International Journal of Engineering & Technology, 9 (2) (2020) 342-349

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Developing a new approach to summarize Arabic text automatically using syntactic and semantic analysis

Amal Alkhudari *

Department of Information Technology, University of Kalamoon, Syria
**Corresponding author E-mail: amalkhudari@gmail.com*

Abstract

Due to the wide spread information and the diversity of its sources, there is a need to produce an accurate text summary with the least time and effort. This summary must preserve key information content and overall meaning of the original text. Text summarization is one of the most important applications of Natural Language Processing (NLP). The goal of automatic text summarization is to create summaries that are similar to human-created ones. However, in many cases, the readability of created summaries is not satisfactory, because the summaries do not consider the meaning of the words and do not cover all the semantically relevant aspects of data. In this paper we use syntactic and semantic analysis to propose an automatic system of Arabic texts summarization. This system is capable of understanding the meaning of information and retrieves only the relevant part. The effectiveness and evaluation of the proposed work are demonstrated under EASC corpus using Rouge measure. The generated summaries will be compared against those done by human and precedent researches.

Keywords: *Abstractive Summarization; Ontology; Semantic Similarity; Syntactic Analysis; Word Sense Disambiguation*.

1. Introduction

There is an ever-increasing need for better automatic systems of Arabic text summarization with the explosion in the amount of information available. We find huge information online daily in the unstructured documents specifically. Information retrieval from unstructured text is more complex than structured or semi-structured text. It is a big challenge to analyze and retrieve Arabic information because of the difficulties in manipulating Arabic language and lacking of researches and tools about Arabic language processing.

Automatic text summarization has many features such as: number of input documents (single or multiple), purpose (generic, domain specific, or query-based), Output (Informative or Indicative). There are two major approaches to summarize a text:

- Extractive Method: This type identifies the important sections of the text depending on statistics like word location and number of its repetition through the text. This method does not provide accurate results because it generates non concise subset of the sentences from the original text. Therefore, the new text content is not trusted because of the less level of importance related information.
- Abstractive Method :This type of summary generates a new brief text which contains accurate and non-duplicate information. It understands the whole text depending on the concepts of the words and its significance. To accomplish this we must know about the science of linguistics. The abstractive summarization methods under semantic based approach rely on semantic representation of the original document text. These methods produce concise, rich information, coherent, and less redundant summary as well as improve the linguistic quality of the summary (1).

Obviously, abstractive summarization is more advanced and closer to human-like interpretation.

The proposed system produces a generic and informative single Arabic document summarization. It depends on the concepts of the words and semantic relations between them.

The main contributions of this paper are as follows: First: It introduces a good study for semantic similarity techniques where they can be used in many applications of NLP. Second, the proposed method is domain independent that does not need any domain-specific knowledge or features. Finally: the proposed method is efficient and precise, and the applied experiments demonstrate them.

The rest of the paper is structured as follows: Section two gives insights into related works for text summarization techniques and especially for Arabic researches. Section three presents Challenges in Arabic NLP (Natural Language Processing). Section four presents features of using ontology. Section five talks about Arabic WordNet Ontology. Section six Studies the measurement techniques of semantic similarity. The proposed method is described in section seven. The data set, experiments and result evaluation are described in section eight. Finally, in section nine, a conclusion and perspectives are presented.

2. Related works

Automatic text summarization gained attraction as early as the 1950s. It is very challenging, because the summary must be concise and fluent while preserving key information content and overall meaning. Luhn et al. (2) introduced a method to extract sentences from the text using features such as word and phrase frequency. They proposed to weight the sentences of a document as a function of high frequency words, ignoring very high frequency common words.

Arabic Text summarization is still in its infancy compared to the literature on English. It has started by work of (Conroy et al., 2006; Douzidia and Lapalme, 2004). Oufaida et al. (3) presented extractive summarization system for both single document and multi-documents; the sentences to be summarized were selected based on the ranks of their terms.

S.Ismail et al. (4) worked on three modules; first they convert the input Arabic text into a semantic graph called Rich Semantic Graph (RSG). The second module is performing graph reduction. The Last module is generating the summary from the reduced graph.

M. A. Alwan et al. (5) proposed a model of four stages, preprocessing, representing the multi-documents by directed weighted graph, traversing the graph and finally applying structural rules to generate summarized sentences. Azmia et al. (6) integrated the advantages of an RST-based system and Frequency computation. They assumed the higher the frequency the more important is the word. Al Breem (7) built an automatic text summarization for large-scale multi Arabic documents using Genetic algorithm and MapReduce mode. In (8) Researchers applied clustering algorithms to group documents into many clusters. Then, they used Key phrase extraction to extract the important Key phrases from each cluster. In (9) Researchers used Ontology for extracting concepts and defining semantic relations between them. Then, they applied decision tree algorithm for generating summary. A. Qaroush et al. (10) considered that sentences which contain cue- words or strong ones must be in the summary, whereas the weak words refer to unimportant sentences. Also. They proposed machine learning based approach which use many statistical features such as sentence’s length and location.

Unlike previous studies that introduced extractive summarizer using various statistical techniques, our work focuses on analyzing words based on their semantic meanings. We use syntactic and semantic analysis in order to retrieve the most relevant sentences whereas the poor one will not be in the summary. We achieve both semantics objectives namely coverage and diversity.

3. Arabic language forms and challenges in Arabic NLP

Arabic Language is the largest group of Semitic languages. It is the native language for more than four hundred million centered in the Arabic region. The Arabic alphabet consists of 25 permanent characters and 3 audio characters that take different forms depending on their position in the text. Semantic processing for Arabic language tends to be more complex than it is for English Language because of: The absence of capitalization in Arabic, makes it hard to identify titles, acronyms, and abbreviations. Also, Arabic is derived, which makes morphological analysis a very complex task (11).

4. Ontology

Ontology is a representation on the level of word meanings, independently of a particular application (General Domain) such as WordNet. WordNet as a lexical resource offers broad coverage of the general lexicon. It has been employed as a resource for many applications in information retrieval. Knowledge of words lies not only in their meanings but also in the context in which they occur. Linking words to appropriate senses provides the desired conceptual information. Terms holding identical meanings are organized around the notion of a synset. Synsets are linked to each other via pre-defined lexical relations (12).

5. Arabic wordnet ontology

Arabic WordNet is currently under construction following a methodology developed for Euro WordNet. It consists of 11,270 synsets, (7,961 nominal, 2,538 verbal, 661 adjectival, and 110 adverbial), containing 23,496 Arabic expressions. This number includes 1,142 synsets that correspond to named entities which have been extracted automatically and are being checked by the lexicographers (13).

6. Measurement techniques of semantic similarity

Semantic is the study of words’ meaning, their structure, and their relationships with other words. Measuring of semantic similarity between texts is considered an important filed in the applications of artificial intelligence and computational linguistics like document summarization, text mining, machine translation and many others. Semantic similarity is a metric defined over a set of documents or terms, which refers to the proximity of two concepts within a given ontology. The distance between two concepts is a numerical representation of how far apart two concepts are in some geometric space, and can be considered the inverse of semantic similarity (i.e. if distance between concepts is ‘0’ then the semantic similarity is ‘1’ and vice versa). If this relationship between distance and semantic similarity holds, having similarity or distance metrics allows the use of the ontology to search efficiently for related items, or to identify associations between concepts that may not be immediately obvious to the user. However, it is a challenging task since it has difficulties in using semantic analysis tools and linguistic resources like WordNet. They require memory for saving the semantic information, and processor capacity for additional linguistic and semantic knowledge processing (14).

We can classify the main methods of measuring the semantic similarity by the type of knowledge representation (sources of information):

6.1. Corpus based measure

Corpus-based measures of word semantic similarity try to identify the degree of similarity between words using information statistically exclusively derived from large corpora. We can also conclude the similarity between the sentences depending on the co-occurrences of words within the corpus (15).

The measure introduced by Resnik (1995) returns the information content (IC) of the (LCS) Longest Common Sequence of two concepts: (16)

$$\mathrm{Sim}_{\mathrm{res}=}\ \mathrm{IC(LCS)} \quad (1)$$

Where IC is defined as:
$\mathrm{IC(c)} = -\log \mathrm{P(c)}$ (2) P(c) is the probability of encountering an instance of concept c in a large corpus.

6.2. Knowledge based measure

This approach is based only on the hierarchy or the edge distances. The taxonomy arcs represent uniform distances, i.e. all the semantic links have the same weight (17).
The researcher Miller used the hierarchical semantic dictionary (WordNet) to measure semantic similarity by identifying the distances between concepts. The smaller the conceptual distance between concepts, the greater the similarity between them. The value of the similarities varies according to the layer of ontology. For example, the words in the upper layer have more abstract concepts and therefore are less similar, unlike words in the lower layers, which have a deeper meaning and thus are more similar. One of the most important and popular knowledge-bases is WordNet. Wu and Palmer method of measuring semantic similarity is one of the most popular methods due to its computational speed. It measures semantic similarity between two nodes in taxonomy. The principle of its computation is based on the depth of nodes (concept1, concept2) from the root node and the distance which separates the LCS (Least–Common–Subsumer) of concept1and concept2 from the root node. Shorter distance between two concepts gives more similarity value. The similarity measure is defined by the following expression (18):

$$\mathrm{Sim(c1,c2)}=\frac{2*\mathrm{depth(LCS)}}{\mathrm{depth(concept1)+depth(concept2)}} \quad (3)$$

6.3. Hybrid measures

This hybrid approach combines the features of the two previous approaches. It brings us better results and higher evaluation. It can use several sources of information and incorporate more than one approach to measure semantic similarities like shortest path between two concepts, information content, semantic density of the concept, edge-counting and link weight.
Li et al. also used WorldNet Ontology. They consider the shortest length between two concepts and the depth of their lowest common subsumer to compute similarity (19). The similarity between concepts c1 and c2 is defined as non-linear function:

$$Sim_{li}(C1,C2) = e^{-\propto * SP} * \frac{e^{\beta*N} - e^{-\beta*N}}{e^{\beta*N} + e^{-\beta*N}} \quad (4)$$

Sp: represents shortest path between two nodes.
N: represents depth of (LCS) in the taxonomy.
α and β refer to parameters scaling the contribution of the shortest path length and depth, respectively. Based on empirical study, the optimal parameters are α=0.2 and β=0.6.

7. Proposed work

We introduce abstractive summaries for Arabic free texts. The role of our system is to generate a summary by picking out sentences which are most relevant and contains the main ideas presented in the document. The system has four main stages which are morphological processing, syntactic analysis, semantic analysis and generating the summary. Morphological processing converts the original text into a structured form. It includes sentence segmentation, word segmentation, stop-words removal, normalization and root extraction. In syntactic analysis stage, we use part of speech to identify which phrases must be chunked and extracted. In semantic analysis stage, we solve word sense disambiguation and measure semantic similarity of all sentences in order to retrieve the important ones. Final stage, we generate the summary based on their scores of similarity and location in the original text. (Figure 1)

7.1. Morphological processing

Next, we will describe steps of morphological processing in more details:

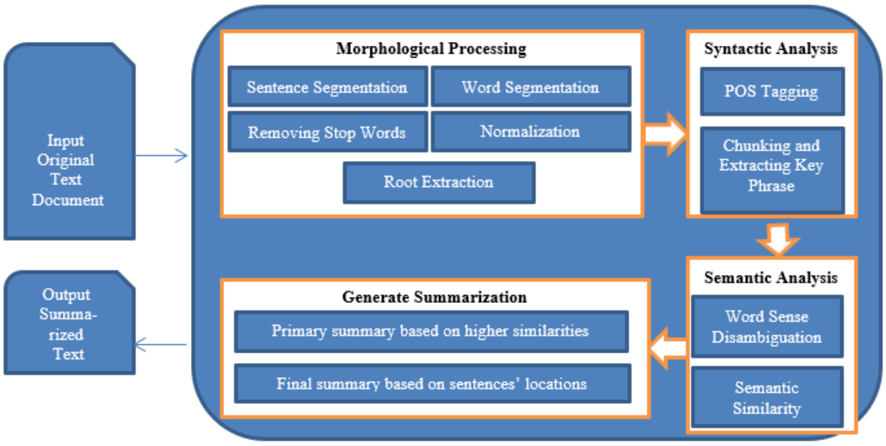


Fig. 1: NLP-Based Stages.

7.1.1. Sentence segmentation

At this step the text is unstructured, we split it into sentences. We extract each sentence from the original text by finding the sentence boundaries. Sentences are assumed to be separated by period, exclamation mark, or commas.

7.1.2. Word segmentation

At this step, we split the sentences into words to get the Tokens.

7.1.3. Removing stop words

It is an essential step to help identifying the most important words; very common words that appear in the text but carry little meaning serve only a syntactic function but do not indicate subject matter. Also, removing stop words help reducing the size of data and time that is required for text processing in next steps.

7.1.4. Normalization

Before further processing, text needs to be normalized. Normalization generally refers to converting all text to the same case such as the normalization of (hamza) (إ) or (أ)in all its forms to (alef (ا)), and (Taa) (ة) to (Haa (ه)). Normalization puts all words on equal footing, and allows processing to proceed uniformly.

7.1.5. Root extraction

Arabic words are classified into three main categories: nouns, verbs and particles. These words are derived from a root word by adding affixes, which are classified into four categories: particles, pronouns, inflectional morphemes, and derivational morphemes. ISRI algorithm is used for rooting. The following table shows an example of this stage:

Table 1: Sample Output of Morphological Processing

Steps	Example Output
Sentence Segmentation	طبقة الأوزون هي الجزء من الغلاف الجوي لكوكب الأرض. (1) وهي متمركزة بشكل كبير في الجزء السفلي من طبقة الستراتوسفير من الغلاف الجوي للأرض. (2)
Word Segmentation	طبقة (1) الأوزون (2) هي (3) الجزء (4) من (5) الغلاف (6) الجوي (7) لكوكب (8) الأرض (9)
Stop Words Removal	طبقة (1) الأوزون (2) الجزء (3) الغلاف (4) الجوي (5) لكوكب (6) الأرض (7)
Normalization	طبقة (1) الاوزون (2) الجزء (3) الغلاف (4) الجوي (5) لكوكب (6) الارض (7)
Root Extraction	طبق (1) ازو (2) جزا (3) غلف (4) جوي (5) كوكب (6) ارض (7)

7.2. Syntactic analysis

The purpose of this stage is to extract the key phrases which conveys the gist of the meaning of the text. We use POS (Part of Speech) tagging followed by pattern-based chunking and extraction. Certain tags are more informative than others. For example, Noun tags (starting with NN) would carry more information than prepositions or conjunctions. Similarly, if we would like to know "what" is being spoken about, Noun words may be more relevant than others. Also, chunks of words would carry more meaning than looking at individual words in isolation. We use (edu.stanford.nlp.ling) package. It contains the different data structures used by JavaNLP for dealing with linguistic objects in general. Tag class for linguistic concepts is used to detect types of all words in the text. For chunking the POS tagged text, we have to define what POS pattern we would consider as a chunk. Noun-Adjective combination (NN||NNS - JJ||JJR||JJS), Noun-Noun combination (NN||NNS - NN||NNS) can be a useful pattern to extract. Also, It is important to chunk and extract proper nouns (NNP||NNPS- NNP||NNPS). Next, we extract chunks matching pattern. Key phrases which consist of two or three words and found in many sentences will be used in the last stage to increase score of those sentences.

7.3. Semantic analysis

Arabic WordNet 2.0 in format of XML is used to represent text. Representations in WordNet are not on the level of individual words or word forms, but on the level of word meanings. A word meaning, in turn, is characterized by simply listing the word forms that can be used to express it in a synonym set (synset). Each node is a synset that represents a concept. As a result, the meaning of the word is determined by its sets of synonyms. This is essentially a recursive definition of word meaning. Hence meaning in WordNet is a structural notion: the meaning of a word is determined by its position relative to the other words.(13). In our proposed work, WordNet is used for extract relationships between concepts by measuring the semantic similarity between them in order to solve the ambiguity of the words' meaning and retrieve the sentences that are the most relevant. We associate the words in context with their most suitable entry in a pre-defined sense inventory (WordNet). To do that, We solve word sense disambiguation by measuring the semantic similarity for each concept of the word with the concepts of three words before it and concepts of three words after it, then we choose the only one closest concept (sense) of the word with the highest similarity value and closest to the meaning of the text. Then, we measure the semantic similarity to compute similarity of each word‘s sense to all words senses in the text. The method of Li measure (19) which depends on the shortest length between two concepts and the depth of their lowest common subsumer is used in our proposed work. Each sentence has a score of the semantic similarity which is equal sum of its words' semantic similarity scores. We increase the score of the sentences that contain important key phrases that are extracted in syntactic analysis stage. Usage of key phrase is very useful for texts that contain un semantically-related proper nouns.

7.4. Summary sentences selection

Eventually, the system selects the most important sentences to produce a summary. The sentences are arranged based on their score of semantic similarity descending from the highest to the lowest. The first 45% of sentences have been chosen, and not more than 50% of words. Finally, the extracted sentences are reordered based on their position in the original document to preserve text coherency and arrangement of ideas in the generated summary.

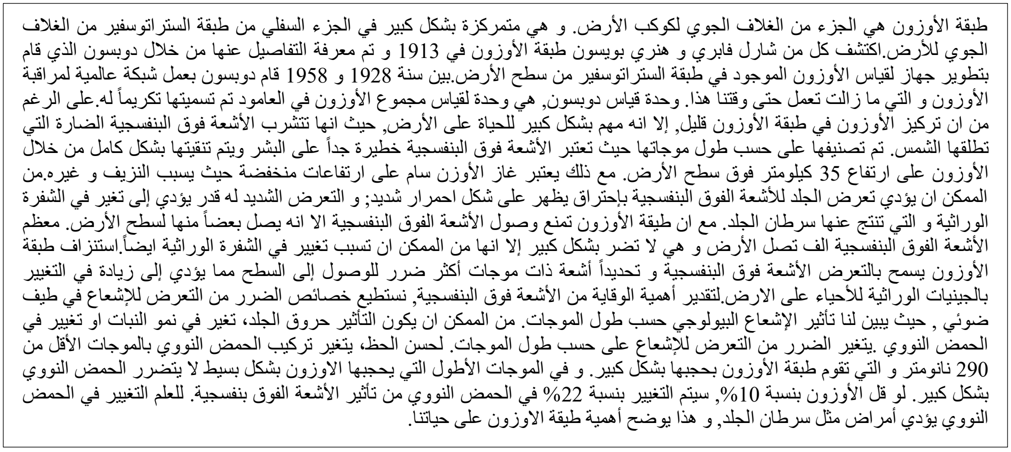
8. Experiments and result evaluation

Evaluation of a summary is a difficult task because there is no ideal summary for both single document and a collection of documents. It has been found that human summarizers have low agreement for evaluating and producing summaries. There has been a set of metrics to automatically evaluate summaries since the early 2000s. Therefore, Essex Arabic Summaries Corpus (EASC) (EL-Haj et al., 2010) has been used for testing and evaluating the proposed method. EASC corpus is a human-generated extractive summary published by a group of researchers at Essex University. It comprises 153 articles on different topics and 765 human-generated extractive summaries of those articles which have been collected from Arabic newspapers and Wikipedia. For each article in the EASC corpus there are five different reference-summaries; each reference summary is generated by a different human. ROUGE is the most widely used metric for automatic evaluation. It introduced a set of metrics called Recall-Oriented Understudy for Gisting Evaluation (ROUGE) to automatically determine the quality of this summary by comparing it to the human (reference) summaries. We used ROUGE 2.0 API which is language independent Java package for summary tasks. (20)
To compare our summaries with those human generated summaries as the benchmark; One thing to note is that the Arabic used in these sample texts is what we currently term Modern Standard Arabic. The results show that the system is able to abstract the most important concepts which are collected from different parts of the text. The proposed work used precision and recall for
evaluation. The recall and precision can be computed as:
Precision is the number of document retrieved that are relevant and Recall is the number of relevant document that are retrieved.
Ra: Number of correctly retrieved documents
A: Total number of documents retrieved
R: Total number of relevant document retrieved
Precision = Ra / A Recall = Ra / R
F-measure is to combine precision and recall into a single measure. This measure usually referred to as F-score.

$$\mathrm{F-measure} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{5}$$

8.1. Evaluation of NLP-based met (HOD)

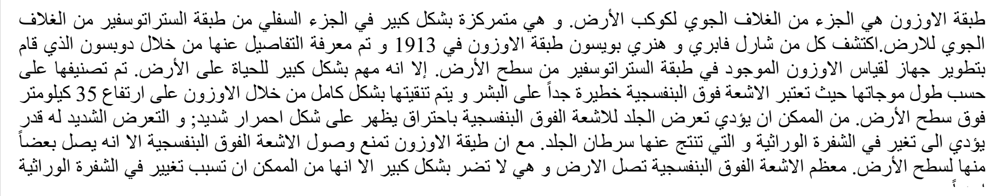
Example: single Arabic document consists of 379 words.



طبقة الأوزون هي الجزء من الغلاف الجوي لكوكب الأرض. و هي متمركزة بشكل كبير في الجزء السفلي من طبقة الستراتوسفير من الغلاف
الجوي للأرض.اكتشف كل من شارل فابري و هنري بويسون طبقة الأوزون في 1913 و تم معرفة التفاصيل عنها من خلال دوبسون الذي قام
بتطوير جهاز لقياس الأوزون الموجود في طبقة الستراتوسفير من سطح الأرض.بين سنة 1928 و 1958 قام دوبسون بعمل شبكة عالمية لمراقبة
الأوزون و التي ما زالت تعمل حتى وقتنا هذا. وحدة قياس دوبسون, هي وحدة لقياس مجموع الأوزون في العامود تم تسميتها تكريماً له.على الرغم
من ان تركيز الأوزون في طبقة الأوزون قليل, إلا انه مهم بشكل كبير للحياة على الأرض, حيث انها تتشرب الأشعة فوق البنفسجية الضارة التي
تطلقها الشمس. تم تصنيفها على حسب طول موجاتها حيث تعتبر الأشعة فوق البنفسجية خطيرة جداً على البشر ويتم تنقيتها بشكل كامل من خلال
الأوزون على ارتفاع 35 كيلومتر فوق سطح الأرض. مع ذلك يعتبر غاز الأوزن سام على ارتفاعات منخفضة حيث يسبب النزيف و غيره.من
الممكن ان يؤدي تعرض الجلد للأشعة الفوق البنفسجية بإحتراق يظهر على شكل احمرار شديد; و التعرض الشديد له قدر يؤدي إلى تغير في الشفرة
الوراثية و التي تنتج عنها سرطان الجلد. مع ان طيقة الأوزون تمنع وصول الأشعة الفوق البنفسجية الا انه يصل بعضاً منها لسطح الأرض. معظم
الأشعة الفوق البنفسجية الف تصل الأرض و هي لا تضر بشكل كبير إلا انها من الممكن ان تسبب تغيير في الشفرة الوراثية ايضاً.استنزاف طبقة
الأوزون يسمح بالتعرض الأشعة فوق البنفسجية و تحديداً أشعة ذات موجات أكثر ضرر للوصول إلى السطح مما يؤدي إلى زيادة في التغيير
بالجينيات الوراثية للأحياء على الارض.لتقدير أهمية الوقاية من الأشعة فوق البنفسجية, نستطيع خصائص الضرر من التعرض للإشعاع في طيف
ضوئي , حيث يبين لنا تأثير الإشعاع البيولوجي حسب طول الموجات. من الممكن ان يكون التأثير حروق الجلد، تغير في نمو النبات او تغيير في
الحمض النووي .يتغير الضرر من التعرض للإشعاع على حسب طول الموجات. لحسن الحظ، يتغير تركيب الحمض النووي بالموجات الأقل من
290 نانومتر و التي تقوم طبقة الأوزون بحجبها بشكل كبير. و في الموجات الأطول التي يحجبها الاوزون بشكل بسيط لا يتضرر الحمض النووي
بشكل كبير. لو قل الأوزون بنسبة 10%, سيتم التغيير بنسبة 22% في الحمض النووي من تأثير الأشعة الفوق بنفسجية. للعلم التغيير في الحمض
النووي يؤدي أمراض مثل سرطان الجلد, و هذا يوضح أهمية طيقة الاوزون على حياتنا.

Fig. 2: Sample Input Text.

Summarized text consists of 165 words:



طبقة الاوزون هي الجزء من الغلاف الجوي لكوكب الأرض. و هي متمركزة بشكل كبير في الجزء السفلي من طبقة الستراتوسفير من الغلاف
الجوي للارض.اكتشف كل من شارل فابري و هنري بويسون طبقة الاوزون في 1913 و تم معرفة التفاصيل عنها من خلال دوبسون الذي قام
بتطوير جهاز لقياس الاوزون الموجود في طبقة الستراتوسفير من سطح الأرض. إلا انه مهم بشكل كبير للحياة على الأرض. تم تصنيفها على
حسب طول موجاتها حيث تعتبر الاشعة فوق البنفسجية خطيرة جداً على البشر و يتم تنقيتها بشكل كامل من خلال الاوزون على ارتفاع 35 كيلومتر
فوق سطح الأرض. من الممكن ان يؤدي تعرض الجلد للاشعة الفوق البنفسجية باحتراق يظهر على شكل احمرار شديد; و التعرض الشديد له قدر
يؤدي الى تغير في الشفرة الوراثية و التي تنتج عنها سرطان الجلد. مع ان طيقة الاوزون تمنع وصول الاشعة الفوق البنفسجية الا انه يصل بعضاً
منها لسطح الأرض. معظم الاشعة الفوق البنفسجية تصل الارض و هي لا تضر بشكل كبير الا انها من الممكن ان تسبب تغيير في الشفرة الوراثية

Fig. 3: Summary Generated by NLP-Based Approach.

The result is compared well with five references summary using Rouge1 metric. Table 2 shows the Recall, Precision and F-Score for summary generated by our approach we have introduced for the text in Figure 3

Table 2: Performance of NLP-Based Summarization Technique

Reference Summary	Rouge-1		
Five reference summaries	Recall	Precision	F-Score
	0.575	0.576	0.575

Our approach uses Natural Language Processing techniques. We focus on syntactic analysis and semantics similarity measurement. It seems to perform well on many types of texts. The results were good after evaluation under Essex Arabic corpus as follow:

Table 3: Performance of NLP-Based Approach Under Essex Arabic Corpus

Five references summary	Rouge-1		
Dataset	Recall	Precision	F-Score
Art	0.513	0.451	0.475
Education	0.675	0.438	0.517
Environment	0.570	0.504	0.526
Politic	0.563	0.451	0.495
Science	0.556	0.452	0.489
Religion	0.542	0.408	0.462
Tourism	0.600	0.364	0.442
Sport	0.576	0.498	0.524
Finance	0.588	0.472	0.516
Health	0.533	0.549	0.530
Total- Average	0.572	0.459	0.498

8.2. Comparing to related works

In this section, results of our proposed method are compared with results of other related Arabic summarization methods based on their published results in terms of recall, precision, and F-Score (10).

إمارة دبي هي ثاني الإمارات المكونة لدولة الإمارات العربية المتحدة وعاصمتها مدينة دبي. تشكل هذه الإمارة مركزاً هاماً للمال والأعمال في العالم، ووجهة سياحية يقصدها الملايين من السياح سنويا. دبي هي العاصمة الاقتصادية للإمارات العربية المتحدة، وقد تطورت تطوراً كبيراً خلال السنوات الماضية. الاقتصاد الحر والنشط في الامارة وعدم وجود نظام ضريبي لعب دوراً كبيراً في جذب المستثمرين من جميع أنحاء العالم.وتقع إمارة دبي بين إمارتي أبو ظبي و الشارقة. وأهل امارة دبي ينحدرون من قبائل عربية متنوعة، على رأسها قبيلة آل بو فلاسه التي تنحدر منها أسرة آل مكتوم الحالكمة. وتقطنها قبائل بني كعب وال بوفلاح وال بو مهير والسودان والشوامس والبلوش والمناصير والرميثات والشحوح وغيرهم. زبها عوائل كثيرة من أصول أفريقية وفارسية. ودين أهالي دبي الاسلام على نهج أهل السنة والجماعة. والمذهب الرسمي في دبي هو المذهب المالكي. آل مكتوم هم حكام دبي. وهم من ال بو فلاسه من بني ياس. حاكمها الآن هو الشيخ محمد بن راشد آل مكتوم.وهو أيضاً نائب لرئيس الدولة و رئيس لمجلس الوزراء في الحكومة الاتحادية. ونائبيه في الحكم هما شقيقه: الشيخ حمدان بن راشد آل مكتوم وزير المالية والصناعة والشيخ مكتوم بن محمد بن راشد آل مكتوم. يرأس المجلس التنفيذي لحكومة دبي الشيخ حمدان بن محمد بن راشد أل مكتوم. ويجمع هذا المجلس في عضويته جميع مدراء الدوائر في حكومة دبي حيث يعقدون اجتماعاتهم الدورية لتسيير شؤون الامارة.

Fig. 4: Sample Input Document.

Golden Reference

إمارة دبي هي ثاني الإمارات المكونة لدولة الإمارات العربية المتحدة وعاصمتها مدينة دبي. دبي هي العاصمة الاقتصادية للإمارات العربية المتحدة، وقد تطورت تطوراً كبيراً خلال السنةات الماضية. وتقع إمارة دبي بين إمارتي أبو ظبي والشارقة. حاكمها الآن هو الشيخ محمد بن راشد آل مكتوم. ونائبيه في الحكم هما شقيقه: الشيخ حمدان بن راشد آل مكتوم وزير المالية والصناعة والشيخ مكتوم بن محمد بن راشد آل مكتوم. يرأس المجلس التنفيذي لحكومة دبي الشيخ حمدان بن محمد بن راشد أل مكتوم.

Summary Generated by F-Score

إمارة دبي هي ثاني الإمارات المكونة لدولة الإمارات العربية المتحدة وعاصمتها مدينة دبي. تشكل هذه الإمارة مركزاً هاماً للمال والأعمال في العالم، ووجهة سياحية يقصدها الملايين من السياح سنويا وأهل امارة دبي ينحدرون من قبائل عربية متنوعة، على رأسها قبيلة بو فلاسه التي تنحدر منها أسرة آل مكتوم الحاكمة. بينما يتولى منصب ولاية العهد بالامارة الشيخ حمدان بن محمد بن راشد آل مكتوم رئيس المجلس التنفيذي للإمارة. يرأس المجلس التنفيذي لحكومة دبي الشيخ حمدان بن محمد بن محمد بن راشد أل مكتوم. ويجمع هذا المجلس في عضويته جميع مدراء الدوائر في حكومة دبي حيث يعقدوا اجتماعاتهم الدورية لتسيير شؤون الإمارة.

Summary Generated by Machine Learning

تشكل هذه الإمارة مركزاً هاماً للمال والأعمال في العالم. ووجهة سياحية يقصدها الملايين من السياح سنويا وأهل امارة دبي ينحدرون من قبائل عربية متنوعة.الاقتصاد الحر والنشط في الامارة وعدم وجود نظام ضريبي لعب دوراً كبيراً في جذب المستثمرين من جميع أنحاء العالم. وتقطنها قبائل بني كعب وال بوفلاح وال بو مهير والسودان والشوامس والبلوش والمناصير والرميثات والشحوح وغيرهم.وبها عوائل كثيرة من أصول أفريقية وفارسية. ودين أهالي دبي الاسلام على نهج أهل السنة والجماعة. والمذهب الرسمي في دبي هو المذهب المالكي. آل مكتوم هم حكام دبي. وهم من ال بو فلاسه من بني ياس. حاكمها الآن هو الشيخ محمد بن راشد آل مكتوم.وهو أيضاً نائب لرئيس الدولة و رئيس لمجلس الوزراء في الحكومة الاتحادية. ونائبيه في الحكم هما شقيقه: الشيخ حمدان بن راشد آل مكتوم وزير المالية والصناعة والشيخ مكتوم بن محمد بن راشد آل مكتوم. يرأس المجلس التنفيذي لحكومة دبي الشيخ حمدان بن محمد بن راشد أل مكتوم.

Summary Generated by NLP-Based Approach

امارة دبي هي ثاني الامارات المكونة لدولة الإمارات العربية المتحدة وعاصمتها مدينة دبي. تشكل هذه الامارة مركزاً هاماً للمال والأعمال في العالم. الاقتصاد الحر والنشط في الامارة وعدم وجود نظام ضريبي لعب دوراً كبيراً في جذب المستثمرين من جميع انحاء العالم.وتقع امارة دبي بين امارتي ابو ظبي و الشارقة. ودين اهالي دبي الاسلام على نهج اهل السنة والجماعة. حاكمها الآن هو الشيخ محمد بن راشد آل مكتوم. ونائبيه في الحكم هما شقيقه: الشيخ حمدان بن راشد آل مكتوم وزير المالية والصناعة والشيخ مكتوم بن محمد بن راشد آل مكتوم. يرأس المجلس التنفيذي لحكومة دبي الشيخ حمدان بن محمد بن راشد أل مكتوم.

Fig. 5: Summary Generated by Score-Based, Machine-Learning And NLP-Based Approach.

As shown in figure 6, our NLP-based method introduces better results:

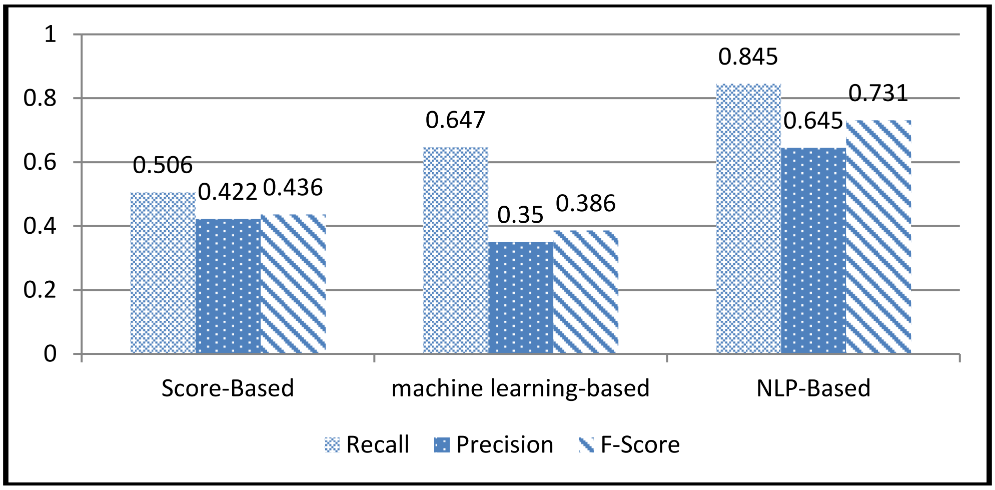


Fig. 6: Comparison of Summary Generated by Score-Based, Machine-Learning and NLP-Based Approach.

The proposed NLP-based method outperforms the others in term of recall, precision and F-score by values of 0.572, 0.459 and 0.498 respectively. This is due to the usage of NLP tasks; syntactic and semantic analysis retrieve a good summary unlike that method depends on statistical features. Our results are compared with score-based and machine learning-based (10) under EASC using rouge-1 and five reference summaries in Table 5.

Table 5: Comparison of NLP-Based Method with Other Related Summarization Methods Using Rouge-1 Under Five References Summary

Reference Summary	System Name	Recall	Precision	F-Score
Five reference Summaries	score-based approach	0.513	0.388	0.442
	machine-learning approach	0.546	0.362	0.405
	NLP-based approach	0.572	0.459	0.498

9. Conclusion

In this paper, we introduced an abstractive Arabic summarization system using syntactic analysis and semantic similarity. The extracted chunks convey some of the key themes presented in the text. The essential sentences of the original document are identified based on their score of semantic similarity. In addition, the meaning of the original document is preserved. The use of Part of Speech (POS) with word senses disambiguation (WSD) and semantic similarity promote quality of automatic text summarization system. It generates more coherent, less redundant and more informative summaries.

Our proposed method is compared well with EASC dataset. The discovered results are interesting. Using ROUGE as a performance measure, our system achieved 0.572, 0.459 and 0.498 for recall, precision and F-score respectively. The highest result was identified in the texts related to health with F-score (0.530) because most of their sentences' length is moderate and have more related concepts. The lowest result was F-score (0.442) which is found in the texts related to tourism that contain too long sentences and use a lot of words semantically unrelated. Therefore, we used key phrase extraction to increase importance of sentences which contain main words combined as key phrase and not related to other words semantically. Future improvements of the summarization system are the generating summary for multi languages, multi documents and using additional features represent the important ideas in the text.

Acknowledgement

The author gratefully acknowledges the support offered by her family.

References

[1] I. F. Moawad and M. Aref, Semantic graph reduction approach for abstractive Text Summarization, in Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on, 2012, pp. 132-138. https://doi.org/10.1109/ICCES.2012.6408498.

[2] Hans Peter Luhn, The automatic creation of literature abstracts, IBM Journal of research and development 2,2(1958), 159–165. https://doi.org/10.1147/rd.22.0159.

[3] H. Oufaida, O. Nouali, and P. Blache, Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization, Journal of King Saud University-Computer and Information Sciences, vol. 26, no. 4, pp. 450–461, (2014). https://doi.org/10.1016/j.jksuci.2014.06.008.

[4] S. S. Ismail, M. Aref, and I. F. Moawad, A model for generating Arabic text from semantic representation, in 2015 11th International Computer Engineering Conference (ICENCO). IEEE, pp. 117–122, (2015). https://doi.org/10.1109/ICENCO.2015.7416335.

[5] Muneer A. Alwan, Hoda M. Onsi, A Proposed Textual Graph Based Model for Arabic Multi-document Summarization, International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, (2016). https://doi.org/10.14569/IJACSA.2016.070656.

[6] Aqil M. Azmia, Suha Al-Thanyyan, A text summarizer for Arabic, SciVerse Science Direct Computer Speech and Language 26 260–273, (2012) https://doi.org/10.1016/j.csl.2012.01.002.

[7] Sulaiman Nasrallah Al Breem, Automatic Arabic Text Summarization for Large Scale Multiple Documents Using Genetic Algorithm and MapReduce, October (2016).

[8] Hamzah Noori Fejer and Nazlia Omar, Automatic Multi-Document Arabic Text Summarization Using Clustering and Keyphrase Extraction, Journal of Artificial Intelligence 8 (1): 1-9, (2015). https://doi.org/10.3923/jai.2015.1.9.

[9] Ibrahim Imam, Alaa Hamouda, Hebat Allah Abdul Khalek. An Ontology based Summarization System for Arabic Documents (OSSAD), International Journal of Computer Applications (0975 – 8887) Vol 74– No.17, July (2013). https://doi.org/10.5120/12980-0237.

[10] A. Qaroush, I. Abu Farha, W. Ghanem et al. An efficient single document Arabic text summarization using a combination of statistical and semantic features, Journal of King Saud University – Computer and Information Sciences, March (2019). https://doi.org/10.1016/j.jksuci.2019.03.010.

[11] Mahmoud El-Haj. Multi-document Arabic Text Summarisation, Computer Science and Electronic Engineering Conference (CEEC), (2012). https://doi.org/10.1109/CEEC.2011.5995822.

[12] Jaap Kamp, Visualizing WordNet structure, Researchgate, January (2002).

[13] Bill Black, Piek Vossen and Adam Pease, Articulate Software Arabic WordNet and the Challenges of Arabic, (2014).

[14] Khan, Atif, A review on abstractive summarization methods, J. Theor. Appl. Inform. Tech. 59, 64–72 (2014).

[15] Mehran Sahami, Timothy D. Heilman. A web-base kernel function for measuring the similarity of short text snippets, Proceedings of the 15th International Conference on World Wide Web, Scotland, (2006). https://doi.org/10.1145/1135777.1135834.

[16] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language, Journal of Artificial Intelligence Research, (1999). https://doi.org/10.1613/jair.514.

[17] Manjula Shenoy.K, Dr.K.C.Shet, Dr. U.Dinesh Acharya. a new similarity measure for taxonomy based on edge counting, International Journal of Web & Semantic Technology, Vol.3, No.4, October (2012). https://doi.org/10.5121/ijwest.2012.3403.

[18] Z.Wu and M. Palmer. Verb semantics and lexical selection, Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, (1994). https://doi.org/10.3115/981732.981751.

[19] D. Mclean, Y. Li & Z.A. Bandar. An approach for measuring semantic similarity between words using multiple information sources, IEEE Transactions on Knowledge and Data Eng., Vol. 15, No. 4. (2003). https://doi.org/10.1109/TKDE.2003.1209005.

[20] Lin C.Y. 2004, Rouge: A package for automatic evaluation of summaries, In: Text Summarization Branches Out, http://www.aclweb.org/anthology/W04-1013.