Comparing Two Novelty Detection Models for Arabic Text Based on Sentence Level Information Patterns

Mohammed Al-Kabi[1], Esra'a AL-Shdaifat[2], Emad Al-Shawakfa[3], Abdullah Wahbeh[4] and Izzat Alsmadi[3]

[1]Zarqa University, [2]Hashemite University, [3]Yarmouk University,
[4]Dakota State University, Jordan

mohammedk@yu.edu.jo, esraa_shdaifat@hu.edu.jo shawakfa@yu.edu.jo,
ahwahbeh@pluto.dsu.edu, ialsmadi@yu.edu.jo

Abstract

Many important applications have used novelty detection in order to reduce redundant and non-relevant information presented to users of the document retrieval systems. In this study, sentence level information patterns are proposed for enhancing the novelty detection for Arabic text documents. Two models based on sentence level information patterns are suggested and compared; the first one is based on sentence length while the second one is based on opinion patterns. Experimental results have showed that both of the proposed models; Length Adjusted (LA) model and Length and Opinion Adjusted (LOA) model, can significantly improve the performance of novelty detection for Arabic texts, in terms of precision at top ranks. Better results were provided by LA model over LOA model. This shows that the sentence length is more important for enhancing the novelty detection than other suggested sentence level information patterns (e.g. opinion patterns).

Keywords: *Novelty detection, information patterns, opinion patterns, information retrieval*

1. Introduction

There is a continuous increase in the data volume that is uploaded and transmitted through the Internet between clients, services and Internet users [1]. People who work in media, security agencies receives a huge amount of stories, essays, reports and articles from a large number of sources. Such difficult situation inspired the researchers to invent new automatic system which is based on novelty detection. The last decade witnessed an increasing interest in the novelty detection which aims to build automatic systems which are capable to ignore old stories, essays, reports and articles already read or known, and notify the users of such systems about any new stories, essays, reports and articles.

Researchers used to use books and libraries to acquire information they need before the Internet era. For instance, in order to facilitate the access to a subject in a university library, a catalog librarian is responsible for classifying the different materials and giving them the code to identify them among others.

Web search engines do not depend on the human effort of cataloguers in building their indexes, while Web directories which represent other tools to help Internet users to access relevant documents they are searching for. Web directories are maintained by human editors. Therefore, the comprehensiveness of those directories is small relative to search engine comprehensiveness. Web directories do not display Search Engine Results Pages (SERPS) which are based on submitted keywords within different queries. Rather, they show lists of URLs categorized by human editors on categories and subcategories. In addition, Web directory results are characterized by their quality and reliability, while Web search engines are characterized by their comprehensive coverage.

The model of the Information Retrieval Systems (IRSs) and digital libraries is based on the assumption that users have some knowledge about the information they are searching for. Significant numbers of evidences are also found which prove that users may obtain information that they have no idea about it at all [2].

With the continuing growth of information, users of IRSs and search engines need to obtain useful information quickly; without the need to examine a lot of redundant information. Using the novelty detection, will reduce the amount of redundant and irrelevant materials presented to users.

The novelty detection has a significant impact on IRSs, where the Selective Dissemination of Information (SDI); which is used primarily in libraries and information science and denotes tools and resources used to inform searchers about new information, and enable them to be updated always when they search for any topic. Common interests of users' community and similarity to a typical profile are used by information filtering systems and most SDIs [3]. The goal of researches on novelty detection is to find techniques to reduce redundant and non-relevant materials produced and presented by document retrieval systems. The ranked list of documents generated by document retrieval systems; which adapt novelty detection, usually eliminates all non-relevant information. The residual essential relevant information will then be scanned searching for old materials to be discarded. Afterward, document retrieval systems present to the user a ranked list of non-redundant relevant sentences. Usually researchers start with a known set of relevant documents to simplify novelty detection, and facilitate the search for novel documents. The assumption is presumably that the process of finding relevant materials can be explored separately [4].

There are different classification schemes for novelty detection that exist at two levels in some research papers: (i.e. event and sentence levels) [5, 6]. Other references supposed three levels (i.e. event, sentence and document levels) [7, 8, 9]. Note that these events are not in the same "category" in comparison with the sentence or document granularity. On one hand, we have event and opinion queries and, on the other hand, we have information structures at document and at sentence level. Hence, events and opinions are used to classify queries, and document and sentence is the granularity we are using in our texts (and, therefore, in our indexes). Obviously it is much easier to deal with sentence retrieval than document retrieval, because one has to deal with much less text. Determining the relevant sentence may need to examine the surrounding context. In order to standardize the task of various passage retrieval approaches as well as simplifying the evaluation, researchers have adopted the sentence as a unit of retrieval. Vector space, language-modeling framework, or other techniques are used to determine relevant sentences to the submitted queries. Afterward, novel sentences are determined through a comparison with old or historical sentences. In case they are sufficiently different they were considered novel [3].

This study compares two methods which use the sentence level novelty detection. The sentence level information patterns; including: sentence length and opinion sentences were used to improve the novelty detection accuracy. In other words, to have sentences that are more relevant to the user information need and not just matching the words appearing in a user's query. The sentence level information patterns, sentence length and opinion sentence that are used in this research are selected depending on the observations described in [10]. The observations are based on data from the text retrieval conference (TREC), 2002-2004 (TREC novelty tracks). The first time a novelty track introduced is in the eleventh TREC conference track, 2002. This track had two main tasks; where the first task aimed to extract relevant sentences from relevant documents to the query, and obviously this task had excluded sentences related to different topics or those that did not contain significant information. The second task had aimed to filter the set of sentences obtained in the first task from sentences [11]. The conclusions obtained from novelty track of (TREC) 2002 were: the first task of isolating the relevant sentences to the query is extremely difficult and is not straight forward, and the value of obtaining novel sentences depends largely on the ability of the system to accomplish the first task (i.e. extracting relevant sentences) [4].

In their research, Li and Croft [10] had the following two observations regarding novel detection of sentences:

1. Relevant sentences, on average, have more words than non-relevant sentences.

2. There are relatively more opinion sentences in relevant (and novel) sentences than in non-relevant sentences.

This research is focused on the above two observations and tries to verify their correctness within the retrieval process of Arabic text.

The rest of the paper is organized as follows: the second section introduces related works, the third section presents the proposed methodology, the fourth section presents the results and evaluation, and section five presents the conclusion and possible future extensions.

2. Related Works

Zhang, et al., have extended an adaptive information filtering system to make decisions about the novelty and redundancy of relevant documents [12]. They have proposed a set of five redundancy measures; with and without redundancy thresholds. The results of the conducted experiments proved that the cosine similarity measure and a redundancy measure based on a mixture of language models were effective techniques to identify redundant documents.

Allan, et al., have presented a paper on Topic Detection and Tracking (TDT) subject which is dedicated to novel online event detection and tracking application [4]. TDT tasks are mainly interested in inter-topic or inter-event novelty detection, in order to determine whether two news stories include the same occasion, event, activity or not. TDT is interested in story-level online evaluation, where the news stories are presented one after the other, in order to be evaluated sequentially, and thus determine the new news stories. This task was presented more comprehensively in a research by Allan, et al., [13] on temporal summarization; where the main concern of their effort was to develop a useful evaluation model.

Yang, et al., has proposed a study about the use of clustering techniques to detect different events. The goal of the proposed system was to automatically detect new events from a temporally ordered stream of news stories, either retrospectively or as the stories arrive [14]. By applying hierarchical and nonhierarchical document clustering algorithms, they found out that temporal distribution patterns of document clusters have provided valuable information for improvement, in both retrospective detection and on-line detection of novel events.

In their study, Allan, et al., [15] have described new event detection and event tracking methods within a stream of news stories. The system has to make decisions about a single story before looking at subsequent stories. This approach uses a single pass clustering algorithm with a novel threshold model that includes the properties of events as a main component.

Discovering new events automatically from chronologically ordered documents is a real challenge to researchers in this field. A particular form of novelty detection; called First Story Detection (FSD), which is known to be difficult in the field of (TDT). FSD aims to detect new stories from on-line news as soon as they arrive in the sequence of documents. Yang et al.

[16] have proposed a new supervised learning algorithm to classify on-line documents by topic and topic-conditioned feature weights, to measure the novelty of documents within a topic at the event level. The researchers have focused on using named-entities for event-level novelty detection and using feature-based heuristics extracted from the topic histories. The results of the tests have showed a substantial improvement over the traditional one-level approach to detect novel documents.

Fernandez and Losada [17] study have addressed Local Context Analysis (LCA) to detect new and relevant sentences within documents related to a certain topic. LCA is beneficial to researchers in different areas of study, such as text summarization, information retrieval, Web search engines, question answering systems, etc. The core idea of this method is based on a common term from the top-ranked relevant documents that tend to co-occur within query terms within the top-ranked documents.

A number of studies exhibit how novelty detection can be used in several applications. A new formula and approach to the Minimal Document Set Retrieval (MDSR) problem was presented by Dai and Srihari [18], where three retrieval and ranking algorithms have been proposed and tested in their work. The three algorithms are: novelty based algorithm, cluster based method, and subtopic extraction based method. The tests have showed that subtopic extraction based method was the most flexible and effective among the three methods under consideration.

One of the approaches used for novelty detection at the sentence level is the new word detection approach (NW). This measure starts from an initial retrieval ranking and keeps sentences with new words that do not appear in the previous sentences as novel sentences, and then removes those sentences; without new words from the list. Similar to NW detection approach; new word threshold approach (NWT) follows the same concept. The NWT approach depends on a predefined threshold to classify the sentence as a novel sentence, another novelty measure; called New Information Degree (NID), calculates the sum of the inverse document frequency (idf) values of new words appearing in a certain sentence and divides it by the sum of all idf values of terms appearing within this sentence. The idf is used extensively in the Information retrieval and text mining systems to determine the importance of a term. By computing the logarithm of the quotient; resulting from dividing the total number of documents by the number of documents containing the subject term, so idf =log (N/n) +1, where (N) is the number of documents in the collection, and (n) is the number of documents that contain the subject term [10].

Another approach was suggested by Carbonell and Goldstein [19] is Maximal Marginal Retrieval approach (MMR). In this approach, every document in the ranked list is selected according to a composite measure of query relevance and novelty of information. This approach aims to measure the degree of dissimilarity between the subject document and previously selected ones that are already existed in the ranked list. Preliminary results have indicated some benefits for MMR diversity ranking in document retrieval and in a single document summarization. The main benefit lies in constructing non-redundant multi-document summaries.

A study by Litkowski [20] inspects every discourse entity in a sentence, instead of counting new words that appear in sentences. Discourse entities represent semantic objects and may have different syntactical forms. A single entity could be represented by a pronoun, a noun phrase, or a person’s name; therefore, these different representations could be treated as the same discourse entity. As a result, the sentence will be considered novel if it had new discourse entities that were not found in the inflating discourse entities that exist in the history list [20].

Li and Croft [21] have exhibited a new novelty detection approach to identify previously unseen query-related patterns (i.e. new named entities and noun phrases) in sentences. According to this approach, the sentence is considered new if the number of new named entities and noun phrases are above a certain threshold.

A significant related work was proposed by Li and Croft [10]. In their study, they provided a new definition for novelty as “new answers to the potential questions representing users’ requests or information needs”. They suggested a new novelty detection approach based on the identification of sentence level information patterns. In the first step, the user query is transformed into potential one or more questions, in order to identify the correspondent query-related information patterns that contain query terms and required answer types [22, 23]. In the second step, the new information is extracted through detecting sentences that include previously unseen answers relevant to the query-related patterns. The information–pattern-based novelty detection (IPND); suggested by Li and Croft [10], depends on three sentence information patterns that include: sentence length, opinion patterns, and named entities. The sentence length represents the number of words in the sentence. Opinion patterns can be identified by searching for such sentence patterns as “Mr T said”, “Miss Y reported”, or as signed by quotation marks. Li and Croft [8] have identified about twenty such opinion-related sentence patterns. A named entity of: Person, Organization, Location and Date (POLD) types were used in the new pattern detection. The formula used to calculate the novelty score compute the number of new terms appearing in a sentence, and do not appear in previous sentences and add this number to POLD-type named entities in the sentence under consideration that did not appear in previous sentences. The sentence is considered new if its novelty score equals to or exceeds the predefined threshold. Results of the conducted tests have showed that this approach has significantly improved the novelty detection for general topics.

3. Methodology

The main title (on the first page) should begin 1 3/16 inches (7 picas) from the top edge of the page, centered, and in Times New Roman 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Please initially capitalize only the first word in other titles, including section titles and first, second, and third-order headings (for example, “Titles and headings” — as in these guidelines). Leave two blank lines after the title.

The ultimate objective of this study is to use the results and conclusions obtained (based on the data and facts gathered above) in a reflective manner, in order to improve learning and teaching of software engineering in large groups, and in particular at UWE.

This section discusses the proposed Information-Pattern-based Novelty Detection (IPND) approach which is illustrated in Figure 1. Two important steps are demonstrated in the approach namely; relevant sentence retrieval and novel sentence extraction.

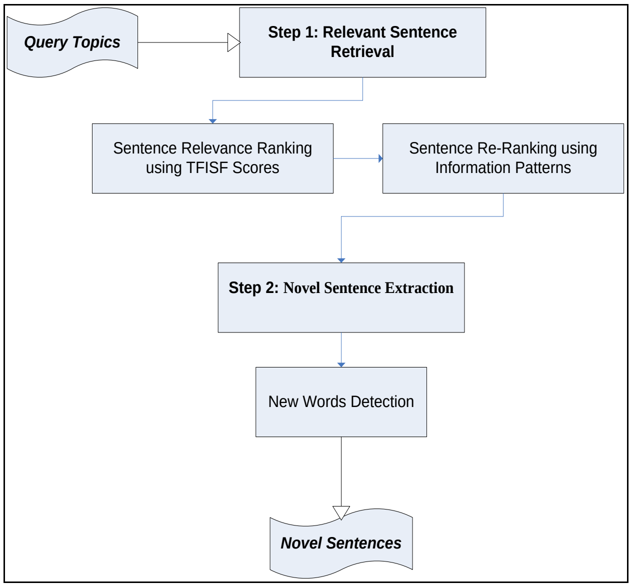

Query Topics
Step 1: Relevant Sentence Retrieval
Sentence Relevance Ranking using TFISF Scores
Sentence Re-Ranking using Information Patterns
Step 2: Novel Sentence Extraction
New Words Detection
Novel Sentences

Figure 1. Information-Pattern-based Novelty Detection (IPND) Approach

3.1 Relevant Sentence Retrieval

The task of the relevant sentence retrieval module is to retrieve sentences that are relevant to the user information query. The first relevant sentence retrieval module extracts query words and starts searching in the data collection to retrieve sentences that are relevant to the query (i.e. initial ranking). It then re-ranks the retrieved sentences using the sentence information patterns; including sentence length and opinion sentences (i.e. adjusted ranking). Sentences that do not satisfy the query are then filtered out because they are unlikely to have potential answers to a user information need. Figure 2 illustrates the overall ranking process.

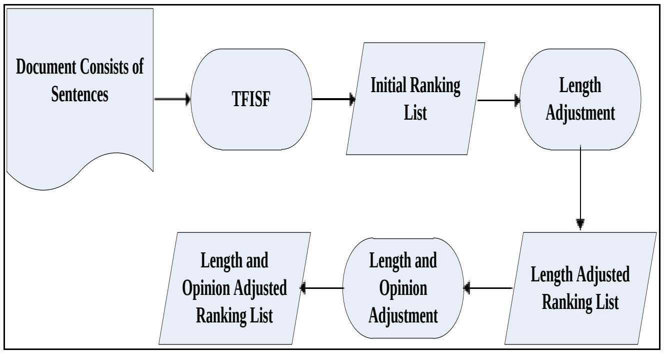

Document Consists of Sentences
TFISF
Initial Ranking List
Length Adjustment
Length Adjusted Ranking List
Length and Opinion Adjustment
Length and Opinion Adjusted Ranking List

Figure 2. Sentence Ranking Process

3.1.1 Initial Ranking Using TFISF Model

In this research, the Term Frequency-Inverse Sentence Frequency (TFISF) model was used as a relevant sentence retrieval model for initial relevance ranking score. This model was also used in other systems and was reported to be able to achieve a performance that is equivalent to, or better than other techniques in sentence retrieval [1]. The initial TFISF relevance ranking score S0 for a sentence, is calculated according to the following formula:

$$S_0 = \sum_{i=1}^{n} [tfs(t_i) \times tfq(t_i) \times (isf(t_i)^2)] \qquad (3.1)$$

Where n is the total number of terms, tfs(ti) is the frequency of term ti in the sentence, and tfq(ti) is the frequency of term ti in the query, isf(ti) is inverse sentence frequency (instead of inverse document frequency in a typical document retrieval system) [1]. The inverse sentence frequency is calculated as:

$$isf(t_i) = \log \frac{N}{N_{ti}} \qquad (3.2)$$

Where (N) is the total number of sentences in the collection, (Nti) which is the total number of sentences that include the term ti.

This research depends on the stop words list which is compiled and published by Yaser Al-Onaizan in his Web page

(http://www.isi.edu/~yaser/arabic/arabic-stop-words.html, [24]. Similar to other Information Retrieval (IR) systems, stop words removal is performed in a preprocessing step for relevant sentence retrieval. Stop words removal is used to filter out common words that do not contain significant or relevant information. There are (117) selected stop words, such as (That,"ذلك"), (In,"في"), (Upon,"على"), (Such,"تلك"), etc., in the used stop words list. They have been removed from all sentences in the relevant sentence step. Nonetheless, this stop words list lacks some important Arabic stop words such as (And,"و"), (OR,"أو"). As a result, such other important lists are added to the list of stop words removed in preprocessing stages.

3.1.2 Adjusted Ranking Using Sentence Level Information Patterns

The initial ranked list; which results from applying the (TFISF), was adjusted in this study using information patterns. The first step in the adjustment process is for the length of the sentence while the second one is for the opinion sentences.

The length adjustment is calculated as:

$$S_1 = S_0 \times (\frac{L}{\bar{L}}) \qquad (3.3)$$

Where S0 denotes the initial TFISF relevance ranking score, (L) denotes the length of a sentence and ($\bar{L}$) denotes the average sentence length.

Finally, the opinion-adjustment is computed as:

$$S_2 = S_1 \times [1 + \beta F\_opinion] \qquad (3.4)$$

Where F_opinion = 1; if a sentence is identified as an opinion sentence with one or more opinion patterns, 0 otherwise. β is a constant equals to 0.5. A number of patterns such as (Said,"قال"), (Added,"أضاف"), etc. (see Table 1) are used to determine whether a sentence is an opinion sentence or not.

Table 1. Arabic Examples of Opinion Patterns

Arabic	English	Arabic	English
صرحت	She stated	أثبتت	She proved
أضاف	Added	اكتشف	He discovered
أضافت	She added	اكتشفت	She discovered
علق	Comment	قال	Said
علقت	She comments	قالت	She said
توقع	Expect	وافق	Agreed
توقعت	She expected	وافقت	She agreed
اعتقد	I think, I thought	أشار	Pointed
اعتقدت	She thinks, she thought	أشارت	She pointed
وضح	Explain	بين	Indicate, show & between
وضحت	She explained	بينت	She indicated
وجد	Found	أظهر	Show
وجدت	She found	أظهرت	She showed
صرح	Declares	أثبت	Proved

3.2 Novel Sentence Extraction

This study has three ranking lists; the initial ranking list; which results from applying (TFISF) model, the Length Adjusted (LA) ranking list, and the "length and opinion" adjusted ranking list (described in Figure 2). For each one of the lists, we found the novelty score depending on the new words in a sentence that do not appear in its previous sentences; as indicated in Figure 3.

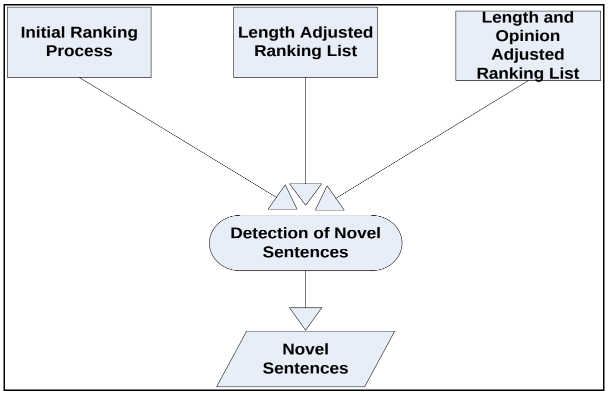

Initial Ranking Process
Length Adjusted Ranking List
Length and Opinion Adjusted Ranking List
Detection of Novel Sentences
Novel Sentences

Figure 3. Novelty Detection Process

The novelty score of a sentence is calculated through the following formula:

$$S_n = N_w \qquad (3.5)$$

Where Sn is the overall novelty score of a sentence S, Nw is the number of new words in S that did not appear in its previous sentences. Note that stop words in our stop words list have been removed from all sentences in the relevant sentence step, and thus, is not considered in the process of novelty score calculation. A sentence is identified as a novel sentence if its novelty score is equal to or greater than a predefined threshold. This threshold is determined according to several statistics and experiments.

4. Results and Evaluation

This section demonstrates and discusses the main experimental results. A corpus was manually collected from five different websites. Stop words elimination was performed on the whole corpus as a preprocessing step, depending on the revised stop words list mentioned in the previous section.

To evaluate the performance of the information pattern based models, the (TFISF) model was used as the baseline for comparing the performance of relevant sentence retrieval for novelty detection. The evaluation measure used for performance comparison is precision at rank N (N = 3, 5, 7, 9 and 11). Precision at rank N was calculated for each ranking list to compare the ranking of relevant sentences retrieved, using the following formula:

Precision= NO. of Novel sentences retrieved/ NO. of sentences retrieved …… (4.1)

Note that precision, at top ranks, is useful in real applications where users only want to go through a small number of sentences [1]. After finding precision at rank N for different queries, the average precision at rank N for each of the three models was calculated.

As described in the previous section, novelty score was measured, then precision at rank N was calculated, after that the average precision at each rank N (N = 3, 5, 7, 9 and 11) was computed as shown in Table 2.

Table 2. Average Precision at Rank N for TFISF Model

Rank #	Average Precision
3	0.7
5	0.56
7	0.57
9	0.48
11	0.5

As shown in Table 2, the TFISF model has achieved the highest precision at rank 3; with a value of 0.70, and then this value is decreased to about 0.56; for rank 5, then is increased to 0.57; for rank 7. The value has also dropped down again to 0.48; for rank 9, then it has increased to 0.50; for rank 11. As a result, the value of the average precision; at rank N, varied between N=3 to N=11.

For the LA Model; which is the second retrieved ranking list for each query, the novelty score was measured then the precision at rank N was also calculated. Later on, the average precision at each rank value N (N = 3, 5, 7, 9 and 11) was computed. Results are shown in Table 3.

Table 3. Average Precision at Rank N for LA Model

Rank #	Average Precision
3	0.66
5	0.76
7	0.82
9	0.84
11	0.79

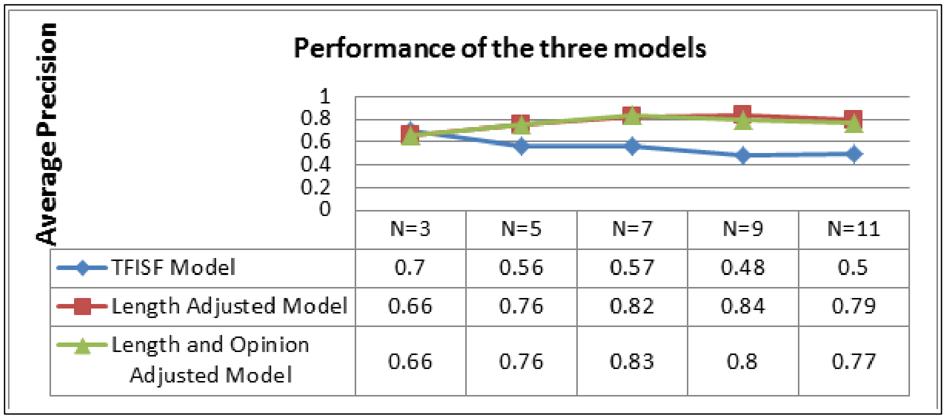
As shown in Table 3, the LA model has achieved the lowest precision at rank 3; with a value of 0.60, and then this value has increased to 0.76, 0.82, and 0.84 for the rank values of 5, 7, and 9 respectively. However, the value has dropped down again to 0.79; for rank 11. As a result, the value of the average precision at rank N was increasing from N=3 to N=9 then has dropped down at rank 11.

Finally, the LOA model was evaluated as before, the novelty score was measured then the precision at rank N was calculated. After that, the average precision at each rank N (N = 3, 5, 7, 9 and 11) was computed. Results of the average precision at rank N for these queries are shown in Table 4.

Table 4. Average Precision at Rank N for LOA

Rank #	Average Precision
3	0.66
5	0.76
7	0.83
9	0.80
11	0.77

As shown in Table 4, the LA model has achieved the lowest precision; at rank 3 with a value of 0.66, and then this value has increased to 0.76 and 0.83 for the rank values of 5 and 7 respectively. However, the value has dropped down again to 0.80 and 0.77 for the rank values of 9 and 11. As a result, the value of the average precision at rank N was increasing from N=3 to N=7 then has decreased from N=7 to N=11.



	N=3	N=5	N=7	N=9	N=11
TFISF Model	0.7	0.56	0.57	0.48	0.5
Length Adjusted Model	0.66	0.76	0.82	0.84	0.79
Length and Opinion Adjusted Model	0.66	0.76	0.83	0.8	0.77

Figure 4. Average Precision at Rank N for the Three Models

Comparing the results of the three models, we can see that the average precision for both the LA Model and the LOA Model was almost the same for different kinds of ranking N. On the other hand, the TFISF model has achieved a precision value; at rank 3 better than the other two models with a value of 0.70. However, the performance of the TFISF model has dropped down significantly for N=5 to N=11. This information is demonstrated in Figure 4. In general, we can notice and compare the improvements between the LA model and the LOA model (i.e. information patterns adjusted models). It can be noticed that the LA model outperforms the LOA model.

5. Conclusion and Future Work

Novelty detection is an important activity used to identify new information, and reduce redundancy and the number of non-relevant information presented to users of systems; such as information retrieval systems, Web search engines, document filters and cross-document summarization. It can also be used by different tasks of natural language processing (NLP); such as machine translation, summarization, and question answering systems. This study has proposed methods to improve the retrieval and novelty detection for Arabic text.

The results showed that sentence level information patterns could be successfully be applied to Arabic text to improve the relevancy and novelty score for Arabic retrieved sentences. The experiments have showed that information patterns have a significant role in novelty detection for general topics and relevance retrieval. Both of the proposed models; LA and LOA can significantly improve the performance of novelty detection for Arabic texts. It was also noticed that the LA model has provided a better performance over the LOA model.

The proposed information-pattern-based approach opens up some further research issues. These issues can be implemented to improve novelty detection for Arabic texts. These issues could be summarized in the following points:

Using other sentence level information patterns; such as named entities of Person, Organization, Location and Date (POLD) types, should be evaluated as other possible sentence level information patterns for novelty detection.

This study presents a pattern-based approach that begins with the results retrieval from TFISF techniques, and tune the believe scores of different sentences according to the length of these sentences and their opinion patterns. There is a need to conduct further research activities to integrate information patterns with other retrieval approaches beside the TFISF.

Test the approach on other corpuses: instead of collecting data from different websites, it is suggested as a future work to use a collection of E-Books as a corpus.

Use other measures for comparing performance; such as computation time was needed for each model and combining the results with the average precision that was also calculated.

Compare the information pattern based approach for novelty detection with other popular approaches that used for novelty detection at sentence level such MMR.

There is also a plan to conduct a comparative study to test the effectiveness of the following two major measures Word Count Measures (Simple New Word Count, Set Difference, Cosine distance) and Language Model Mesaures (TREC_KL, LMDiri, LMShrink, LMMix) to identify novel Arabic sentences.

References

[1] E. Greengrass, "Information Retrieval: A Survey, DOD Technical Report TR-R52-008-001", (**2000**).

[2] E. Toms, "Serendipitous Information Retrieval", In proceedings of the First DELOS Network of Excellence Workshop on Information Seeking, Searching and Querying in Digital Libraries, Zurich, Switzerland: European Research Consortium for Informatics and Mathematics, (**2000**).

[3] I. Soboroff and D. Harman, "Novelty detection: the TREC experience", In proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, (**2005**), pp. 105-112.

[4] J. Allan, C. Wade and A. Bolivar, "Retrieval and novelty detection at the sentence level", In proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada, (**2003**), pp. 314-321.

[5] F. S. Tsai, Y. Zhang, A. T. Kwee and W. Tang, "Multilingual novelty detection", Expert Systems with Applications, vol. 38, no. 1, (**2011**), pp. 652-658.

[6] L. Zhao, M. Zhang and S. Ma, "The nature of novelty detection", Information Retrieval Journal, vol. 9, no. 5, (**2006**), pp. 527-541.

[7] X. Li and W. B. Croft, "An Answer Updating Approach to Novelty Detection", CIIR Technical Report, IR-359, Department of Computer Science, University of Massachusetts Amherst, (**2004**).

[8] X. Li and W. B. Croft, "Sentence level information patterns for novelty detection", Ph.D. dissertation, University of Massachusetts Amherst, (**2006**).

[9] R. T. Fernández, "The effect of smoothing in Language Models for novelty detection", In proceedings of the BCS IRSG Symposium: Future Directions in Information Access 2007, (**2007**), pp. 11-16.

[10] X. Li and W. B. Croft, "Improving novelty detection for general topics using sentence level information patterns", In proceedings of the 15th ACM international conference on Information and knowledge management, Arlington, Virginia, USA, (**2006**), pp.238 - 247.

[11] D. Parapar and A. Barreiro, "Sentence Retrieval with LSI and Topic Identification", In proceedings of 28th European Conference on IR Research, ECIR 2006, London, UK, Lecture Notes in Computer Science 3936 Springer, ISBN 3-540-33347-9, (**2006**), pp. 119-130.

[12] Y. Zhang, J. Callan and T. Minka, "Novelty and redundancy detection in adaptive filtering", In proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland, (**2002**), pp. 81-88.

[13] J. Allan, R. Gupta, V. Khandelwal, "Temporal summaries of new topics", In proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States, (**2001**), pp. 10-18.

[14] Y. Yang, T. Pierce and J. Carbonell, "A study of retrospective and on-line event detection", In proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, (**1998**), pp. 28-36.

[15] J. Allan, R. Papka and V. Lavrenko, "On-line new event detection and tracking", In proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, (**1998**), pp. 37-45.

[16] Y. Yang, J. Zhang, J. Carbonell and C. Jin, "Topic-conditioned novelty detection", In proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada, (**2002**), pp. 688-693.

[17] R. T. Fern ández and D. E. Losada, "Novelty detection using local context analysis", In proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, (**2007**), pp. 725-726.

[18] W. Dai and R. Srihari, "Minimal document set retrieval", In proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, (**2005**), pp. 752-759.

[19] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries", In proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, Melbourne, Australia, (**1998**), pp. 335-336.

[20] K. C. Litkowski, "Use of metadata for question answering and novelty tasks", In Voorhees, E. M., Buckland, L. P., eds.: Proceedings of the Twelfth Text REtrieval Conference TREC, Gaithersburg, MD, (**2004**), pp.161-170.

[21] X. Li and W. B. Croft, "Novelty detection based on sentence level patterns", In proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, (**2005**), pp. 744–751.

[22] X. Li and W. B. Croft, "Evaluating question-answering techniques in Chinese", In proceedings of the first international conference on Human language technology research, San Diego, (**2001**), pp.1-6.

[23] X. Li, "Syntactic features in question answering", In proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada, (**2003**), pp. 383-384.

[24] Y. Al-Onaizan, "Arabic Stop Words List", California Information Science Institute, http://www.isi.edu/~yaser/arabic/arabic-stop-words.html, (**2002**).

Authors

Mohammed Al-Kabi is born in Baghdad/Iraq in 1959. He obtained his Ph.D. degree in Mathematics from the University of Lodz/Poland (2001), his Masters degree in Computer Science from the University of Baghdad/Iraq (1989), and his bachelor degree in statistics from the University of Baghdad/Iraq (1981). Mohammed Naji AL-Kabi is an assistant Professor in the Faculty of Sciences and IT, at Zarqa University. Prior to joining Zarqa University, he worked many years at Yarmouk University in Jordan, Nahrain University and Mustanserya University in Iraq. He also worked as a lecturer in Jordan University of Science and Technology. AL-Kabi's research interests include Information Retrieval, Web search engines, Data Mining, Software Engineering & Natural Language Processing. He is the author of several publications on these topics. His teaching interests focus on information retrieval, Web programming, data mining, DBMS (ORACLE & MS Access).

Esraa Shdaifat is a full time lecturer in the computer information systems department in the IT faculty at Hashemite University in Jordan. She got her master in computer information systems from Yarmouk University in Jordan. Her main research focuses in information retrieval and natural language processing.

Dr. Emad Shawakfa is an assistant professor in the computer information systems department in the IT faculty at Yarmouk University in Jordan. He got his phd in computer science from Illinois Institute of Technology, USA in 2000, his master in computer engineering from Orta Doğu Teknik Üniversitesi, Turkey, in 1989 and Bsc degree in computer science from Yarmouk University in Jordan in 1986. Research Interests are: Data Mining, Computer Networks, and Information Retrieval.

Abdullah Hamdi Wahbeh is a graduate assisstnat at Dakota State University; he is a doctoral student in the Information System (IS) program, specializing in decision support system, knowledge and data management. He has got his Master degree in Computer Information System (CIS) from Yarmouk university, Irbid-Jordan, in 2009; he also obtained his Bachelor degree in Compuetr Information System (CIS) from Yarmouk university, Irbid-Jordan, in 2007. His research interest is concentrated on information security, health informatics, data mining, web mining and information retrieval. Address: 1051 N Summit Ave. Apt 13, Madison, SD 57042 [email: ahwahbeh@pluto.dsu.edu]

Izzat Mahmoud Alsmadi is an associate professor in the department of computer information systems at Yarmouk University in Jordan. He obtained his Ph.D degree in software engineering from NDSU (USA). His second master in software engineering from NDSU (USA) and his first master in CIS from University of Phoenix (USA). He had B.sc degree in telecommunication engineering from Mutah university in Jordan. Before joining Yarmouk University he worked for several years in several companies and institutions in Jordan, USA and UAE. His research interests include: software engineering, software testing, e-learning, software metrics and formal methods. Address: IT faculty, Yarmouk university, 21163 Irbid, Jordan [email: ialsmadi@yu.edu.jo]