Automatic Headline Generation using Character Cross-Correlation

Fahad A. Alotaiby
Department of Electrical Engineering,
College of Engineering, King Saud University
P.O.Box 800, Riyadh 11421, Saudi Arabia
falotaiby@hotmail.com

Abstract

Arabic language is a morphologically complex language. Affixes and clitics are regularly attached to stems which make direct comparison between words not practical. In this paper we propose a new automatic headline generation technique that utilizes character cross-correlation to extract best headlines and to overcome the Arabic language complex morphology. The system that uses character cross-correlation achieves ROUGE-L score of 0.19384 while the exact word matching scores only 0.17252 for the same set of documents.

1 Introduction

A headline is considered as a condensed summary of a document. It can be classified as the acme of text summarization. The necessity for automatic headline generation has been raised due to the need to handle huge amount of documents, which is a tedious and time-consuming process. Instead of reading every document, the headline can be used to decide which of them contains important information.

There are two major disciplines towards automatic headline generation: extractive and abstractive. In the work of (Douzidia and Lapalme, 2004), and extractive method was used to produce a 10-words summary (which can be considered as a headline) of an Arabic document, and then it was automatically translated into English. Therefore, the reported score reflects the accuracy of the generation and translation which makes it difficult to evaluate the process of headline generation of this system. Hedge Trimmer (Dorr *et al.*, 2003) is a system that creates a headline for an English newspaper story using linguistically-motivated heuristics to choose a potential headline. Jin and Hauptmann (2002) proposed a probabilistic model for headline generation in which they divide headline generation process into two steps; namely the step of distilling the information source from the observation of a document and the step of generating a title from the estimated information source, but it was for English documents.

1.1 Headline Length

One of the tasks of the Document Understanding Conference of 2004 (DUC 2004) was generating a very short summary which can be considered as a headline. The evaluation was done on the first 75 bytes of the summary. Knowing that the average word size in Arabic is 5 characters (Alotaiby *et al.* 2009) in addition to space characters, the specified summary size in Arabic words was roughly equivalent to 12 words. In the meantime, the average length of the headlines was about 8 words in the Arabic Gigaword corpus (Graff, 2007) of articles and their headlines. In this work, a 10-words headline is considered as an appropriate length.

1.2 Arabic Language

Classical Arabic writing system was originally consonantal and written from right to left. Every letter in the 28 Arabic alphabets represents a single consonant. To overcome the problem of different pronunciations of consonants in Arabic text, graph-

Proceedings of the ACL-HLT 2011 Student Session, pages 117–121,
Portland, OR, USA 19-24 June 2011.

ical signs known as diacritics were invented in the seventh century. Currently in the Modern Standard Arabic (MSA), diacritics are omitted from written text almost all the time. As a result, this omission increases the number homographs (words with the same writing form). However, Arab readers normally differentiate between homographs by the context of the script.

Moreover, Arabic is a morphologically complex language. An Arabic word may be constructed out of a stem plus affixes and clitics. Furthermore, some parts of the stem may be deleted or modified when appending a clitic to it according to specific orthographical rules. As a final point, different orthographic conventions exist across the Arab world (Buckwalter, 2004). As a result of omitting diacritics, complex morphology and different orthographical rules, two same words may be regarded as different if compared literally.

2 Evaluation Tools

Correctly evaluating the automatically generated headlines is an important phase. Automatic methods for evaluating machine generated headlines are preferred against human evaluations because they are faster, cost effective and can be performed repeatedly. However, they are not trivial because of various factors such as readability of headlines and adequacy of headlines (whether headlines indicate the main content of news story). Hence, it is hard for a computer program to judge. Nevertheless, there are some automatic metrics available for headline evaluation. F1, BLEU (Papineni *et al.* 2002) and ROUGE (Lin, 2004*a*) are the main metrics used.

The evaluation of this experiment was performed using Recall-Oriented Understudy for Gisting Evaluation (ROUGE). ROUGE is a system for measuring the quality of a summary by comparing it to a correct summary created by human. ROUGE provides four different measures, namely ROUGE-*n* (usually *n* = 1,2,3,4), ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU. Lin (2004*b*) showed that ROUGE-1, ROUGE-L, ROUGE-SU, and ROUGE-W were very good measures in the category of short summaries.

3 Preparing Data

The dataset used in this work was extracted from Arabic Gigaword (Graff, 2007). The Arabic Gigaword is a collection of text data extracted from newswire archives of Arabic news sources and their titles that have been gathered over several years by the Linguistic Data Consortium (LDC) at the University of Pennsylvania. Text data in the Arabic Gigaword were collected from four newspapers and two press agencies. The Arabic Gigaword corpus contains almost two million documents with nearly 600 million words. For this work, 260 documents were selected from the corpus based on the following steps:

- 3170 documents were selected automatically according to the following:
  i. The length of the document body is between 300 to 1000 words
  ii. The length of the headline (hereafter called original headline) was between 7 to 15 words.
  iii. All words in the original headline must be found in the document body.
- 260 documents were randomly selected from the 3170 documents.

After automatically generating the headlines, 3 native Arabic speaker examiners were hired to evaluate one of the generated headlines as well as the original headline. Also, they were asked to generate 1 headline each for every document. These new 3 headlines will be used as reference headlines in ROUGE to evaluate all automatically generated headlines and the original headline.

4 Headline Extraction Techniques

The main idea of the used method is to extract the most appropriate set of consecutive words (phrase) from a document body that should represent an adequate headline for the document. Then, evaluate those headlines by calculating ROUGE score against a set of 3 reference headlines.

To do so, first, a list of nominated headlines was created from the document body. After this, four different evaluation methods were applied to choose the best headline that reflects the idea of the document among the nominated list. The task of these methods is to catch the most suitable headline that matches the document. The idea here is to

choose the headline that contains the largest number of the most frequent words in the document taking into account ignoring stop words and giving earlier sentences in documents more weight.

4.1 Nominating a List of Headlines

A window of a length of 10-words was passed over the paragraphs word by word to generate chunks of consecutive words that could be used as headlines. Moving the widow one word step may corrupt the fluency of the sentences. A simple approach to reduce this issue is to minimize the size of paragraphs. Therefore, the document body was divided into smaller paragraphs at new-line, comma, colon and period characters. This step increased the number of nominated headlines with proper start and end. The resulting is a nominated list of headlines of a length of 10 words. In the case of a paragraph of a length less than 10, there will be only one nominated headline of the same length of that paragraph.

Table 1 shows an example of nominating headline list where *a* is the selected paragraph, *b* is the first nominated headline and *c* is the second nominated headline. Nominated headlines *b* and *c* are word-by-word translated.

a	ارتبطت نشأة المخطوطات العربية في السودان ببروز معالم الثقافة العربية الإسلامية،
	The emerging of the Arabic manuscripts in Sudan was associated with the rise of the formation of Arabic-Islamic culture,
b	ارتبطت نشأة المخطوطات العربية في السودان ببروز معالم الثقافة العربية
	Associated emerging manuscripts Arabic in Sudan with-rise formation culture Arabic
c	نشأة المخطوطات العربية في السودان ببروز معالم الثقافة العربية الإسلامية
	Emerging manuscripts Arabic in Sudan with-rise formation culture Arabic Islamic

Table 1: An example of headlines nomination.

4.2 Calculating Word Matching Score

The very basic process of making a matching score between every two words in the document body is to give a score of 1 if the two words exactly match or 0 if there is even one mismatch character. This basic step is called the Exact Word Matching (EWM). Unfortunately, Arabic language contains clitics and is morphologically rich. This means the same word could appear with a single clitic attached to it and yet to be considered as a different word in the EWM method. Therefore, the idea of using Character Cross-Correlation (CCC) method emerged. In which a variable score in the range of 0 to 1 is calculated depending on how much characters match with each other. For example, if the word "وكتبها" "*and he wrote it*" is compared with the word "كتب" "*he wrote*" using the EWM method the resulting score will be 0, but when using the CCC method it will be 0.667. The CCC method comes from signals cross-correlation which measures of similarity of two waveforms. In the CCC method the score is calculated according to the following equation:

$$CCC_{w_i,w_j} = \frac{2\max_n c[n]}{M+N} \quad (1)$$

and

$$c[n] = \sum_{m=-(N-1)}^{M-1} w_i[m] * w_j[n+m] \quad (2)$$

where w_i is the first word containing M characters, w_j is the second word containing N characters and the operation * result *1* if the two corresponding characters match each other and *0* otherwise.

4.3 Calculating Best Headline Score

After preparing the two tables of words matching score, now they will be utilized in the selection of the best headline. Except stop-words, every word in the document body (w_d) will be matched with every word in the nominated headline (w_h) using the CCC and the EWM methods and a score will be registered for every nominated sentence. A simple stop-word list consisting of about 180 words was created for this purpose. Calculating matching score for every sentence is also performed in two ways. The first way is the SUM method which is defined in the following equation:

$$SUM_p = \sum_{i=1}^{L} \sum_{j=1}^{K} CCC_{w_d,w_j} \quad (3)$$

where SUM_p is the score using SUM method for the nominated headline p, K is the size of unique words in the document body and L is the size of words in the nominated headline (except stop-words).

In this method the summation of the cross-correlation score of every word in the document body and every word in the headline is added up.

In a similar way, in the other method MAX_p the maximum score between every word in the document body and the nominated headline is added up. Therefore, for every word in the document, its maximum matching score will be added in either cases, CCC or EWM. And it can be defined in the following equation:

$$MAX_p = \sum_{i=1}^{L} \max_j CCC_{w_d,w_j} \quad (4)$$

SUM_p and MAX_p were calculated using EWM and CCC method resulting four different variation of the algorithm namely SUM-EWM, SUM-CCC, MAX-EWM and MAX-CCC.

4.4 Weighing Early Nominated Headlines

In the case of news articles usually the early sentences absorb the subject of the article (Wasson, 1998). To reflect that, a nonlinear multiplicative scaling factor was applied. With this scaling factor, late sentences are penalized. The suggested scaling factor is inspired from sigmoid functions and described in the following equations.

$$SF = -\left(\frac{e^z-1}{e^z+1} - 1\right)/2 \quad (5)$$

where

$$z = 5\left(\frac{2r}{S} - 1\right) \quad (6)$$

and r is the rank of the nominated headline and S is the total number of sentences.

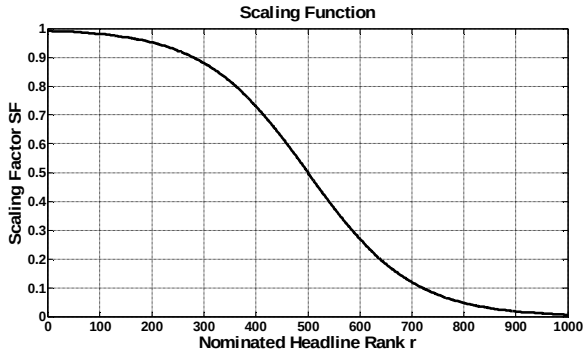

Figure 1: Scaling function of a 1000 nominated headline document.

According the nominating mechanism hundreds of sentences could be nominated as possible headlines. Figure 1 shows the scaling function of a one thousand nominated headlines. After applying the scaling factor, the headline with the maximum score was chosen.

5 Results

Table 2 shows the ROUGE-1 and ROUGE-L scores on the test data. ROUGE-1 measures the co-occurrences of unigrams where ROUGE-L is based on the longest common subsequence (LCS) of an automatically generated headline and the reference headlines.

It is clear that the MAX-CCC scores the highest result in the automatically generated headlines. Unfortunately there are no available results on an Arabic headline generation system to compare with and it is not right to compare these results with other systems applied on other languages or different datasets. So, to give ROUGE score a meaningful aspect, the original headline was evaluated in addition to randomly selected 10 words (Rand-10) and the first 10 words (Lead-10) in the document.

Method	ROUGE-1 (95%-conf.)	ROUGE-L (95%-conf.)
Rand-10	0.08153	0.07081
Lead-10	0.18353	0.17592
SUM-EWM	0.11006	0.10624
SUM-CCC	0.18974	0.17944
MAX-EWM	0.18279	0.17252
MAX-CCC	0.20367	0.19384
Original	0.37683	0.36329

Table 2: ROUGE scores on the test data.

From the registered results it is clear that the MAX-CCC has overcome the problem of the rich existence of clitics and morphology.

6 Conclusions

We have shown the effectiveness of using character cross-correlation in choosing the best headline out of nominated sentences from Arabic document. The advantage of using character cross-correlation is to overcome the complex morphology of the Arabic language. In the comparative experiment, character cross-correlation got ROUGE-L=0.19384 and outperformed the exact word match which got ROUGE-L= 0.17252. Therefore, we conclude that character cross-correlation is effective when com-

paring words in morphologically complex languages such as Arabic.

Acknowledgments

I would like to thank His Excellency the Rector of King Saud University Prof. Abdullah Bin Abdulrahman Alothman for supporting this work by a direct grant. I would also like to thank Dr. Salah Foda and Dr. Ibrahim Alkharashi, my PhD supervisors, for their help in this work.

References

Bonnie Dorr, David Zajic and Richard Schwartz. Hedge Trimmer: A Parse-and-Trim Approach to Headline Generation. In Proceedings of the HLT-NAACL 2003 Text Summarization Workshop and Document Understanding Conference (DUC 2003), Edmonton, Alberta, 2003.

Chin-Yew Lin, ROUGE: a Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out, pages 56-60, Barcelona, Spain, July, 2004*a*.

Chin-Yew Lin, Looking for a few Good Metrics: ROUGE and its Evaluation, In Working Notes of NTCIR-4 (Vol. Supl. 2), 2004*b*.

Document Understanding Conference, http://duc.nist.gov/duc2004/tasks.html, 2004.

Fahad Alotaiby, Ibrahim Alkharashi and Salah Foda. Processing large Arabic text corpora: Preliminary analysis and results. In Proceedings of the Second International Conference on Arabic Language Resources and Tools, pages 78-82, Cairo, Egypt, 2009.

Fouad Douzidia and Guy Lapalme, Lakhas, an Arabic summarization system. In Proceedings of Document Understanding Conference (DUC), Boston, MA, USA, 2004.

David Graff. Arabic Gigaword Third Edition. Linguistic Data Consortium. Philadelphia, USA, 2007.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002.

Mark Wasson. Using Lead Text for news Summaries: Evaluation Results and Implications for Commercial Summarization Applications. In Proceedings of the 17th International Conference on Computational liguistics, Montreal, Canada, 1998.

Rong Jin, and Alex G. Hauptmann, A New Probabilistic Model for Title Generation, The 19th International Conference on Computational Linguistics, Academia Sinica, Taipei, Taiwan, 2002.

Tim Buckwalter. Issues in Arabic Orthography and Morphology Analysis. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, Geneva, Switzerland, 2004.

Zajic. D., Dorr. B. and Richard Schwartz. Automatic Headline Generation for Newspaper Stories. In Workshop on Automatic Summarization, pages. 78-85, Philadelphia, PA, 2002.