Lakhas, an Arabic summarization system

Fouad Soufiane Douzidia
RALI-DIRO Université de Montréal
CP 6128, Succ Centre-ville
Montréal, Québec, Canada, H3C 3J7
douzidif@iro.umontreal.ca

Guy Lapalme
RALI-DIRO Université de Montréal
CP 6128, Succ Centre-ville
Montréal, Québec, Canada, H3C 3J7
lapalme@iro.umontreal.ca

Abstract

This paper describes the Arabic summarization system that we have developed and evaluated on the very short summary of noisy text task of DUC2004. We describe the structure of the system and the various compaction techniques we developed in order to produce 10 words summaries of news articles. We also present the score we obtained using two different machine translation systems.

1 Introduction

The goal of a summary is to produce a short representation of a long document. This problem can be solved by building an abstract representation of the whole document and then generating a shorter text or by selecting a few relevant sentences of the original text. Given that the former method, called extraction, has already proven its success in the past for English texts, we decided to adapt it to Arabic.

This paper describes a new Arabic text summarization system using extraction techniques. It is the first Arabic summarization system to be formally evaluated and compared with English competitors in an evaluation competition.

Section 2 describes *Lakhas*[1], an Arabic summarization system that we developed for our participation to *Task 3* of the Document Evaluation Conference (DUC2004). Section 3 describes the modules of *Lakhas* and section 4 shows the compaction techniques we developed for Arabic. Section 5 presents different steps we used to translate our Arabic summaries with commercial web translation system. Section 6 gives the evaluation results we obtained at DUC2004 and show some translations errors. Section 7 shows the new better results obtained by the English translation produced by one of the MT systems used by other participants and we explain the reasons behind the rise in the Rouge evaluation. We conclude by indicating future work needed to obtain a complete Arabic summarization system.

2 Source and target documents

Document Understanding Conferences is an annual evaluation in the area of text summarization organized by the American National Institute of Standards and Technology (NIST). In its 2004 evaluation, NIST decided to include two tasks (3 and 4) to explore summarization from noisy input produced by Arabic to English machine translation. So the proposed scenario by the DUC organizers was the following: the original Arabic texts are first translated to English by a Machine Translation (MT) system. The resulting English text is used as a source for the summarization system in order to obtain a very short summary (≤ 75 bytes) of the document in English. Two MT systems were used: one from the Information Science Institute (ISI) of the University of Southern California and another one developed by an IBM team.

A preliminary study of a small sample of examples of translated documents revealed a number of shortcomings:

- English texts are hardly understandable without the corresponding Arabic texts
- MT systems often ignored important information for example in *The Agency said that Ibrahim, in the event at the level of cooperation and trade between Iraq and Saudi Arabia*[2] the verb *appreciated* has been omitted after the word *event*; and even with it the text is still hard to understand.

[1] Corresponding roughly to *summarize* in Arabic

[2] Our literal translation of the original Arabic text is the following: *And the agency said that Ibrahim at this occasion appreciated the level of cooperation and commercial exchange between Iraq and Saudi Arabia*.

- MT systems often translated the same Arabic word into different English words for example منازل was translated by *home* in one sentence and by *workers* in another.

This is why we decided to follow another path: summarize the Arabic text directly and only translate the summarized text. We thus have less text to translate but more importantly we work directly with the original documents, which are then less noisy in the hope of getting better results. During the DUC2004 evaluation we analyzed 240 documents and produced the corresponding very short summaries. Given that the evaluation of DUC2004 was done on the first 75 bytes of the English text, we had to rely on various heuristics (see section 4) in order to produce an Arabic summary that would fit within this size constraint once translated.

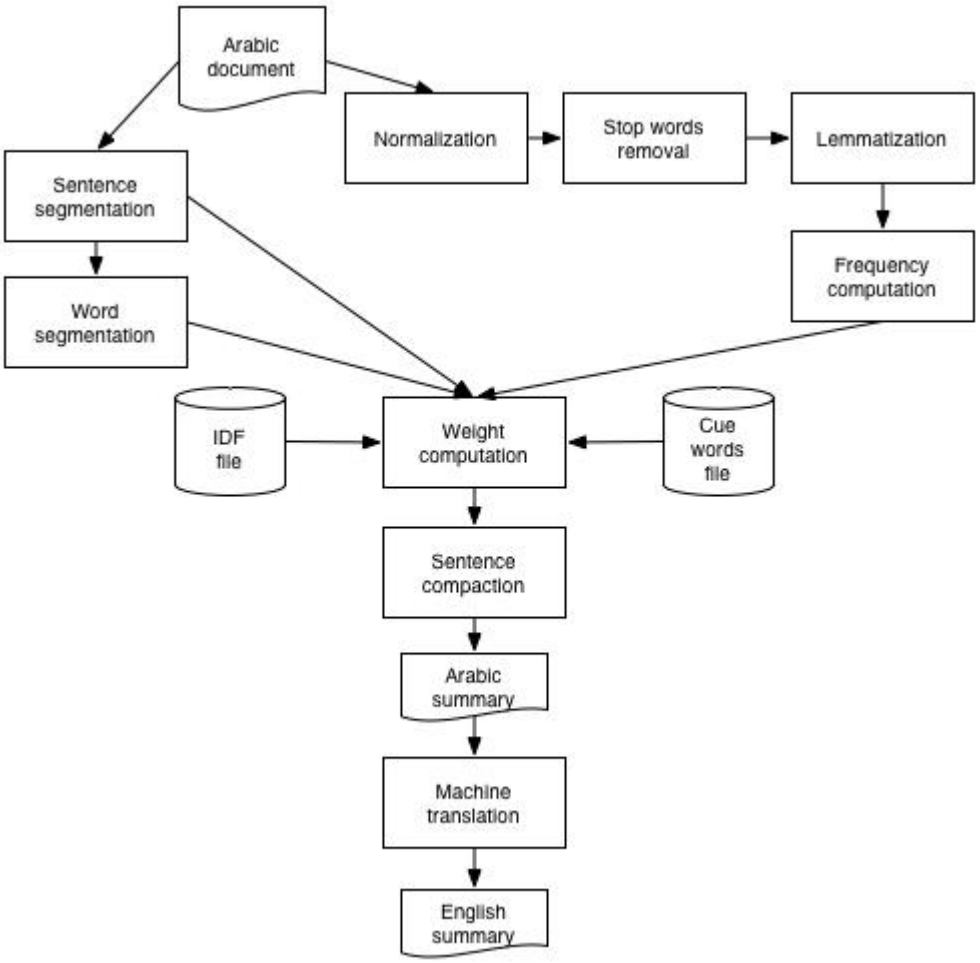

Figure 1: Modules of Lahkas used in the DUC competition.

3 Architecture of the System

Figure 1 gives a functional view of Lakhas in terms of the modules that we now briefly describe:

Sentence segmentation extracts each sentence and headline from the original XML documents and assign preliminary scores according to their position in the document.

Word segmentation (tokenization) determines the boundaries of each word and computes the frequency of each word in a sentence.

Normalization replaces some variants of characters by a single one (examples إ, آ and أ were replaced by ا, ى was replaced by ي, …)

Stop words removal for identifying the most frequent important words.

Lemmatization is a daunting and delicate task in Arabic because it is highly inflectional and derivational language (Attia 2000); irregular plurals are common and do not obey normal morphological rules; the absence of diacritics creates ambiguity and therefore complex morphological rules are required to identify the tokens. Moreover capitalization is not used in Arabic which makes it hard to identify proper names, acronyms, and abbreviations (Xu & al. 2002). For resolving ambiguity, (Aljlayl & al. 2002) show that the light stemming (approach based on suffix and prefix removal) significantly outperforms the root-based in information retrieval. Although we could have thought that root finding might be better because its is more semantically motivated, its also induces more noise and is thus less precise for Information Retrieval. The root-based algorithm merges many different related terms into a single one (eg. writer, book, desk, write, …). (كاتب , كتاب , مكتب , كتب). This is why for summarization, we also decided to use simple prefix and suffix removal (Darwish 2002). Given the fact that most Arabic words have a three or four letter root, we make sure that at least three characters are kept in each word to preserve the integrity.

Frequency computation determines which words are significant in the document.

Indicative expressions (Cue words) increase the weight of certain sentences that might bring some useful information.

Weight computation of each sentence *S* is obtained by combining the value of 4 scores:

$$Sc = \alpha_1 Sc_{lead} + \alpha_2 Sc_{title} + \alpha_3 Sc_{cue} + \alpha_4 Sc_{tf \cdot idf}$$

where

Sc_{lead} is 2 if the sentence is the first one and 1 otherwise

$$Sc_{title} = \sum_{w \in S} a(w).tf(w)$$

$a(w)$ is 2 if word *w* appears in the title of the article and 0 otherwise
$tf(w)$ is the frequency of *w* in the sentence.

$$Sc_{cue} = \sum_{w \in S} c(w).tf(w)$$

$c(w)$ is 1 when *w* appears in the list of indicative words and 0 otherwise.

$$Sc_{tfi \cdot df} = \frac{1}{|S|} \sum_{w \in S} \frac{tf(w) - 1}{tf(w)} \log \frac{DN}{df(w)}$$

$tf(w)$ is the frequency of *w* in *S*.
DN is the total number of documents in the corpus.
$df(w)$ is the number of documents in which *w* occurs.

For DUC2004, we set all α_i to 1 but we intend to experiment with different values.

Sentence extraction and compaction. The above steps are sufficient for short summaries of a few sentences, the table 1 gives an idea on the number of words and compression ratio generated by Lakhas.

As we can see in the 3rd column of table 1, summaries generated by extraction have 29 words on average, but DUC evaluation called for summaries of only about 10 words. So we had to develop further special purpose compression procedures, described in the next section, in order to satisfy this constraint.

DOCSET	Source	Summary	Very short summ	Summary/Source	VS Summ/Source	Ajeeb Translation	ISI Translation
D1001	59	26	14	44%	23%	20	16
D1003	130	29	12	22%	9%	18	15
D1005	180	29	15	16%	8%	24	22
D1011	297	29	17	10%	6%	28	24
D1012	207	32	13	15%	6%	18	14
D1014	180	26	11	15%	6%	17	13
D1016	167	34	18	20%	11%	27	24
D1018	149	31	13	21%	9%	19	16
D1019	294	27	15	9%	5%	24	19
D1023	239	29	15	12%	6%	22	18
D1038	100	29	14	29%	13%	20	19
D1043	201	28	13	14%	6%	20	16
D30002	146	32	16	22%	11%	21	18
D30003	174	30	17	17%	10%	28	22
D30033	223	36	17	16%	8%	25	20
D30040	291	27	15	9%	5%	20	16
D30042	177	30	14	17%	8%	18	17
D30053	194	27	15	14%	8%	22	18
D31001	222	31	13	14%	6%	21	16
D31009	197	27	16	14%	8%	24	20
D31016	90	27	16	30%	18%	25	19
D31022	175	30	15	17%	9%	20	20
D31029	148	31	14	21%	10%	22	17
D31043	256	29	12	11%	5%	18	17
Mean	**187**	**29**	**15**	**16%**	**8%**	**21**	**18**

Table 1: Per docset average of number of Arabic words for source and summaries.
The last two columns give the number of words in the corresponding English.

4 Sentence reduction methods

Four methods were applied in order to reduce the length the summary generated by extraction methods.

Name substitution by removing the position name, for example within *the United Nations Secretary-General, Kofi Anan* we only keep *Kofi Anan.*

Removal of some type of words such as days of the week or months, numbers written in full, adverbs, some subordination conjunctions, etc… since they do not add substantial information.

Removal of part of sentence following some boundaries, such conjunctions of coordination or subordination (*and, with, as,*…), or adjective like (*since, during, which, that*,….)

Removal of indirect discourse construction by keeping only the informative fact using the patterns given in table 2.

English Pattern	Arabic pattern
X declared … that **R**	أعلن X… ان **R**
X reported … that **R**	أفاد X… ان **R**
X declared … that it is **R**	أعلن X… انه **R**

Table 2: Some reduction patterns with their English translation, only **R** is kept.

In some cases, instead of naming an entity, we refer to the speaker. For example in the case of a country as an entity, instead of saying *Iraqi Vice President declared today, Sunday that Iraq rejects cooperating with…* we instead have … *Iraqi Vice President declared today, Sunday that his country rejects cooperating with* … the use of the 1st pattern in this case will keep *his country rejects cooperating with* … in which an important information (whose country?) was omitted.

The words انه / بانه used as border of the 3rd pattern in table 2 can have two interpretations:

i. As neutral demonstrative pronoun (this/it, that this/it), which doesn't influence the results.

ii. As subordinating conjunction (that) with the personal pronoun (he) where ه indicate (he) in which case, this would leave the sentence incomplete after the reduction.

Example:

initial text	اعلن رئيس الحكومة الجزائرية مولود حمروش لوكالة فرانس برس انه مرشح للانتخابات الرئاسية
English translation of initial text	The Algerian Prime Minister Mouloud Hamrouche declared to the agency France Presse that he is a candidate for the presidential elections
after compaction	مرشح للانتخابات الرئاسية
after compaction in English	is a candidate for the presidential elections

Table 3: example of incomplete information by using the 3rd pattern in table 2.

By applying above reduction methods, we were able to reduce the summaries by approximately 50% further, as shown in column 5 of Table 1. These four reduction methods gave us summaries of around 15 words.

5 Arabic to English Translation

In order to compare our results with the ones of other teams at DUC, we had our Arabic summaries translated by Ajeeb (http://english.ajeeb.com) a commercial web translation system. In the next section we will describe the results we obtained after the competition with the ISI machine translation of our Arabic summaries.

For DUC, Lakhas produced a file regrouping the summaries of the documents of each docset, separated by tags using the same references as those of NIST.

To translate the file with Ajeeb MT, we had to:

- Encode the documents in Windows CP-1256 format.
- Generate a web page and send its URL to Ajeeb
- Transform the translated web page in a single XML file conforming to the DUC DTD

6 Results for Task 3

NIST evaluated the English summaries by using ROUGE (Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics)

ID	ROUGE-1		ROUGE-2		ROUGE-3		ROUGE-4		ROUGE-L		R-W-1.2	
model	*0.395*		*0.147*		*0.064*		*0.027*		*0.344*		*0.200*	
142	0.218	6	0.076	1	0.029	1	0.010	1	0.201	5	0.126	3
134	0.259	1	0.047	9	0.011	12	0.002	15	0.220	1	0.129	1
LKS	**0.236**	**5**	**0.052**	**6**	**0.016**	**8**	**0.003**	**8**	**0.207**	**2**	**0.125**	**5**
8	0.255	3	0.075	2	0.026	2	0.009	2	0.207	3	0.127	2
59	0.255	2	0.071	3	0.023	3	0.006	3	0.206	4	0.126	4
…												
3	0.137	24	0.029	20	0.009	17	0.002	14	0.116	24	0.074	24

Table 4: Rouge scores for some systems and their ranks according to each score.

As we can see Lakhas (**LKS**) results are very good (ranked about 5th or 6th) compared to other systems even though we followed a totally different track as the one followed by others. Looking at the results, we conjectured that translation errors were mainly responsible for some of our relatively bad scores.

For unknown or badly spelled words, hoping it is a proper noun, Ajeeb tries to translate it in English while maintaining the same pronunciation (see Table 5 for examples)

text word	Translation	Correct word	Correct translation
لشلروع Permutation of " ل "	Lshlroa	للشروع	to begin
الفلسطينين Lack " ي "	Alflstinin	الفلسطينيين	the Palestinians

Table 5: Some Ajeeb translation errors of transliterated words.

Like other MT, Ajeeb cannot avoid the problem of ambiguous words. For example, تطلق *divorce* in the sentence *The bombardment operations against Iraq take place by passing missiles that divorce from the American aircraft carrier* whereas the correct word is *launched*. This ambiguity is caused by the absence of vowel in Arabic texts. The Arabic word تطلق can have two interpretations (to launch, divorce). The use of vowel تُطْلَقْ=launch تُطَلَّقْ=divorce could give the good interpretation without considering the context.

The absence of capital letters in the Arabic language can also influence on the translation in particular for the proper nouns. For example, بون which indicates *Bonn* in Germany was translated as *difference*.

By the nature of composition of its words, Arabic often increases the number of words within the translation, because the articles, the possessive adjectives, the adverbs, the conjunctions and the personal pronouns are concatenated to the word, Table 1 gives the number of words for the original summary in Arabic and the corresponding translation in number of words. The fact that summaries were truncated to 75 bytes before evaluation was also instrumental in lowering our scores. Our summaries had 15 Arabic words which were expanded to 21 English words many of which were then badly truncated before evaluation (see Table 7).

7 Post competition evaluation with ISI translation

After the competition, Franz Och kindly accepted to translate our Arabic summaries using the same system as one used for the original texts. When we ran the ROUGE scoring on the translations of our Arabic summaries

with ISI MT engine, we got higher scores by far the best compared with other systems. The following table gives the score and the rank of the new translation

ID	ROUGE-1		ROUGE-2		ROUGE-3		ROUGE-4		ROUGE-L		R-W-1.2	
model	*0.395*		*0.147*		*0.064*		*0.027*		*0.344*		*0.200*	
LKS-ISI	**0.297**	**1**	**0.084**	**1**	**0.029**	**1**	**0.009**	**3**	**0.256**	**1**	**0.153**	**1**
142	0.218	7	0.076	2	0.029	2	0.010	1	0.201	6	0.126	4
134	0.259	2	0.047	10	0.011	13	0.002	16	0.220	2	0.129	2
LKS	**0.236**	**6**	**0.052**	**7**	**0.016**	**9**	**0.003**	**9**	**0.207**	**3**	**0.125**	**6**
8	0.255	4	0.075	3	0.026	3	0.009	2	0.207	4	0.127	3
59	0.255	3	0.071	4	0.023	4	0.006	4	0.206	5	0.126	5
…												
3	0.137	25	0.029	21	0.009	18	0.002	15	0.116	25	0.074	25

Table 6: Rouge scores for some systems and their ranks adding the new translation (**LKS-ISI**)

From table 6, we can observe that although **LKS** give good results, **LKS-ISI** seems much better, mainly due to two reasons:

- ISI translations generated an average of 3,5 words less than Ajeeb, and ignores unknown words, while Ajeeb retains them and tries to guess a word by decomposition and often generates worthless words, the 7th and 8th column in table 1 show the difference in number of words between the two systems.
- The words translated by ISI are usually the same to the words used in reference models while those of Ajeeb are often synonymous.

The following table gives a few examples of sentences with their relative Rouge score in which we can clearly see the influence of some words related to the two translations

Model Translation	Ajeeb Translation ISI Translation	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L	R-W-1.2
King Hussein nearly **finished** *chemotherapy* **treatments** at American Mayo Clinic	Al-Malik Hussain ended the fourth stage from a *chemotherapy* from origin six|stages	0.13	0.03	0.00	0.00	0.13	0.09
	King Hussein finished the fourth phase of the chemical **treatment** of the six|stages	0.42	0.15	0.03	0.00	0.39	0.22
Nelson Mandela arrives to participate in **annual** Gulf States **summit**	Nelson Mandela arrived to the United Arab Emirates for the participation in| the annual summit to the Gulf countries	0.35	0.11	0.06	0.04	0.33	0.20
	Nelson Mandela arrived in the Emirates to participate in the **annual summit**| of the Gulf	0.60	0.28	0.16	0.00	0.55	0.32
Cohen confident Gulf *countries* will support **"appropriate action" against Iraq**	The Gulf Arab *countries* will offer the support to the doing of a suitable w|ork against Iraq	0.37	0.06	0.00	0.00	0.34	0.21
	Gulf Arab states will support for **"appropriate action" against Iraq,**	0.55	0.27	0.13	0.04	0.53	0.32

Table 7: Rouge scores for some Ajeeb and ISI sentences translation. The underlined words were not taken in consideration during the evaluation because of truncation. Italic/Bold is for words found in the Ajeeb/ISI translation and also in the model.

As we can see in table 7, on top of the frequency of the same word between ISI and W model, the 2-gram are also are present more often in ISI when the words exist in Ajeeb (example *to_participate, will_support,*…)

ROUGE seems very interesting tool for summaries evaluation, but it depends a lot on the reference models used. The scores can vary by the use of synonyms of words used and would be perhaps more relevant if more variations of the words could be used. Currently, it seems that the four model summaries used most often the same words produced by ISI but synonymous with the ones used by Ajeeb thus lowering its score.

8 Conclusion

In this work, we took part in DUC 2004 but following another approach that the one proposed by the NIST. We used extractions techniques, which combine four methods applied to Arabic documents and to which we added four reduction processes. This first experiment seems interesting by its approach and its results, because it is more practical to treat data which respect some rules of sentence structure.

We will continue to improve our system by developing other techniques than extraction by exploring the semantics and conceptual aspect in order to investigate text regeneration.

9 Acknowledgements

Our thanks go to Paul Over of NIST for his collaboration and help in getting the original Arabic text corresponding to the English text. We also thank Franz Och of ISI for providing English translations of our Arabic summaries using the ISI machine translation system.

References

(Ajeeb) Ajeeb Translator, 2004, http://arabic.ajeeb.com/ , Sakhr Technologies.

(Aljlayl & al. 2002) Mohammed Aljlayl and Ophir Frieder. On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach, *In 11th International Conference on Information and Knowledge Management (CIKM)*, November 2002.

(Attia 2000) Ahmed, Mohamed Attia, A Large-Scale Computational Processor of the Arabic Morphology, and Applications. *A Master's Thesis*, *Faculty of Engineering, Cairo University, Cairo*, Egypt, 2000

(Darwish 2002) Darwish, K. and D. Oard. CLIR Experiments at Maryland for TREC 2002: Evidence Combination for Arabic-English Retrieval. *in TREC. 2002. Gaithersburg, MD*.

(DUC 2004) *Document Understanding Conference* http://duc.nist.gov/duc2004/tasks.html

(Ishikawa 2001) Ishikawa, K., Ando, S., Okumura, A. Hybrid Text Summarization Method based on the TF Method and the Lead Method. *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization. Tokyo. Japan. March 2001*. p.5-219-5-224.

(Larkey 2002) Larkey L.S, Ballesteros L et Connell M.E., Improving Stemming for Arabic Information Retrieval, Light Stemming and Co-occurrence Analysis, *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002), Tampere, Finland*, August 11-15, 2002, pp. 275-282.

(Xu 2002) Xu, Jinxi, Fraser, Alexander and Weischedel, Ralph, Empirical Studies in Strategies for Arabic Retrieval, *Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR 2002),* August 11-15, 2002, pp. 269-274.