Extractive summarization methods – subtitles and method combinations

Prof. Nikitas N. Karanikolas
Technological Educational Institute (TEI) of Athens
Department of Informatics

What is and approaches

- 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 consider the meaning . They statistically select the most promising (as being relevant) sentences for quick understanding.

Extraction-based summarization methods

- Sentence weighting. It is based on the terms importance. It combines two factors:
  - importance of term inside a document
  - the ability of the term to discriminate among documents in the collection.
- position of sentences
  - Baxendale 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.
  - the “News Articles” algorithm 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
- Title words

Tems importance

- TF/IDF
- TF/ISF
- TF/RIDF

The simple term frequency depends on other text characteristics, e.g. the text size, and consequently it is not a valid yardstick. Therefore, different term weighting functions have been introduced. Most of them are trying to normalize the term weights and make them comparable across different documents.

TF functions

$$t_{ij} = \frac{F_{ij}}{\max F_i}$$

$$t_{ij} = 0.5 + 0.5 \cdot \frac{F_{ij}}{\max F_i}$$

$$t_{ij} = \frac{F_{ij}}{\sum F_i}$$

- where t_{ij} is the weight of term j in document D_i, F_{ij} is the frequency of term j in document D_i, $max\ F_i$ is the frequency of the most frequent term in document D_i and ΣF_i is the sum of frequencies of the index terms existing in document D_i.

IDF function

$$q_j = \log_2\left(\frac{N}{DocFreq_j}\right) = -\log_2\left(\frac{DocFreq_j}{N}\right)$$

- *qj :* weight of term *j in the collection,*
- *N :* number of documents existing in the collection
- $DocFreq_j$ *:* number of documents where the term *j occurs.*

TF / IDF

$$T(S_{ik}) = \sum_j t_{ij} \cdot q_j$$

T(S$_{ik}$) : the weight of the k$_{th}$ sentence existing in document D$_i$,
T$_{ij}$ calculated (using one of the TF functions) for each term j existing in the kth sentence of document D$_i$,
q_j : calculated according to the IDF function.

TF / ISF

$$T'(S_{ik}) = \sum_j t_{ij} \cdot isf_{ij}$$

$$isf_{ij} = \log_{10}(ns_i / ns_{ij})$$

isf_{ij} *:* inverse sentence frequency of term j in document D_i,
ns_i *:* number of sentences in document D_i,
ns_{ij} *:* number of sentences of document D_i that contain the term j.

TF / RIDF

$$T''(S_{ik}) = \sum_j t_{ij} \cdot ridf_j$$

$$ridf_j = idf_j - \text{expected}(idf_j) =$$

$$-\log_{10}\left(\frac{DocFreq_j}{N}\right) + \log_{10}\left(1 - e^{-\frac{TotFreq_j}{N}}\right)$$

$ridf_j$ *:* Residual IDF of a term j in a given document,
$TotFreq_j$ is the cumulative frequency of term j across all documents.

The rest of the variables are as in the previous equations.

Position of Sentences

News Articles Algorithm

- Assigns 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 using:

((SP - P + 1) / SP) * ((SIP - SPIP + 1) / SIP)

  - *SP* : number of paragraphs in the document,
  - *P* : serial number of the paragraph under investigation,
  - *SIP* : number of sentences in the paragraph under investigation
  - *SPIP* : sentence position inside the paragraph.

Title method

- Edmundson 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 "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.

Title weight adaptation and methods combination

- In a first trial of us, 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.
- Combination of methods:

$$w1 * ST + w2 * SL + w3 * TT$$

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

Title words – linear vs non-linear

- 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.
- It is a linear function for sentence weighting according to the inclusion of title terms.
- Vesus the previous is the idea that even a single title word existing in some sentence make this sencence eligible for summarization.
- Two title words existing in some sentence increase this plausibility but they do not double it. Thus a non linear function should be invented.

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

x	$\text{Log}_2(x+1)$	$\frac{\text{Log}_2(x+1)}{\max(\text{Log}_2(x+1))}$	$\text{Log}_3(x+2)$	$\frac{\text{Log}_3(x+2)}{\max(\text{Log}_3(x+2))}$
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

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

Ensuring uniformity of the Title Method

- There exist documents with different length of titles. Consequently with the linear approach, the *TT* factor has different influence to the sentence weighting schema
- 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.

Exploit words from the medially titles

- In our present approach we are not aiming to create a method for automatic document structure detection. 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.
- 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 simply approach we can assume that words from all medially titles constitute a second glossary, the "**Global medially title glossary**".

- Next we can apply the non-linear function and assign a sentence weight against
  - title words (“front Title Terms”, shortly **fTT**)
  - 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:

$$TT = \alpha * fTT + \beta * mTT$$

or

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

Evaluation – source documents

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».

Evaluation - approch

- 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.
- The same documents are next 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).
- For the computation of *TT* factor, we have used the max version.
- 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.

Evaluation - Results

- The average percent is 54% which is a very promising (remind that we have excluded the ST factor).
- 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 is decreased to 46%.
- So, medially titles has influence in the result.
- A third experiment is conducted using our previous system . Now, the average percent of matching sentences is more decreased to 41%.
- So, the non-linear version of sentence weighting based on title terms has better results.

Conclusions – Future work

- The results in our experiments suppose that medially titles should be considered in order to get better mechanically extracted summaries.
- The *TT* factor contributes in a better way to the summarization when equation *max (fTT, mTT)* is used
- In our plans we have to repeat our experiments with a larger document set (the current is constituted with only 21 documents).
- We also have to consider all factors together (enable the *ST* factor).
- Moreover alternative approaches for the *TT* factor (e.g. equation $TT = \alpha * fTT + \beta * mTT$) should be evaluated.

Based on

- **Nikitas N. Karanikolas**, Eleni Galiotou and Christodoulos Tsoulloftas, *A workbench for extractive summarizing methods.* PCI'2012: 16th Panhellenic Conference on Informatics, October 5-7, 2012, Piraeus, Greece. IEEE CPS.
- **Nikitas N. Karanikolas**, *Extractive summarization methods – subtitles and method combinations.* RTA-CSIT 2016, November 18 - 19, 2016, Tirana, Albania.

Closing

- Thank you for your attention!
- Questions can be asked.

- nnk@teiath.gr
- http://users.teiath.gr/nnk/