Web page automatic abstracting : towards a revised extract

Mehdi Belguith[1], Mariem Ellouze[1] and Yannick Estève[2]

[1] ANLP Research Group, MIRACL Lab, FSEGS, University of Sfax, Tunisia
mehdi.belguith2017@gmail.com mariem.ellouze@planet.com
[2] LIA, University of Avignon, France
yannick.esteve@univ-avignon.fr

Abstract.
In this paper, we propose an original method that allows to summarize Web pages automatically. Our method is numerical and differs from other methods in that the generated summary can include text as well as images and graphics. Moreover, the proposed method can, in addition to the reordering of the sentences of the summary, detect those that are similar and therefore avoid redundancy, which allows to revise the generated extract and improve its quality. Our method has been implemented and the results of its evaluation are very interesting.

Keywords: Automatic summarization, Web pages, extraction approach, sentence redundancy detection, extract revision.

1 Introduction

With the increase of the number of websites (i.e. 1.4 billion sites in the world, according to statistics dating from 2018[1]), summarizing Web pages automatically has become an important research area in the field of NLP. Indeed, when entering a query to a search engine, the user will have as a result hundreds or even thousands of URLs. Navigating through a long list of web pages is such a tedious task that some users just visit the first links returned by the search engine. In such situations, Web page automatic summarization systems are the key factor in the everyday use of the Web, since they are useful for providing a global overview of the Web site content.

In this paper, we propose an original method for Web page automatic summarization based on an extraction approach. This method allows to rapidly provide a summary, without going through a deep analysis of the content of the web page or a real understanding of its content. In addition, our method has the advantage of integrating a semantic similarity analysis between the sentences to avoid redundancy and consequently improve the quality of the generated summary.

2 State of the art

The state of the art distinguishes two main approaches for automatic summarization: the abstraction approach (or understanding approach), the extraction approach (also called numerical approach) which does not rely on any in-depth analysis but only on a shallow analysis of the document to summarize.

2.1 Abstraction approach

The abstraction approach consists of producing summaries based on a total or a partial comprehension of the document to be summarized. It is therefore an in-depth document analysis.

Among the research works based on this approach, we can cite (Maaloul, 2012) who proposed a system of Arabic text automatic summarization based on the RST technique (Rhetorical Structure Technique) to determine the Semantic relationships between sentences. Then, sentences with important rhetorical relations are selected for the summary.

In the same context, (Keskes, 2015) proposed an automatic summarization method for Arabic documents that consists in segmenting the text into discourse segments. Then, it determines the semantic relations between these segments according to the Segmented Discursive Representation Theory (SDRT). Thus, the final summary consists of the set of segments having relevant discursive relations.

Other summarizing methods are based on graphs. In this context, (Khushboo et al., 2010) proposed a method that allows to build a graph from the text. The graph nodes are represented by the text sentences. The edge of the graph represents the connection (similarity) between the sentences. The weight of each node is calculated using the COS function. The summary is generated by taking the shortest path that begins with the first sentence of the original text and ends with the last sentence. SUMGRAPH (Patil and Brazdil, 2007) and Time stamped Graph (Lin, 2006) are two graph-based summary systems.

2.2 Extraction approach

The extraction approach allows to quickly produce an extract, without in-depth analysis or understanding of the document content. It is generally based on a set of criteria allowing the shallow analysis of the document to be summarized. The idea is to identify and extract the most important sentences from the text in order to build an extract. We can classify the methods of this extraction approach, in two main classes: numerical methods and machine learning based methods.

Numerical methods. They usually consist of computing scores for the textual segments (often sentences) of the document to be summarized. These scores are calculated according to several criteria ((Oufaida et al., 2014), (Bois et al., 2014), (Bharti and Babu, 2017), (Elvys Linhares, 2018)).

The main criteria used to evaluate the relevance ofrc a sentence are keywords frequency, sentence position, title words, "bonus"/"stigma" phrases, etc. Note that these numerical methods are based on numerical values calculated using scores that are either arbitrarily given, calculated, or dependent on machine learning. Thus, for example (Liu et al., 2012) have proposed an automatic summarization method of Chinese which allows first to recognize the compound words in a document, determines the parts of speech (i.e. words categories such as verb, noun, adjective, etc.) and revises the words segmentation. Then, it determines the keywords and calculates the sentences weights according to the keywords that they contain. The obtained evaluation measures are 68.31% for precision and 66.72% for recall.

Other methods consist in analyzing the Webpage contexts based on its links ((Zhang et al., 2010), (Porselvi1 and Gunasundari, 2013)). Some research works proposed the use of Latent Semantic Analysis (LSA), which is an algebraic-statistical method that extracts and represents the semantic knowledge of the text based on the words co-concurrency. This method constructs a semantic space with a very large dimension from the statistical analysis of all the co-occurrences in a text corpus. The starting point of LSA consists of a lexical table that contains the occurrences of each word in each document. In (Yeh, et al, 2005) a summarizing method using LSA has been proposed. Its main objective is to represent the document as a graph of relations between sentences. A ranking algorithm is then applied to the resulting graph to generate a summary.

Machine learning based methods. These methods try to analyze how a corpus of pairs (document / summary), usually associated manually, can be used to automatically learn rules or techniques for summary generation (Boudin 2018).

(Motta et al., 2011) have proposed a method for the selection of the summary sentences that uses a function that classifies sentences into two groups: important or non-important. Important sentences then form the summary. The analysis of the obtained results showed that the algorithms that produce best results are Naïve Bayes and Support Vector Machine.

(Zhang et al., 2010) proposed a method based on machine learning and NLP techniques to automatically summarize entire websites. The proposed method is based on four steps: retrieving URLs and texts, classifying narrative paragraphs, extracting relevant sentences.

(Parth and Majumder, 2018) studied the roles of three main components of an extraction summary technique: the sentence classification algorithm, the sentence similarity metric and the text representation schema. They showed that using a combination of several similarity measures of different sentences significantly improves the performance of the resulting meta-system. Other researchers have proposed the use of neural models to generate automatic summaries. They consist of "sequence-to-sequence" models in which recurrent neural networks (RNNs) are used for both reading and generating texts (Chopra et al., 2016; Nallapati et al., 2016; Rush et al., 2015, Zeng et al., 2016). The resulting systems are promising even though their performance in terms of ROUGE has not yet reached those of systems based on extraction approach (See et al, 2017).

3 Difficulties of automatic Web page summarization

Basically, Web page summarization techniques are inspired from text summarization ones. However, it is a big challenge to summarize Web pages automatically and effectively because they differ from textual documents both in structure and content. Besides specific problems of automatic text summarization, Web page summarization presents other problem related to:

- The page structure: the web pages contain images, frames, animations, etc. in addition to the textual content.
- The page linguistic form: presence of incomplete sentences, or sentences which do not respect a good linguistic form.

The current search engines (Google, Yahoo, Excite, etc.) generate results that change according to the user query. Indeed, most of these search engines display, as a summary, the first text segment (of the Web page) that contains the most query keywords. Or to be relevant, a summary must not change according to the user query but should be produced by the search engines offline (for example while indexing the web page).

In addition, these search engines consider only the textual content of the web pages and do not consider the non-textual one (images, graphics, videos, etc.).

4 Relevant sentence selection

To select the relevant sentences that will form the summary, we propose to use six numerical criteria. Note that some of these criteria are inspired from (Belguith et al. 2015). However, we suggest in this section revised formulas for computing the sentence scores according to six criteria.

4.1 Title words based criterion (C_1)

Titles are important because they could contain important words. Thus, this criterion favors sentences containing words belonging to titles. This is based on the content of some tags such as <title>, <h1>, <h2>, <h3>. Therefore, the score assigned to a sentence according to this criterion represents the number of the titles words occurring in this sentence:

$$C_1\ (p)= \frac{\text{Number of tittlewordsin sentencep}}{\text{Number of wordsin sentencep}}$$

4.2 Position sentence criterion (C_2)

In the literature, the sentences occurring at the beginning (of the text, the paragraph, etc.) are considered to be more important than the ones occurring at the end. Indeed, one usually pays more attention when writing the beginnings of the texts and the paragraphs. Thus, we propose to calculate the sentence score as follows :

$$\mathbf{C_2}\ (p)= \frac{\text{Number of sentancesin theWeb page}}{\text{Position(p)}}$$

where p is the sentence and position(p) is the position of p in the Web page.

4.3 Keywords based criterion (C_3)

This criterion favors sentences that contain keywords of the Web page. To determine the web page keywords, we propose to use the *tf.idf* technique (Term Frequency Inverse Document Frequency) to assign weights to the terms (words) of a document. According to the tf.idf technique a word is important if it is frequent in the web page and relatively rare in the large collection of web pages linked by hypertext links to the web page in question. We propose to ignore the so-called "empty" words (such as conjunctions, personal pronouns, prepositions, ...) then calculate the *tf.idf* of the remaining terms according to the following formula :

$$w_{ij} = tf_{ij} \times \log \frac{N}{n}$$

where w_{ij} is the weight of term T_j in the page Pg_i
tf_{ij} id the term frequency of T_j in page Pg_i
N is the number of pages linked by hypertext links to the page Pg_i
n is the number of pages where the term T_j occurs at least once.

We calculate for each word of the web page to be summarized, its tf.idf. The only keywords used are those whose tf.idf is greater than the average tf.idf.
Keywords generated from the web page are then used in addition to the keywords that are in the Keyword meta-tag. The keyword meta-tag generally includes terms that are considered, by the Web page author, as the keywords of the web page. The score assigned to a sentence, according to this criterion, is therefore the number of keywords that occurs in this sentence.

$$\mathbf{C_3}\text{ (p)} = \frac{\text{Number of keywords in sentence p}}{\text{Number of words in sentence p}}$$

4.4 Positive/negative terms based criterion (C4)

A positive term (or "Bonus phrase") is a term that represents a word or group of words considered important such as "the main objective", "in conclusion", "it is important to emphasize".
On the other hand, a negative term (or "stigma phrase") represents a word or a group of words considered not important such as "It is not important", "it is difficult to conclude that", "it is impossible to say that ".
Thus, this criterion makes it possible to penalize sentences containing negative terms and to favor sentences containing positive terms.

$$\mathbf{C_4}\text{ (p)} = \frac{\text{Number of positive terms in sentence p - number of negative terms in sentence p}}{\text{Number of terms in sentence p}}$$

4.5 Sentence length based criterion (C5)

Generally, we prefer to put short sentences in the summaries. Thus, this criterion favors short sentences. We determine the average length (in term of words) of the sentences according to the following formula:

AverageLength = Sum of sentences lengths / number of sentences

AverageLength is used as a threshold to calculate the score of a sentence according to the following formula:

if Length(p) ≤ AverageLength then $C_5(p) = 1$ else $C_5(p) = 0$

4.6 Formatting based criterion (C6)

Sentences containing formatting such as a different color, size, style, underlining, etc. will have a higher level of importance than a normal sentence (without particular formatting). Thus, we use three levels of importance: Level 1 (very important, score = 6), Level 2 (moderately important, score = 4) and Level 3 (important, score = 2) The importance level depends on the tags used in a sentence (<b>, <big>, <strong>, <font>, <p>, <div>, <span > tags, ...).

Note that these values and these levels were chosen on the basis of an empirical study that we conducted on a set of web pages.

5 Image selection

Given that images and graphics can be important and can even play the role of a summary when they are expressive, we will consider them in the generated summary. We propose two criteria for selecting images (or graphics): the **Image referring sentence criterion**, and the **Expressive image criterion**.

5.1 Image referring sentence based criterion (C7)

For this criterion, we propose to calculate the score of an image as follows: a sentence that refers to an image (i.e., it contains a linguistic index that points to an image such as "the following image shows") and that is followed by this image is advantaged to others and will have as a score:

C_7 (p) = 1 otherwise C_7 (p) = 0.

Note that if this sentence is retained for the summary, it will be included with the correspondent image/graphic.

5.2 Important image based criterion (C8)

This criterion concerns important images that are not introduced by sentences. In this case, the score of the image will be determined according to the number of key words which appear in the image description. Note that an image in an HTML page is usually described by the Alt attribute of the corresponding IMG tag. In the case where the Alt attribute of the IMG tag is empty, we rely on the hypertext link of the image and we determine the number of keywords contained in the link. In addition, if the image refers to a web page, we take into account its title since we use it to determine the number of keywords (i.e. the title is used instead of the Alt content when this latter is empty). In the three cases, one will obtain a total number of key words Number-Keywords (img) describing the image. The score of the image will be given by the following formula:

C8 (img) = 1 if Number-Keywords (img) (img)> 0 otherwise C8 (img) = 0

6 Sentence score normalisation

In order to allow a homogeneous comparison across the different criteria that have different measurement units, it is necessary to standardize the sentences scores. The goal is to make these scores between 0 and 1.

We propose to normalize by dividing the score of the sentence (according to a given criterion) on the maximum sentences score (according to this same criterion).

$$N_{ij} = \frac{C_{ij}}{Maximum(C_j)} \quad i=1,\ldots, n \; ; j=1,\ldots, q$$

where C_{ij} : the score of sentence i according to criterion j (before normalization)
N_{ij} the score of sentence i according to criterion j (after normalization)

7 Main steps of the proposed method

In order to select the important sentences / images that will appear in the summary of the web page, we use 5 steps (see Fig. 1)

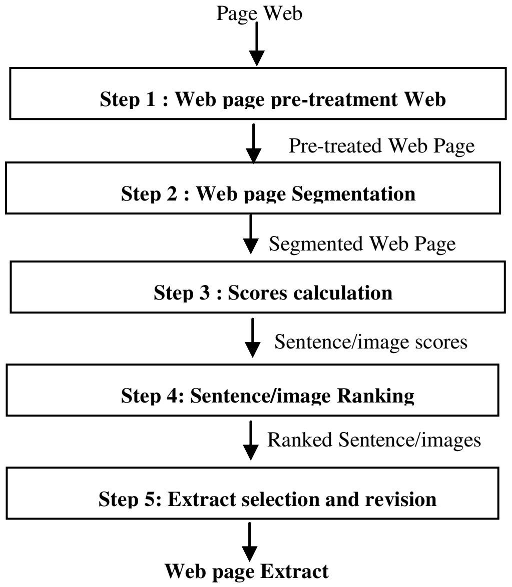
Page Web
Step 1 : Web page pre-treatment Web
Pre-treated Web Page
Step 2 : Web page Segmentation
Segmented Web Page
Step 3 : Scores calculation
Sentence/image scores
Step 4: Sentence/image Ranking
Ranked Sentence/images
Step 5: Extract selection and revision
Web page Extract

Fig. 1. Etapes de la méthode de résumé proposée

Step 1 : Web page pre-treatment.
This pre-treatment step of the Web page consists of deleting Meta tags, applets, and HTML tags that will not be used in the calculation of scores proposed by our method.
Step 2 : Web page segmentation. The pre-treated web page will be segmented into titles, paragraphs, and sentences based on a set of segmentation rules. These rules are based on punctuation (dot, colon, semicolon, etc.) that mark the end of a sentence and on some specific HTML tags: the beginning and ending tags of paragraphs <p> and </ p > or <div ...> and </ div>, <BR>, the title tags <title> tags, <h1> ... <h5>, the list tags<ul>, etc.
Step 3 : Score calculation. For each sentence resulting from the segmentation step, we calculate an overall score which represents the sum of the scores of the weighted criteria. Note that we have proposed 6 criteria to select important sentences (the sentence position criterion, the title-based criterion, the keyword frequency criterion, the positive / negative terms based criterion, the sentence length based criterion and the sentence formatting based criterion) and 2 non-textual criteria for images and graphics: the criterion of a sentence pointing to an image and the criterion of an important image.
Step 4 : Sentence/image ranking. This step consists of ranking sentences and images in descending order of their overall scores. So the one with the highest score will be in first place and so on. An acceptance threshold S is then defined for sentences and images. If the scores of the latter are greater than the threshold, they will be retained at this stage, otherwise they will not be retained and will not be considered in the next step.
Step 5 : Extract selection and revision. This step consists in eliminating the redundant sentences and reordering the remaining ones. We first determine the semantic relations between the sentences retained in step 4. If two or more sentences have a strong semantic relation that exceeds a given threshold, we consider them to be identical and we retain only one (having the best score) to be included in the summary. The eliminated sentence, will be replaced by the sentence whose rank is n + 1 (if we consider that we have retained n sentences in step 4) provided that the latter is not redundant as well. Then, we reorder the list of non-redundant sentences according to their order of appearance in the Web page for a better flow of ideas in the summary.

8 Realisation et evaluation

Our proposed method has been implemented with the JAVA language. Note that in (Belguith et al., 2015) we have proposed a first version of our system. This version has been improved by adopting the new criteria proposed in this paper. It also includes a revision step of the extract.

In order to evaluate our system, we have undertaken manual and automatic evaluation by computing the recall, precision and F-measure (for the manual evaluation) and the ROUGE-2 measurement (for the automatic evaluation). We used a test corpus consisting of 60 Web pages in French, with different themes. We compared the summaries generated by our system to summaries developed by an expert. Table 1. presents the obtained results.

Table 1**.** Evaluation results

Recall	65%
Precision	68%
F-mesure	66.46%
ROUGE-2	72%

Note that the achieved F-measure (i.e. 66.46%) is better than the one obtained with the first version of our system (i.e. 65.38%) proposed in (Belguith et al. 2015). Moreover, the measure of 72% for ROUGE-2 represents the average measurement for the 60 summaries of our test corpus. This value is also very encouraging as the best automatic text abstraction systems have achieved a value around 74%.

9 Conclusion et perspectives

In this paper, we have proposed an original method for automatic summarization of Web pages which has the advantage of generating, for a Web page, a summary in the form of an extract containing the most important sentences and images without processing a deep analysis or understanding. For the sentence extraction we have proposed 6 numeric criteria which are the sentence position, the title words, the key words, the positive/negative terms, the sentence length and its formatting. As for the image selection, we proposed two criteria: the criterion of sentence pointing to an image and the criterion of an important image.

The proposed method has been implemented and evaluated. The obtained evaluation results are very encouraging. Indeed, the Recall, Precision, F-measure and ROUGE-2 measurements are respectively 65%, 68%, 66.46% and 72%. As perspectives, we plan to align sentences according to the time and to solve anaphora to improve the summary quality. We also plan to consider adding an "external image" to the summary from the Internet in case of non-presence of internal images in the web page to be summarized.

References

Belguith, M., Touati, I., Mâaloul, M.H., Keskes, I. : A multi-criteria method for automatic web page summarization. In: Proceedings of the workshop on NLP applications; Completing the puzzle, WNACP@NLDB2, Germany, 2015WNACP@NLDB, (2015).

Bois, R. Leveling, J., Goeuriot, L., Gareth J. F., Jones: Porting a Summarizer to the French Language. In: 21ème Traitement Automatique des Langues Naturelles, Marseille(2014).

Boudin, F: Unsupervised Keyphrase Extraction with Multipartite Graphs. NAACL-HLT (2), 667-672, (2018).

Chopra S., Auli M., and Rush A. M. Abstractive sentence summarization with attentive recurrent neural networks. In North American Chapter of the Association for Computational Linguistics, (2016).

Elvys Linhares P., Huet S. , Torres-Moreno, J.M., Carneiro Linhares, A: Cross-Language Text Summarization Using Sentence and Multi-Sentence compression. In: NLDB, (2018).

Keskes, I.: Discourse Analysis of Arabic Documents and Application to Automatic Summarization. Thèse de Doctorat en Informatique, Université de Sfax (Tunisie) et université Paul Sabatier, France (2015).

Khushboo, R. V. Dharaskar, D. and Chandak, M. B.: Graph-based algorithms for text summarization. In: Third International Conference on Emerging Trends in Engineering and Technology, (2010).

Lin, Z. : Graph-based methods for automatic text summarization. Ph.D. Thesis, School of Computing National University of Singapore (2006).

Liu, X., Zheng, Q., Ma, Q., Lin, G.: A Novel Automatic Summarization Method from Chinese Document. IJCSI International Journal of Computer Science Issues, 9(3), (2012).

Luhn, H. P.: The Automatic Creation of Literature Abstracts. IBM Journal of Research Development 2(2), 159-165, (1958).

Maâloul, M. H: Approche hybride pour le résumé automatique de textes. Application à la langue arabe. Thèse de Doctorat en Informatique, Université de Sfax (Tunisie) et Université de Provence, Aix-Marseille I (2012).

Motta, J., Capus, L. and Tourigny, N.: Insertion of Ontological Knowledge to Improve Automatic Summarization Extraction Methods. Journal of Intelligent Learning Systems and Applications 3(3), 131-138, (2011).

Nallapati, R., Zhou B., Santos C D., Gulçehre C., and Xiang B.. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Computational Natural Language Learning, (2016).

Oufaida, H., Nouali, O. Blache, P.: Résumé Automatique Multilingue Expérimentations sur l'Anglais, l'Arabe et le Français. In: 21ème Traitement Automatique des Langues Naturelles, Marseille (2014).

Patil, K. and Brazdil, P.: Sumgraph: Text summarization using centrality in the pathfinder network. IADIS International Journal of Computer Science Information System 2(?) 18–32, (2007).

Parth, M., Majumder, P.: Effective aggregation of various summarization techniques. Inf. Process. Manage. 54(2): 145-158, (2018).

Porselvi1, A., Gunasundari, S.: Survey on Web page visual summarization. International Journal of Emerging Technology and Advanced Engineering 3(1), (2013).

Rush, A. M., Chopra S. and Weston J.. A neural attention model for abstractive sentence summarization. In Empirical Methods in Natural Language Processing (2015).

Santosh B. and Korra, B. : Automatic Keyword Extraction for Text Summarization: A Survey. http://arxiv.org/abs/1704.03242. (2017)

See A., Liu P.J. , Manning C. D. Get To The Point: Summarization with Pointer-Generator Networks 5th Annual Meeting of the Association for Computational Linguistics Volume 1, Canada (2017).Yeh, J.-Y. Ke, H.-R. , Yang, W.-P. and Meng, I.-H.: Text summarization using a trainable summarizer and latent semantic analysis. Special Issue of Information Processing and Management on An Asian Digital Libraries Perspective 41(?),75–95, (2005).

Zhang, Y., Evangelos Milios, E. Nur Zincir-Heywood, A.: Topic-based web site summarization. IJWIS 6(4), 266-303, (2010).

Zeng W., Luo W., Fidler S., and Urtasun, R. Efficient summarization withread-again and copy mechanism, (2016).