Effects of the College Major on Assessments of Arabic Text Summaries

Bassam Hammo[1], Martha Evens[2], and Hani Abu-Salem[3]
[1]Department of Computer Information Systems, the University of Jordan, Jordan
[2]Department of Computer Science, Illinois Institute of Technology, USA
[3]Department of Mathematics and Computer Science, University of South Carolina-Aiken, USA

Abstract: *We set out to discover whether or not the summaries produced by our Arabic text summarization software were potentially useful to a wide range of people. 1200 students at the University of Jordan were each given a copy of a newspaper article and a system-generated summary and asked to classify the summary as Rejected (R), Not-Related (N), Satisfactory (S), Good (G) or Accepted (A). 76.92% of the summaries were judged to be G or A and 92.34% were judged to be S, G or A. These students came from four different majors*: *300 from Arabic studies, 300 from humanities, 300 from Information Technology (IT) and 300 from a one-year program designed to help K-12 teachers to learn how to use computers effectively in the classroom. To our surprise, students from these four different majors differed significantly in their assessments; the teachers rated the summaries significantly more favourably; the IT students rated them significantly lower than did the students in Arabic and the humanities.*

Keywords: *Arabic natural language processing, arabic text summarization, extraction, software testing, evaluation.*

Received June 18, 2014; accepted December 16, 2014; published online June 11, 2015

1. Introduction

The goal of this paper is to report and to try to explain some unforeseen results of an experiment in software testing. We developed a system that produces summaries of Arabic articles and set out to test it on students at the University of Jordan. Since, we wanted to make sure that the summaries output by our system were acceptable to a wide range of students, we chose students from four different majors: Information Technology (IT), Arabic studies, humanities and also, students enrolled in a one-year special program for teachers who returned to the university to learn how to use computers in the classroom. Because we were eager to obtain robust statistical results, inspired by [29] we decided to look for a large number of human evaluators. We asked our colleagues among the faculty at the University of Jordan to give us fifteen minutes of class time with their students. Many students were generous enough or curious enough to agree to help us and we wound up with three hundred students from each of the four majors. As we had hoped, the majority of the students found the summaries useful to some degree. We were very surprised, however, to discover that there were large systematic differences between the responses of students in different majors. The main question here is why should students in different majors at the same university differ significantly in their response to a text summary.

The remainder of this paper is organized as follows: We briefly describe some of the literature in the sociology of education that describes systematic differences among majors in different areas. Then, we give a short nontechnical explanation of how the summarization process works with an example. A complete technical description can be found in [12]. We then describe the process of administering this experiment and collecting the data and present our results. Finally, we give a brief discussion of these results and present our conclusions.

2. Some Previous Studies of Personality Types and College Majors

When we looked for sociological studies of characteristics, attitudes and opinions of students in different college majors we discovered that much of the available literature on the subject is really a by-product of research in career counselling, so the primary focus of many studies is the characteristics, attitudes and opinions of people in different careers. Hearn [14] argued for generalizing from one group to another: "College students tend to major in areas that are relevant to their career plans and major departments appear to be the most significant college-level influence on their vocation-related values, attitudes and aspirations." He looked especially at factors in the major department that determine the well-being of their students and affect their attitudes.

Thistlethwaite [33] did one of the first large-scale studies and showed that the departmental environment is a major factor in student decisions to go on for further graduate training in the major area as a step towards a career in the area. He found that encouragement and positive evaluations from

departmental faculty and a vibrant intellectual climate were important in convincing students to continue. Vreeland and Bidwell [34] found that departmental goals, whether technical or moral and the degree of consensus among faculty members and the degree of commitment to those goals by faculty had a strong influence as well. These studies also, tend to support the assumption that students choose a college major as a pathway to an already chosen career, which underlies many of the studies in this area.

Strong [30, 31] started a stream of research in career counselling designed initially to help men leaving the military in the years after the first, World War. His focus was on determining interests and talents. He spent his career on sharpening his inventory, extending it to women and establishing norms for it. Hansen [13] carried out extended studies of the application of Strong's inventory [31] in career counselling.

Another approach to career counselling was developed by Myers and Briggs [23]. Their myers-briggs type indicator is based on Jung's theory of personality types, with questions that focus on perception and judgment. The myers and briggs foundation provides personality tests and training in giving the tests and interpreting the results.

Holland [15, 16, 17, 18] carried out a series of studies of personality differences of people in different careers with the goal of developing personality tests to help people make appropriate career choices. He came up with 6 different basic types: The doers (the realistic type) like to solve practical hands-on problems. The thinkers (the Investigative type) are analytical, intellectual and scientific. The artists (the creative type) are original, independent, disorganized and creative. The helpers (the social type) are nurturing, healing and supportive people. The persuaders (the enterprising type) are leaders, sellers and concerned with status. The organizers (the conventional type) are orderly, precise and attentive to detail. When it comes to the students who participated in our study, Holland includes IT (or actually, computer science) in the lists of careers given for both the realistic and investigative types. Teachers are listed as helpers or thinkers. Majors in Arabic and in the humanities may have a variety of careers in mind. Lawyers, politicians, journalists and advertisers, are all classified as persuaders. Retailers and administrators are classified as organizers [17, 18]. Campbell and Holland [4] combined Holland's work [17, 18] with Strong's work [30, 31] to produce a revised interest inventory, which is still in use today.

Focquaert *et al.* [9] in contrast, did not consider career choices but look directly at current students and attempted to identify their cognitive traits. They found that individuals majoring in sciences possess a cognitive style that is more systemizing-driven than empathizing-driven, whereas individuals majoring in humanities are much more empathizing-driven than systemizing-driven.

3. Our Summarization System

Text summarization has been identified as one of the central issues in natural language processing [28]. A summarization program is designed to take a document such as a newspaper or journal article and reduce it to a few paragraphs while retaining the most important concepts. State-of-the-art summarization techniques are applied to text files. Recent work also, considers summarizing XML documents [32]. People use a summary to get an overall picture of the article and decide whether it is relevant to the task at hand and whether they need to read the entire article. Since, users may sometimes want very short summaries and sometimes longer ones; we allow them to specify what percentage of the original article length they want to see in the summary. For this experiment, we set that percentage at 40%; thus, each summary is roughly 40% as long as the original article.

There are two main approaches to text summarization [12]. The first and more widely used approach is text extraction. In this approach, the summarization program identifies important sentences in the original article and uses these sentences to construct the summary. The second approach is based on text generation. The system parses the input text and translates the core ideas into a graph made up of logic forms. Then, it prunes the graph down to a few core ideas and generates new text from the reduced set of logic forms. Since, there has been very little research on Arabic text generation, we chose the extraction strategy described in [12]. What is new about our system is that we combine several different strategies to identify those sentences important enough to include in the summary.

The simplest of the strategies that we use involves the identification of the keywords in the original article. This strategy has a long history in information retrieval and automatic indexing [2, 24, 25]. We also, use a more sophisticated technique developed by Matsuo and Ishizuka [21] which is based on word co-occurrences. The keywords obtained in this way, along with the words in the title and the list of keywords supplied with the article, if there are any, are matched against the sentences in this article. Sentences that contain any of these keywords are given a bonus score for later use in a text extraction scheme [12]. Douzidia and Lapalme [5] described an Arabic text summarization system named "Lakhas" (to summarize) that combines sentence extraction and machine learning algorithms.

Another important strategy is to identify the core topics involved in the original article. The first and most important step in the identification process is to classify the paper-does this article discuss politics, economics, the arts, athletic events or what? Abuleil *et al.* [1] built a classifier that starts with a database of proper names gleaned from a database of newspaper articles and uses this information to classify articles.

Techniques developed for question answering by Hammo *et al.* [11] helped us to find the focus of the article. Sobh *et al.* [27] built an Arabic summarization system that combines Bayesian and genetic programming methods to classify articles and then uses knowledge about articles in each category to identify the relevant sentences and extract them. Schlesinger *et al.* [26] constructed a system named CLASSY to perform document classification and then used the classification results in Arabic/English multi-document summarization. Mesleh [22] developed a classifier for Arabic articles using a support vector machine approach. The features used in the support vector machine are chosen by collecting detailed term frequency information and then applying the χ^2 statistic to determine which are the most important ones. Our text summarization system uses a classification program developed by Mesleh [22] to carry out the classification phase.

Our third strategy uses sentence and discourse analysis for articles in different categories. For a long time the dearth of Arabic language corpora was a major stumbling block to research in Arabic language processing. Fortunately, recent cooperation between researchers organized into a global WorldNet effort has led to the development of major resources, especially the Arabic WorldNet [3, 8, 10] based on an Arabic newspaper corpus. These resources have allowed us to carry out sentence and discourse analyses for different categories of articles. Our third strategy makes use of these analyses to identify the focus of the articles and find the sentences that present the major topics and concepts. It starts by examining the sentences in the introduction and the conclusion and then looks at the other major components of the article. Our Arabic text summarization algorithm combines these strategies into seven steps:

1. The system divides the document into separate paragraphs, sentences and words.
2. The system identifies and removes the stop words. Stop words were originally defined by Luhn [19] as the words to be avoided in indexing or searching in an information retrieval system. In English ordinary stop words include articles (a, an, the), conjunctions (and, or, because, etc.,), auxiliary verbs (am, is, are, has, have, may, might, etc.,), pronouns (he, she, they, etc.,) and prepositions (of, in, on, etc.,). The Arabic equivalents of many of these English words are affixed to verbs or nouns. Whether affixed or free-standing they are removed from the text at this stage. But, the original text is also, stored in case this sentence is chosen for use in the final version of the summary.
3. The system identifies the keywords using several different techniques and uses them to assign scores to sentences.
4. The system uses a text classifier developed by Mesleh [22] to identify the central ideas (i.e., the thematic goals) of the article and increases the scores of sentences that express these goals using the sentence and discourse analysis derived from [8].
5. The system removes any redundant sentences (sentences that repeat the information in an earlier sentence).
6. The system identifies the sentences with the highest scores and restores the stop words. It systematically discards the sentences with the lowest scores until it has cut the article to approximately the length specified by the user.
7. It rearranges the sentences in the order in which they appeared in the original article and produces the summary.

For purposes of the evaluation, the summaries were specified to be 40% of the length of the original article. One of the articles used in our experiment appears in Figure 1; the summary of that article produced by our system appears in Figure 2.

حاسوب لكل صف

تطلق وزارة التربية والتعليم مشروعا رياديا لإدخال الحاسوب كوسيلة تعليمية في جميع الغرف الصفية في المدارس الحكومية، يقوم على أساس توفير جهاز حاسوب وجهاز عرض "Data show" في جميع الغرف الصفية البالغ عددها "35" ألف غرفة في كافة مدارس المملكة الأساسية والثانوية.

وتخطط الوزارة لتعميم مشروع "حاسوب لكل صف" على ثلاث مراحل على جميع مدارس المملكة، على ان يتم الانتهاء من تنفيذه مع نهاية عام 2007، في وقت تتوقع فيه الوزارة الانتهاء من حوسبة 95% من مدارس المملكة مع نهاية شهر ايار.

ويعد المشروع المرحلة الإستراتيجية الثانية في خطة منهجية تعمل الوزارة بموجبها على إدخال الحاسوب كوسيلة تعليمية حديثة الى مدارسها، بما يمكن المعلمين من الاستعانة بأحدث التقنيات العالمية في تدريس جميع المباحث، حيث جرى حتى الآن تجهيز 2935 مختبرا حاسوبيا في 2553 مدرسة زودت بـ 60 ألف جهاز حاسوب، في الوقت الذي تسعى فيه الوزارة إلى زيادة عدد أجهزة الحاسوب في المدارس لتصل إلى 120 ألف جهاز بحلول عام 2006 ، لتصل بذلك نسبة عدد الطلبة إلى الأجهزة 8 طلاب لكل جهاز.

وباشرت الوزارة بإعداد عطاء لتوريد 20 ألف جهاز حاسوب تمثل المرحلة الأولى من مشروع حاسوب لكل صف، بينما كانت الوزارة بدأت منذ عام 2002 الماضي بتنفيذ مشروع "التعليم نحو الاقتصاد المعرفي" بكلفة 280 مليون دينار تقريبا، الذي يقوم على أساس التكامل بين الخطط التربوية، وخطط التنمية الاقتصادية والاجتماعية المستدامة".

وبموازاة خطط الحوسبة التي تنفذها تعمل الوزارة على تدريب المعلمين وتأهيلهم حاسوبا ليكونوا قادرين على استخدام أحدث تقنيات التعليم الالكترونية، وأوضح طوقان أن "نحو 52 ألف معلم التحقوا بدورات الرخصة الدولية لقيادة الحاسوب "ICDL"، أنهى المتطلبات منهم حتى الآن 18 ألف معلم"، مشيراً إلى أن "حوالي 25 ألف معلم التحقوا بدورات التعليم للمستقبل وإدارة الشبكات وصيانة الحواسب، فيما التحق قرابة ألف معلم ببرنامج تدريب المعلمين الذي ينفذه البنك الدولي".

Source: http://moe.gov.jo

Figure 1. An Arabic article about Information and Communications Technology (ICT).

جهاز حاسوب لكل صف

تطلق وزارة التربية والتعليم مشروعا رياديا لإدخال الحاسوب كوسيلة تعليمية فى جميع الغرف الصفية فى المدارس الحكومية، يقوم على أساس توفير جهاز حاسوب وجهاز عرض "Data show" في جميع الغرف الصفية البالغ عددها "35" ألف غرفة في كافة مدارس المملكة الأساسية والثانوية. وتخطط الوزارة لتعميم مشروع "حاسوب لكل صف" على ثلاث مراحل على جميع مدارس المملكة. ويعد المشروع المرحلة الإستراتيجية الثانية في خطة منهجية تعمل الوزارة بموجبها على إدخال الحاسوب كوسيلة تعليمية حديثة الى مدارسها، في الوقت الذي تسعى فيه الوزارة إلى زيادة عدد أجهزة الحاسوب في المدارس لتصل إلى 120 ألف جهاز بحلول عام 2006. وبموازاة خطط الحوسبة التي تنفذها تعمل الوزارة على تدريب المعلمين وتأهيلهم حاسوبا ليكونوا قادرين على استخدام أحدث تقنيات التعليم الالكترونية، وأوضح طوقان أن "نحو 52 ألف معلم التحقوا بدورات الرخصة الدولية لقيادة الحاسوب "ICDL"، مشيراً إلى أن "حوالي 25 ألف معلم التحقوا بدورات التعليم للمستقبل وإدارة الشبكات وصيانة الحواسب.

Figure 2. Summary of the Arabic article about ICT produced using a hybrid text summarization approach (adapted from [12]).

4. Data and Methodology

The goal of our experiment was to provide a convincing case for the utility of our summarization technique to a variety of people with different perspective; we did not expect to see consistent and significant differences between majors. Our system evaluation plan was influenced by the ideas of El-Haj *et al.* [6, 7], the SUMMAC evaluation by Mani *et al.* [20, 29] monograph on the evaluation of natural

language processing programs. The first author selected 40 articles in Arabic (in HTML) on sports, religion, economics, education and politics, ranging in size from a few paragraphs to one full letter-sized page from a local Jordanian newspaper. One reason for choosing newspaper articles as a source is that they tend to be intelligible to most university students. He then downloaded the articles in their HTML format and fed them to our summarization program.

Then, the first author set out to find student volunteers. Several of his colleagues at the University of Jordan agreed to help us carry out this experiment and we wound up with 300 students in each of four different major areas: Arabic, humanities, IT and a new one-year program designed to help K-12 teachers learn to use technology in teaching. The evaluation process was designed to take up the first fifteen minutes of a lecture session. The first author attended all 8-10 sessions in each major program, explained the purpose of the experiment, asked for volunteers and handed around a box of article/summary pairs. Each student who agreed to take part in the evaluation process was asked to pick one pair from the box. Before each session the box was filled with one copy of each of the 40 article/summary pairs, plus any leftovers from previous sessions, in case the session had more than 40 students. As a result the number of ratings for a particular document ranges from 24 to 42. Some of the students majoring in Arabic and in humanities were sophomores and some were seniors. The students in IT were all seniors. The students in the K-12 group were all experienced teachers who had chosen to participate in a one-year program on using IT in the classroom and were therefore somewhat older than the students in the other groups.

The first author asked the evaluators to read carefully through the article and the summary and then rate the summary on a five point scale with values: 0=Rejected (R), 1=Not-Related (N), 2=Satisfactory (S), 3=Good (G) and 4=Accepted (A).

5. Results

Our first goal was to discover whether the system works, whether the majority of students felt that the summarization system produced useful results. The results for each of the 40 articles appear in Table 1.

Table 1. Results of the evaluation of 40 text summaries from participants with four different majors [R, N, S, G, A].

Article # and Summary	Majors of Participants																				Total Ratings
	Arabic Major					Humanities Major					IT Major					K-12 Teachers					
	R	N	S	G	A	R	N	S	G	A	R	N	S	G	A	R	N	S	G	A	
1				5	2			1	5	3			1	2	3				2	4	28
2			1		4		1	1	3	2			2	4					3	3	24
3				6	2	2		1	6			1	1	5	3				2	5	34
4	2	3	3				1	1	3	3	1	1	2	5	1	1	3	4	3	5	42
5				2	4			2	4	4				5	2				3	1	27
6				5	2			1	4			1		5	2			1	2	4	27
7				5	2			1	5	2		1	2	1	3			2	4	5	33
8				2	3				3	4			2	5	2				2	2	25
9				2	4		1	1	6			1	3	2	1				2	2	25
10				2	3			1	5	2			1	4	3				4	5	30
11	4	5	2			5	2	2	2	1	1	2	4	1	1	1	2	2	3	4	44
12					6			1	4	4			1	4	3			1	3	2	29
13				2	2			2	5	2			3	5					4	3	28
14			1	2	2				7				2	2	3			1	3		23
15	1			5				2	5	1				5	2			1	2	5	29
16				2	4				3	3			2	3	1		1		3	2	24
17				2	4				3	2			2	4	4			1	4		26
18		1		5	2	2	1	3	2			1	4	3					4	4	32
19			2	2	2			3	2			3	2	2				3	4	1	26
20			3	2	2		1	2	1	1	1	2	4	1	1		2	1	2	3	29
21			2	5	2			1	6	1			3	4	1				5	5	35
22		3	4	2				3	7	3	1	2	1	2				1	4	3	36
23			1	6	4				5	3			2	1	3				2	2	29
24				6	4				6	1			1	2	4				5	3	32
25				2	5				5	3			2	4	1			1	3	4	30
26			4	2	2				5	1			2	3	3				4	1	27
27		3	2		4				3	2		2		2	4			1	5	2	30
28				5	4			1	5	2			2	5					4	5	33
29				7	2				2	2				4	2			1	4	2	26
30				8	2		1	1	5			1	3	4			1	2	3	5	36
31				3	4				4	5				3	3				4	4	30
32				2	4		1		6				1	3	4			2	2	2	27
33				5	4		1	1	3				3	1				2	4	1	25
34				8	2			3	3	1		1	2	4	1				3	1	29
35			2	2	2		2	1	1			1	4	2				2	3	3	25
36				4	4		2	1	6	1	1		3	2	3		1		4	4	36
37		1	3	2	4	1	1	2	4				4	3				1	3	4	33
38				5	3			2	3	3		1	4	3					4	5	33
39			2	2	4		1		4	2				4	3		1	1	3	4	31
40		3	3	2				1	3	4		1	2	5					4	4	32
Total	7	19	35	129	110	10	16	42	164	68	5	22	77	129	67	2	11	31	132	124	1200

If we put aside our interest in different majors for the moment and consider our original goal for this experiment, these results are certainly encouraging. Suppose we count a summary as a success, if the human evaluators evaluate it as G or as A. Then, the overall performance of the system is approximately (76.92%), as shown in Table 2. If instead we count a summary as a success, when the human evaluators evaluate it as S or G or A, then the overall performance of the system is (92.33%) as can be also seen from Table 2.

Table 2. Ratings from participants grouped by major area assuming success is defined as: A rating of G or A and a rating of S or G or A [n is the number of human evaluators from a given major].

Major	Sample Size	What did the Testers Say?										Overall Success [G+ A]	Overall Success [S+G+A]
		R		N		S		G		A			
		N	%	N	%	N	%	N	%	N	%	%	%
Arabic	300	7	2.33	19	6.33	35	11.67	129	43.00	110	36.67	79.67	91.34
Humanities	300	10	3.33	16	5.33	42	14.0	164	54.67	68	22.67	77.34	91.34
IT Majors	300	5	1.67	22	7.33	77	25.67	129	43.00	67	22.33	65.33	91.00
Teachers	300	2	0.67	11	3.67	31	10.33	132	44.00	124	41.33	85.33	95.67
Total	1200	24	2.0	68	5.67	185	15.41	554	46.17	369	30.75		
Overall Performance [G + A]												79.92%	
Overall Performance [S + G + A]													92.33%

The first author was entering the data into the computer when he first noticed that the students from different majors seemed to be expressing quite different reactions to the system output. Table 2 shows the results from Table 1 grouped by major field under the assumption that success is defined as a rating of G or A. The Arabic majors rated (79.67%) of the summaries as successful; the humanities majors rated (77.34%) of the summaries as successful; the IT majors rated far fewer (65.33%) as successful. The teachers rated far more (85.33%) as successful. Table 2 shows those same results assuming that success is defined as a rating of S or G or A.

Once we had obtained these results we needed to determine whether these differences are statistically significant. We carried out a two-tailed Student's t-Test on a matrix with the 1200 student ratings. We found a significant difference, $p< 0.01$, between the IT students and the Arabic and humanities majors, whether we assumed equal or unequal variance. As can be seen from a careful study of Table 2, this difference is especially marked in the distribution of S ratings. Carrying out a two-tailed Student's t-Test comparing the K-12 teachers with the Arabic and Humanities majors we again found a significant difference, $p<0.001$. Again we get this result whether we assume equal or unequal variances in the two populations.

We also, carried out χ^2 tests on the relevant 2×2 tables. Beginning with the IT majors and comparing them with the majors in Arabic and the humanities, using the data in Table 2 and assuming that success is defined as ratings of G or A, we constructed Table 3. For this table the value of $\chi^2=17.39$.

Table 3. Comparing ratings of participants majoring in IT with ratings of participants majoring in Arabic and humanities where success is defined as: the ratings of G or A only.

Ratings\Majors	IT Majors	Arabic and Humanities	Combined
G+A	196	471	667
R+N+S	104	129	233
Total Number of Students	300	600	900

Since, we have only one degree of freedom, this means that the probability that this difference occurs by chance $p<0.001$. In other words, there is a significant difference between the behaviour of the IT majors and the students majoring Arabic and the humanities. The ratings by the IT majors were significantly less positive.

If instead, we define success as ratings of S, G or A and again focus on the IT majors we get Table 4.

Table 4. Comparing ratings of participants majoring in IT with ratings of participants majoring in Arabic and humanities where success is defined as: The ratings of G or A or S.

Ratings\Majors	IT Majors	Arabic and Humanities	Combined
G + A + S	273	548	821
R + Not-Related	27	52	79
Total Number of Students	300	600	900

Here, $\chi^2=0.00173$ and so the probability of obtaining a distribution like this by chance is rather large and the difference is not significant. We could have expected this result, if we had just looked more carefully at the numbers in Table 1. The number of summaries classified as R or N by the Arabic and humanities majors was 26 in both cases, whereas the IT majors between them placed 27 summaries in the R and N categories. That is, the IT students behaved a lot like the others when it came to classifying summaries as R or not-related, but they behaved quite differently when it came to the S designation, including in this category many papers that other students rated more highly.

When we carried out the same analysis for the school teachers enrolled in the program, we again constructed a 2×2 table, our Table 5. For this table, $\chi^2=5.58$ and so the probability $p< 0.02$ that values like these could happen by chance.

Table 5. Comparing ratings of school teachers with ratings of participants majoring in Arabic and humanities where success is defined as: The ratings of G or A.

Ratings\Majors	Teachers	Arabic and Humanities	Combined
G+A	256	471	727
R+N+S	44	129	173
Total Number of Students	300	600	900

To repeat this analysis when the S category is included with the G and A categories, we constructed Table 6. For this 2×2 table, $\chi^2=4.98$ and thus the probability of this distribution occurring by chance, $p< 0.05$. In other words, there is a significant difference between the behaviour of the school teachers and the Arabic and humanities majors, with both definitions of success. The school teachers showed a much more positive attitude.

Table 6. Comparing ratings of school teachers with ratings of participants majoring in Arabic and humanities where success is defined as: The ratings of G or A or S.

Ratings\Majors	Teachers	Arabic and Humanities	Combined
G+ A+ S	287	548	835
R+ N	13	52	65
Total Number of Students	300	600	900

6. Discussion

Since, we did not expect these results; we did not make any formal attempt to interview our human evaluators or circulate a questionnaire, so the only explanations of these results that we can give are based on brief comments addressed to the first author as the students handed in their assessments. We believe that the K-12 teachers were delighted to see a piece of software that they could imagine using in classes with their own older students and that this enthusiasm played a part, at least, in their ratings of the summaries.

We suspect that the IT majors, on the other hand, were influenced by two very different factors. Their classes are taught in English with textbooks and papers in English, as well, whereas the Arabic and the humanities majors are taught in Arabic with assignments that require reading large quantities of Arabic text. Thus, the IT majors do not have an immediate need for an Arabic text summarization program. A few of them also expressed disappointment in the extraction-based software they were expecting text generation. They consider text extraction to be "low-tech."

If we accept Holland's [17, 18] categories and his characterization of each, we could say that the teachers felt that our summarization system could help them to become better helpers for their students, while the IT students did not find that our system helped solve a problem for them. The remaining questions is do these points really explain significant differences of opinion on what is basically a fairly simple cognitive task.

7. Conclusions

In this paper, we have described an experiment in which we asked 300 students from each of four different majors to rate a summary of an Arabic newspaper article on a five-point scale (R, N, S, G, A). If we defined success as getting one of the top two ratings, then almost 79% of the ratings classified the summary a success. If we define success as getting one of the top three ratings then more than 92% of the summaries were rated as successful. We can conclude that our Arabic summarization system works, sometimes at least, but we have many remaining questions about the surprising differences we see in responses from students in different major programs. The IT students showed a significantly less positive response to the summaries, while the K-12 teachers back on campus for a year-long program on using computers in the classroom showed a significantly more positive response to the same summaries.

Future research could lead us in several different directions. We certainly want to carry out an experiment with more academic articles. We need to be sensitive to two issues in the choice of materials. One issue is copyright; another is intelligibility. Perhaps we could use student term papers, with, of course, permission from the students involved. Whenever we carry out an experiment like this one involving majors from different departments, we want to give the students a short questionnaire and also, ask them for comments, hoping to discover whether these significant differences in ratings from students in different departments show up again and if they do, whether we can provide a more authoritative explanation.

Acknowledgments

The authors wish to thank Prof. Nadim Obeid, Mahmoud Al-Haj and the experimental subjects from JU. We would also, like to thank several anonymous referees for their valuable comments.

References

[1] Abuleil S., Alsamara K., and Evens M., "Tagging Proper Nouns and Keywords to Classify Arabic Newspaper Text," *in Proceedings of Midwest Artificial Intelligence and Cognitive Science Conference*, Chicago, USA, pp. 137-142, 2002.

[2] AbuSalem H., Al-Omari M., and Evens M., "Stemming Methodologies over Individual Query Words for an Arabic Information Retrieval System," *Journal of the American Society for Information Science*, vol. 50, no. 6, pp. 524-529, 1999.

[3] Black W., Elkateb S., Rodriguez H., Alkhalifa M., Vossen P., Pease A., and Fellbaum C., "Introducing the Arabic WordNet Project," available at: http://nlp.lsi.upc.edu/papers/fellbaum-alkhalifa-206.pdf, last visited 2005.

[4] Campbell P. and Holland J., "A Merger in Vocational Interest Research: Applying Holland's Theory to Strong's Data," *Journal of Vocational Behaviour*, vol. 2, no. 4, pp. 353-376, 1972.

[5] Douzidia F. and Lapalme G., "Lakhas, an Arabic Summarization System," *in Proceedings of the Document Understanding Conference (DUC'04)*, Boston, pp. 128-135, 2004.

[6] El-Haj M. and Hammo B., "Evaluation of Query-Based Arabic Text Summarization System," *in Proceedings of International Conference on Natural Language Processing-Knowledge Engineering*, Beijing, pp. 1-7, 2008.

[7] El-Haj M., Kruschwitz U., and Fox C., "Experimenting with Automatic Text Summarisation for Arabic," *in Proceedings of the*

4th Language and Technology Conference, Poznan, Poland, pp. 490-499, 2011.

[8] Elkateb S., Black W., Vossen P., Farwell D., Pease A., and Fellbaum C., "Arabic WordNet and the Challenges of Arabic Natural Language Processing," available at: http://www.mt-archive.info/BCS-2006-Elkateb.pdf, last visited 2006.

[9] Focquaert F., Steven M., Wolford G., Colden A., and Gazzaniga M., "Empathizing and Systemizing Cognitive Traits in the Sciences and Humanities," *Personality and Individual Differences*, vol. 43, no. 3, pp. 619-625, 2007.

[10] Global WordNet/ Arabic WordNet., available at: http://www.globalwordnet.org/AWN, last visited 2014.

[11] Hammo B., Abuleil S., Lytinen S., and Evens M., "Experimenting with a Question Answering System for the Arabic Language," *Computers and the Humanities*, vol. 38, no. 4, pp. 397-415, 2004.

[12] Hammo B., Abusalem H., and Evens M., "A Hybrid Arabic Text Summarization Technique Based on Text Structure and Topic Identification," *International Journal of Computer Processing of Languages*, vol. 23, no. 1, pp. 39-65, 2011.

[13] Hansen C., *Career Development and Counselling: Putting Theory and Research to Work*, New York: Wiley and Sons, 2005.

[14] Hearn C., "Major Choice and the Well-Being of College Men and Women: An Examination from Developmental, Organizational and Structural Perspectives," *Sociology of Education*, vol. 53, no. 2, pp. 164-178, 1980.

[15] Holland L., "A Personality Inventory Employing Occupational Titles," *Journal of Applied Psychology*, vol. 42, no. 5, pp. 332-336, 1958.

[16] Holland L., "A Theory of Vocational Choice," *Journal of Counselling Psychology*, vol. 6, no. 1, pp. 35-45, 1959.

[17] Holland L., *Dictionary of Holland Occupational Codes*. Lutz, FL: Psychological Assessment Resources, Inc., 1996.

[18] Holland L., *Making Vocational Choices: A Theory of Vocational Personalities and Work Environments*, Lutz, FL: Psychological Assessment Resources Inc., 1997.

[19] Luhn P., "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159-162, 1958.

[20] Mani I., Klein G., House D., Hirschman L., Firmin T., and Sundheim B., "SUMMAC: A Text Summarization Evaluation," *Natural Language Engineering*, vol. 8, no. 1, pp. 43-68, 2002.

[21] Matsuo Y. and Ishizuka M., "Keyword Extraction from a Single Document Using Word Co-occurrence Statistical Information," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 11, pp. 57-69, 2004.

[22] Mesleh A., "Chi Square Feature Extraction Based SVM Arabic Language Text Categorization System," *Journal of Computer Science*, vol. 3, no. 6, pp. 430-435, 2007.

[23] Myers B. and Briggs K., *The Myers-Briggs Type Indicator*, Palo Alto, CA: Consulting Psychologists Press, 1962.

[24] Salton G. and Buckley C., "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513-523, 1998.

[25] Salton G., Wong A., and Yang C., "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613-620, 1975.

[26] Schlesinger J., O'Leary D., and Conroy J., "Arabic/English Multi-document Summarization with CLASSY-the Past and the Future," *in Proceedings of CICLING*, Haifa, pp. 568-581, 2008.

[27] Sobh I., Darwish N., and Fayek M., "An Optimized Dual Classification System for Arabic Extractive Generic Text Summarization," *in Proceedings of the 7th Conference on Language Engineering*, Cairo, Egypt, pp. 149-154, 2007.

[28] Spärck K., "Automatic Summarizing: The State of the Art," *Information Processing and Management*, vol. 43, no. 6, pp. 1449-1481, 2007.

[29] Jones K. and Galliers R., *Evaluating Natural Language Processing Systems*, Berlin: Springer Verlag, USA, 1996.

[30] Strong J., *Strong Vocational Interest Blank for Men*, Stanford, CA: Stanford University Press, 1927.

[31] Strong J., "Norms for Strong's Vocational Interest Tests," *Journal of Applied Psychology*, vol. 35, no. 1, pp. 50-56, 1951.

[32] Teng L. and Ping Y., "A Framework of Summarizing XML Documents with Schemas," *the International Arab Journal of Information Technology*, vol. 10, no. 1, pp. 18-27, 2013.

[33] Thistlethwaite D., "Fields of Study and Development of Motivation to Seek Advanced Training," *Journal of Educational Psychology*, vol. 53, no. 2, pp. 53-64, 1962.

[34] Vreeland R. and Bidwell C., "Classifying University Departments: An Approach to the Analysis of Their Effects upon Undergraduate Values and Attitudes." *Sociology of Education*, vol. 39, no. 3, pp. 237-254, 1966.

Bassam Hammo received a BSc in computer science from the University of Jordan in 1987. His MSc in computer science was from Northeastern University and his PhD in computer science was from DePaul University, Chicago, IL 2002. He is an associate professor at the University of Jordan since 2003. His research interests are arabic natural language processing and its applications, databases and e-learning. He leads the ANLP research group at the University of Jordan.

Martha Evens received an AB in Mathematics from Bryn Mawr College with a minor in Greek in 1955; received an AM from Radcliffe in Mathematics in 1957; received a PhD from Northwestern University in 1975 in computer science and has taught at Illinois Institute of Technology ever since. In 1984 she was President of the Association for Computational Linguistics, an international organization for people interested in natural language processing.

Hani Abu-Salem received a PhD from Illinois Institute of Technology in 1992 in computer science and has taught at Mutah University in Jordan and DePaul University in USA. Currently, he is a full professor with the University of South Carolina Aiken. His research interests include natural Language processing (NLP), information retrieval, databases and data mining systems. Now, he is focusing on the development and implementation of IT strategic plans, Multi-lingual retrieval systems and IT security.