Extractive summarization methods – subtitles and method combinations

Nikitas N. Karanikolas
Technological Educational Institute of Athens
Ag. Spyridodos street, Aigaleo 12243, Greece
nnk@teiath.gr

Abstract

In some previous work, we have presented a software tool for experimenting with well known methods for text summarization. The methods offered are belonging to the extractive summarization direction. These methods do not understand the meaning in order to condense the text but simply extract a subset of the original sentences which are the most (promising as being) relevant for expressing shortly the text meaning. However, in order to pay attention to the whole idea (a workbench for testing available extractive summarization), we have avoided to concentrate to some potential improvements or we have made some simplification assumptions of the existing extractive summarization methods. Here, we remove the simplifications and also examine some improvements to the existing methods, in order to achieve better summarizations.

1. Introduction

Summarization is technology for the reduction of a text's length in order to be easily and quickly understandable. The reduction can be based either on shallow processing methods or on semantic oriented ones. The semantic oriented methods understand – somehow – the text and try to combine the meanings of similar sentences and generate generalizations. Shallow processing methods do not actually take into account the meaning of the text but they statistically select the most promising (as being relevant) sentences for quick understanding. Such an extraction-based summary is not necessarily coherent. In some previous work, we have presented a software tool for experimenting with well known shallow processing (extraction-based) methods for text summarization. One of these methods is the Title Method proposed by Edmundson [Edm69]. In our consideration of method we made the simplification assumption that documents have only a title (something that is in general correct) but they don't have other titles (like chapter, section, subsection titles; in the following medially titles). Here, we are going to resolve this simplification and consider how the existence of words from the medially titles in some sentence can adapt the likelihood of sentence to be relevant for expressing the meaning of the document. Moreover we suppose and consider using a non-linear function for measuring the likelihood of some sentence that contains more than one from the (front and medially) title words. Also some other issues regarding the uniformity of the Title Method and the competition and also combination of the Title Method with other extraction-based summarization methods are examined.

In the following we present some extraction-based summarization methods. We provide a simple, user configurable, combination schema. Next we invent and consider using a non-linear function for measuring the likelihood of sentences having more than one from the title words. The proposed function also ensures the uniformity of the Title Method. Next we consider how the existence of words from the medially titles in some sentence can adapt the likelihood of sentence to be included in the extraction-based summary. An evaluation of the adapted Title method is conducted. Conclusions and Future work is the last section.

2. Extraction-based summarization methods

The extraction-based summarization methods follow the idea that some sentences are more important than others for expressing the meaning of the document. Consequently, the summarization can be based on some weighting function that assigns weights to sentences and extract the sentences having the greater weighting. We can mention three main Sentence weighting ideas: based on the terms importance, based on sentence location and based on the inclusion of title terms.

The Sentence weighting based on the terms importance has to combine two factors: what is the importance of term inside a document and what is the ability of the term to discriminate among documents in the collection. There are three schemas that combine these two factors. These are: Sentence weighting based on *TF*IDF*, Sentence weighting based on *TF*ISF* and Sentence weighting based on *TF*RIDF*. *TF* (Term Frequency) and *IDF* (Inverse Document Frequency)

are basic ideas coming from the past and from the Information Retrieval discipline [Kar07]. *ISF* (Inverse Sentence Frequency) [Cho09] and *RIDF* (Residual IDF) [Mur07] are newer ideas.

Baxendale [Bax58] examined the position of sentences as a feature for selecting sentences for summarization. He concluded that in 85% of the paragraphs the topic sentence came as the first one and in 7% of paragraphs the last sentence was the topic sentence. Thus, a naive but fairly accurate way to select a topic sentence would be to choose one of these two [Das07]. Another more sophisticated sentence weighting based on sentence location is the "News Articles" algorithm [Har10]. It utilizes a simple equation in order to assign a different weight to each sentence in a text, based on the position of the sentence inside the document as a whole and inside the host paragraph:

Edmundson [Edm69] has proposed the "Title Method" which supposes that an author conceives the title as circumscribing the subject matter of the document. According to this method, sentences that include words from the document's title are more relevant for expressing the meaning of the document. The suggested "final Title weight" for each sentence is the sum of the "Title weights" of its constituent words. Edmundson also defined the "Title glossary" which is the set of words existing in the title and subheadings, with different weights for title and subheading words.

In our previous work [Kar12] we made the simplification assumption that documents have only a title (something that is in general correct) but they don't have other medially titles (like chapter, section, subsection titles/subheadings). This assumption is because our system was designed in order to work with articles available through the internet, blog posts, and other similar sources. According to this assumption, our previous system assigns a predefined constant for each title word. Thus, in our previous system, the "final Title weight" for each sentence is the product of the predefined constant multiplied by the number of title words occurring in the examined sentence. In the above, we talk about words but we actually mean valid word stems.

3. Combination of methods

During the design phase of our summarization methods benchmarking system (our previous work [Kar12]), we decided to provide all above discussed sentence weighting approaches. Both sentence location (Baxendale's and News Articles) approaches, the Edmundson's Title Method, together with the alternative Sentence weightings based on the terms importance are provided to the user. Regarding the contribution of these three categories of factors, we decided to use a simple linear relation, but leave the user to decide on the weight of each factor. The following equation is implemented in our system:

$$w1 * ST + w2 * SL + w3 * TT \quad (1)$$

where *ST* is the sentence weighting based on terms, *SL* is the sentence location factor, and *TT* is the title terms factor.

4. Non-linear combination of title words

As it is already stated, our previous system assigns a predefined constant for each title word that exists in a sentence. Thus, the "final Title weight" for each sentence is the product of the predefined constant multiplied by the number of title words occurring in the examined sentence. In other words we have a linear function for sentence weighting according to the inclusion of title terms. However, another idea says that even a single title word existing in some sentence, the plausibility of sentence to express the meaning of document is very high. Two title words existing in some sentence increase this plausibility but they do not double it. Thus a non linear function should be invented. In table 1 we present two such non linear functions. We assume a title having sixteen words. Third and fifth (last) columns of table 1 represent these functions and contain the result (the sentence weight) for a sentence containing x (out of 16) title words. It is a matter of experimentation for selecting one of the functions.

5. Ensuring uniformity of the Title Method

Our previous linear approach for assigning weights to sentences according to their title words had also a negative consequence. The proportion of contribution of each factor (*ST*, *SL* and *TT*) in the overall sentence weight (see equation 1) varied. In documents with long title, the *TT* factor had greater contribution than the contribution of *TT* factor in a document with short title.

In order to explain, we assume that the values of *SL* range from 0.0 to 1.0 (this is the actual range of values in the "News Articles" algorithm). We also assume that the constant weight of a term title is *C*. Thus a sentence

having x title terms gets a *TT* factor as defined in next equation.

$$TT = x*C \quad (2)$$

Because of these, documents with different length of titles have different range of their *TT* factor while their *SL* factor remains in the same range of values. For example, any sentence from an 8-words-title document gets a *TT* factor value in the range 0.0 to *8*C* while any sentence from a 4-words-title document gets a *TT* factor value in the range 0.0 to *4*C*. In both cases (both title lengths) the range of *SL* remains from 0.0 to 1.0.

This problem is resolved with our non linear (logarithmic) function. The range of *TT* is always from 0.0 to 1.0.

Table 1. Sentence weight for sentence having x (out of 16) title terms

x	$Log_2(x+1)$	$\frac{Log_2(x+1)}{max(Log_2(x+1))}$	$Log_3(x+2)$	$\frac{Log_3(x+2)}{max(Log_3(x+2))}$
1	1,00	0,24	1,00	0,38
2	1,58	0,39	1,26	0,48
3	2,00	0,49	1,46	0,56
4	2,32	0,57	1,63	0,62
5	2,58	0,63	1,77	0,67
6	2,81	0,69	1,89	0,72
7	3,00	0,73	2,00	0,76
8	3,17	0,78	2,10	0,80
9	3,32	0,81	2,18	0,83
10	3,46	0,85	2,26	0,86
11	3,58	0,88	2,33	0,89
12	3,70	0,91	2,40	0,91
13	3,81	0,93	2,46	0,94
14	3,91	0,96	2,52	0,96
15	4,00	0,98	2,58	0,98
16	4,09	1,00	2,63	1,00

Table 2. Sentence weight for sentence having x (out of 8) title terms

x	$\frac{Log_2(x+1)}{max(Log_2(x+1))}$	$\frac{Log_3(x+2)}{max(Log_3(x+2))}$
1	0,32	0,48
2	0,50	0,60
3	0,63	0,70
4	0,73	0,78
5	0,82	0,85
6	0,89	0,90
7	0,95	0,95
8	1,00	1,00

6. Exploit words from the medially titles

In our present approach we are not aiming to create a method for automatic document structure detection. Something like this demand to identify the diferent parts of the document (such as chapters, sections, subsections, articles and paragraphs), identify how each one of these (narrower structure) nests inside other (broader structure) and then add markups for these parts. A parser for automatic mark-up of such a document structure is a very demanding process. However, it is simply enough to create parser that identifies titles in between paragraphs. In other words, we are expecting from our parser to return a list of items where the first item is the front title while the rest items can be either paragraphs or medially titles.

Having identified a front title and medially titles we can apply the previous non-linear function and assign a sentence weight against title words and a sentence weight against the words of the medially-title coming before the sentence. In a simpler approach we can assume that words from all medially titles constitute a second glossary, the "Global medially title glossary". In the later case we can apply the previous non-linear function and assign a sentence weight against title words ("front Title Terms", shortly *fTT*) and a sentence weight against the "Medially title glossary" ("medially Title Terms", shortly *mTT*). In our evaluation we assume the second (Global medially title glossary) approach. The final weight for a sentence based on the inclusion of terms can be:

$$TT = \alpha * fTT + \beta * mTT \quad (3)$$

where α=0.6 and β=0.4

(in general, α is set in range 0.1 .. 0.9 and β=1-α)

or

$$TT = max\ (fTT, mTT) \quad (4)$$

Since "Global medially title glossary" consists of words from many subtitles/subheadings, we suppose that *mTT* should be computed with the $Log_3(x+2)$ based function and *fTT* should be computed with the $Log_2(x+1)$ based function.

7. Evaluation

In order to evaluate our approach, we have selected a small subset of documents from the Greek language corpora. All the selected documents have a front title and few (usually 2 to 5) medially titles. One such document is presented in figure 1.

For each document, we have asked text retrieval experts to extract the most promising (20%) subset of sentences for shortly expressing the document meaning. These extractions are the manually selected summaries. Then the same documents are given in our system to mechanically extract summaries. For this reason we have excluded the *ST* factor and given equally weights for the *SL* and *TT* factors (*w1*=0, *w2*=1 and *w3*=1 in the first (1st) equation). For the computation of *TT* factor, we have used the fourth (4th) equation. The number of sentences for the mechanic summarization is set to the same percentage (20%). Next, for each document, we have measured the percent of sentences in the mechanically extracted summary that exist in the manually extracted summary. The average percent is 54% which is a very promising result since in the automatic summarization we have excluded the *ST* factor (terms-based sentence weighting). In order to evaluate if the medially titles has influence in the result, we conducted the experiment again but now considering the medially titles as simple single-sentence paragraphs. In this experiment the average percent of matching sentences (between manual and mechanical summary) is decreased 46%. A third experiment is conducted but now using our previous system. We remind that in our previous system the "final Title weight" (*TT* factor) for each sentence is the product of the predefined constant (*C*) multiplied by the number of title words occurring in the examined sentence). Again we set *w1*=0 and moreover we set *C*=0.5. Now, the average percent of matching sentences is more decreased to 41%.

Με φαντασία και δυναμισμό *front title*

medially title

Χιλιάδες έδωσαν το «παρών» στα μεγάλα αντιπολεμικά συλλαλητήρια Αθήνας και Θεσσαλονίκης

Με συμβολικές αυτοσχέδιες θεατρικές παραστάσεις για την φρίκη του πολέμου και μουσική από μουσικά συγκροτήματα, με φαντασία και δυναμισμό πολλοί νέοι έδωσαν τον δικό τους τόνο στο μεγάλο αντιπολεμικό συλλαλητήριο του Σαββάτου στην Αθήνα. Και στη Θεσσαλονίκη, όμως, χιλιάδες πολίτες έδωσαν το «παρών» στα μεγάλα αντιπολεμικά συλλαλητήρια προχθές.

Στη Θεσσαλονίκη. Παρά το τσουχτερό κρύο πολίτες όλων των ηλικιών διαδήλωσαν κατά του πολέμου.

Στην Αθήνα, από τις 11 το πρωί οι διαδηλωτές-μέλη του ΠΑΜΕ συγκεντρώθηκαν στα Προπύλαια. Το κεντρικό πανό είχε παραστάσεις από την Γκερνίκα και οι συγκεντρωμένοι κρατούσαν πανό με συνθήματα: «Όχι στον πόλεμο», «Όχι στην βαρβαρότητα του πολέμου», «Όχι αίμα για το πετρέλαιο».

-- 2 more paragraphs hidden --

Στη Θεσσαλονίκη *medially title*

Αμερικανική πρεσβεία. Με αυτοσχέδιες θεατρικές παραστάσεις διαδήλωσαν πολλοί νέοι στην Αθήνα δίνοντας τον δικό τους τόνο στο μεγάλο αντιπολεμικό συλλαλητήριο.

Και στη Θεσσαλονίκη, παρά το τσουχτερό κρύο πολίτες όλων των ηλικιών ανταποκρίθηκαν στο κάλεσμα των οργανώσεων ΕΔΥΕΘ, ΠΑΜΕ, Αντιπολεμική Επιτροπή Θεσσαλονίκης, «Δράση 2003», «Πρωτοβουλία αγώνα 2003» και «Σαλόνικα 2003» συγκροτώντας δύο μεγάλες πορείες μετά τις συγκεντρώσεις τους σε τρία διαφορετικά σημεία (Λιμάνι, Άγαλμα Βενιζέλου και Καμάρα).

-- 2 more paragraphs hidden --

«Χρόνο και χώρο στην ειρήνη» *medially title*

Την πεποίθηση ότι έστω και την ύστατη στιγμή υπάρχουν περιθώρια για ειρήνη με τον αφοπλισμό του Ιράκ, εξέφρασε χθες, σε δηλώσεις του στην Άρτα, ο γραμματέας του ΠΑΣΟΚ Κ. Λαλιώτης. «Πρέπει να δώσουμε χρόνο και χώρο στις πρωτοβουλίες για ειρηνικές λύσεις», είπε και επισήμανε ότι η ελληνική κυβέρνηση έχει πάρει πρωτοβουλίες για να διαμορφώσει ένα κοινό πλαίσιο αναφοράς όλων των ευρωπαϊκών χωρών.

Τόσο ο κ. Λαλιώτης όσο και ο υπουργός Ανάπτυξης Άκης Τσοχατζόπουλος, σε δηλώσεις του στη Θεσσαλονίκη, χαιρέτισαν τα αντιπολεμικά συλλαλητήρια στην Ελλάδα. Ο κ. Τσοχατζόπουλος επισήμανε επιπλέον πως «αν επιθυμία είναι ο ειρηνικός αφοπλισμός για την προστασία της διεθνούς κοινότητας, υπάρχει λύση».

Figure 1. Example document (#3644) taken from http://www.greek-language.gr/

8. Conclusions and Future Work

The results in our experiments suppose that medially titles should be considered in order to get better mechanically extracted summaries. Also the *TT* factor contributes in a better way to the summarization when equation 4 is used (versus equation 2). In our plans we have to repeat our experiments with a larger document set (the current is constituted with only 21 documents) and also have to consider all factors together (enable the *ST* factor). Moreover alternative approaches for the *TT* factor (e.g. equation 3) should be evaluated.

References

[Cho09] L. H. Chong, and Y. Y. Chen. Text Summarization for Oil and Gas News Article. *World Academy of Science, Engineering and Technology*, 53, 2009.

[Mur07] G. Murray and S. Renals. Term-Weighting for Summarization of Multi-Party Spoken Dialogues. In A. Popescu-Belis, S. Renals, and H. Bourlard (eds), *Machine Learning for Multimodal Interaction IV*. Lecture Notes in Computer Science, 4892: 155-166. Springer, 2007.

[Kar07] N. N. Karanikolas, *The measurement of similarity in stock data documents collections*. eRA-2: 2nd Conference for the contribution of Information Technology to Science, Economy, Society and Education, September 22-23, 2007, Athens, Greece.

[Edm69] H. P. Edmundson. New Methods in Automatic Extracting. *Journal of the ACM*, 16 (2): 264–285, 1969.

[Das07] D. Das and A.F.T. Martins. *A Survey on Automatic Text Summarization*. Carnegie Mellon University, 2007.

[Har10] S. Hariharan. Multi Document Summarization by Combinational Approach. *International Journal of Computational Cognition*, 8 (4), December 2010.

[Bax59] P. B. Baxendale. Machine-Made Index for Technical Literature—An Experiment. *IBM Journal of Research and Development*, 2: 354-361, 1958.

[Kar12] N. N. Karanikolas, E. Galiotou and C. Tsoulloftas. *A workbench for extractive summarizing methods*. PCI'2012: 16th Panhellenic Conference on Informatics, October 5-7, 2012, Piraeus, Greece. IEEE CPS.