

Identifying the Nuclei Sentences within a Piece of Text

Amir Ali

Department of Computer Science

Lund University

amir.ali.429@student.lu.se

Abstract

This document suggests a method of identifying the nuclei sentences within a piece of text and provides some results from an implemented prototype.

1 Credits

This document was motivated by the Criterion system created by ETS, software aimed at evaluating the text in student essays (<http://criterion4.ets.org/cwe/>). The methodology and implementation discussed in this document was produced with the guidance and suggestions of Pierre Nugues, Computational Linguistics professor at Lund University in Sweden.

2 Introduction

With the now overwhelming amount of textual information stored throughout the World Wide Web, automatic text summarization has quickly become an essential tool. Text summarization tackles the problem of extracting the most important ideas from source text, and producing an abridged version (Mani and Maybury, 1999).

While the goal of text summarization is always consistent: to summarize a larger piece of text into a smaller piece of text, the desired level of brevity can vary depending on the application. For example, in the summarization of news articles it may be desirable to summarize a piece of text into the main events, dates and participants. In particular, the summarization technique described in this paper attempts to simply identify the nuclei, or most significant, sentences in each paragraph.

A prototype of the suggested algorithm has been implemented, and is described in Section 4.

3 Nuclei Detection

The proposed algorithm is quite simple and logical. It follows from the following assumptions:

- 1) The most significant sentences will contain the most significant words
- 2) The most significant words will appear more often in a piece of text than insignificant words (with special consideration given to stop words, discussed in 3.2)

Given these assumptions, a means of scoring each sentence is proposed below.

3.1 Sentence Scoring

Sentence scoring is completed in two passes of the original text.

The first pass involves a counting the unigrams. In terms of the algorithm, the count of a particular unigram could be considered a score for that word.

The second pass assigns a score to each sentence. The score assigned to a sentence is defined as:

$$f(S) = \sum_{w \in S} W(w)$$

where w is a word, S is the sentence being scored and $W(w)$ is the unigram count of word w .

Clearly a sentence containing words that appear often in the text will have a higher sentence score, as desired. However, using this algorithm sentence scores are highly sensitive to:

- 1) Words that appear very frequently, but do not hold significance in the text (i.e. stop words)
- 2) The number of words in the sentence

These two dependencies are discussed in Sections 3.2 and 3.3.

3.2 Addressing Stop Words

In order to remove words that are not significant in the text, we borrow the concept of stop words used by most search engines.

Stop words in English include words like “the”, “of” and “a” which appear very often in text but play little to no role in the meaning of a piece of text.

If we consider a news article, for instance, the need for removing stop words becomes apparent. A given news article may mention the main participant’s name four or five times throughout the text, while the word “the” may appear four times in a given sentence. In this case, “the” would be assigned a much higher score than the main participant’s name, which clearly holds more significance.

While there is no definitive list of stop words available, multiple lists were tested on the prototype described in Section 4. In general, the larger the list of stop words, the better the performance. It is most beneficial to remove as many of the insignificant words as possible.

3.3 Addressing Sentence Length

The sentence scoring algorithm appears to favour sentences with more words. Each word in a sentence has a unigram count of at least 1, thus an extra word always implies a higher score (with the exception of stop words, which are ignored).

One potential solution is to normalize each sentence score by the number of words in the sentence; that is, calculate the average word score within a sentence. However, this in turn strongly favours short sentences. For instance, if one considers that most words in a piece of text probably appear only once, we would expect the average of an insignificant sentence to be close to 1. However, consider a sentence like “Bob agreed.” is present in a piece of text, where Bob is the main participant so his name has appeared 5 times. This sentence would be assigned a score of at least 3 ($6 \text{ points} \div 2 \text{ words}$), which is a high score considering the sentence is probably not significant and that more significant sentences are likely to have 1-count unigrams that pull their normalized score down.

Another, more appropriate means of addressing sentence length would be to modify the first pass described in Section 3.1. That is, rather than perform a basic unigram count, use a non-linear means of scoring words. The design

of such a system should try to relate the significance of multiple occurrences of the same word to the significance of having an extra word in the sentence. Using the linear unigram count described in Section 3.1 suggests that an additional occurrence of a given word is equivalent to the addition of any non-stop word to a sentence (both increase sentence score by 1). A more intelligent system may assume that no word is significant until it has been mentioned at least twice. Furthermore, more significance could be attached to each reoccurrence, either by using a steeper step (i.e. +2 for each occurrence) or using an exponential function.

4 Implemented Prototype

The interface of the implemented prototype is similar to the Criterion system created by ETS that is used to analyze student essays. The application accepts a piece of text and outputs the same piece of text with the nuclei sentences identified via coloured font. The nuclei of the first paragraph and the final paragraph are identified to be the main thesis and main conclusion, respectively. The nuclei sentences in the remaining paragraphs are identified as main points.

The prototype implements the exact sentence scoring algorithm described in Section 3.1.

In order to address stop words, a list of 319 stop words, borrowed from the University of Glasgow Department of Computing Science (<http://www.dcs.gla.ac.uk/>) has been used. Originally a list of 35 stop words was used, but this was not comprehensive enough, as words with no significance emerged with the high unigram counts.

The implemented prototype achieves good results without addressing sentence length; no normalization or non-linear word scoring is used.

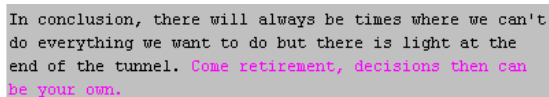
In addition, the prototype also extracts the position within the paragraph of the identified nuclei sentence.

4.1 Prototype Results

The following section attempts to analyze the results of the implemented prototype described above. The full pieces of text can be found in the included appendices; however, the specific paragraphs analyzed are included inline.

It should be noted that the *most significant sentence* can be ambiguous, so this section simply assesses whether a *good* choice was made.

We see that in Text 1, entitled Making Decisions, the prototype does a good job selecting sentences in the thesis and body paragraphs. All of these selected sentences reflect what could be the authors main point of the paragraph. However, in the final paragraph, shown in Figure 4.1.1, the prototype does a poor job selecting the main conclusion.



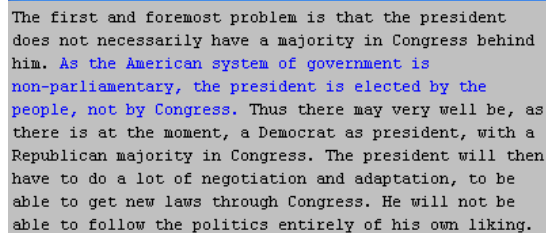
In conclusion, there will always be times where we can't do everything we want to do but there is light at the end of the tunnel. Come retirement, decisions then can be your own.

Figure 4.1.1 – Concluding paragraph of Text 1 – Making Decisions

Clearly, the author makes his conclusion in the first sentence of the paragraph, and the second sentence is a form of elaboration. The poor selection occurs because the author has decided to use completely new wording in stating his conclusion. For instance, he refers to making decisions as “doing everything we want to do” and uses the expression “light at the end of the tunnel,” which he elaborates on in the final sentence (using the words “retirement” and “decisions” which appear frequently in the text). The sentence scoring algorithm is not strong enough to address these conditions; however, it is clear that identifying key expressions, such as “in conclusion” could be used in order to make better selections.

We see, again, in Text 2, entitled The Presidency, that the prototype does a good job in most cases. An appropriate thesis and conclusion is identified; however, in body paragraphs 1, 2 and 3 the prototype selects sentences that are more likely intended as supporting material or elaborations.

In paragraph 1, , shown in Figure 4.1.2, the selected sentence is likely intended to support the first sentence, which is the main point of the paragraph. The poor selection could be avoided if the algorithm had cross-referencing capability. The word ‘him’ in the first sentence has adds zero to the sentence score because it is included in the list of stopwords. In reality, ‘him’ refers to ‘the president’, which is the most significant word within the essay. Also, using ‘first and foremost’ as a key expression for identifying a likely main point. Coreferencing and key-words would also apply to body paragraphs 2 and 3.



The first and foremost problem is that the president does not necessarily have a majority in Congress behind him. As the American system of government is non-parliamentary, the president is elected by the people, not by Congress. Thus there may very well be, as there is at the moment, a Democrat as president, with a Republican majority in Congress. The president will then have to do a lot of negotiation and adaptation, to be able to get new laws through Congress. He will not be able to follow the politics entirely of his own liking.

Figure 4.1.2 – Body Paragraph 1 of Text 2 – The Presidency

5 Further Work

There are several improvements that could be made to the algorithm and implementation described.

One obvious improvement would be cross-reference capability, such that words referring to the same object/person are counted as the same unigram. Cross-referencing is a thoroughly researched area, so such an addition should be possible. Similarly, embedding a part of speech (POS) tagger and performing the counting and sentence scoring using unigram-POS tuples would improve performance, but likely not significantly.

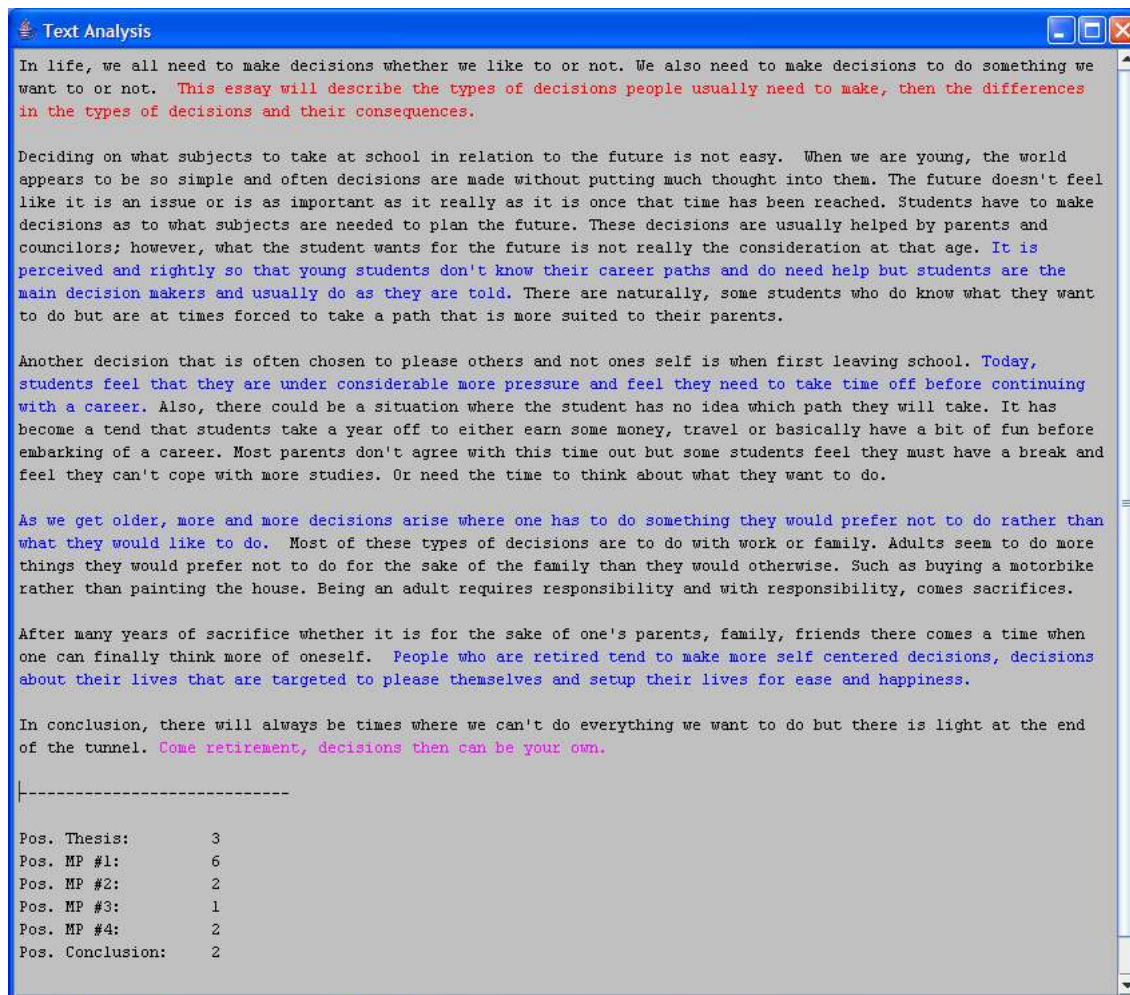
A valuable extension to the described implementation would be identifying further discourse structure within the paragraphs, such as concessions, elaborations and contrasts. This could be done by searching and recognizing the appropriate cue phrases/words (Corston-Oliver, 1998). An alternative possibility is to embed a dependency parser, such as the McDonald Parser (see <http://ryanmcd.googlepages.com>) and use the dependency relations to identify satellite sentences.

Reference

- Indrjeet Mani and Mark T. Maybury. 1999. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA.
- Simon Corston-Oliver. 1998. *Computing Representations of the Structure of Written discourse*. Technical Report, Microsoft Research, WA.

Appendix 1

Text 1 – Making Decisions



In life, we all need to make decisions whether we like to or not. We also need to make decisions to do something we want to or not. **This essay will describe the types of decisions people usually need to make, then the differences in the types of decisions and their consequences.**

Deciding on what subjects to take at school in relation to the future is not easy. When we are young, the world appears to be so simple and often decisions are made without putting much thought into them. The future doesn't feel like it is an issue or is as important as it really is once that time has been reached. Students have to make decisions as to what subjects are needed to plan the future. These decisions are usually helped by parents and counselors; however, what the student wants for the future is not really the consideration at that age. **It is perceived and rightly so that young students don't know their career paths and do need help but students are the main decision makers and usually do as they are told.** There are naturally, some students who do know what they want to do but are at times forced to take a path that is more suited to their parents.

Another decision that is often chosen to please others and not ones self is when first leaving school. **Today, students feel that they are under considerable more pressure and feel they need to take time off before continuing with a career.** Also, there could be a situation where the student has no idea which path they will take. It has become a trend that students take a year off to either earn some money, travel or basically have a bit of fun before embarking of a career. Most parents don't agree with this time out but some students feel they must have a break and feel they can't cope with more studies. Or need the time to think about what they want to do.

As we get older, more and more decisions arise where one has to do something they would prefer not to do rather than what they would like to do. Most of these types of decisions are to do with work or family. Adults seem to do more things they would prefer not to do for the sake of the family than they would otherwise. Such as buying a motorbike rather than painting the house. Being an adult requires responsibility and with responsibility, comes sacrifices.

After many years of sacrifice whether it is for the sake of one's parents, family, friends there comes a time when one can finally think more of oneself. **People who are retired tend to make more self centered decisions, decisions about their lives that are targeted to please themselves and setup their lives for ease and happiness.**

In conclusion, there will always be times where we can't do everything we want to do but there is light at the end of the tunnel. **Come retirement, decisions then can be your own.**

Pos. Thesis:	3
Pos. MP #1:	6
Pos. MP #2:	2
Pos. MP #3:	1
Pos. MP #4:	2
Pos. Conclusion:	2

Text Source:

Julienne Sandgren. English Department. ETS.

Text 2 – The Presidency

E. Text Analysis

Although it is often said that the President of the United States holds the most powerful office in the world, this does not mean that he is able to decide very much for himself. The American Constitution, which was adapted in 1789, clearly states the Separation of Powers. Thus, the president makes up only one third of the government, namely the executive branch. He is also controlled by a complex system of checks and balances, which makes sure that he (or any of the other branches, for that matter) does not become too powerful. We will now have a look at the different problems which may be facing a recently elected president, and then discuss to what extent his powers are important.

The first and foremost problem is that the president does not necessarily have a majority in Congress behind him. As the American system of government is non-parliamentary, the president is elected by the people, not by Congress. Thus there may very well be, as there is at the moment, a Democrat as president, with a Republican majority in Congress. The president will then have to do a lot of negotiation and adaptation, to be able to get new laws through Congress. He will not be able to follow the politics entirely of his own liking.

But even if the president is supported by a majority in Congress, this does not mean that everything is necessarily fine. Since there are only two important parties in the USA, the representatives from each group make up a far from homogenous mass. Conservative Democrats may very well support the Republicans in many cases, and liberal Republicans may support the Democrats. President Clinton experienced the trouble connected to this in the years 1993-94, when he faced a Democrat, but nevertheless reluctant Congress. In fact, many people claimed that he co-operated better with the Republican Congress 1995-96, than with the people of his own party.

Congress is of course often the most serious problem to a new president. It is after all the legislative branch of government, and passes both ordinary and tax bills. Another important element worth mentioning here, is the huge amount of lobbying which goes on in America. Private associations and companies, that officially do not have any power, gain a lot of influence through the "persuasion" of congressmen (and women) or/and senators. One example is the National Rifle Association (NRA), which has so far succeeded in destroying any attempt to introduce gun control in America. Another one is how the powerful insurance companies helped to kill President Clinton's welfare reform a couple of years ago.

In the US system of checks and balances, the Supreme Court is also very powerful. They are able to create a lot of trouble for the president if they want to, as they have the so-called power of judicial review. This means that they can declare any law or presidential act unconstitutional. The Supreme Court Justices are appointed for life, and the president has therefore no means of controlling them, unless they resign voluntarily or die. The balance between liberal and conservative justices here, is therefore of course subject to a lot of interest, and the president has good reasons in maintaining a harmonious relation to the Supreme Court.

In addition to the different branches of government, there is another aspect of the American Constitution which often will give the president trouble. Federal government has only limited powers, the so-called reserved powers ensure that each state may have very different laws and policies on such areas as education, crime and health care. If the president wants to achieve something in these areas (President Bush notably wanted to reform the educational system), he has very restricted possibilities, and is often left to encouraging the states to co-operate.

As we have seen, many elements of the US government have the possibility to create trouble for the president. They often seize this opportunity, and this is what makes the job as President of the US so hard. As Jan Mønstad put it: "The power of the President is great if he can use it; but it is a moral power, a power exercised by persuasion and discussion." The president will always have to co-operate in order to achieve something. If conflicts arise between him and Congress, for example, trouble is in the horizon. This happened in 1995, when President Clinton refused to sign the national budget proposed by Congress. The entire government came to a stand still for a couple of weeks, and then Congress had to back off. They were not strong enough to override Clinton's veto (they would have needed a 2/3 majority). And of course, as already mentioned, the president exercises an immense influence on political life. Therefore, despite all the elements which may threaten his existence, the president could rightly be called the most powerful man of the US, and thus, in today's situation, of the world.

For. Thesis:	1
For. HF #1:	2
For. HF #2:	4
For. HF #3:	6
For. HF #4:	1
For. HF #5:	1
For. Conclusion:	3

Text Source:

Martin Bratt. University of Oslo. 1997. URL: <http://home.online.no/~helhoel/elev.htm>.