Digital Learning for Summarizing Arabic Documents

Mohamed Mahdi Boudabous[1], Mohamed Hédi Maaloul[2],
and Lamia Hadrich Belguith[1]

[1] ANLP Research Group, MIRACL Laboratory, Faculty of Economic Sciences
and Management of Sfax (FSEGS) - B.P.1088, 3018 Sfax, Tunisia
[2] LPL Laboratory, CNRS-Université de Provence - 5 Avenue Pasteur
13604 Aix en Provence – France
mahdiboudabous@gmail.com, mohamed.maaloul@lpl-aix.fr,
l.belguith@fsegs.rnu.tn

Abstract. We present in this paper an automatic summarization method of Arabic documents. This method is based on a numerical approach which uses a semi-supervised learning technique. The proposed method consists of two phases. The first one is the learning phase and the second is the use phase. The learning phase is based on the Support Vector Machine (SVM) algorithm. In order to evaluate our method, we conducted a comparative study that involves the results generated by our system AIS (Arabic Intelligent Summarizer) with that realized by a human expert. The obtained results are very encouraging and we plan to extend our evaluation on a larger corpus to ensure the performance of our system.

Keywords: Automatic summarization, Arabic documents, Machine Learning, Numerical approaches.

1 Introduction

In the current context, we have to deal with a huge mass of electronic textual documents available through the net. We need tools offering fast visualization of the documents (so that the user can evaluate its relevance). Automatic summarization provides a solution which makes it possible to extract interesting information for an advantageous reuse. Indeed, the summary helps the reader to decide whether the original document contains the required information or not. Moreover, in some cases the reader does not need to read the totality of the original document, simply because the required information is in the summary [1].

Automatic summarization approaches are inspired by various orientations. Some approaches rely on symbolic techniques (based on the analysis of the discourse and its discursive structure), some others are based on numerical treatments (based on statistical, or even on learning) [2].

In addition, the majority of automatic summarization systems mainly treat documents in Indo-European languages such as English and French. To our knowledge, there are only few implementations of these methods on Arabic language, such as LAKHAS [3] and Al Lakas El'eli [4]. Thus, there is an increasing need to develop

H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (Eds.): IceTAL 2010, LNAI 6233, pp. 79–84, 2010.

automatic summarization systems dedicated to Arabic to handle the increasing amount of electronic documents written in Arabic [1].

Thus, the achievements in the field of automatic summarization are generally set out again according to the used approaches. Mainly three approaches are distinguished: numerical, symbolic and hybrid. Our contribution is in the context of numerical approach and we propose a system for the automatic summarization of Arabic documents which is based on a purely Machine learning (ML) technique: ML technique within the framework classification, is shown to be a promising way to combine automatically sentence features [5]. In our method, a classifier is trained to distinguish between two classes of sentences: summary and non-summary ones.

Statistical features that we consider in this work are partly from the state-of-art, and they include cue-sentencess and positional indicators [6], title-keyword similarity [7], and other features.

This paper is structured around four sections: Section 1 presents most related works to ours. Section 2 exposes the proposed method and the summarizing workflow and Section 3 describes the implementation of our approach and the primary results. Section 4 presents the conclusion and the future works.

2 Related Work

Three approaches have been proposed to the summarizing of documents: Linguistic approaches based on a formal representation of knowledge contained in documents or on a reformulation technique. Indeed, these approaches are usually a formal representation of knowledge contained in documents or on reformulation techniques. Numerical approaches are based on calculating a score associated for each sentence to estimate its importance relative to other sentences of the document. This score is calculated by using various statistical methods, probabilistic and learning. Hybrid approaches combine the previous approaches to improve the quality of the summary.

In this paper, we explore a numerical approach and present some examples. Numerical approaches are essentially based on calculating a score associated for each sentence to estimate its importance. The final summary will only keep the sentences that have the highest scores.

There are two main techniques: statistical and learning techniques. Recently, various authors have explored Machine Learning techniques to summarize documents [7]. This is thanks to the best performance of these techniques.

The learning techniques are classified into three classes. The first class is the supervised learning, this class is based on two phases: the learning phase that use a training corpus of a very large size and the validation phase that use another corpus called validation corpus [8]. The second class is the semi-supervised learning that has only a learning phase; this phase requires a training corpus of small size [9] [10]. The third one is the non-supervised learning, which does not require either a training corpus or a validation corpus.

The numerical approaches can be applied to all types of corpus and can operate in a big number. The most important systems which are based on the numerical approaches are: LAKHAS system [3] which summarizes Arabic documents in XML format. CBSEAS “Clustering-Based Sentence Extractor for Automatic Summarization” system

[11] treats the case of multi-document summary. Its principle is that the more redundant information are the more important they will be.

Our method treats the numerical approaches that have proven their effectiveness in other languages. More precisely, we use Machine learning techniques based on semi-supervised learning; this choice is justified by the fact that it allows involving a system with only a small number of labeled sentences and a large number of not labeled ones.

3 Proposed Method

In this section, we present an overview of the proposed method and the summarizing workflows for the HTML documents.

3.1 An Overview of Our Method

We propose a new method for the automatic summarization of the newspaper articles in Arabic language. It is based on a Machine learning technique. More precisely it is based on the semi-supervised learning technique which is composed of two phases: the first one is the learning phase which allows the system to learn how to extract summary sentences. We use Support Vector Machines algorithm (SVM) for this phase. The second phase is the use phase which allows users to summarize a new document. Fig. 1 presents the details of the proposed method and the two phases.

3.2 Summarization Workflow

3.2.1 The Learning Phase

In this phase, the system designer should provide the training corpus and the extraction features to perform the learning.

The training corpus is composed of the source documents and their summaries. All the documents are initially pretreated to prepare their segmentation in titles, sections, paragraphs and sentences. This segmentation is based on the criteria of punctuation and HTML tags. After the segmentation step, each sentence of the segmented document will be notified according to some features. This step leads to the construction of a set of the vectors V corresponding to the values of the specific features to the sentence. These vectors are called extraction vectors or score vectors. Each vector is associated with a Boolean criterion which indicates the sentence class: summary or non-summary.

The extraction vector has the following structure: V1 (*S1, S2, S3... Sn*), where *Si* is the score of the criterion *i* and *n* is the number of the criteria.

In the learning phase, extraction vectors are combined to associate a score with each feature and generate rules.

3.2.2 The Use Phase

In this phase, the user provides a HTML document as an input for the system. This document is segmented and notified in order to generate a set of extraction vectors. The system uses the generated rules to classify each sentence.

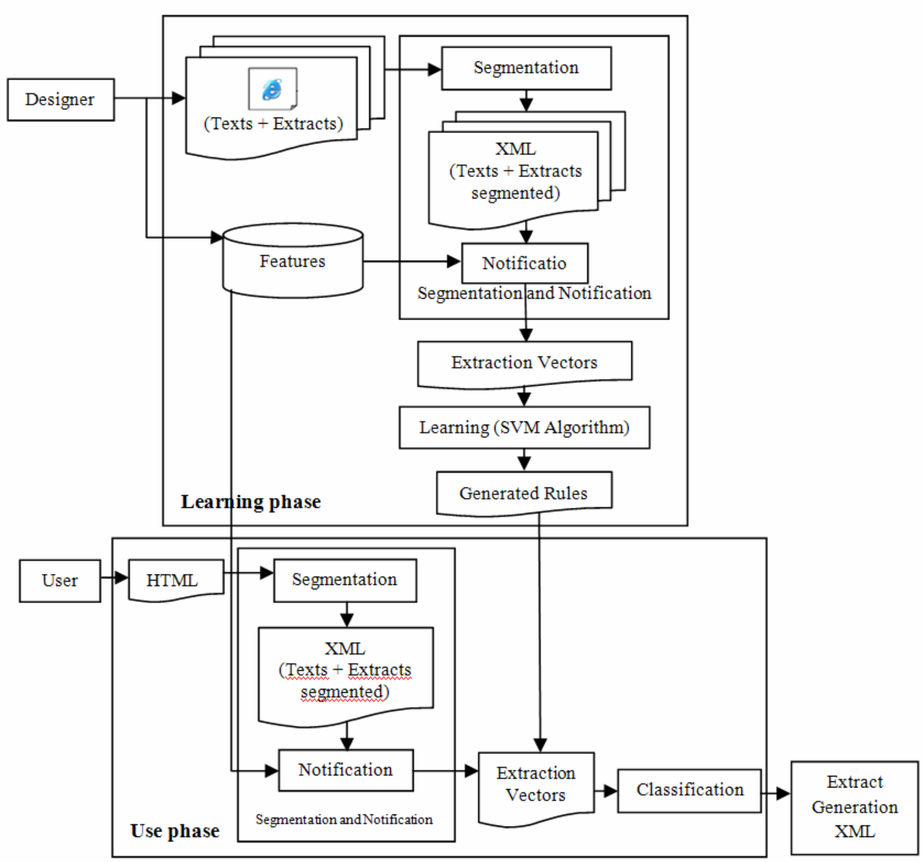

Fig. 1. The principle of the proposed method

4 The AIS System

The method that we proposed for automatic summarization of Arabic documents has been implemented through the AIS (Arabic Intelligent Summarizer) system. In this section, we present the implementation details and the preliminary results.

4.1 Implementation Details

Our corpus is composed of 500 Arabic documents collected from the web. These documents represent newspaper articles selected according to various orientations (sport, economy, education, etc.). The newspaper articles are of HTML type with a UTF-8 coding. The summaries of these documents are produced by three human experts. Then, we use the index of kappa[1] to calculate the similarity between human experts and generate one summary for each document.

[1] http://kappa.chez-alice.fr/

After the segmentation step, we use 15 features to classify each sentence. Some of these features are detailed in Table 1.

Table 1. Features details

Features	Details
Position in the text	Indicates the position of the sentence in the text.
First sentence in the section	Indicates if the sentence is the first in the section or not.
First sentence in the paragraph	Indicates if the sentence is the first in the paragraph or not.
Range of the paragraph	Indicates the range of paragraph that contains the sentence.
Tf_idf score	Calculates the tf*idf of the score.
Tf score	Calculates the Tf of the score.
Title keywords	Presents the number of title keywords in the sentence.
Indicative expressions	Presents the number of indicative expressions in the sentence.

Finally, we obtain a file that contains the set of extraction vectors which constitute the input of the learning phase. In the learning phase, we use the SVM algorithm to learn how to classify the summary and non-summary sentences. At the end of the learning phase, a score is associated with each feature. Some features can have a score of zero. The SVM algorithm generates a rule by summing scores associated with each feature.

The system uses the generated rules to calculate the score of each sentence. If the score is positive, the sentence will be considered as a summary sentence, otherwise the sentence is considered as a non-summary sentence. Finally, the system combines summary sentences to obtain the summary.

4.2 Preliminary Results

We used 60 documents of our corpus to experiment our system (i.e. 50 documents for the learning phase and 10 documents for the evaluation phase). The obtained summaries are compared to the human summaries. The average measures for Precision, Recall and F-measure are respectively 0.992, 0.991and 0.991 (see Table 2).

Table 2. Evaluation results

	Precision	Recall	F-measure
Weighted Avg	0.992	0.991	0.991

5 Conclusion and Future Work

In this paper, we have proposed a method for automatic summarization of Arabic documents. Our method is implemented by AIS system and is based on the Machine learning technique. Our work focuses on a particular type of documents (i.e., the newspaper articles in HTML format). We believe that the preliminary results are very encouraging. Indeed the F-measure is equal to 0.991. We note that we used a small corpus for the evaluation but as perspectives, we plan to extend the evaluation on a larger corpus.

We also intend to apply the proposed method for other types of documents such as XML and TXT.

References

1. MaÃloul, M.H., Ellouze Khemakhem, M., Belguith Hadrich, L.: Al Lakas El'eli /ÇááÎÇÕ ÇáÂáí: Un système de résumé automatique de documents arabes. International Business Information Management Association (IBIMA 2008) (2008)
2. Amini, M.R., Gallinari, P.: Apprentissage numérique pour le résumé de texte. Les Journées d'Étude de l'ATALA, Le résumé de texte automatique: solutions et perspectives (2003)
3. Douzidia, S., Lapalme, G.: Lakhas, an Arabic summarization system. In: Proceedings of the HLT-NAACL Workshop on Text Summarization DUC 2004 (2004)
4. Maâloul, M.H., Ellouze Khemakhem, M., Belguith Hadrich, L.: Proposition d'une méthode de résumé automatique de documents arabes. GEI 2006 (2006)
5. Kupiec, J., Pedersen, J., Chen, F.: A trainable document summarizer. In: Proceedings of the 18th ACM SIGIR Conference (1995)
6. Luhn, H.P.: The automatic creation of literature abstracts. IBM Journal of Research and Development (1958)
7. Alrahabi, M., Mourad, G., Djioua, B.: Filtrage sémantique de textes en arabe en vue d'un prototype de résumé automatique. In: JEP/TALN 2004 (2004)
8. Mani, I., Bloedorn, E.: Machine Learning of Generic and User-Focused Summarization. In: Proceedings of the Fifteenth National Conference of Artificial Intelligence, AAAI 1998 (1998)
9. Amini, M.R.: Apprentissage automatique et recherche de l'information: application à l'extraction d'information de surface et au résumé de texte. Thèse de doctorat (2001)
10. Amini, M.R., Gallinari, P.: The Use of Unlabeled Data to Improve Supervised Learning for Text Summarization. In: SIGIR (2002)
11. Bossard, A., Généreux, M., Poibeau, T.: CBSEAS, a Summarization System Integration of Opinion Mining Techniques to Summarize Blogs (2009)