2-2014

The Inner Workings of Text Summarization Systems

Hope Cotton
Graduate Center, City University of New York

The Inner Workings of Automatic Text Summarization Systems: Mead and SweSum

By

Hope Cotton

A master’s thesis submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Master of Arts, The City University of New York.

2014

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the requirement for the degree of Master of Arts.

Dr. Virginia Teller **Date**
Thesis Adviser

Approved

Dr. Gita Martohardjono **Date**
Executive Officer

THE CITY UNIVERSITY OF NEW YORK

Abstract

THE INNER WORKINGS OF AUTOMATIC TEXT SUMMARIZATION SYSTEMS: MEAD AND SWESUM

by

Hope Cotton

Adviser: Professor Virginia Teller

This paper investigates automatic text summarization systems (ATS). To begin, a general description of how some ATS systems operate is provided. Following is a report on the testing of two summarization systems, Mead and SweSum. News articles from online sources were used for testing. Analyzing each system’s output under different parameter settings helped to uncover (to some extent) the architecture of these summarizers. Both systems were compared as to how user-focused they are, how informative the summaries are in relation to the source document, and how punctuation and format are reconstructed in a summary.

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF APPENDICES

A Brief Overview of Sentence Extraction and Features

Unlike our traditional understanding of a summary, many automatic text summarizers summarize by extracting sentences from the source document rather than by creating a new condensed version of the source document, containing different language and new information. The methods involved in automatic text summarization are often shallow and do not incorporate deep analysis of language. As mentioned by Inderjeet Mani in *Automatic Text Summarization*, part of the appeal of automatic summarizers is that deep analysis of language is not necessary, eliminating the need for an ontology which would otherwise require more time and resources to create. Documents in a summarizer are analyzed at the sentence level for features.

Current summarization software programs today continue to adhere to early work by Edmundson with some or little modification of it. The process of sentence extraction based on Edmundson's framework entails looking for certain features in a document that are indicative of a sentence's importance. According to Mani (2001), Edmundson determined that these features had to do with a sentence's position in a document, or certain words in a document which he called cue words, title words and keywords. Hand-assigned weights were applied to each feature based on evaluations of summaries he made in the testing phase. Edmundson concluded that the combination of cue-title-location yielded the best extracts and that location was the best individual feature while keywords was the worst individual feature.

Cue words are words that Edmundson took from the training corpus which were basically used as information regarding a sentence's selectability. Words like *significant*, *impossible*, and *hardly* allude to the importance of a sentence. Cue words above a certain corpus frequency cut-off were bonus words and those under it were stigma words. Mani explains that "according to Edmundson, the bonus terms consisted of 'comparatives, superlatives, adverbs of conclusion, value terms relative interrogatives, [and] causality terms' while the stigma terms consisted of 'anaphoric expressions, belittling expressions, insignificant detail expressions, [and] hedging expressions' p. 48."

Title words are words from titles, subtitles or section headings that are taken from the document to be summarized. Edmundson assigned weights to title words based on which title words provided the best output. Sentences containing title words from titles, subtitles and headings were deemed relevant to a summary extract under the assumption that title words are generally informative with regard to what the topic of the document is. Keywords are high-frequency content words above a corpus frequency cut-off; these words were also taken from the document to be summarized.

With sentence location, sentences under specific section headings like *introduction* or *conclusion* were given a positive weight for location; also, the ordinal position of sentences was used to assess a sentence's relevance in a summary extract. For example, the first and last sentences of a document were highly considered for selection.

Training and Learning rules

A summarization system, as explained by Mani, extracts relevant sentences from a document with a program that can learn rules. To learn these rules the summarizer first needs to be trained on how and which sentences to extract. Training involves two sets of corpora: a training corpus and a test corpus. The training set consists of source documents and their human-written summary extracts or abstracts. If training involves the use of abstracts rather than extracts, then the training phase becomes much more involved. New information not present in the source document requires a more complex method of matching the source document with the extracted sentences. The ideal type of summary to train a summarizer is with a human written summary extract, not an abstract. An abstract is a summary containing new information. The ideas in an abstract represent the general ideas of a source document using different words. The sentences in an abstract are not extracted from a document; rather, an abstract is the more traditional notion of what a summary is. The test corpus consists of just documents used to test the effectiveness of rules and the quality of the extracts.

In the training phase, the summarizer is programmed to extract the sentences of the source documents containing certain features. Vectors carrying the extracted sentences containing the features of interest interact with a vector labeler; this vector labeler matches the sentences extracted by the feature extractor with the human written summary extracts. Those sentences extracted by the feature extractor not matching the sentences in the human written summary extract are put aside or discarded. These two different summaries are then transmitted over to a learning algorithm program, which creates rules based on the matched sentences between the two different summary extracts and the features of interest.

After the rules are generated by the learning algorithm, the summarizer can now summarize documents from the test corpus. In the testing phase, the feature extractor extracts sentences from the various documents in the test corpus carrying the relevant features. These sentences are then put through the rule applicator which discards sentences incompatible with the rules. My guess is that there are high ranking and low ranking rules which allow for sentence ranking. The more features in a sentence, the more likely it will be ranked highly and appear in the extract.

Summarization Phases, from the training corpora to generating a summary

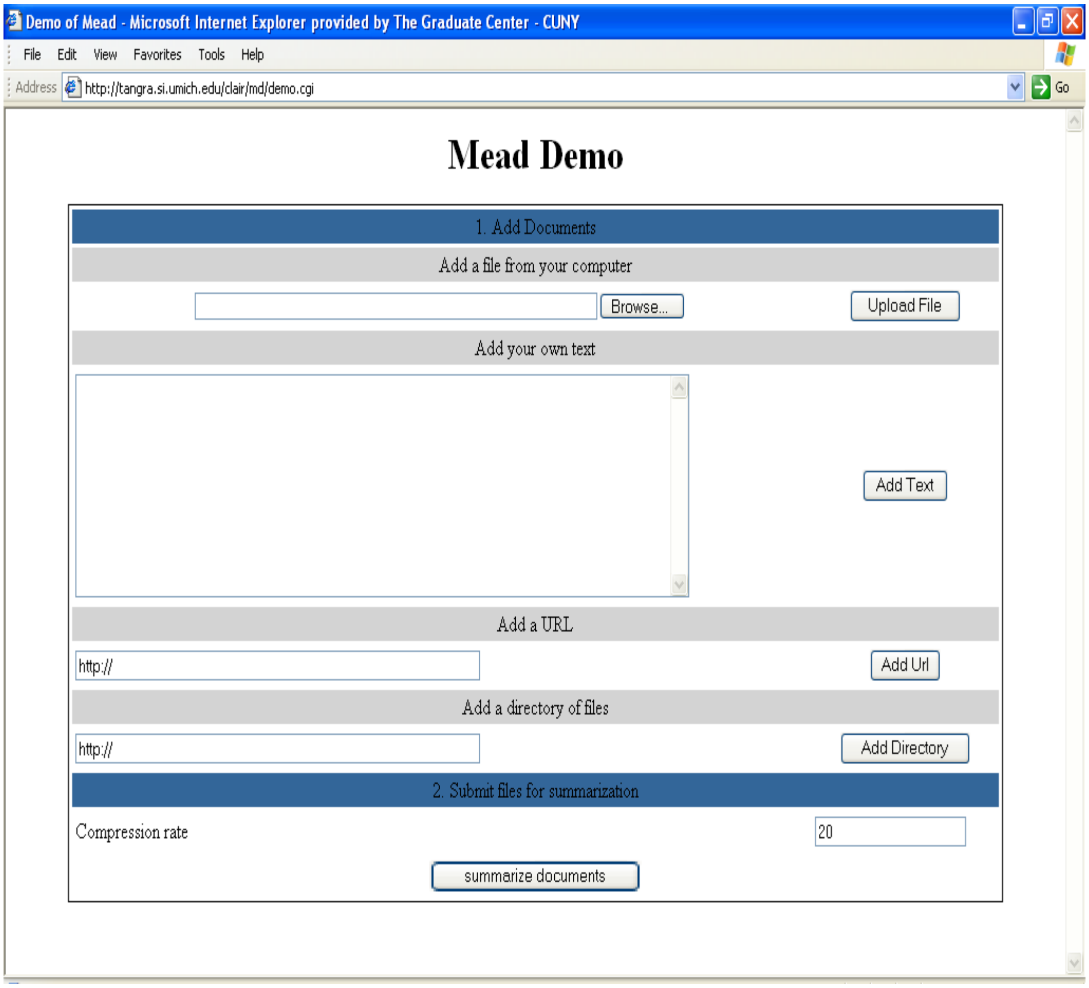
1. Training
    a. Corpora =>
    b. Feature Extractor =>

   c. Vector Labeler =>
   d. Rule Generator =>
2. Testing
   a. Corpora =>
   b. Feature Extractor =>
   c. Rule Applicator =>
3. Summarizing
   a. Text Document =>
   b. Feature Extractor =>
   c. Rule Applicator =>

Report on Mead and SweSum Testing

A Brief Description of the Summarizers' User-Interface. Narrowing down the list of summarizers found on the Internet was done by choosing from ones which are free and allow unlimited use. Brevity, Mead, Zentext, and SweSum provide unlimited use of their systems. However, testing was confined to Mead and SweSum being that they were more user oriented compared to the other summarizers, allowing for the manipulation of various parameters. Mead was created by the MEAD team at the Center for Language and Speech Processing at the John Hopkins University and SweSum was developed by Martin Hassel and Hercules Dalianus in Sweden. The first SweSum summarizer was originally in Swedish, but now it has expanded to summarize in many different languages.

Mead allows for parameters like compression rate to be controlled by the user. It summarizes single documents and multi-documents. To summarize a document one can paste the document in the, "add your own text section" of the user-interface, browse for a file and upload it, or type in a URL or a directory of files. The types of files, which it will upload, are limited to html files.



Demo of Mead - Microsoft Internet Explorer provided by The Graduate Center - CUNY
File Edit View Favorites Tools Help
Address
Go
Browse...
Upload File
Add Text
Add Url
Add Directory
summarize documents

Address: http://tangra.si.umich.edu/clair/md/demo.cgi

Mead Demo

1. Add Documents

Add a file from your computer

Add your own text

Add a URL

http://

Add a directory of files

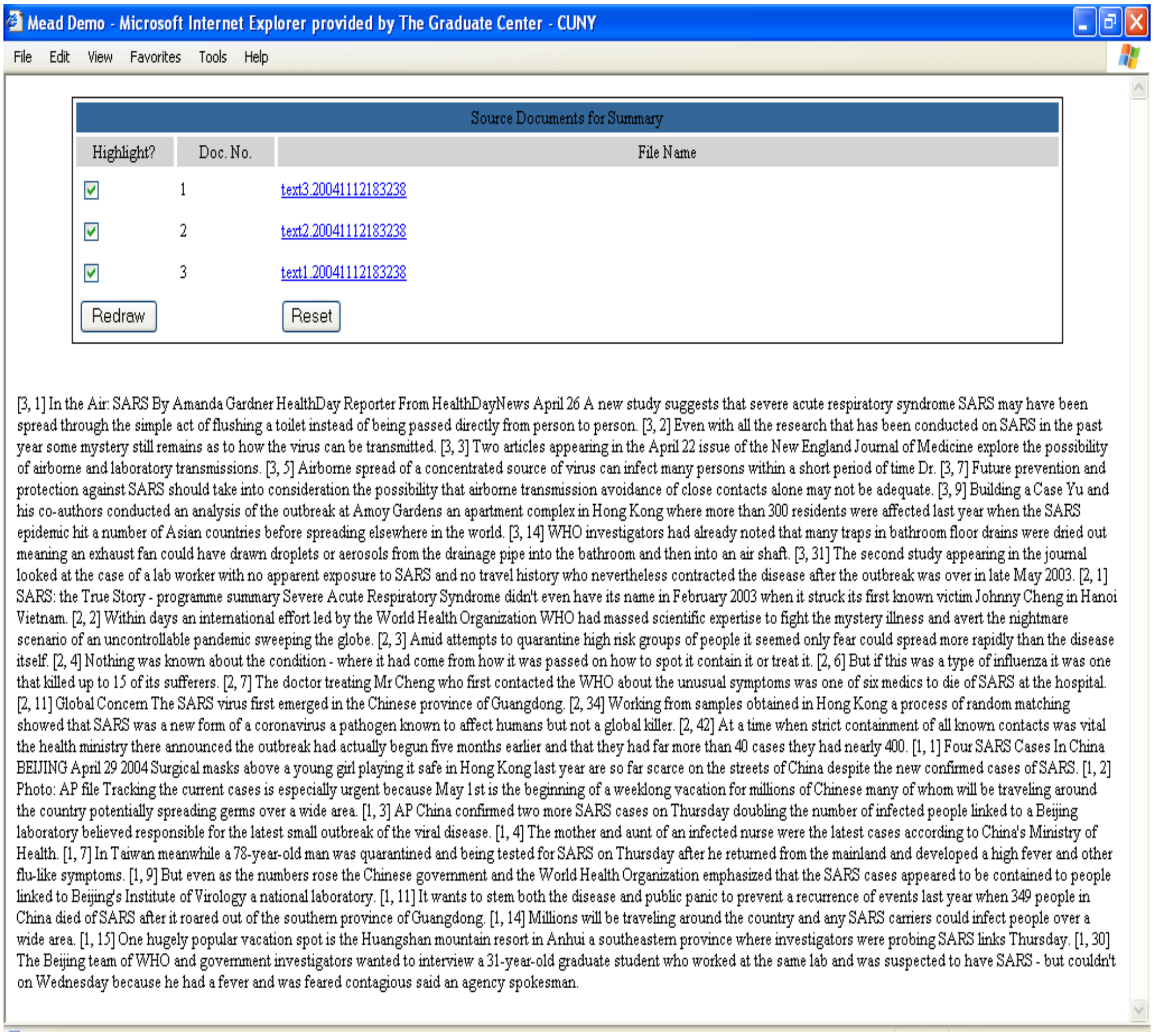
http://

2. Submit files for summarization

Compression rate: 20

Figure 1: Image of Demo of Mead

In seconds, Mead can upload a document, process it, and assemble the extracted sentences into one summary extract. The summary extract can include the URL of the source document if the documents are uploaded from websites, along with the summary.


Mead Demo - Microsoft Internet Explorer provided by The Graduate Center - CUNY
File Edit View Favorites Tools Help
Redraw
Reset

Source Documents for Summary

Highlight?	Doc. No.	File Name
☑	1	text3.20041112183238
☑	2	text2.20041112183238
☑	3	text1.20041112183238

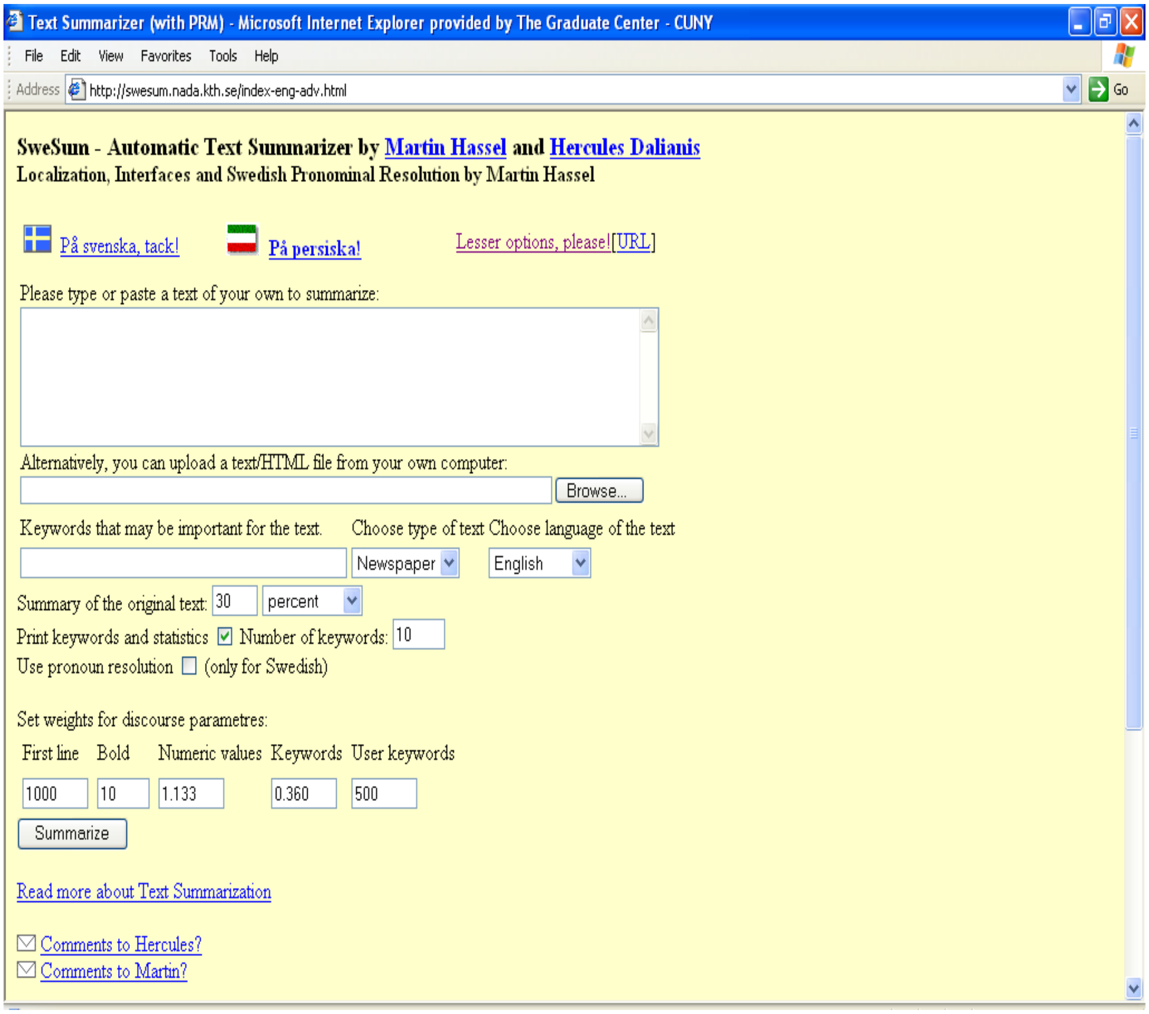
[3, 1] In the Air: SARS By Amanda Gardner HealthDay Reporter From HealthDayNews April 26 A new study suggests that severe acute respiratory syndrome SARS may have been spread through the simple act of flushing a toilet instead of being passed directly from person to person. [3, 2] Even with all the research that has been conducted on SARS in the past year some mystery still remains as to how the virus can be transmitted. [3, 3] Two articles appearing in the April 22 issue of the New England Journal of Medicine explore the possibility of airborne and laboratory transmissions. [3, 5] Airborne spread of a concentrated source of virus can infect many persons within a short period of time Dr. [3, 7] Future prevention and protection against SARS should take into consideration the possibility that airborne transmission avoidance of close contacts alone may not be adequate. [3, 9] Building a Case Yu and his co-authors conducted an analysis of the outbreak at Amoy Gardens an apartment complex in Hong Kong where more than 300 residents were affected last year when the SARS epidemic hit a number of Asian countries before spreading elsewhere in the world. [3, 14] WHO investigators had already noted that many traps in bathroom floor drains were dried out meaning an exhaust fan could have drawn droplets or aerosols from the drainage pipe into the bathroom and then into an air shaft. [3, 31] The second study appearing in the journal looked at the case of a lab worker with no apparent exposure to SARS and no travel history who nevertheless contracted the disease after the outbreak was over in late May 2003. [2, 1] SARS: the True Story - programme summary Severe Acute Respiratory Syndrome didn't even have its name in February 2003 when it struck its first known victim Johnny Cheng in Hanoi Vietnam. [2, 2] Within days an international effort led by the World Health Organization WHO had massed scientific expertise to fight the mystery illness and avert the nightmare scenario of an uncontrollable pandemic sweeping the globe. [2, 3] Amid attempts to quarantine high risk groups of people it seemed only fear could spread more rapidly than the disease itself. [2, 4] Nothing was known about the condition - where it had come from how it was passed on how to spot it contain it or treat it. [2, 6] But if this was a type of influenza it was one that killed up to 15 of its sufferers. [2, 7] The doctor treating Mr Cheng who first contacted the WHO about the unusual symptoms was one of six medics to die of SARS at the hospital. [2, 11] Global Concern The SARS virus first emerged in the Chinese province of Guangdong. [2, 34] Working from samples obtained in Hong Kong a process of random matching showed that SARS was a new form of a coronavirus a pathogen known to affect humans but not a global killer. [2, 42] At a time when strict containment of all known contacts was vital the health ministry there announced the outbreak had actually begun five months earlier and that they had far more than 40 cases they had nearly 400. [1, 1] Four SARS Cases In China BEIJING April 29 2004 Surgical masks above a young girl playing it safe in Hong Kong last year are so far scarce on the streets of China despite the new confirmed cases of SARS. [1, 2] Photo: AP file Tracking the current cases is especially urgent because May 1st is the beginning of a weeklong vacation for millions of Chinese many of whom will be traveling around the country potentially spreading germs over a wide area. [1, 3] AP China confirmed two more SARS cases on Thursday doubling the number of infected people linked to a Beijing laboratory believed responsible for the latest small outbreak of the viral disease. [1, 4] The mother and aunt of an infected nurse were the latest cases according to China's Ministry of Health. [1, 7] In Taiwan meanwhile a 78-year-old man was quarantined and being tested for SARS on Thursday after he returned from the mainland and developed a high fever and other flu-like symptoms. [1, 9] But even as the numbers rose the Chinese government and the World Health Organization emphasized that the SARS cases appeared to be contained to people linked to Beijing's Institute of Virology a national laboratory. [1, 11] It wants to stem both the disease and public panic to prevent a recurrence of events last year when 349 people in China died of SARS after it roared out of the southern province of Guangdong. [1, 14] Millions will be traveling around the country and any SARS carriers could infect people over a wide area. [1, 15] One hugely popular vacation spot is the Huangshan mountain resort in Anhui a southeastern province where investigators were probing SARS links Thursday. [1, 30] The Beijing team of WHO and government investigators wanted to interview a 31-year-old graduate student who worked at the same lab and was suspected to have SARS - but couldn't on Wednesday because he had a fever and was feared contagious said an agency spokesman.

Figure 2*:* Image of Mead summary of multiple documents

At the very top of the summary extract it says *Source Document from Summary* with various columns below it. The first column reads *highlight?* which allows the user to highlight or de-highlight various sections of the summary. It is useful as far as allowing the user to determine which sections of the summary extract refer to which document. The second column entitled *Doc. No.* is simply the summarizer's numbering of the documents which match the first number of the number set in the actual summary extract. The buttons found under the *highlight?* and *file name* columns of the diagram operate with the *highlight?* feature. The user may redraw the documents to be highlighted and then reset everything to its default state so that everything is highlighted. Finally, the third column entitled *File Name* contains numbers used to identify the documents. The number next to the word *text* represents the order in which the source documents were uploaded. The rest of the numbers in that name represent the date the source document was uploaded, starting with the year, month and day. In the above illustration, because the remaining six-digit sequence of numbers is identical for every source document, these numbers could be a code for the identification of the batch of source documents to be summarized.

Each sentence of the summary extract begins with a set of numbers in brackets which are separated by a comma. For example, the first sentence of the sample summary extract above begins with [3,1] . The first number is the number assigned by the summarizer to the document; it refers to the order in which a document is summarized from a set of documents. Note, this is different from the number in file text which tells the user the order in which the documents were uploaded. The number following the comma in [3,1] informs the user of a sentence's ordinal position from the document it comes from. With this information the user can tell whether a sentence is the first, second or third sentence of the source document.

Like Mead, SweSum allows for text to be either uploaded or pasted and it is limited to summarizing html files or just plain text. SweSum, however, only performs single document summarizations (SDS); such limitation may allow SweSum to use more of its resources to making the system more user-focused. There are several parameters which the user may manipulate in SweSum: compression rate; type of text or genre; language; key words; number of key words; and various weights for discourse parameters. The compression rate in SweSum may be controlled by specifying a certain percentage, number of words or even number of characters of the original document to be summarized. The genre or type of text to be summarized has two settings: newspaper and academic. The user may also change the number of keywords which the summarizer uses to extract sentences, and even the weights of the discourse parameters (or features) may be manipulated by the user as well. SweSum can summarize in seven different languages (Swedish, Danish, English, Spanish, French, German and Norwegian).


Text Summarizer (with PRM) - Microsoft Internet Explorer provided by The Graduate Center - CUNY
File Edit View Favorites Tools Help
Address
Go
Browse...
Summarize

Address: http://swesum.nada.kth.se/index-eng-adv.html

SweSum - Automatic Text Summarizer by Martin Hassel and Hercules Dalianis
Localization, Interfaces and Swedish Pronominal Resolution by Martin Hassel

På svenska, tack! På persiska! Lesser options, please![URL]

Please type or paste a text of your own to summarize:

Alternatively, you can upload a text/HTML file from your own computer:

Keywords that may be important for the text. Choose type of text: Newspaper. Choose language of the text: English

Summary of the original text: 30 percent

Print keywords and statistics ☑ Number of keywords: 10

Use pronoun resolution □ (only for Swedish)

Set weights for discourse parametres:

First line	Bold	Numeric values	Keywords	User keywords
1000	10	1.133	0.360	500

Read more about Text Summarization

✉ Comments to Hercules?
✉ Comments to Martin?

Figure 3: Image of SweSum Automatic Text Summarizer

How Mead generates multi-document summary (MDS) extracts. An analysis of the results of multi-document summarization was performed to learn about the summarizer's extraction methods and capability. Three news articles were chosen to produce a multi-document summarization from the science and nature section of BBC news online ("SARS: the True Story"), the health section of ABC news online ("In the Air: SARS"), and the health section of CBS news online ("Four SARS Cases in China"). All of the articles are roughly two pages long and about the same topic and genre. The articles in general give an account of where and how this disease started to spread and why.

The documents summarized were chosen for their similarity of content. They are all news articles from online newspapers or magazines. Of particular interest was to look at which sentences the summarizer extracted and why to determine the level of intelligence at which the summarizer is able to extract and assemble sentences from various documents to create one, coherent summary. That is, did the summarizer manage to extract sentences from the various documents in a complimentary fashion or was there a great deal of overlap and redundancy among the sentences chosen? An analysis of the ordering of the sentences extracted in a multi-document summarization would shed light as to how this MDS extract was created. I expected the ordering of the sentences in a summary extract to be based in part on how well the sentences connect with each other; perhaps, cohesion constraints direct the ordering of sentences in multi-document summarizations.

Essentially, summarization of the documents happen in the order they are received and what it yields are what strongly appear to be three single summaries put together, starting with the first document that was summarized and ending with the last document summarized. To verify this, I did a single document summarization (SDS) with a 10% compression rate of each source document and compared the sentences of those single summaries with what is in the multi-document summarization using the same source documents. I looked to see whether or not the number of sentences from the single-document summaries matched those present in the multi-document summarization. The results show that there is little difference in content among the SDS extracts of the source documents and the summarization of the same documents in a MDS extract. The extracted sentences in multi-document summarization are not inter-weaved in any way. That is, multi-document summarization extracts are not a product of a process by which the Mead summarizer puts different documents together in such a way so as to link and connect sentences in an informative, non-redundant manner. The small difference I found between the MDS extract and the SDS extracts is the absence of two sentences in the MDS extract present in the single-document summary. For example, as shown in Appendix A, sentence [1,16], and [1,6] are missing from the MDS extract. The fact that these sentences are missing from the multi-document summarization may have to do with the compression rate of 10% constraining the length of the multi-document summary extract. The two sentences which were omitted from the MDS extract are likely to be low-ranking sentences. To determine whether or not these omitted sentences appeared in a multi-document summary under a lower compression rate, I summarized the source documents using a compression rate of 15%. The results confirm that sentences 1,16 and 1,6 are low ranking sentences, but more importantly, it shows that the reason why the MDS omitted these sentences is due to compression rate.

Manipulating compression rate to learn about sentence ranking in Mead. One can get a sense of which sentences rank higher than others within a summary extract by summarizing three different summaries, using compression rates of 10 to 5 to 1% percent. In comparing a summary extract with a 10 percent compression rate with one that has a 5 percent compression rate, the absence of sentences from the 5 percent extract is indicative of which sentences the summarizer deems to be more content-filled and informative than others. The absence of sentences [3,2] and [3,9] from the summary extract shows that the presence of sentence [3,1] in both extracts ranks higher than [3,2] and [3,9]. Further, only sentences [2,1], [2,2], [1,1], [1,2], [1,3] and [1,30] remain in the summary extract with a 5% compression rate. Interestingly, the 1% compression rate contains only the first sentences in total from documents one and two, leaving out any sentences from document three, indicating that sentence position may be the highest ranking feature in Mead (See Appendix B for summaries with varying compression rates).

Word Frequency, Keywords and Sentence Ranking in SweSum. Keywords include content words that have a high frequency word count, but not all high-frequency content words appear on the keywords list found at the end of every extract. Besides frequency, the summarizer must use other criteria to decide which words will become the keywords of a document. A word count of 17 may just be too high for a document of only 25 sentences. If part of the main job of the summarizer is to extract sentences that are important and relevant and not too redundant, it must filter out content words that are too frequent. Choosing content words that are frequent and that also meet other criteria may help filter out words that will not point to important sentences. For example, in SweSum content words appearing in the titles of a document and/or words in the beginning of a document which are relatively frequent are likely to become the keywords used to extract sentences. Sentence position, therefore, may also play an important role in determining which content words will be deemed relevant and extract worthy.

To test the hypothesis above, the Georgetown Linguistics Web Frequency Indexer was used to yield a word frequency count of the article *Body Building: Growing replacement organs is still a long way off* (shown in Appendix C1). As can be observed, words like *cells*, *tissue*, *kidney*, *type*, and *blood* are more frequent than or just as frequent as the keywords in the summarized document, yet they fail to be part of the keywords used to extract sentences. The word *cells* occurs 17 times in this document and it occurs in the second sentence of the document, but it does not become a keyword. I include the Georgetown word count list, along with SweSum's summary of the article to illustrate the frequency of the keywords of the summarized article. The most frequent words on the Web Frequency Indexer are at the top and the least frequent at the bottom. Because the list would have been too long to include in this paper, I eliminated from the list all words occurring just once. The highlighted words represent the keywords of the summarized document. The locations of the keywords come from various critical parts of the document. Frequency and sentence location of a word appear to combine to determine what the keywords will be. For example, *building,* with a word count of *three* can be found in the main title of the article; *growing* and *organ* each have a word count of three are located in the subtitle. *Sefton* and *heart* have a word count of three each and *human* has four; they all appear in the first sentence. The word *embryonic* also has a word count of three and it occurs in the second sentence of the source document. The word *living* has a frequency of three. The first letter of *Living* is capitalized as it is part of a title within the document, making it worthy of becoming a keyword. It is not evident how it is that *patient*, occurring only three times, made it to the keywords list since it doesn't seem to occur in any important sentence position.

To test whether or not the position of a sentence has an effect on a frequent content word's likelihood of becoming a keyword, I modified the original article source document by removing the word "building" in the main title and placing it in the last sentence of the document. This test shows that indeed keywords are ranked in SweSum according to the position of the sentences they're in. The results show that moving the word "building" into the last sentence of the document caused it to no longer be a keyword for summarization of this document as shown in the keywords section at the bottom of that summary, allowing for a different word ("patient") to become a keyword. In this test, however, having a different set of keywords did not cause the summarizer to produce a different summary extract (see Appendix D).

It is unclear why SweSum would include in the keywords list words like *simple* or *another*, especially when the keywords parameter weight is as low as .360. In general, most of the keywords present in the list are nouns save for the word *another* and *simple*. Perhaps the presence of these words can be explained if the summarizer's dictionary does not have these words listed as function words, so that in the extraction process such a word is categorized as a content, open class word. An alternative explanation is that *another* is categorized as cue word, indicating that the information following *another* is important, making the sentence in which *another* is in worthy of extraction. Other function words such as *would*, *could* and *simple* show up in the keywords list in extracts where the keywords parameter weight has been increased. A few more verbs do enter the keywords list, but just a few. For example, *admit* appears in the keywords list of an extract that uses 34 keywords, with the weight of the keywords parameter at 2. At the very maximum, SweSum can base its summary extracts on 34 different keywords even though it allows you to change the keyword rate to a number exceeding the 34 limit to sixty or one hundred. That is, changing the keyword number to 500 does not mean that the summarizer will use 500 keywords from which to extract sentences. It is misleading for SweSum's user-interface to allow for this manipulation when really the extent of keywords the summarizer will use is limited to 34.

In order to determine the effects of changing the number of keywords, two different summary extracts were produced with keyword amounts of 34, and 10 (See Appendix E for results of extracts). The compression rate was left at the default setting (30%). When the weight of the keywords parameter is as low as 0.360, altering the keyword feature does nothing to change the summary extract, especially in relation to the other parameters such as *first* *line* parameter (weight, 1000), *bold* parameter (weight, 10), *numeric value* parameter (weight, 1.133) and *user* keywords *parameter* (weight, 500).

Sentence Position and Sentence Ranking in SweSum

Both the first and second sentences of the text are ranked the highest even without any of the keywords. Summarization of documents under SweSum's default setting allows sentence position parameter to have more weight than the keywords parameter. The first line parameter has a weight of 1000 while the keywords parameter has a weight of only .360. However, to determine which sentence positions were most critical, two documents were summarized showing that indeed under the default settings in SweSum, the first two sentences of a document consistently appear in the extract. Moreover, the first sentence of a document will be in the extract even after changing the weight of the first line parameter to 0. These results can be seen in Appendix F.

Surprisingly, the last sentence of the source document (a low ranking sentence) crops up in both summary extracts (the newspaper and academic) even without having any keywords, but the fact that it is a quotation may be what makes it extract worthy. Dalianis (2000) does not mention quotation marks or what makes sentences extract worthy. To test whether or not quotation marks weigh heavily in SweSum, I altered the original document by placing quotation marks in a sentence occurring towards the end of the source document and deleted all keywords from it. Indeed, a sentence with quotations without any keywords, is deemed extract worthy in SweSum in academic mode. To further test this, I altered the original source document by removing quotation marks from that same last sentence; the summary extract did not have such sentence.

Genre and Corpora

Can knowledge of genre improve summarization systems? The more knowledge we have about what genre is composed of, the better can a corpus be compiled for the development of summarization systems. The type of corpus used in training a summarization system can play an important part in explaining why certain sentences are extracted over others. Programs that learn rules in training a summarization system learn by creating algorithms that are based on the distribution and organization of features in a corpus. The distribution and organization of these features vary depending on the genre of the text. If genre were not a factor at all in the design of a summarization system, it would be possible to summarize any type of document, and having such capability would enhance any summarization product. Developers of automatic summarization systems would be able to sell their product to users who may summarize a range of documents (scientific reports, financial reports, newspaper articles etc.) Maximizing the use of their product would draw in more users, and more business.

However, given that many shallow summarizers like Mead and SweSum are designed to be genre-dependent must, therefore, mean that they simply do not have the capacity at this point in technology to be genre independent. (That’s not to say that summarizers attempting to be genre-independent do not exist at some level.) Genre poses limitations on the type of corpora that summarization systems can be trained on, and as a result, limits the type of documents a user can summarize. The position of relevant or topical sentences, for example, is a feature that varies depending on the genre of a document.

Before suggesting two possible solutions to the problems introduced by genre, some definitional information on genre is necessary.

In my search for a definition of genre I learned that, indeed, genre is a very vague notion to define. However, David Lee (2001) defines genre in the following way: "Genre is used when we view the text as a member of a category: a culturally recognized artifact, a grouping of texts according to some conventionally recognized criteria, a grouping according to purposive goals, culturally defined. Here, the point of view is more dynamic and, as used by certain authors, incorporates a critical linguistic (ideological) perspective: Genres are hence subject to change as generic conventions are contested/challenged and revised, perceptibly or imperceptibly, over time (P. 46)." Lee points out, however, that genre labels "…can have many different levels of generality." For example, some genres such as 'academic discourse' are actually very broad, and texts within such a high-level genre category will show considerable internal variation: that is, individual texts within such a genre can differ significantly in their use of language… p.48" Steen's work (1999) (as mentioned by Lee) where he uses prototype theory to help create a taxonomy of different genres might be a solution to the problem of excessive generality within genres. Steen differentiates genres from sub- or super-genres in that actual genres have many different characteristics from one another while sub-genres and super-genres do not. Genres are considered basic level categories that fall in the middle of a hierarchy. For example, literature is a super-genre; a novel is a genre under literature (dramas and poetry are other examples), falling in the middle of the hierarchy, so that western, romance and adventure are sub-genres, prototypes or exemplars of the genre, novel. More importantly, genre according to Steen has a set of seven characteristics or attributes: domain, medium, content, form, function, type, and language. Of interest to me, for the purpose of possibly enhancing automatic text summarization systems, are the attributes of form, type, and language. Form involves the organization and structure of a text; type has to do with whether the text is in the rhetorical category of narration, argumentation, description, or exposition; and the language attribute has to do with linguistic characteristics used in a given genre such as register and style.

Having stated that genre is problematic in automatic text summarization, I'd like to propose two possible approaches to reducing the restrictions imposed by genre.

(1) One such solution would be to search for universal-linguistic markers of a sentence's importance in a document without the need to refer to genre. The objective here is to be able to develop summarizers that are genre-independent.

(2) Another solution would be to investigate the form and language of texts of a given genre for the presence of indicators of critical textual areas of a document useful in automatic text summarization.

On the one hand, (1) searches for a solution that looks for universal linguistic cues that are genre-independent that point to the importance of a sentence for automatic summarization. On the other hand, (2) opts for a solution that utilizes knowledge of the specific structure and language of a given genre that are informative and important and useful in summaries. The summarizers in this case will not be genre-independent.

Muresan et al.'s work on the summarization of emails supports the possibility of solution (1) above. Salient noun phrases representing the summary of an email were extracted using machine learning techniques and linguistic knowledge. Their collection of emails varied in genre, length and topic. 2500 noun phrases were extracted from 51 emails in the training set and 324 noun phrases were extracted from 8 emails in the test set. One judge was used to tag the NPs for saliency. (Their results would have been more interesting had they used more judges). Noun phrases like conference workshop announcement or international conference were the types of noun phrases considered to be salient in emails. Their study found that simple noun phrases were useful in the task of gisting or summarizing emails.

Genre: Academic versus Newspaper in SweSum, is there a difference?

The empirical aspect of genre in this paper looks at how SweSum handles newspaper and Academic summarization modes. The purpose of this test is to determine whether there are observable differences in the summarization of a single document under two different genres and to investigate to some degree how these differences are made to surface.

While testing documents in SweSum using different parameter settings, I discovered that an academic summary can look very much like the summary extract of a document in newspaper mode and whose keywords parameter weight has been increased to 1000 and all other parameter settings reduced to a zero weight (see Appendix G). The default settings under the newspaper mode include the “first line” (1000), “bold” (10), “numeric values” (1.133), “keywords” (0.360) and “user keywords” (500) with “first line” having the heaviest weight out of all of the features. Perhaps, the meaningful feature in academic mode under which to extract sentences is the keywords feature.

Further testing comparing the summary extracts in newspaper and academic modes of the same document (see Appendix H) shows that indeed the academic summary contains more keywords from the keywords list, including the low ranking keywords. The summary extract contains low ranking keywords such as *conference*, *government*, and *study*, along with high ranking keywords such as *passive*, *smoking*, *people*, and *work*. The summary in academic mode contains nine of the ten keywords while the summary in newspaper mode contains only seven, reinforcing the hypothesis that summarization in academic mode is relies on mostly, if not exclusively, on keywords features.

According to one of the developers of SweSum (Dalianus, 2000), sentences in the beginning of a newspaper text score higher than sentences occurring at the end. Sentences with keywords are given higher scores than sentences without keywords. SweSum summarization in academic mode breaks the rule that sentences at the beginning of a source document will become sentences in the summary being that they are ranked highly. Because the second sentence of the source document appears in the summary in academic mode shown in Appendix H, I tested whether the second sentence in a document scores highly instead for summaries generated in academic mode. To do this, the second sentence of the source document was repositioned to the middle of the text and put through the summarizer. The sentence still appears in the summary extract. Sentence position is not ranked highly in academic mode; rather, the presence of keywords is. Testing proves that it is the keywords feature that determines which sentences are extract-worthy in academic mode.

Sentence position is not ranked highly in academic mode; rather, the presence of keywords is. Testing proves that it is the keywords feature with SweSum that determines which sentences are extract-worthy in academic mode in contrast to the newspaper genre that considers sentence position critical for a good summarization of a newspaper text.

Are there Good and Bad Summaries?

In the article, *Summarization Evaluation: an Overview,* Mani explains that there are two methods by which to evaluate summaries: extrinsic and intrinsic. Intrinsic evaluations determine the quality of a summary based on informativeness or coherence. The evaluations are based on measuring the quality of a summary by determining how much agreement there is between a human-generated summary and its machine-generated counterpart. Extrinsic evaluations determine whether a summary is good or not based on certain tasks. For example, subjects may be given reading comprehension tests on machine-generated extracts. Whether a summary is good or not will depend on how well the subjects score on the test. The focus in this paper will be on intrinsic evaluations. Coherence will be discussed briefly only insofar as it sheds light on how an ATS operates internally. More attention will be paid to the informativeness aspect of a summary extract.

In determining whether a summary is good or not with regard to informativeness, one main problem crops up: evaluating a summary extract is not a clear-cut task. The process of extracting important sentences or good sentences for a summary extract lends itself to a high degree of subjectivity and variability. What one person may consider a sentence to be summary worthy because it is “relevant” or “important”, may differ from another’s judgment. Mani points out a study on humans extracting 20-sentences from a number of *Scientific American* articles showing that people can produce very different summaries based on the same source and that even the same person may produce an entirely different summary extract eight weeks later, using the same source he used originally. In addition, evaluations based on comparing a machine-generated summary to a human generated reference are incomplete in nature. A machine-generated summary extract may have as a reference a multitude of human-generated summaries without any of them matching yet still being fully coherent and informative.

A comparison of Mead and SweSum: Which one is better?

Evaluation: Punctuation and Coherence. In comparing the summarization of a two-page news article in the medical domain, containing 998 words, Mead and SweSum produced slightly different summary extracts. The article used for this comparison (*Body Building: Growing replacement organs is still a long way off*) comes from an online popular science magazine (scientificamerican.com). The document in both Mead and SweSum was pasted and not uploaded being that the summarizers are generally unable to filter out extraneous material from web pages such as text relating to an ad from the web page of a given article.

The summary extract serves as a window into the design of a summarization system. To begin, there are certain problems Mead has with summarizing text. How it interprets punctuation is generally the cause for mistakes entering the summary extract. Punctuation like quotation marks are often missing from Mead's summary extracts, as shown below. Parenthesis and quotation marks around the word LIFE are absent in sentence [1, 3]; and, quotation marks and commas are also missing entirely from sentence [1,4].

Original

> Now Sefton admits that the deadline on his Living Implants from Engineering ("LIFE") initiative was naive, and he thinks it will be at least another 10 to 20 years. "We need to be able to walk before we can run," he says, "and the worry today is, Can we make a vascularized piece of tissue or a tissue with two or three cell types in a controlled way?"

Mead

> [1, 3] Now Sefton admits that the deadline on his Living Implants from Engineering LIFE initiative was naive and he thinks it will be at least another 10 to 20 years. [1, 4] We need to be able to walk before we can run he says and the worry today is Can we make a vascularized piece of tissue or a tissue with two or three cell types in a controlled way?

SweSum

> Now Sefton admits that the deadline on his Living Implants from Engineering ("LIFE") initiative was naive, and he thinks it will be at least another 10 to 20 years. "We need to be able to walk before we can run," he says, "and the worry today is, Can we make a vascularized

> piece of tissue or a tissue with two or three cell types in a controlled way?"

It is unclear why Mead would remove important punctuation from a summary extract; punctuation that doesn't appear to pose a problem to the extraction process. Interestingly, however, in the sentence, "The idea of seeding an organ and letting the body do the rest of the construction might work for a kidney, because the patient could be treated with dialysis while the new organ was being generated, according to Jeffrey L. Platt, director of transplantation biology at the Mayo Clinic.", Mead would have benefited from removing the period from the proper name Jeffrey L. Platt, but didn't. Instead, the presence of this period in the name causes the summarizer to extract everything before and including the period of the middle name initial, L.:

"The idea of seeding an organ and letting the body do the rest of the construction might work for a kidney because the patient could be treated with dialysis while the new organ was being generated according to Jeffrey L." (Sentence extract from Mead)

Mead interprets the period after the initial L. as the marker for the end of the sentence, causing an incomplete sentence to be placed in the summary extract.

The absence and presence of certain punctuation could reveal how it is that the summarizer processes documents. A copy of the original document may be temporarily stored somewhere in the summarizer's memory. As it begins to extract the sentences, it does so without commas, quotation marks, or parenthesis, but preserving periods, hyphens and the possessive marker, '. The mechanism used for the identification of sentences for extraction looks for periods, any period unfortunately; hence, permitting errors produced by periods in proper names to enter the summary extract. Deletion of punctuation other than periods, hyphens and the possessive marker from the summary extract seems unnecessary, and pointless for Mead to do. It only achieved to degrade the quality of the summary extract.

As far as punctuation goes, SweSum seems to do a better job at preserving the original punctuation as it appears in the source. Preserving punctuation makes SweSum better in that it allows the extract to represent the sentences as they appear in the source. However, SweSum is not perfect with regard to punctuation. Both Mead and SweSum have difficulties distinguishing between a period after the first initial of a middle name and a period which marks the end of a sentence; as a result; both summarizers extract just fragments of whole sentences. It is apparent from this error that both Mead and SweSum require some sort of an algorithm to capture the rule of, if a period occurs after a middle name initial then such period is not an end of a sentence period. A parser, of course, would be ideal for a summarization system to have, but it is really not required. A parser would make a summarizer less shallow, producing more coherent summaries but it would be more costly for developers.

Separating titles or subtitles from sentences at first appeared to be problematic for the Mead summarizer to do. The summarizer conflated both the subtitle and the sentence that follows it, in sentence [1,9]:

> "…[1, 9] Building a Case Yu and his co-authors conducted an analysis of the outbreak at Amoy Gardens an apartment complex in Hong Kong where more than 300 residents were affected last year when the SARS epidemic hit a number of Asian countries before spreading elsewhere in the world."

Being that title words are for many summarizers a high-ranking feature, sentences with title words are more likely to be extracted than sentences without them. Therefore, if a summarizer mistakenly extracts a sentence only because it processed the title words as part of the extracted sentence, then the quality of the summary extract is likely to degrade. Low ranking sentences are entering the summary extract due to an error on the part of the summarizer. Sentence [1,9] may not have been part of the extract if the summarizer hadn’t considered the actual sentence in [1,9] as being part of the subtitle. To test this hypothesis, the titles and subtitles from the original document were removed as a way to determine whether or not such sentences which were just below the title words would have still been extracted without the subtitle (see Appendices J and K). The result was that Mead included the sentence in [1,9] in the summary extract. Mead did not incorrectly extract sentence [1,9] because of its proximity to the subtitle: Mead knows what is a subtitle and what is a sentence; it just doesn’t separate titles or subtitles from the sentences when organizing and assembling the summary extract.

Even though Mead correctly distinguishes subtitles and titles from the article itself, the way in which it presents or reconstructs the original formatting of the article is problematic. Paragraphs aid in organizing the author’s ideas in an article. A summary extract that does away with all of the original organization may also be misrepresenting what the author intends to communicate. Moreover, Mead not only reorganizes the source document after summarization, it disrupts ease of readability by not separating enough a subtitle from a paragraph, for example. SweSum on the other hand preserves all of the original paragraphing even though the summary is essentially just extracts of a source document.

Evaluation: Informativeness. To determine whether a summary is good or bad, one would need to clearly define its function and purpose. Is the summary designed to meet the needs of the user by being topic-focused or query-focused? Does the user want an indicative type of summary or an informative one? In *Automatic Text Summarization,* indicative summaries are defined as summaries that provide just enough information to indicate to the reader to read the source document or not. Informative summaries have more details in comparison to an indicative one. Also, the output of summaries that are query-based must be evaluated by whether or not the query was answered. If it is topic-focused, the summary must be evaluated as to whether it contains the user's topics.

Both SweSum and Mead have default settings that the user may summarize documents in, but these summarizers also allow for their parameters (compression rate, number of keywords used, and user specified keywords) to be manipulated. User- manipulation of parameters, namely compression rate, makes it possible for summaries to be indicative or informative. The higher the compression rate is, the more indicative the summary; the lower the compression rate is, the more informative the summary.

Mani speaks of evaluating summaries by looking at their information content and measuring the semantic informativeness of a summary in information theoretic terms. What this entails is being able to predict what the source document is about or "reconstructing the source document" from just the summary.   A semantically informative summary should allow for the salient ideas or propositions of the source to be reconstructed.

Evaluation of summaries which were summarized under Mead and SweSum's default settings will be based on how informative the summaries are.   That is, can I get a strong sense from the summaries what the salient idea(s) of the source document are? A summary extract with sentences carrying too many details, preventing the user from attaining a gestalt view of the article will be graded as being a poor summary. Both SweSum and Mead in their default settings produce summaries intended to substitute news articles. The summary extracts in Appendix L belong to SweSum and Mead respectively.

On the one hand, SweSum offers a summary extract with few sentences containing scientific terminology while at the same time introducing the topic of the document, expressing the complexity of this procedure and its limitations to how this tissue engineering can be applied. On the other hand, Mead's extract, which has only three sentences that differ from that of SweSum's extract, has more technical language also helps convey the function of the article: to explain to some degree the complexity and process of tissue engineering and its applications. Notice, for example, that the language in the sentences below is more technical with words like *neural liver* and *cartilage cells, formation of a 3D vessel-like network*, or *biodegradable polymer scaffold*. "[1,6] By mimicking the natural 3-D shape in which an organ grows tissue engineers are trying to get adjacent cells to talk to one another and complete the task of building the desired tissues. [1, 14] Last fall for example researchers from the Massachusetts Institute of Technology and the Technion-Israel Institute of Technology reported generating tissues of neural liver and cartilage cells as well as formation of a 3D vessel-like network on a biodegradable polymer scaffold seeded with human embryonic stem cells."

In evaluating other summaries by Mead and SweSum for informativeness, determining which summary is better was not such a clear-cut task. Even though evaluating for informativeness is separate from evaluating for coherence according to Mani, it seems that SweSum contains sentences with coherence problems that interfere with the level of informativeness that the reader can attain. Moreover, the sentences extracted by SweSum on average are shorter than the one's by Mead. Length of sentences may be contributing in making Mead's summaries more informative and consequently better (see Appendix M).

Conclusion

The following chart illustrates graphically how these two systems compare. To sum up, Mead has problems with punctuation, not just with mis-processing periods, resulting in the extraction of fragments, but with not preserving enough of the original punctuation such as quotation marks and parentheses from the source document. Mead does not format the summary in an easy to read fashion. For example, the subtitles and titles or date of an article all appear to be part of a sentence. Mead proved to have its strength in the degree of informativeness that its summaries have in relation to its source document. SweSum seems to perform better in those areas that Mead does not. Punctuation and format are maintained in SweSum, but SweSum also is extremely user-focused. Many parameters in SweSum can be manipulated, from compression rate, to keywords, to genre, to the manipulation of the weight of the bold-face-words. However, evaluation of its summary extract based on informativeness shows that Mead may provide the user with higher quality summaries, but again, making an assessment on informativeness is a very vague notion. One concrete observation made about Mead and the sentences it extracts is that compared to SweSum, its sentences are longer and its sentences are more coherent as a whole whereas with SweSum, many sentences begin with pronouns rather than proper names, causing coherence problems to surface in its summaries. However, with that said, I can’t emphasize enough that evaluation of a summary is very subjective and differs from person to person.

The graph below uses a point system to indicate which summarizer performs best in a given area. It starts at 0 and goes up to 3 points, three being the most points the summarizer has earned given an area. Mead has three points in informativeness, one point each in punctuation and format, and two points in how user-focused it is compared to SweSum. SweSum earned three points each in format and how user-focused it is compared to Mead. SweSum only earns two points each in punctuation and informativeness.

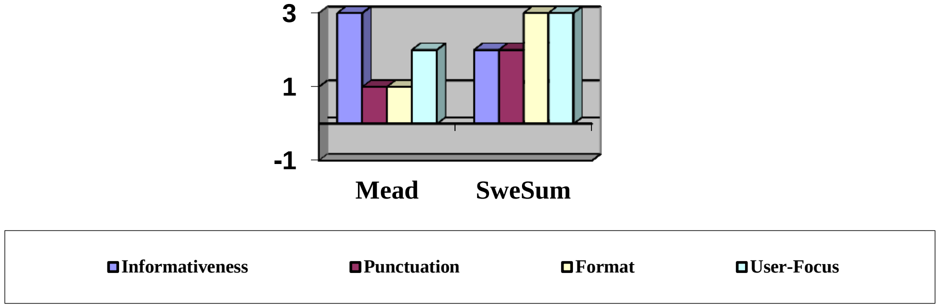


Figure 4: Mead & SweSum Performance Chart

Future Work

Exploring whether it is feasible to rely on solely linguistics and machine learning techniques in developing genre-independent summarizers is of particular interest to me. Do universal linguistic patterns, informative of a sentence’s saliency for summarization purposes exist? Muresan et al. found that certain noun phrases can do just that; however, are there other linguistic features that may also inform us of a sentence’s extract-worthiness?

Appendix A
MDS extract of all three documents

[3, 1] ABCNEWS.com : Flushing Out the Spread of SARS Search the Web and ABCNEWS.com Print This Story Email This Story See Most Sent May 19 2004 HOMEPAGE NEWS SUMMARY US INTERNATIONAL MONEYScope WEATHER LOCAL NEWS ENTERTAINMENT ESPN SPORTS SCI/TECH POLITICS HEALTH TRAVEL VIDEO AUDIO FEATURED SERVICES RELATIONSHIPS NEW! [3, 2] INSURANCE SHOPPING WIRELESS E-MAIL CENTER BOARDS FREE HEADLINE FEED Two Chinese women wear masks before entering Ditan Hospital in Beijing where a suspected SARS patient is being treated. [3, 3] Guang Niu/Reuters In the Air: SARS By Amanda Gardner HealthDay Reporter From HealthDayNews April 26 A new study suggests that severe acute respiratory syndrome SARS may have been spread through the simple act of flushing a toilet instead of being passed directly from person to person. [3, 4] Even with all the research that has been conducted on SARS in the past year some mystery still remains as to how the virus can be transmitted. [3, 11] Building a Case Yu and his co-authors conducted an analysis of the outbreak at Amoy Gardens an apartment complex in Hong Kong where more than 300 residents were affected last year when the SARS epidemic hit a number of Asian countries before spreading elsewhere in the world. [2, 2] You are here: BBC Science Nature TV Radiofollow-up Horizon BBC Two Thursday 29 May 2003 9pm SARS: the True Story Classic Horizon Programme summary Transcript BBC Health BBC News Online See the award-winning Horizon Life Story at the National Film Theatre in London. [2, 4] SARS: the True Story - programme summary Severe Acute Respiratory Syndrome didn't even have its name in February 2003 when it struck its first known victim Johnny Cheng in Hanoi Vietnam. [2, 5] Within days an international effort led by the World Health Organization WHO had

massed scientific expertise to fight the mystery illness and avert the nightmare scenario of an uncontrollable pandemic sweeping the globe. [2, 7] Nothing was known about the condition - where it had come from how it was passed on how to spot it contain it or treat it. [2, 23] We were dealing with something unknown... incredibly scary Julie Hall World Health Organization Unprecedented action Two weeks since the Vietnam case people were falling victim across Asia - and then on 13 March Toronto Canada went on SARS alert after a suspected fatality. [1, 1] CBS News Four SARS Cases In China April 29 200406:42:44 Home U.S. [1, 2] AP China confirmed two more SARS cases on Thursday doubling the number of infected people linked to a Beijing laboratory believed responsible for the latest small outbreak of the viral disease. [1, 6] In Taiwan meanwhile a 78-year-old man was quarantined and being tested for SARS on Thursday after he returned from the mainland and developed a high fever and other flu-like symptoms. [1, 8] But even as the numbers rose the Chinese government and the World Health Organization emphasized that the SARS cases appeared to be contained to people linked to Beijing's Institute of Virology a national laboratory. [1, 10] It wants to stem both the disease and public panic to prevent a recurrence of events last year when 349 people in China died of SARS after it roared out of the southern province of Guangdong. [1, 29] The Beijing team of WHO and government investigators wanted to interview a 31-year-old graduate student who worked at the same lab and was suspected to have SARS - but couldn't on Wednesday because he had a fever and was feared contagious said an agency spokesman.

SDS of Document 3

[1, 1] ABCNEWS.com : Flushing Out the Spread of SARS Search the Web and ABCNEWS.com Print This Story Email This Story See Most Sent May 19 2004 HOMEPAGE NEWS SUMMARY US INTERNATIONAL MONEYScope WEATHER LOCAL NEWS ENTERTAINMENT ESPN SPORTS SCI/TECH POLITICS HEALTH TRAVEL VIDEO AUDIO FEATURED SERVICES RELATIONSHIPS NEW! [1, 2] INSURANCE SHOPPING WIRELESS E-MAIL CENTER BOARDS FREE HEADLINE FEED Two Chinese women wear masks before entering Ditan Hospital in Beijing where a suspected SARS patient is being treated. [1, 3] Guang Niu/Reuters In the Air: SARS By Amanda Gardner HealthDay Reporter From HealthDayNews April 26 A new study suggests that severe acute respiratory syndrome SARS may have been spread through the simple act of flushing a toilet instead of being passed directly from person to person. [1, 11] Building a Case Yu and his co-authors conducted an analysis of the outbreak at Amoy Gardens an apartment complex in Hong Kong where more than 300 residents were affected last year when the SARS epidemic hit a number of Asian countries before spreading elsewhere in the world. [1, 16] WHO investigators had already noted that many traps in bathroom floor drains were dried out meaning an exhaust fan could have drawn droplets or aerosols from the drainage pipe into the bathroom and then into an air shaft.

SDS of Document 2

[1, 2] You are here: BBC Science Nature TV Radiofollow-up Horizon BBC Two Thursday 29 May 2003 9pm SARS: the True Story Classic Horizon Programme summary Transcript BBC Health BBC News Online See the award-winning Horizon Life Story at the National Film Theatre in London. [1, 4] SARS: the True Story - programme summary Severe Acute Respiratory Syndrome didn't even have its name in February 2003 when it struck its

first known victim Johnny Cheng in Hanoi Vietnam. [1, 5] Within days an international effort led by the World Health Organization WHO had massed scientific expertise to fight the mystery illness and avert the nightmare scenario of an uncontrollable pandemic sweeping the globe. [1, 6] Unless you know what you're looking for you've no tools to find it Mike Leahy virologist Amid attempts to quarantine high risk groups of people it seemed only fear could spread more rapidly than the disease itself. [1, 7] Nothing was known about the condition - where it had come from how it was passed on how to spot it contain it or treat it. [1, 23] We were dealing with something unknown... incredibly scary Julie Hall World Health Organization Unprecedented action Two weeks since the Vietnam case people were falling victim across Asia - and then on 13 March Toronto Canada went on SARS alert after a suspected fatality.

SDS of Document 1

[1, 1] CBS News Four SARS Cases In China April 29 200406:42:44 Home U.S. [1, 2] AP China confirmed two more SARS cases on Thursday doubling the number of infected people linked to a Beijing laboratory believed responsible for the latest small outbreak of the viral disease. [1, 6] In Taiwan meanwhile a 78-year-old man was quarantined and being tested for SARS on Thursday after he returned from the mainland and developed a high fever and other flu-like symptoms. [1, 8] But even as the numbers rose the Chinese government and the World Health Organization emphasized that the SARS cases appeared to be contained to people linked to Beijing's Institute of Virology a national laboratory. [1, 10] It wants to stem both the disease and public panic to prevent a recurrence of events last year when 349 people in China died of SARS after it roared out of the southern province of Guangdong. [1, 29] The Beijing team of WHO and government investigators wanted to interview a

31-year-old graduate student who worked at the same lab and was suspected to have SARS - but couldn't on Wednesday because he had a fever and was feared contagious said an agency spokesman.

Appendix B

Mead 10 Percent Compression Rate

[3, 1] In the Air: SARS By Amanda Gardner HealthDay Reporter From HealthDayNews April 26 A new study suggests that severe acute respiratory syndrome SARS may have been spread through the simple act of flushing a toilet instead of being passed directly from person to person. [3, 2] Even with all the research that has been conducted on SARS in the past year some mystery still remains as to how the virus can be transmitted. [3, 9] Building a Case Yu and his co-authors conducted an analysis of the outbreak at Amoy Gardens an apartment complex in Hong Kong where more than 300 residents were affected last year when the SARS epidemic hit a number of Asian countries before spreading elsewhere in the world. [2, 1] SARS: the True Story - programme summary Severe Acute Respiratory Syndrome didn't even have its name in February 2003 when it struck its first known victim Johnny Cheng in Hanoi Vietnam. [2, 2] Within days an international effort led by the World Health Organization WHO had massed scientific expertise to fight the mystery illness and avert the nightmare scenario of an uncontrollable pandemic sweeping the globe. [2, 3] Amid attempts to quarantine high risk groups of people it seemed only fear could spread more rapidly than the disease itself. [2, 4] Nothing was known about the condition - where it had come from how it was passed on how to spot it contain it or treat it. [2, 7] The doctor treating Mr Cheng who first contacted the WHO about the unusual symptoms was one of six medics to die of SARS at the hospital. [1, 1] Four SARS Cases In China BEIJING April 29 2004 Surgical masks above a young girl playing it safe

in Hong Kong last year are so far scarce on the streets of China despite the new confirmed cases of SARS. [1, 2] Photo: AP file Tracking the current cases is especially urgent because May 1st is the beginning of a weeklong vacation for millions of Chinese many of whom will be traveling around the country potentially spreading germs over a wide area. [1, 3] AP China confirmed two more SARS cases on Thursday doubling the number of infected people linked to a Beijing laboratory believed responsible for the latest small outbreak of the viral disease. [1, 9] But even as the numbers rose the Chinese government and the World Health Organization emphasized that the SARS cases appeared to be contained to people linked to Beijing's Institute of Virology a national laboratory. [1, 11] It wants to stem both the disease and public panic to prevent a recurrence of events last year when 349 people in China died of SARS after it roared out of the southern province of Guangdong. [1, 30] The Beijing team of WHO and government investigators wanted to interview a 31-year-old graduate student who worked at the same lab and was suspected to have SARS - but couldn't on Wednesday because he had a fever and was feared contagious said an agency spokesman.

Mead 5 Percent Compression Rate

[3, 1] In the Air: SARS By Amanda Gardner HealthDay Reporter From HealthDayNews April 26 A new study suggests that severe acute respiratory syndrome SARS may have been spread through the simple act of flushing a toilet instead of being passed directly from person to person. [2, 1] SARS: the True Story - programme summary Severe Acute Respiratory Syndrome didn't even have its name in February 2003 when it struck its first known victim Johnny Cheng in Hanoi Vietnam. [2, 2] Within days an international effort led by the World Health

Organization WHO had massed scientific expertise to fight the mystery illness and avert the nightmare scenario of an uncontrollable pandemic sweeping the globe. [1, 1] Four SARS Cases In China BEIJING April 29 2004 Surgical masks above a young girl playing it safe in Hong Kong last year are so far scarce on the streets of China despite the new confirmed cases of SARS. [1, 2] Photo: AP file Tracking the current cases is especially urgent because May 1st is the beginning of a weeklong vacation for millions of Chinese many of whom will be traveling around the country potentially spreading germs over a wide area. [1, 3] AP China confirmed two more SARS cases on Thursday doubling the number of infected people linked to a Beijing laboratory believed responsible for the latest small outbreak of the viral disease. [1, 30] The Beijing team of WHO and government investigators wanted to interview a 31-year-old graduate student who worked at the same lab and was suspected to have SARS - but couldn't on Wednesday because he had a fever and was feared contagious said an agency spokesman.

Mead 1% Compression Rate

[2, 1] In the Air: SARS By Amanda Gardner HealthDay Reporter From HealthDayNews April 26 A new study suggests that severe acute respiratory syndrome SARS may have been spread through the simple act of flushing a toilet instead of being passed directly from person to person. [1, 1] Four SARS Cases In China BEIJING April 29 2004 Surgical masks above a young girl playing it safe in Hong Kong last year are so far scarce on the streets of China despite the new confirmed cases of SARS.

Appendix C1

Word frequency count results of the article, *Body Building: Growing replacement organs is still a long way off, using the* Georgetown Linguistics *Web Frequency Indexer.*

```
Word count:     718
Unique words:   381
Sort order:     descending
```

34	THE	4	ALL	3	IT	2	HIS	2	STILL
28	A	4	AS	3	JUST	2	INSTITUTE	2	TECHNOLOGY
28	OF	4	BEEN	3	LIVING	2	IS	2	THEY
19	AND	4	HAVE	3	ORGANS	2	KIND	2	TRANSPLANTED
19	TO	4	HUMAN	3	PATIENT	2	LARGER	2	UNIVERSITY
12	**CELLS**	4	**KIDNEY**	3	SAYS	2	LAYER	2	UP
12	WITH	4	ON	3	SEFTON	2	LIVER	2	USING
11	IN	4	**TYPES**	3	SIMPLE	2	MIGHT	2	VESSELS
7	BE	4	WE	3	SOME	2	MULTIPLE	2	WAKE
7	FOR	3	10	3	**YEARS**	2	NATURAL	2	WAS
7	INTO	3	AN	2	3-D	2	NATURE'S	2	WAY
7	ORGAN	3	ANOTHER	2	ALREADY	2	NEED	2	WELL
6	OR	3	ARE	2	AT	2	NETWORK	2	WILL
6	**TISSUE**	3	**BLOOD**	2	COMPLEX	2	NEW	2	WORK
5	**BODY**	3	BUILDING	2	COULD	2	NOW	2	WORKING
5	**CELL**	3	BY	2	COW	2	ONE	2	WOULD
5	FROM	3	CAN	2	ENGINEERING	2	PLATT	2	YIELDED
5	MORE	3	EMBRYONIC	2	FOREST	2	PRINT		
5	**STEM**	3	GROWING	2	FORMATION	2	PUT		
5	THAT	3	HAS	2	GROW	2	SEEDING		
5	**TISSUES**	3	HEART	2	HE	2	STARTING		

Table 1

Appendix D

Full article (*Body Building: Growing replacement organs is still a long way off*)of the SweSum summary above with key words highlighted:

May 10, 2004
Body **Building**
Growing replacement organs is still a long way off

By Christine Soares

Six years ago Michael Sefton of the University of Toronto challenged his colleagues in the fledgling field of tissue engineering to build a functioning human heart within 10 years. With the isolation of human embryonic stem cells later that year, Sefton's challenge seemed all the more relevant: stem cells, after all, are nature's starting point for building working organs.

Now Sefton admits that the deadline on his Living Implants from Engineering ("LIFE") initiative was naive, and he thinks it will be at least another 10 to 20 years. "We need to be able to walk before we can run," he says, "and the worry today is, Can we make a vascularized piece of tissue or a tissue with two or three cell types in a controlled way?"

Thin sheets of skin and single blood vessels have been grown in the laboratory, and some versions have already been put through human clinical trials. Yet any whole organ would be a complex three-dimensional edifice comprising specialized cells, nerves and muscle, all interwoven with a dense web of veins and capillaries diffusing oxygen and nutrients. The main hurdles have been just getting multiple cell types to grow and work in harmony and spurring formation of the blood vessels required to nourish tissues more than a few hundredths of a millimeter thick.

By mimicking the natural 3-D shape in which an organ grows, tissue engineers are trying to get adjacent cells to "talk" to one another and complete the task of building the desired tissues. This approach has yielded "ink-jet"-dispensed dollops of cell aggregates "printed" in simple patterns that flow together, linking up into larger pieces of tissue. The next step will be to "print" designs using multiple cell types and eventually to print them layer on layer to create larger structures. A similar technique suspends living cells in a clear hydrogel matrix that can be layered or molded into 3-D shapes. Neither tactic has yielded the all-important vascular network needed to sustain thicker tissues.

More progress has been made by seeding stem cells onto a variety of simple scaffolds impregnated with growth-promoting chemicals. Last fall, for example, researchers from the Massachusetts Institute of Technology and the Technion-Israel Institute of Technology reported generating tissues of neural, liver and cartilage cells, as well as formation of a "3D vessel-like network" on a biodegradable polymer scaffold seeded with human embryonic stem cells. When transplanted into a mouse, the constructs remained intact and appeared to connect with the animal's blood supply.

Still, scientists working with stem cells, embryonic or otherwise, admit that they are just beginning to learn tricks for controlling the kind of tissue the cells become and just starting to discern the cues cells give to one another as well as take from their natural environment during the course of organ development. "We don't have anything like [nature's] exquisite repertoire of tools," Sefton says. And so most models for growing entire organs involve using some kind of living "bioreactor." In some cases, it could be the same patient in need of the organ. Anthony Atala of Wake Forest University, who once grew a simple bladder in a beaker and transplanted it into a dog, teamed up more recently with Robert P. Lanza, also now with Wake Forest, and others to grow a mini kidney inside a cow. Kidney progenitor cells were taken from a fetal clone of the cow in question, then implanted into the cow's body, where they developed into proto-organs with all the cell types of a normal kidney. These "renal units" even produced a urinelike liquid.

The idea of seeding an organ and letting the body do the rest of the construction might work for a kidney, because the patient could be treated with dialysis while the new organ was being generated, according to Jeffrey L. Platt, director of transplantation biology at the Mayo Clinic. For a patient suffering from lung or heart failure, however, growing a new organ would put too much strain on an already weak body. But every advance toward creating ever more complex tissues might yield a lifesaving patch for a moderately damaged heart or liver, Platt says, along with fresh insight into how nature builds bigger body parts.

SDS by SweSum of the article, *Body Building: Growing replacement organs is still a long way off*

May 10, 2004
Body Building
Growing replacement organs is still a long way off
By Christine Soares

Six years ago Michael Sefton of the University of Toronto challenged his colleagues in the fledgling field of tissue engineering to build a functioning human heart within 10 years. With the isolation of human embryonic stem cells later that year, Sefton's challenge seemed all the more relevant: stem cells, after all, are nature's starting point for building working organs.

Now Sefton admits that the deadline on his Living Implants from Engineering ("LIFE") initiative was naive, and he thinks it will be at least another 10 to 20 years. Yet any whole organ would be a complex three-dimensional edifice comprising specialized cells, nerves and muscle, all interwoven with a dense web of veins and capillaries diffusing oxygen and nutrients.

The idea of seeding an organ and letting the body do the rest of the construction might work for a kidney, because the patient could be treated with dialysis while the new organ was being generated, according to Jeffrey L. For a patient suffering from lung or heart failure, however, growing a new organ would put too much strain on an already weak body.

Lexicon: English
Words before 743
Words after 211
Summary length: 28%
Type of text: newspapertext

Keywords: *organ human heart simple building embryonic another sefton living growing*

SDS by SweSum of the article, *Body Building: Growing replacement organs is still a long way off* with the removal of the keyword "building"

May 10, 2004
Body Growing replacement organs is still a long way off
By Christine Soares

Six years ago Michael Sefton of the University of Toronto challenged his colleagues in the fledgling field of tissue engineering to build a functioning human heart within 10 years. With the isolation of human embryonic stem cells later that year, Sefton's challenge seemed all the more relevant: stem cells, after all, are nature's starting point for building working organs.

Now Sefton admits that the deadline on his Living Implants from Engineering ("LIFE") initiative was naive, and he thinks it will be at least another 10 to 20 years. Yet any whole organ would be a complex three-dimensional edifice comprising specialized cells, nerves and muscle, all interwoven with a dense web of veins and capillaries diffusing oxygen and nutrients.

The idea of seeding an organ and letting the body do the rest of the construction might work for a kidney, because the patient could be treated with dialysis while the new organ was being generated, according to Jeffrey L. For a patient suffering from lung or heart failure, however, growing a new organ would put too much strain on an already weak body.

Lexicon: English
Words before 743
Words after 210
Summary length: 28%
Type of text: Newspaper text
Keywords: *organ human simple patient embryonic another heart sefton living growing*

Appendix E

SDS by SweSum of the article, *Body Building: Growing replacement organs is still a long way off with* Keywords 34

May 10, 2004
Body Building
Growing replacement organs is still a long way off

By Christine Soares

Six years ago Michael Sefton of the University of Toronto challenged his colleagues in the fledgling field of tissue engineering to build a functioning human heart within 10 years. With the isolation of human embryonic stem cells later that year, Sefton's challenge seemed all the more relevant: stem cells, after all, are nature's starting point for building working organs.

Now Sefton admits that the deadline on his Living Implants from Engineering ("LIFE") initiative was naive, and he thinks it will be at least another 10 to 20 years. Yet any whole organ would be a complex three-dimensional edifice comprising specialized cells, nerves and muscle, all interwoven with a dense web of veins and capillaries diffusing oxygen and nutrients.

The idea of seeding an organ and letting the body do the rest of the construction might work for a kidney, because the patient could be treated with dialysis while the new organ was being generated, according to Jeffrey L. For a patient suffering from lung or heart failure, however, growing a new organ would put too much strain on an already weak body.

Lexicon: English
Words before 773
Words after 212
Summary length: 27%
Type of text: newspapertext
Keywords: organ human embryonic growing patient living building sefton another heart simple would piece kidney admit using formation layer whole engineering technology might institute transplanted build liver starting university forest natural complex working multiple could

SDS by SweSum of the article, *Body Building: Growing replacement organs is still a long way off with* Keywords 10

May 10, 2004
Body Building
Growing replacement organs is still a long way off

By Christine Soares

Six years ago Michael Sefton of the University of Toronto challenged his colleagues in the fledgling field of tissue engineering to build a functioning human heart within 10 years. With the isolation of human embryonic stem cells later that year, Sefton's challenge seemed all the more relevant: stem cells, after all, are nature's starting point for building working organs.

Now Sefton admits that the deadline on his Living Implants from Engineering ("LIFE") initiative was naive, and he thinks it will be at least another 10 to 20 years. Yet any whole organ would be a complex three-dimensional edifice comprising specialized cells, nerves and muscle, all interwoven with a dense web of veins and capillaries diffusing oxygen and nutrients.

The idea of seeding an organ and letting the body do the rest of the construction might work for a kidney, because the patient could be treated with dialysis while the new organ was being generated, according to Jeffrey L. For a patient suffering from lung or heart failure, however, growing a new organ would put too much strain on an already weak body.

Lexicon: English
Words before 773
Words after 212
Summary length: 27%
Type of text: newspapertext
Keywords: organ human embryonic growing patient living building sefton another heart

Appendix F

SDS by SweSum of the article, *Body Building: Growing replacement organs is still a long way off with* Keywords Parameter weight at 1000

May 10, 2004
Body Building
Growing replacement organs is still a long way off
By Christine Soares

Six years ago Michael Sefton of the University of Toronto challenged his colleagues in the fledgling field of tissue engineering to build a functioning human heart within 10 years. Yet any whole organ would be a complex three-dimensional edifice comprising specialized cells, nerves and muscle, all interwoven with a dense web of veins and capillaries diffusing oxygen and nutrients.

And so most models for growing entire organs involve using some kind of living "bioreactor." In some cases, it could be the same patient in need of the organ.

The idea of seeding an organ and letting the body do the rest of the construction might work for a kidney, because the patient could be treated with dialysis while the new organ was being generated, according to Jeffrey L. For a patient suffering from lung or heart failure, however, growing a new organ would put too much strain on an already weak body. But every advance toward creating ever more complex tissues might yield a lifesaving patch for a moderately damaged heart or liver, Platt says, along with fresh insight into how nature builds bigger body parts.

Lexicon: English
Words before 773
Words after 218
Summary length: 28%
Type of text: newspapertext
Keywords: *organ human embryonic growing patient living building sefton another heart*

1st line 0, Bold 0, Numeric value 0, Keywords .360

May 10, 2004
Body Building
Growing replacement organs is still a long way off
By Christine Soares

years ago of the University of challenged his colleagues in the fledgling field of tissue engineering to a functioning within . Yet any whole organ would be a complex three-dimensional edifice comprising specialized cells, nerves and muscle, all interwoven with a dense web of veins and capillaries diffusing oxygen and nutrients. And so most models for growing entire organs involve using some kind of living "bioreactor." In some cases, it could be the same patient in need of the organ.

The idea of seeding an organ and letting the body do the rest of the construction might work for a kidney, because the patient could be treated with dialysis while the new organ was being generated, according to Jeffrey L. For a patient suffering from lung or heart failure, however, growing a new organ would put too much strain on an already weak body. But every advance toward creating ever more complex tissues might yield a lifesaving patch for a moderately damaged heart or liver, Platt says, along with fresh insight into how nature builds bigger body parts.

Lexicon: English
Words before 735
Words after 210
Summary length: 28%
Type of text: newspapertext
Keywords: *organ growing simple another patient kidney starting formation forest complex*

1st line 1000, Bold 0, Numeric value 0, Keywords .360

May 10, 2004
Body Building
Growing replacement organs is still a long way off
By Christine Soares

Years ago of the University of challenged his colleagues in the fledgling field of tissue engineering to a functioning within . With the isolation of stem cells later that year, challenge seemed all the more relevant: stem cells, after all, are nature's starting point for working .

Now admits that the deadline on his Implants from Engineering initiative was naive, and he thinks it will be at least another 10 to 20 years. Yet any whole organ would be a complex three-dimensional edifice comprising specialized cells, nerves and muscle, all interwoven with a dense web of veins and capillaries diffusing oxygen and nutrients.

The idea of seeding an organ and letting the body do the rest of the construction might work for a kidney, because the patient could be treated with dialysis while the new organ was being generated, according to Jeffrey L. For a patient suffering from lung or heart failure, however, growing a new organ would put too much strain on an already weak body.

Lexicon: English
Words before 734
Words after 197
Summary length: 26%
Type of text: newspapertext
Keywords: *organ growing simple another patient kidney starting formation forest complex*

Appendix G

The above summaries are identical under different parameter settings

Newspaper
First Line 0, Keywords 1000, bold 0, numeric values 0

May 10, 2004
Body Building
Growing replacement organs is still a long way off
By Christine Soares

Six years ago Michael Sefton of the University of Toronto challenged his colleagues in the fledgling field of tissue engineering to build a functioning human heart within 10 years. Yet any whole organ would be a complex three-dimensional edifice comprising specialized cells, nerves and muscle, all interwoven with a dense web of veins and capillaries diffusing oxygen and nutrients.

And so most models for growing entire organs involve using some kind of living "bioreactor." In some cases, it could be the same patient in need of the organ.

The idea of seeding an organ and letting the body do the rest of the construction might work for a kidney, because the patient could be treated with dialysis while the new organ was being generated, according to Jeffrey L. For a patient suffering from lung or heart failure, however, growing a new organ would put too much strain on an already weak body. But every

Academic
Default Settings

May 10, 2004
Body Building
Growing replacement organs is still a long way off
By Christine Soares

Six years ago Michael Sefton of the University of Toronto challenged his colleagues in the fledgling field of tissue engineering to build a functioning human heart within 10 years. Yet any whole organ would be a complex three-dimensional edifice comprising specialized cells, nerves and muscle, all interwoven with a dense web of veins and capillaries diffusing oxygen and nutrients.

And so most models for growing entire organs involve using some kind of living "bioreactor." In some cases, it could be the same patient in need of the organ.

The idea of seeding an organ and letting the body do the rest of the construction might work for a kidney, because the patient could be treated with dialysis while the new organ was being generated, according to Jeffrey L. For a patient suffering from lung or heart failure, however, growing a new organ would put too much strain on an already weak body. But every

advance toward creating ever more complex tissues might yield a lifesaving patch for a moderately damaged heart or liver, Platt says, along with fresh insight into how nature builds bigger body parts.

Lexicon: English
Words before 749
Words after 218
Summary length: 29%
Type of text: newspapertext
Keywords: *organ human heart simple building embryonic another sefton living growing*

advance toward creating ever more complex tissues might yield a lifesaving patch for a moderately damaged heart or liver, Platt says, along with fresh insight into how nature builds bigger body parts.

Lexicon: English
Words before 749
Words after 218
Summary length: 29%
Type of text: rapporttext
Keywords: *organ human heart simple building embryonic another sefton living growing*

Appendix H

Highlighted sections represent the sentences both summaries have in common. Keywords are in bold face.

Newspaper

Passive **smoking** 'killing **workers**' Passive **smoking** at work kills three **people** every day, according to research.

The **study** found that around 900 office **workers**, 165 bar **workers** and 145 manufacturing **workers** die each year as a direct result of breathing in other **people's** tobacco **smoke** at work.

It also found that there are three times as many **deaths** a year from passive **smoking** at work as there are from **workplace** injuries.

This would clarify how existing health and safety law applies to passive **smoking**, effectively banning **smoking** from the vast majority of **workplaces**.

"Relying on weak voluntary arrangements will simply not have the desired effect."

Lexicon: English
Words before 428
Words after 100

Academic

Passive **smoking** 'killing **workers**' Passive **smoking** at work kills three **people** every day, according to research.

It also found that there are three times as many **deaths** a year from passive **smoking** at work as there are from **workplace** injuries.

Study was carried out by James Repace, who has previously conducted research into passive **smoking** for the California Department of **Health.**

The **conference** will call on the government to implement a legally binding Code of Practice for **workplace smoking,** proposed over two years ago by the **Health** and Safety Commission.

This would clarify how existing **health** and safety law applies to passive **smoking**, effectively banning **smoking** from the vast majority of **workplaces**.

"Relying on weak voluntary arrangements will simply not have the desired effect."

Lexicon: English

Summary length: 23%
Type of text: newspapertext
Keywords: *smoking health workplace people worker government study death smoke conference*

Words before 428
Words after 121
Summary length: 28%
Type of text: rapporttext
Keywords: *smoking health workplace people worker government study death smoke conference*

Appendix I

SweSum’s summary extract based on a modified source document with missing quotation marks

Passive smoking 'killing workers'

Passive smoking at work kills three people every day, according to research. The study found that around 900 office workers, 165 bar workers and 145 manufacturing workers die each year as a direct result of breathing in other people's tobacco smoke at work.

It also found that there are three times as many deaths a year from passive smoking at work as there are from workplace injuries.

The conference will call on the government to implement a legally binding Code of Practice for workplace smoking, proposed over two years ago by the Health and Safety Commission.

This would clarify how existing health and safety law applies to passive smoking, effectively banning smoking from the vast majority of workplaces.

Appendix J

Mead Summary Extract of document with titles and subtitles

[1, 1] In the Air: SARS By Amanda Gardner HealthDay Reporter From HealthDayNews April 26 A new study suggests that severe acute respiratory syndrome SARS may have been spread through the simple

act of flushing a toilet instead of being passed directly from person to person. [1, 2] Even with all the research that has been conducted on SARS in the past year some mystery still remains as to how the virus can be transmitted. [1, 3] Two articles appearing in the April 22 issue of the New England Journal of Medicine explore the possibility of airborne and laboratory transmissions. [1, 7] Future prevention and protection against SARS should take into consideration the possibility that airborne transmission avoidance of close contacts alone may not be adequate. [1, 9] Building a Case Yu and his co-authors conducted an analysis of the outbreak at Amoy Gardens an apartment complex in Hong Kong where more than 300 residents were affected last year when the SARS epidemic hit a number of Asian countries before spreading elsewhere in the world. [1, 14] WHO investigators had already noted that many traps in bathroom floor drains were dried out meaning an exhaust fan could have drawn droplets or aerosols from the drainage pipe into the bathroom and then into an air shaft. [1, 26] However aerosolization of the virus source is not uncommon inside hospitals and the ventilation systems inside many general hospitals or wards are not particularly helpful in removing the virus-laden aerosols once they are generated Yu added. [1, 31] The second study appearing in the journal looked at the case of a lab worker with no apparent exposure to SARS and no travel history who nevertheless contracted the disease after the outbreak was over in late May 2003.

Appendix K

Mead Summary extract of document with no title or subtitles

[1, 1] April 26 A new study suggests that severe acute respiratory syndrome SARS may have been spread through the simple act of

flushing a toilet instead of being passed directly from person to person. [1, 2] Even with all the research that has been conducted on SARS in the past year some mystery still remains as to how the virus can be transmitted. [1, 3] Two articles appearing in the April 22 issue of the New England Journal of Medicine explore the possibility of airborne and laboratory transmissions. [1, 7] Future prevention and protection against SARS should take into consideration the possibility that airborne transmission avoidance of close contacts alone may not be adequate. [1, 9] Yu and his co-authors conducted an analysis of the outbreak at Amoy Gardens an apartment complex in Hong Kong where more than 300 residents were affected last year when the SARS epidemic hit a number of Asian countries before spreading elsewhere in the world. [1, 14] WHO investigators had already noted that many traps in bathroom floor drains were dried out meaning an exhaust fan could have drawn droplets or aerosols from the drainage pipe into the bathroom and then into an air shaft. [1, 26] However aerosolization of the virus source is not uncommon inside hospitals and the ventilation systems inside many general hospitals or wards are not particularly helpful in removing the virus-laden aerosols once they are generated Yu added. [1, 31] The second study appearing in the journal looked at the case of a lab worker with no apparent exposure to SARS and no travel history who nevertheless contracted the disease after the outbreak was over in late May 2003.

Appendix L
SweSum

May 10, 2004
Body Building
Growing replacement organs is still a long way off
By Christine Soares

Six years ago Michael Sefton of the University of Toronto challenged his colleagues in the fledgling field of tissue engineering to build a functioning human heart within 10 years. With the isolation of human embryonic stem cells later that year, Sefton's challenge seemed all the more relevant: stem cells, after all, are nature's starting point for building working organs.

Now Sefton admits that the deadline on his Living Implants from Engineering ("LIFE") initiative was naive, and he thinks it will be at least another 10 to 20 years. Yet any whole organ would be a complex three-dimensional edifice comprising specialized cells, nerves and muscle, all interwoven with a dense web of veins and capillaries diffusing oxygen and nutrients.

The idea of seeding an organ and letting the body do the rest of the construction might work for a kidney, because the patient could be treated with dialysis while the new organ was being generated, according to Jeffrey L. For a patient suffering from lung or heart failure, however, growing a new organ would put too much strain on an already weak body.

Lexicon English
Words before 749
Words after 212
Summary length: 28%
Type of text: newspapertext
Keywords: *organ human heart simple building embryonic another sefton living growing*

Mead

[1, 1] May 10 2004 Body Building Growing replacement organs is still a long way off By Christine Soares Six years ago Michael Sefton of the University of Toronto challenged his colleagues in the fledgling field of tissue engineering to build a functioning human heart within 10 years. [1, 2] With the isolation of human embryonic stem cells later that year Sefton's challenge seemed all the more relevant: stem cells after all are nature's starting point for building working organs. [1, 4] We need to be able to walk before we can run he says and the worry today is Can we make a vascularized piece of tissue or a tissue with two or three cell types in a controlled way? [1, 6] Yet any whole organ would be a complex three-dimensional edifice comprising specialized cells nerves and muscle all interwoven with a dense web of veins and capillaries diffusing oxygen and nutrients. [1, 8] By mimicking the natural 3-D shape in which an organ grows tissue engineers are trying to get adjacent cells to talk to one another and complete the task of building the desired tissues. [1, 14] Last fall for example researchers from the Massachusetts Institute of Technology and the Technion-Israel Institute of Technology reported generating tissues of neural liver and cartilage cells as well as formation of a 3D vessel-like network on a biodegradable polymer scaffold seeded with human embryonic stem cells.

Appendix M

SweSum

In the Air: SARS
By Amanda Gardner
HealthDay Reporter
From HealthDayNews

April 26 — A new study suggests that severe acute respiratory syndrome (SARS) may have been spread through the simple act of flushing a toilet instead of being passed directly from person to person.

Two articles appearing in the April 22 issue of the New England Journal of Medicine explore the possibility of airborne and laboratory transmissions. Both scenarios point to new public health measures that should be taken to contain the disease.

"Airborne spread of a concentrated source of virus can infect many persons within a short period of time," Dr. The prevention of aerosolization of the virus source should take priority," he added. The authors conclude that an earlier World Health Organization (WHO) report was at least partly correct.

When they constructed a model of the drainage system, the study authors found that large number of aerosol particles were drawn out by the flushing of the toilet. However, aerosolization of the virus source is not uncommon inside hospitals, and the ventilation systems inside many general hospitals or wards are not particularly helpful in removing the virus-laden aerosols once they are generated," Yu added.

The second study appearing in the journal looked at the case of a lab worker with no apparent exposure to SARS and no travel history who nevertheless contracted the disease after the outbreak was over in late May 2003.

Lexicon: English
Words before 1036
Words after 253
Summary length: 24%
Type of text: newspapertext
Keywords: *virus study author building source health aerosol which could drain*

Mead

[1, 1] In the Air: SARS By Amanda Gardner HealthDay Reporter From HealthDayNews April 26 A new study suggests that severe acute respiratory syndrome SARS may have been spread through the simple act of flushing a toilet instead of being passed directly from person to person. [1, 2] Even with all the research that has been conducted on SARS in the past year some mystery still remains as to how the virus can be transmitted. [1, 3] Two articles appearing in the April 22 issue of the New England Journal of Medicine explore the possibility of airborne and laboratory transmissions. [1, 7] Future prevention and protection against SARS should take into consideration the possibility that airborne transmission avoidance of close contacts alone may not be adequate. [1, 9] Building a Case Yu and his co-authors conducted an analysis of the outbreak at Amoy Gardens an apartment complex in Hong Kong where more than 300 residents were affected last year when the SARS epidemic hit a number of Asian countries before spreading elsewhere in the world. [1, 14] WHO investigators had already noted that many traps in bathroom floor drains were dried out meaning an exhaust fan could have drawn droplets or aerosols from the drainage pipe into the bathroom and then into an air shaft. [1, 26] However aerosolization of the virus source is not uncommon inside hospitals and the ventilation systems inside many general hospitals or wards are not particularly helpful in removing the virus-laden aerosols once they are generated Yu added. [1, 31] The second study appearing in the journal looked at the case of a lab worker with no apparent exposure

to SARS and no travel history who nevertheless contracted the disease after the outbreak was over in late May 2003.

References

Anthony, T. (April 29, 2004). Four SARS Cases in China. CBS News. Retrieved April 29, 2004), from http://www. cbsnews.com/stories/2004/04/29/ health/main614642.shtml

Ball, C. N. (1996-2003) Georgetown Linguistics Web Frequency Indexer [Software]. Available from http://www.georgetown. edu/faculty/ballc/webtools/web_freqs.html

Dalianis, H. (2000, October). SweSum, a text summarizer for Swedish. Retrieved April 29, 2004, from ftp://ftp.nada.kth.se/IPLab/TechReports/IPLab-174.pdf

Gardner, A. (2004, April 26). In the Air: SARS. ABC NEWS. Retrieved April 26, 2004, from http://abcnews.go.com/

Hassel, M and Dalianis, H. (1999-2003) SweSum – Automatic Text Summarizer [Online Software] http://swesum.nada.kth.se/index-eng-adv.html

Lee, D. (September 2001) Genres, Registers, Text Types, Domains, and Styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology. Vol. 5, Num. 3. pp. 37-72*. Retrieved October 2004, from httsp://llt.msu.edu/vol5num3/lee/default.html

Mani, I. (2001). Automatic Summarization. Amsterday/Philadelphia: John Benjamins Publishing Company

Mani, I. Summarization Evaluation: An Overview. Retrieved July 2004, from http://research.nii.ac.jp /ntcir/workshop/OnlineProceedings2/sum-mani.pdf

Mead [Software]. http://tangra.si.umich.edu/clair/md/demo.cgi

Mureson, S., Tzoukermann, E., Klavans, J. (n.d) Combining Linguistic and Machine Learning Techniques for Email Summarization. Retrieved September 2004, from http://acl.ldc.upenn.edu/W/W01/W01-0719.pdf Passive smoking 'killing workers'. Retrieved April 10, 2004, from http://news.bbc.co.uk /1/hi/health /2925633.stm

SARS: the True Story. Retrieved April 10, 2004, from http://www.bbc.co.uk /science/horizon /2003/sars.shtml

Soares, C. (2004, May 10). Body Building: Growing replacement organs is still a long way off. *Scientific American*. Retrieved April 10, 2004, from Scientific American Body Building.htm