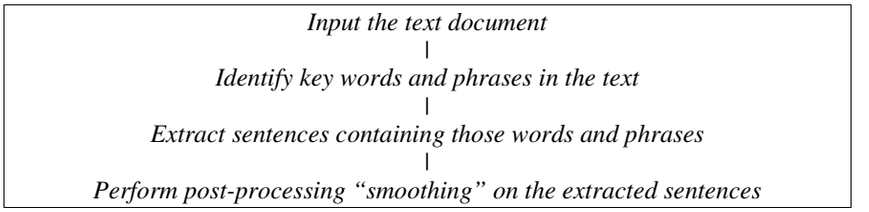

Fig. 1. Standard model of summary production

What if the summarizer is able to identify key concepts and not just key words and phrases? Not only can the key concepts by themselves stand as an encapsulated summary of the document, concepts can provide a better basis for selecting sentences to be extracted. An enhanced model results, as depicted in Figure 2.

A summary that provides information not immediately evident from a surface reading of the text is of potentially great value in data mining. To test the assumption that concepts can provide a better basis for selecting sentences to be extracted, and to understand the significance of this enhanced model, the action of three different commercially available summarizers on the news article in Appendix I is considered. The first summarizer simply produces sentences. The second additionally displays key words and phrases from the text along with the extracted sentences. The third, ConverSpeech's Whiskey, abstracts concepts and uses those concepts to extract sentences.

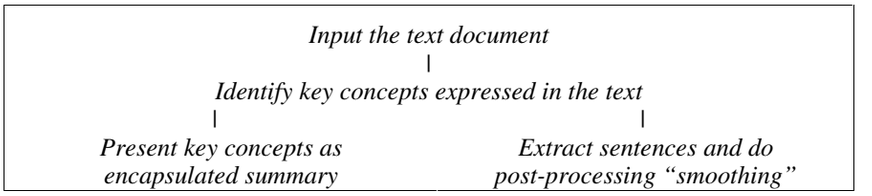

Fig. 2. Enhanced, concept-based, model of summary production

The first summarizer produced the following sentences extracted from the text:

- About 20 Bay Area companies are performing so badly that they are in danger of being booted off the Nasdaq, the stock exchange that lists most of the area's high-tech companies.
- Five local companies were already bumped off last year, and a sixth – PlanetRx.com Inc., a former South San Francisco health care company – was just delisted.
- The whole Internet market crashed down, and we're rolling with it," says Peter Friedman, CEO of Talk City Inc., a company that could get kicked off Nasdaq if it doesn't boost its stock price soon.
- With stock prices down and the economy slowing, companies are falling short of the standards Nasdaq sets for its some 3,802 companies.
- While the listing standards are arcane, the most obvious cardinal sin in the eyes of Nasdaq's regulators is simple: The fall of a company's stock price below $1 for 30 consecutive trading days.

The second summarizer produced the following key words and phrases and sentences extracted from the text:

- Nasdaq, stock, delisting, firms, investors, stock price, stock exchange, San, officer, Edison
- It's the company version of the pink slip in the mail – get your act together, or you're fired from Nasdaq.
- About 20 Bay Area companies are performing so badly that they are in danger of being booted off the Nasdaq, the stock exchange that lists most of the area's high-tech companies.
- Five local companies were already bumped off last year, and a sixth – PlanetRx.com Inc., a former South San Francisco health care company – was just delisted.
- While the delisting doesn't have to mean the game is over, it relegates companies to the junior and less reputable leagues of the stock exchange world, where it's much harder to raise money.
- "The whole Internet market crashed down, and we're rolling with it," says Peter Friedman, CEO of Talk City Inc., a company that could get kicked off Nasdaq if it doesn't boost its stock price soon.
- Once booted, companies usually end up in the netherworlds of the stock market, where only a few brave investors venture.
- This exchange doesn't require firms to register with the Securities and Exchange Commission or even file financial statements.
- "We're working on strategic partnerships that will have a major impact on the stock," says Nadyne Edison, chief marketing officer for the company.

The third summarizer produced a high-level abstraction, a listing of the key concepts expressed in the text, and a list of extracted sentences. The number after each sentence is its score, calculated on the basis of how many occurrences of the words in the concept list appear in the sentence, optionally normalized for sentence length. Those sentences that receive the top 75% scores are selected for inclusion. This percentage is set as a parameter.

Notice that the concept of business has been extracted from the text even though the word "business" appears only once in the text, in the last sentence. Note also that

the word “time” also does not occur frequently in the text but the concept of time does.

time	business	capital
company	Nasdaq	stock
day	working	share

- About 20 Bay Area companies are performing so badly that they are in danger of being booted off the Nasdaq, the stock exchange that lists most of the area’s high-tech companies. (378)
- Five local companies were already bumped off last year, and a sixth—PlanetRx com Inc., a former South San Francisco health care company—was just delisted. (352)
- Nationwide, Nasdaq has either sent notices or is close to notifying at least 200 other companies, many of whom offered stocks to the public for the first time last year. (368)
- While the delisting doesn’t have to mean the game is over, it relegates companies to the junior and less reputable leagues of the stock exchange world, where it’s much harder to raise money. (400)
- “The whole Internet market crashed down, and we’re rolling with it,” says Peter Friedman, CEO of Talk City Inc., a company that could get kicked off Nasdaq if it doesn’t boost its stock price soon. (438)
- While the listing standards are arcane, the most obvious cardinal sin in the eyes of Nasdaq’s regulators is simple: the fall of a company’s stock price below $1 for 30 consecutive trading days. (404)
- Autoweb.com Inc., a Santa Clara Internet company that specializes in auto consumer services, has about 40 days left under the 90-day rule, but is busy scrambling to avoid a hearing. (390)

The three summaries have three sentences in common, and the third summary has one additional sentence in common with each of the first and second summaries.

6 A New Evaluation Method

How good are these three summaries and the summarizers that produced them? Any of the evaluation methods mentioned in §3 could be applied to assess the value of the extracted sentences. However, in the context of data mining, A new evaluation method for summarizers is proposed here. It asks the following question: *How sensitive is a summarizer to surface perturbations in the text, such as in word choice or sentence order*?

Specifically, this method asks what happens if synonyms are substituted for words and phrases in the text. Does the summarizer give a different summary, selecting sentences that differ markedly in content from the previously selected ones? Similarly, the order of some of the sentences is changed, does that markedly alter what gets identified as key sentences?

This test gives a good indication of the robustness of the summarizer and the soundness of the methods used to identify the content of the document. If simple

changes in word choice or sentence order produce different summaries, it could be argued that the summarizer is not getting at the core of the document's content.

The news article in Appendix I uses the words "firm" and "company" interchangeably, with 23 occurrences of "company" (the more familiar word) and four occurrences of "firm. If we substitute "firm" for "company" in key sentences in the text, what happens? Two tests were performed. In the first two substitutions were made: *About 20 Bay Area companies are performing so badly that they are in danger of being booted off the Nasdaq, the stock exchange that lists most of the area's high-tech ~~companies~~ firms. Five local ~~companies~~ firms were already bumped off last year, and a sixth—PlanetRx.com Inc., a former South San Francisco health care company—was just delisted.*

In the second test there were three additional substitutions: *"The whole Internet market crashed down, and we're rolling with it," says Peter Friedman, CEO of Talk City Inc., a ~~company~~ firm that could get kicked off Nasdaq if it doesn't boost its stock price soon. …With stock prices down and the economy slowing, ~~companies~~ firms are falling short of the standards Nasdaq sets for its some 3,802 ~~companies~~ firms.*

The results obtained were as follows:

First Summarizer. With the first round of substitutions, one sentence from the original summary was removed and a different sentence was inserted.

- *Out*: Five local firms were already bumped off last year, and a sixth– PlanetRx.com Inc., a former South San Francisco health care company was just delisted.
- *In*: When that happens, Nasdaq sends a notice giving the company 90 calendar days to get the stock price up again.

With the second round of substitutions, another sentence from the original summary was removed and a different sentence from the text inserted.

- *Out*: With stock prices down and the economy slowing, firms are falling short of the standards Nasdaq sets for its some 3,802 firms.
- *In*: If a company sold things on the Web—cars, pet food, you name it—it was almost guaranteed a spot on the stock exchange.

Second Summarizer. The two rounds of substitutions produced only one change—the removal of the following sentence after the second round of substitutions:

- *Out*: This exchange doesn't require firms to register with the Securities and Exchange Commission or even file financial statements.

Third Summarizer. The two rounds of substitutions produced the same sentences. The word "firm" was added to the list of concepts for both rounds.

To further test the second and third summarizers, which appeared somewhat equally robust, they were run on two more versions of the article with several further substitutions of "firm" for "company." Both summarizers produced stable sets of sentences for these changes: the second summarizer retained the same altered set of sentences as for the other substitutions, and the third summarizer continued to select the same sentences throughout.

These two summarizers were also run on the Clinton/Putin test article give earlier, and on two variations of that article. The first variation was obtained by substituting

"anti-missile device" for the following four phrases which, in the context, were synonymous with "anti-missile device": "missile shield," "shield," and "system."

High on the agenda will be the United State's plans to build ~~a missile shield~~ an anti-missile device in Alaska. Russia opposes the ~~shield~~ anti-missile device as it contravenes a pact signed by the two countries in 1972 that bans any anti-missile devices. … The Russian leader has said that he will suggest an alternative to the US ~~system~~ anti-missile device.

A second variation was obtained by further substituting the presidents' last names ("Putin" and "Clinton" respectively) for the referring expressions "the Russian leader" and "the President" in the following sentence:

Putin ~~The Russian leader~~ has said that he will suggest an alternative to the US anti-missile device. … Accompanying Clinton ~~the President~~ is US Secretary of State Madeline Albright.

These were the results obtained.

Second Summarizer. For the original article, the following words and phrases and extracted sentences had been produced: *Clinton, Putin, Russia, president, Moscow, STAR WARS PLAN, missile shield, business, informal dinner, heads*

CLINTON TAKES STAR WARS PLAN TO RUSSIA

- US president Bill Clinton has arrived in Moscow for his first meeting with Russia's new president Vladmir Putin.
- The two heads of state will meet on Saturday night for an informal dinner before getting down to business on Sunday.
- High on the agenda will be the United State's plans to build a missile shield in Alaska.
- Russia opposes the shield as it contravenes a pact signed by the two countries in 1972 which bans any anti-missile devices.
- Clinton—in his last few months of office and keen to make his mark in American history—will be seeking to secure some sort of concession from Putin.

The following lists of key words and phrases were produced for the two altered versions of the article:

- Clinton, Putin, Russia, president, anti-missile device, Moscow, STAR WARS PLAN, business, informal dinner, heads
- Clinton, Putin, Russia, anti-missile device, Moscow, president Bill Clinton, STAR WARS PLAN, business, informal dinner, heads

For both of the two altered versions, the following sentence was dropped from the summary, with no other sentence being substituted:

- *Out*: High on the agenda will be the United State's plans to build an anti-missile device in Alaska.

Third Summarizer. The same set of sentences was extracted for the original article and the two variations. The following listing of abstracted concepts preceded the original summary. It is notable that the concept of country was identified as significant in the article even though the word "country" does not itself appear in the text.

state	business	president
Putin (Vladmir Putin)	US	Clinton (Bill Clinton)
country	missile (missile shield, missile devices)	system

The concepts abstracted for both of the two altered versions of the article were the following:

state	business	president
Putin (Vladmir Putin)	US	Clinton (Bill Clinton)
device	missile (missile shield, missile devices)	

The sentences extracted for all three versions of the article were as follows (with scores omitted):

- US president Bill Clinton has arrived in Moscow for his first meeting with Russia's new president Vladmir Putin.
- The two heads of state will meet on Saturday night for an informal dinner before getting down to business on Sunday.
- High on the agenda will be the United State's plans to build a missile shield in Alaska.
- Russia opposes the shield as it contravenes a pact signed by the two countries in 1972 which bans any anti-missile devices.
- Clinton – in his last few months of office and keen to make his mark in American history – will be seeking to secure some sort of concession from Putin.
- Kremlin officials said Putin would propose a system that would shoot down the missiles with interceptors shortly after they were fired rather than high in their trajectory.
- "What's new is that Putin is signalling that he is open to discuss it, that he is ready for talks," she said.
- US National Security Adviser Sandy Berger said last week Clinton would raise human rights and press freedom.

Once again, the second summarizer, while not as stable as the third, concept-based, summarizer, did perform with relative robustness. Only one sentence was eliminated from the summaries for the two versions of the article containing substitutions for synonymous terms.

However, the second and third summarizers differed markedly in their behavior when they were tested on articles that had re-ordered sentences. To illustrate, the same Clinton/Putin article.was used (Similar results were obtained with the news story on the Nasdaq delistings.) The Clinton/Putin article was rearranged to begin at the following sentence, with the displaced first two paragraphs tacked on at the end. (See Appendix II.) *Clinton—in his last few months of office and keen to make his mark in American history—will be seeking to secure some sort of concession from Putin.*

These were the results obtained.

Second Summarizer. It selected the following key words and phrases for the permuted article. There were seven in common with the original summary, three that were different ("Albright," "Russian leader," "concession").

- Clinton, Putin, Russia, president, STAR WARS PLAN, missile shield, State Madeline Albright, Moscow, Russian leader, concession.

The real limitation of the summarizer, however, is revealed in the sentences it selected for extraction. It had only four sentences in common with the original summary, eliminating two and adding three different ones:

- *Out*: The two heads of state will meet on Saturday night for an informal dinner before getting down to business on Sunday.
  Russia opposes the shield as it contravenes a pact signed by the two countries in 1972 which bans any anti-missile devices.
- *In*: The Russian leader has said that he will suggest an alternative to the US system.
  "We'll talk about it in Russia," Clinton told reporters before leaving Berlin for Moscow. Accompanying the President is US Secretary of State Madeline Albright.

This summarizer most likely uses a heuristic that is commonly employed in summarizing algorithms. The heuristic gives greater weight to a sentence the nearer it is to the beginning of the article. (A variation of this heuristic assigns a greater weight only to the first sentence of the article or the sentences in the first paragraph.) However, there is something fundamentally mistaken about over-reliance on this heuristic even though it may improve results under some of the other evaluation methods. A sentence is placed at the beginning of an article because it is important. It is not important because it is at the beginning of an article. Over-reliance on the heuristic confuses these two points.

Third Summarizer. In marked contrast, the concept-based summarizer produced exactly the same results for the permuted article (and all other permuted articles it was tested on.)

The alterations in the summaries produced by the second summarizer, resulting simply from sentence reordering, suggest that the summarizing technique lacks robustness. Similarly, the alterations in the summaries produced by the first summarizer, resulting simply from synonym substitution, also suggest lack of robustness. What is essentially different about the third summarizer is that it abstracts from the words and phrases that appear in the text, and relies on those abstracted concepts to extract sentences.

7 Conclusion

To capture the essence of a document, regardless of authorial intent, a summarizer must do more than identify key words and phrases in the text and extract sentences on that basis. It must also identify concepts expressed in the text. Summarizers that offer this level of abstraction appear to get at the essence of a text more reliably, showing a greater tolerance for superficial changes in the input text. Such summarizers are potentially powerful tools in data mining, uncovering information that lies beneath the surface of the words and phrases of the text.

References

1. H.P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2 (2), 1958.
2. Marti Hearst. Untangling Text Data Mining. Proceedings of ACL 99. *37th Annual Meeting of the Association for Computational Linguistics*, University of Maryland, June 1999.

3. Kathleen R. McKeown, et al. PERSIVAL, a System for Personalized Search and Summarization over Multimedia Healthcare Information, In *Proceedings of The First ACM+IEEE Joint Conference on Digital Libraries*. Roanoke, W. Va., June 2001.
4. Don R. Swanson and N.R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence, 91. 183-203 (1977)
5. C.E. Crangle. What words mean: some considerations from the theory of definition in logic. *Journal of Literary Semantics*, Vol. XXI, No. 1, 17-26, 1992.
6. C.E. Crangle and P. Suppes. *Language and Learning for Robots*. Stanford University, Stanford. CSLI Press. Distributed by Cambridge University Press, 1994
7. Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. *Proceedings of the ANLP/NAACL-2000 Workshop on Automatic Summarization*, pp. 21-30, Seattle, WA., 2000.
8. Robert L. Donaway, Kevin K. Drummey, and Laura A. Mather. A Comparison of Rankings Produced by Summarization Evaluation Measures. *Proceedings of the ANLP/NAACL-2000 Workshop on Automatic Summarization*, pp. 69-78, Seattle, WA., May 2000.

Appendix I: News Article

It's the company version of the pink slip in the mail—get your act together, or you're fired from Nasdaq.

About 20 Bay Area companies are performing so badly that they are in danger of being booted off the Nasdaq, the stock exchange that lists most of the area's high-tech companies. Five local companies were already bumped off last year, and a sixth—PlanetRx.com Inc., a former South San Francisco health care company—was just delisted.

Nationwide, Nasdaq has either sent notices or is close to notifying at least 200 other companies, many of whom offered stocks to the public for the first time last year.

While the delisting doesn't have to mean the game is over, it relegates companies to the junior and less reputable leagues of the stock exchange world, where it's much harder to raise money. For shareholders, a Nasdaq delisting sounds like a chilling death knoll – the value of their stock could all but implode. Some delisted companies, like Pets.com, simply close their doors.

"The whole Internet market crashed down, and we're rolling with it," says Peter Friedman, CEO of Talk City Inc., a company that could get kicked off Nasdaq if it doesn't boost its stock price soon. "The emotion was too much. Things just snapped."

This round of delistings is the ignominious end to a year of decadence now coming back to haunt us.

Most of these companies had no profits, and many had hardly any sales, when investor enthusiasm created a wave of new stock offerings last year. If a company sold things on the Web—cars, pet food, you name it—it was almost guaranteed a spot on the stock exchange.

But in less than a year, many of the same investors have abandoned their former darlings. With stock prices down and the economy slowing, companies are falling short of the standards Nasdaq sets for its some 3,802 companies.

While the listing standards are arcane, the most obvious cardinal sin in the eyes of Nasdaq's regulators is simple: The fall of a company's stock price below $1 for 30 consecutive trading days.

When that happens, Nasdaq sends a notice giving the company 90 calendar days to get the stock price up again. If it fails to do so—for 10 consecutive days—the firm has one last resort: an appeal to Nasdaq.

That involves a trek to Washington, D.C., and a quick hearing at a room in the St. Regis Hotel, where Nasdaq's three-person panel grills executives. Unless there's good reason to prolong the struggle, the company's Nasdaq days are over.

Once booted, companies usually end up in the netherworlds of the stock market, where only a few brave investors venture.

First, it's the Over The Counter Bulletin Board, which is considerably more risky and yields lower return to investors. However, even the OTCBB has requirements.

Failing that, the next step down is the so-called Pink Sheets, named for the color of the paper they used to be traded on. This exchange doesn't require firms to register with the Securities and Exchange Commission or even file financial statements.

"They're the wild, wild West," says Nasdaq spokesman Mark Gundersen.

Autoweb.com Inc., a Santa Clara Internet company that specializes in auto consumer services, has about 40 days left under the 90-day rule, but is busy scrambling to avoid a hearing.

"We're working on strategic partnerships that will have a major impact on the stock," says Nadyne Edison, chief marketing officer for the company. On Tuesday, Edison was in Detroit, busy opening a new office near the nation's auto capital. Edison says the firm is considering moving its headquarters to Detroit to be nearer its clients.

Other companies that got delisting notices are trying layoffs. Take Mountain View-based Network Computing Devices, which provides networking hardware and software to large companies. Its sales have been pinched as the personal computer industry slows down, so it has laid off people.

"We've had to downsize, downsize, downsize," says Chief Financial Officer Michael Garner.

Women.com, a San Mateo-based Internet site devoted to women, has laid off 25 percent of the workforce recently to avoid delisting. Becca Perata-Rosati, vice president of communications, says the site isn't being fairly rewarded by Wall Street. The company is the 29th most heavily visited Web site in the world, she says.

One trick that doesn't seem to work is the so-called "reverse stock split," which PlanetRx.com tried on Dec. 1. By converting every eight shares into one, PlanetRx.com hoped each share price would be boosted eightfold. But the move was seen by investors as a sign of desperation, and the stock plunged from $1 to 53 cents.

Out of alternatives, PlanetRx didn't even show up for its hearing with Nasdaq. It is now trading on the OTCBB after a recent move to Memphis and faces an uncertain future.

At least one executive says he doesn't mind the prospect of going to the OTCBB.

Talk City's Friedman says his company is growing, and expects its 9 million in service fee revenue to double this year. Even if he's forced off the Nasdaq, he has hopes of returning.

"I'd like to stay on the Nasdaq," he says. ``If we get off, we'll build a business. Then we'll go back on."

Contact Matt Marshall at mmarshall@sjmercury.com or (408)920-5920.

Appendix II: Permuted Clinton/Putin News Article

CLINTON TAKES STAR WARS PLAN TO RUSSIA

Clinton—in his last few months of office and keen to make his mark in American history—will be seeking to secure some sort of concession from Putin.

The Russian leader has said that he will suggest an alternative to the US system.

Kremlin officials said Putin would propose a system that would shoot down the missiles with interceptors shortly after they were fired rather than high in their trajectory.

"We'll talk about it in Russia," Clinton told reporters before leaving Berlin for Moscow. "It won't be long now." Accompanying the President is US Secretary of State Madeline Albright. "What's new is that Putin is signalling that he is open to discuss it, that he is ready for talks," she said. "We will discuss it."

Arms control will not be the only potentially troublesome issue. US National Security Adviser Sandy Berger said last week Clinton would raise human rights and press freedom.

US president Bill Clinton has arrived in Moscow for his first meeting with Russia's new president Vladmir Putin. The two heads of state will meet on Saturday night for an informal dinner before getting down to business on Sunday.

High on the agenda will be the United State's plans to build a missile shield in Alaska. Russia opposes the shield as it contravenes a pact signed by the two countries in 1972 that bans any anti-missile devices.