# Developing and Pre-Processing a Dataset using a Rhetorical Relation to Build a Question-Answering System based on an Unsupervised Learning Approach

**Ashit Kumar Dutta[1] Abdul Rahaman Wahab sait [2] Ismail Mohamed Keshta [3] and Abheer Elhalles [4]**
adotta@mcst.edu.sa asait@kfu.edu.sa imohamed@mcst.edu.sa 181220067@student.mcst.edu.sa
[134] Department of Computer Science and Information Systems, College of Applied Sciences, AlMaarefa University, Riyadh,13713, Kingdom of Saudi Arabia
[2] Center of Documents and Archive, King Faisal University, Al Ahsa, Kingdom of Saudi Arabia

**Summary**
Rhetorical relations between two text fragments are essential information and support natural language processing applications such as Question - Answering (QA) system and automatic text summarization to produce an effective outcome. Question - Answering (QA) system facilitates users to retrieve a meaningful response. There is a demand for rhetorical relation based datasets to develop such a system to interpret and respond to user requests. There are a limited number of datasets for developing an Arabic QA system. Thus, there is a lack of an effective QA system in the Arabic language. Recent research works reveal that unsupervised learning can support the QA system to reply to users queries. In this study, researchers intend to develop a rhetorical relation based dataset for implementing unsupervised learning applications. A web crawler is developed to crawl Arabic content from the web. A discourse-annotated corpus is generated using the rhetorical structural theory. A Naïve Bayes based QA system is developed to evaluate the performance of datasets. The outcome shows that the performance of the QA system is improved with proposed dataset and able to answer user queries with an appropriate response. In addition, the results on fine-grained and coarse-grained relations reveal that the dataset is highly reliable.

***Key words:***
*Arabic dataset, rhetorical relation, discourse relation, rhetorical structure theory, Question-Answering system, natural language processing.*

## 1. Introduction

In recent times, there has been a great deal of effort to make computer-aided annotations of complicated linguistic elements, which are currently unreliable to automated annotation techniques. Morphological annotation of specific corporate terms is usually implemented automatically. However, annotation of the complicated intertextual entities is often done manually[1-5]. There is a lack of global annotation standards that cause challenges even for experienced language researchers. At present, the annotations cannot be performed algorithmically. One of the essential annotation studies of intertextual elements is the Penn Discourse Treebank (PDTB). An extensive collection of annotated, explicit and implicit speech relations found in a million-word Wall Street Journal article.

Manuscript received November 5, 2021
Manuscript revised November 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.11.28

A comprehensive text is not a chain of discrete parts of texts; and instead, a coherent text has a structure of speech linking its components to represent their whole meaning. Discourse relationships describe the relations between these textual components and contribute to the creation and interpretation of the text's discourse structure[6-8].
Identification of discourse relations is a critical stage in discourse analysis to reveal the discourse structure behind the text that is useful for many Natural Language Processing (NLP) applications, such as translation software, information extraction, question answering, and automatic text summarization. The discourse relation indicate that the individual part of the most consistent text usually has a rationale, a specific function, to the whole semantics of the text. Rhetorical Relation (RR) is an instance of discourse relation, unlike grammatical relationships that are usually visible explicitly in the language[9-11].
Therefore, the purpose of the discourse analysis is to identify RR and their scope, limitations, and operations. There are numerous research projects in the natural language text, both descriptive and predictive models of the rhetorical structure and discourse analysis.
For instance, projects on annotation achieved essential advances towards the development of semantic and annotated corpora. Some of these annotation efforts have already had a computational impact that enables semantic roles to be inducted automatically and rhetorical relations to be identified, thereby obtaining near-human performance levels for specific tasks[12-16].
Applications of discourse analysis to automated tasks like a summarization of languages further show that rhetorical relations can improve the performance of well-qualified natural language systems[17-21]. In interpreting the text as a whole, explicit connections are crucial because they show that discourse relationships between vast text chunks successfully interlink the text together. Thus, semantic annotation of connectives is a critical factor for NLP and machine learning techniques. Several investigations have proven that the integration of annotated discourse link labels can improve the overall performance of information collection[22-28].

In this study, researchers proposed a RR based Arabic dataset for unsupervised learning approach. Researchers construct a highly reliable Rhetorical Structural Theory (RST) annotated corpus and allow access to the research community to implement this study viable. Some current state-of-the-art characteristics are evaluated and reused, and innovative ideas are contributed to support machines automatically understand Arabic RR. The study evaluates the performance of the proposed and existing datasets. Research outcome shows that it performs better than the current one by a significant margin.

Here are the research's contributions:

1. Applying pre-processing and annotation technique on an Arabic web content.
2. Generating a RR based Arabic dataset for unsupervised learning approach.

The remainder of the paper is structured as follows: review of literature is presented in section 2. Section 3 offers the research methodology of the proposed study. Section 4 includes the outcome of the study. Finally, section 5 concludes the study with its future direction.

## 2. Literature review

This section summarizes the literature based on RST and discourse relations. The RR approach offers the research community to employ both supervised and unsupervised learning methodologies for NLP research studies. Authors[1] proposed a sentence-level discourse parser for the implementation of lexical and syntactic features retrieved from the lexicalized syntactic tree of a sentence. It has been empirically shown that syntax and discourse structure are correlated. In another study, authors developed an advanced discourse parser which was based on support vector machines. A multi-class SVM classifier was utilized for relations labelling. Some shallow structural, lexical, and organizational aspects were evaluated. This work has been enhanced by including additional linguistic elements that show the relatedness between distinct discourse relations, as well as semantic similarities for verbs and nouns. When deciding which qualities to include in each sentence level, an essential issue was the differentiation between intra-sentential and inter-sentential interactions. Convergence in discourse parsing was additionally discussed by the study[1-3] in their use of conditional random fields for CODRA. The success of deep learning-based NLP models is exponentially increasing[4-7]. The approaches developed by other researchers, moving toward deep neural networks and related feature representation methods, are being employed by researchers today. A shallow convolutional neural network (SCNN) was proposed by the study [8] that uses one convolution layer. Using two recursive neural networks, researchers [9] developed distributed representations of arguments and entity spans. Authors [10] suggested a convolutional neural network that incorporates multi-task learning embedded in a larger system, including PDTB, RST, and the New York Times, to synthesize diverse discourse analysis tasks by implementing the solution. In addition, they presented a neural network model in which the maximum margin is incorporated.

Likewise, the authors [12] suggest learning distributed features representations from words, arguments, and syntax structures to sentences. When it comes to evaluating English discourse relation recognition, the amount of literature in this field is significantly less. The suggested algorithms identified discourse connectors and explicit linkages in the PDTB model that hold between adjacent elementary units, as developed by the study [13]. In order to obtain English implicit relations recognition features, the writers employed what they already had at their disposal. It was found that experimental rules did not have a favorable influence on categorization, since total accuracy was reduced due to these rules.

Extending previous research by answering both explicit and implicit relationships that hold between neighboring and non-adjacent units, authors [14] furthered their research. Instead of the same features employed by the study [15], however, the feature set related to production rules and punctuation, contextual, lexicon-semantic, and lexical features, which were all previously shown to work in the case of English relations recognition. All relevant features from the Arabic Treebank corpus were retrieved automatically. The proposed model successfully calculated results with more accuracy on discourse relations that are minutely detailed. In the final stages, authors [16] found a Causal Relationship Model, designed to identify whether phrases have causal relationships. Linguistic patterns are modeled to encompass roughly 700 specific syntactic elements. Including tests on a small number of samples, it has been discovered that these particles play a central role in distinguishing Arabic causes.

In scenario-based context QA, researchers [17] study the impact of discourse processing and query expansion. They consider a series of queries as a mini-discourse. Evaluating three discourse theoretic models demonstrates that their discourse-based approach can significantly increase QA performance compared to an essential reference resolution baseline. On the other hand, authors [18] evaluate Web user forum threads to detect the discourse interdependence between postings to improve information access via Web forum archives. Their work suggests that three distinct techniques for classifying the discourse relationships between postings outperform an informed baseline.

Using the discourse structure of a document, authors [19] conduct document sentiment analysis. A sentiment classifier's performance can be improved by breaking it down into sections that transmit different kinds of sentiment.

These sections can then be given different weights based on how essential they are in the text as a whole.

The Opinion Corpus for Arabic was generated by a study [20], and it is a balanced dataset that includes 500 online Arabic movie reviews collected from the Web. AWATIF, a multi-genre corpus for Modern Standard Arabic sentiment analysis, is another attempt to generate an Arabic dataset [21]. The data for this corpus was gathered from three distinct places: Newswires contain 2855 sentences, Wikipedia talk pages contain 5342 sentences, and Web forums contain 2532 threaded dialogues. On the other hand, this corpus is not publicly accessible and has not been used in any other work. In order to conduct sentiment analysis on Arabic book reviews, Nabil et al. have built LABR, a relatively large and widely available dataset [1]. An extensive analysis of more than 63,000 book reviews, ranging in star rating from one to five, was published in a study[22]. As previously mentioned, Elnagar et al. suggested the BRAD 1.0 dataset [23] and HARD (Hotel Arabic-Reviews Dataset) [24].

A number of initiatives have lately been reported to offer Arabic-based sentiment analysis systems with a deep Arabic vocabulary. The language of the news domain was employed in one example of Modern Standard Arabic, as in [25]. By manually annotating the text, the lexicon's data is categorized according to its degree of subjectivity. Arabic morphological features and tagging on Arabic social media content have been integrated to study the system's success rate as an addition to the original vocabulary. The findings of their approach demonstrated a 95% success rate [26]. A random graph walk approach was used to extend a manually built lexicon in a study [27], and the results were comparable.

It is possible to determine the discourse structure by applying rhetorical structure theory at the sentence level. Compared to a non-discourse baseline, the study found a 4.5% increase in sentiment categorization accuracy when considering discourse. Other similar studies show improvements in opinion polarity classification when discourse is included, and still, other studies show that classification performance is positively correlated with discourse characteristics [28]. Authors [29] provide an overview of discourse analysis for opinion detection. The existing research gaps enable us to develop a RR based dataset for improving the performance of unsupervised learning techniques.

## 3. Research Methodology

This section includes the methodology of the study to generate a RR dataset. Figure 1 illustrates the process involved in the proposed study. Initially, a crawler is used to crawl Arabic web content. In addition, Arabic WordNet[12] is acted as a primary source. The extracted documents are pre-processed to remove unwanted terms. For instance, a skimming process extracts a general term in a sentence. It reduces the size of the sentences without modifying the meaning of the sentences. The feature extraction process assists to extract segments from the documents. Finally, the RR is generated between the terms.

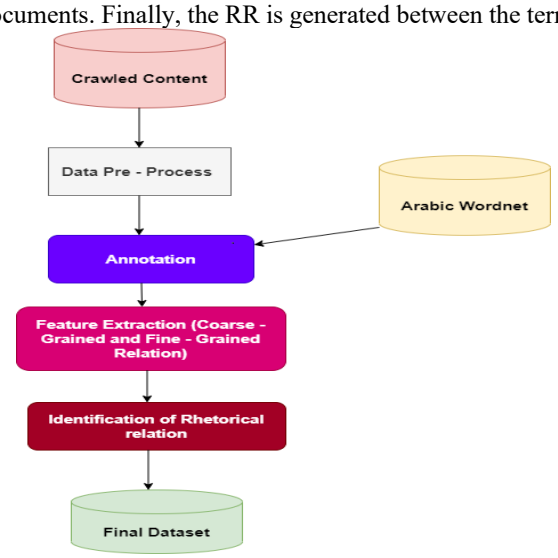


Figure 1: Steps in generating RR based dataset

### 3.1 Dataset and Pre-processing

In order to generate a RR based dataset, a crawler is developed to crawl Arabic web content. In addition, the websites include www.arabic.rt.com, www.albayan.ae, and www.limaza.com are used in the study. Table 1 shows the number of pages and their classifications, respectively. A total of 5420 pages were collected from websites and crawler. Researchers selected these websites due to their familiarity among Arabic communities. Furthermore, the crawler is employed to collect random Arab websites. The crawler is designed to crawl any Arabic websites. Table 2 contains some sample sentences in Arabic and English language, accordingly. In addition, it shows the nouns and verbs used in the sentences. These features are utilized in forming the RR in the proposed dataset. Thus, text summarizer and QA system able to respond a meaningful response.

**Table 1:** Number of pages and documents

| Sources | Pages | Documents |
|---|---|---|
| www.arabic.rt.com | 546 | 756 |
| www.albayan.ae | 325 | 649 |
| www.limaza.com | 1245 | 2145 |
| Crawler | 3304 | 5364 |

**Table 2:** sample sentences from web

| Sentence | Arabic | Verb | | Noun | |
|---|---|---|---|---|---|
| | | Arabic | English | Arabic | English |
| 1. Why we learn Arabic language? | لماذا نتعلم اللغة العربية؟ | نتعلم | Learn | اللغة | Language |
| | | | | العربية | Arabia |
| 2. Learning Arabic language help us to identify the culture and civilization of the Arab world. | يساعدنا تعلم اللغة العربية على التعرف على ثقافة وحضارة العالم العربي | يساعدنا | Help | تعلم | Learn |
| | | | | اللغة | language |
| | | | | العربية | Arabia |
| | | | | التعرف | Identify |
| | | | | ثقافة | culture |
| | | | | حضارة | civilization |
| | | | | العالم | World |
| | | | | العربي | Arabic |
| 3. Ministry of Health releases the status of the vaccine, frequently. | وزارة الصحة تطلق حالة التطعيم بشكل متكرر | تطلق | Release | وزارة | Ministry |
| | | | | الصحة | health |
| | | | | التطعيم | Vaccination |
| | | | | شكل | appearance |
| | | | | متكرر | frequent |
| 4. Saudi Vision 2030 encourages industries to implement artificial intelligence to promote business activities. | تشجع رؤية المملكة العربية السعودية 2030 الصناعات على تطبيق الذكاءالاصطناعي لتعزيز الأنشطة التجارية. | تشجع | encourage | رؤية | Vision |
| | | | | المملكة | The kingdom |
| | | | | العربية | Arabia |
| | | | | السعودية | Saudi Arabia |
| | | | | الصناعات | Industries |
| | | | | تطبيق | Application |
| | | | | الذكاء | intelligence |
| | | | | الاصطناعي | artificial |
| | | | | تعزيز | Strengthen |
| | | | | الأنشطة | Activities |
| | | | | التجارية | commercial |
| 5. Globally, the rate of vaccinated individuals are raising. | على الصعيد العالمي ، فإن معدل الأفراد الذين تم تلقيحهم آخذ في الارتفاع. | تم | completing | الصعيد | level |
| | | | | العالمي | Global |
| | | | | معدل | Rate |
| | | آخذ | raising | الأفراد | Individuals |
| | | | | تلقيحهم | Vaccination |
| | | | | الارتفاع | Height |
| 6. Olympic 2020 was successfully completed in the month of August 2021. | تم الانتهاء من الألعاب الأولمبية 2020 بنجاح في شهر أغسطس 2021. | تم | completed | الانتهاء | Completed |
| | | | | الألعاب | Games |
| | | | | الأولمبية | Olympic |
| | | | | نجاح | success |
| | | | | شهر | Month |
| | | | | أغسطس | August |
| 7. The regular classes for students began in the Kingdom of Saudi Arabia. | بدأت الفصول النظامية للطلاب في المملكة العربية السعودية. | بدأت | begin | الفصول | seasons |
| | | | | النظامية | systemic |
| | | | | طلاب | students |
| | | | | المملكة | The kingdom |
| | | | | العربية | Arabia |
| | | | | السعودية. | Saudi Arabia. |
| 8. The Delta variant was the reason for the increased mortality in America. | كان متغير دلتا هو السبب في زيادة معدل الوفيات في أمريكا. | كان | increasing | متغير | variable |
| | | | | دلتا | Delta |
| | | | | السبب | Reason |
| | | | | زيادة | More |

| | | | | | |
|---|---|---|---|---|---|
| | | | | معدل | Rate |
| | | | | الوفيات | Deaths |
| | | | | أمريكا. | America. |
| 9. Many Airlines have started their regular flights for multiple destinations. | بدأت العديد من شركات الطيران رحلاتها المنتظمة إلى وجهات متعددة. | بدأت | started | العديد | Many |
| | | | | شركات | company |
| | | | | الطيران | Aviation |
| | | | | رحلاتها | Flights |
| | | | | المنتظمة | regular |
| | | | | وجهات | destinations |
| | | | | متعددة | Multiple |

Table 3 presents the Arabic rhetorical relationships with relations classes. These relation classes are included in RST. In the proposed study, researchers adopted few relation classes and generated a coarse-grained and fine-grained relations [10-11].

**Table 3**: Taxonomy of Arabic Rhetorical relations

| Coarse – grained Relations | Fine – grained Relations |
|---|---|
| Background | Environment, Background, Circumstance, Noun position |
| Cause | Outcome, cause – effect, purpose, consequence |
| Joint | Relation, connection, rectification |
| Explanation | Evidence, argumentation, raison, Noun action |
| Topic – comment | Problem – solution, question – answer, Request – Response |
| Attribution | Negation, Attribution |
| Elaboration | Exemplification, Temporal, example, sample, Elaboration - additional |

### 3.1.1 Annotating corpus

Researchers engaged the services of five Arabic academics to annotate the corpus with their insights and observations. Each rhetorical relation is clearly specified in a complete annotation manual, including a list of possible markers that signal each relationship and its significance. There is further discussion of the nuclearity principle regarding the RR framework and how the relative importance of the two text segments connected by a given discourse relation can be determined. Initially, annotators were required to go through a training phase, during which they were encouraged to debate any differences of opinion with one another and provide feedback on the manual as needed. A number of 450 documents are then annotated blindly to determine inter-annotator agreement. Five annotators who do not have information about other's annotation carried out the annotation of each document.

The degree of agreement was examined for the identification of relations and the assignment of nuclearities. When it came to relation identification, the study had kappa values as high as 0.95 for Cohen's and as low as 1.3 for nuclearity assignment, indicating unique agreement for the task of discourse annotation. Annotators were instructed to resolve their differences and establish a unique corpus after debating the significant issues of agreement and disagreement. Using the whole set of discourse relations, the corpus is enhanced with nuclearity annotation. Each document is labelled with 25 labels, one for each document.

## *3.2 Feature Extraction*

This section presents the process of extracting features from the corpus. The annotation process supports to divide the sentences into meaningful terms that can be employed by a Machine Learning (ML) method. However, the ML techniques demand features to improve its predicting performance. Thus, there is a need for an effective feature extractor process. Features are part of the relations in the theory of discourse relation. Recent studies[10][11] proposed some vital features of Arabic corpus. Therefore, the study utilized the same features for the reliability of content. The following features / patterns are employed in the proposed study.

Pattern 1 (P1): Al Masdar

Al Masdar feature was employed by the study[10]. It assists to find the Al Masdar noun in the sentences. Furthermore, noun is one of the fine – grained relations in the proposed RR taxonomy. Thus, the noun analyzer performs the extraction of such patterns.

Pattern 2 (P2): Connectivity

Arabic is one of the languages that contains a larger number of connectivity. A small change in the connection makes a different meaning. Thus, this feature plays a significant role in the RR dataset. As shown in the Table 3.2, there are numerous sentences with multiple connectives. Researchers developed a connective finder that extracts the connective patterns from the documents.

Pattern 3 (P3): Entities

Entity is a noun part in the sentences. It can be utilized to investigate the RR in the sentences. An entity finder is developed in order to extract entity from the complex sentences.

Pattern 4 (P4): Actions

Likewise, entity patterns, action is also a significant relation in sentences. Based on the action, a RR can be extracted from the sentences. The entity extractor is customized for extracting the action patterns from the documents.

Pattern 5 (P5): Numeric

Numeric is a common patterns found in a document. Date, time, and monetary values can be used to find the RR. Thus, it should be treated as a key feature of a corpus.

Pattern 6 (P6): Punctuation
Arabic language contains various punctuation marks. Thus, ";", ",", and """ are tested in the corpus.
Pattern 7 (P7): Annotation markers
Finally, the annotation markers are analyzed using a pattern matcher. The previous section discussed the process of annotation. Using the patterns, the annotations are analyzed from the corpus.

*3.3 Performance Evaluation*

Let C be a corpus, q a query and R a rhetorical relation in the corpus. Thus, $\sum_R P(R|C) = 1$. In the probability of information retrieval, each c in C can be ranked using $P(c|q)$ of being relevant to q.
Using Bayes' law,

$$P(c|q) = \frac{P(q|C)P(C)}{P(q)} = P(q|c) \quad (1)$$

Equation (1) represents the ranking methodology of the corpora according to the query (q). The absence of q leads to drop P(c) due to the uniform structure of a corpora without any prior knowledge. According to the language modeling, P(q|c) indicates the probability of generating the words in q from a system induced by d.
Let include a RR (R) ∈ C in Eq. (1) as follows:

$$P(q|c) = \sum_R P(q|c,R)P(R|C) \quad (2)$$

The first item $P(q|c,R)$ reflects the probability of generating the responses based on q from a system depend on d and R. Thus, $P(q|c,R)$ can be estimated using the probability of from c and R.

$$P(q|c,R) = (1-m).P(q|c) + m.P(q|R) \quad (3)$$

Where m be a free parameter that minimize the total number of responses. The other item of Eq. (2), P(R|C) represents the probability of the RR in a corpora. It can be inferred as the probability of generating the terms in R from a model by c. Using the RR, both memory – based and model – based information retrieval system can be developed, respectively. A predicator can be built as shown in Eq. (4) in order to find the similarity of any two responses for an user query.

$$P_{v,i} = \frac{\sum_w (rr_{v,i} * S_{v,w})}{\sum_w S_{v,w}} \quad (4)$$

$rr_{v,i}$ is the RR in the corpus and $S_{v,w}$ is the similarity between the queries v and w, respectively. Maximum likelihood estimation for query m based on the RR can be expressed as shown in Eq. (5).

$$\log P(q|R) = \sum_{i=1}^{|q|} \frac{f(q_i,R)}{|R|} \quad (5)$$

$$\log P(R|C) = \sum_{i=1}^{|R|} \frac{f(R,C)}{|C|} \quad (6)$$

$$R = \arg\max_{r \in R} E[S|q,C] \quad (7)$$

## 4. Results and Discussion

Researchers experimented their dataset with three evaluation phases. In the first phase, a ranking method is employed using a retrieval model. The retrieval model is developed based on the ranking methods discussed in the previous section. DAWQAS is the recent Arabic dataset proposed by the study[5]. Researchers utilized this dataset in order to compare the performance of the proposed dataset. There are 3205 pages found in the DAWQAS dataset. Similarly, top 3205 pages from the 5420 pages of the proposed dataset is extracted. According to the ranking method, the re-ranking of items in the datasets are performed using a discourse parser. The discourse parser is developed based on the SPADE discourse parser[10]. It is used to identify the RR in the dataset.
Moreover, Mean Average Precision (MAP) and Binary Performance (BPer) are used to measure the average of ranked list as shown in Table 4. Table 4 shows the outcome of the execution multiple queries passed to retrieve a response. The insight from the outcome describes that the proposed dataset able to score a high score rather than the average score for Exemplification and Elaboration relation classes are higher whereas Noun position is low. These scores are depending on the query passed by the user. However, it is evident that the performance of the proposed dataset is better than the existing one.

**Table 4:** Fine – Grained Relations and Datasets

| Fine – Grained relations | DAWQAS | | Proposed Corpus | |
|---|---|---|---|---|
| | MAP | BPer | MAP | BPer |
| Environment | 0.2 | 0.3 | 0.3 | 0.7 |
| Elaboration | 0.1 | 0.4 | 0.9 | 0.9 |
| Noun Position | 0.4 | 0.2 | 0.4 | 0.5 |
| Exemplification | 0.1 | 0.1 | 0.7 | 0.6 |
| Evidence | 0.3 | 0.3 | 0.8 | 0.7 |
| Attribution | 0.4 | 0.1 | 0.2 | 0.7 |
| Noun action | 0.5 | 0.5 | 0.5 | 0.6 |
| Background | 0.4 | 0.4 | 0.3 | 0.9 |
| Negation | 0.3 | 0.6 | 0.7 | 0.4 |
| Sample | 0.1 | 0.4 | 0.1 | 0.8 |
| Connection | 0.7 | 0.5 | 0.6 | 0.7 |
| Raison | 0.6 | 0.4 | 0.5 | 0.9 |
| Argumentation | 0.2 | 0.5 | 0.6 | 0.8 |
| Purpose | 0.3 | 0.4 | 0.9 | 0.7 |

In the second evaluation phase, based on the study[10-11], researchers selected the 14 most fine – grained relations with 4625 documents. The selection is based on the coverage of the whole corpus. These 14 relations have covered more than 85% of the corpus. A Naïve Bayes (NB) is developed to classify both coarse – grained and fine – grained relations. A multinomial logistic regression proposed by the study[10] is used for comparing the performance with the NB classifier. Table 5 shows the F – Score and accuracy of the NB classifier is superior to the Samira L et al.[10] model.

**Table 5**: Performance of Retrieval Models

| Model / Corpus | DAWQAS | | Proposed Corpus | |
|---|---|---|---|---|
| | F-Score | Accuracy | F-Score | Accuracy |
| Samira L et al | 79.2 | 81.1 | 81.2 | 86.9 |
| NB Classifier | 81.6 | 84.6 | 87.3 | 88.5 |

Finally, in the third evaluation method, the NB classifier is trained with the extracted patterns from P1 to P7, exponentially. The trained classifier is tested with coarse – grained relations that contains 4600 documents. Table 6 includes the outcomes of the testing phase. It shows that the classifier able to find more documents with number of patterns / features. Moreover, it indicates the patterns are highly reliable in finding the relations in the corpus.

**Table 6**: Fine – Grained relations with features / patterns

| Fine – Grained relations | P1 % | P2 % | P3 % | P4 % | P5 % | P6 % | P7 % |
|---|---|---|---|---|---|---|---|
| Environment | 61 | 65 | 64.8 | 66.4 | 67.8 | 69.4 | 79.4 |
| Elaboration | 63 | 64 | 67.9 | 68.4 | 69.5 | 75.4 | 77.5 |
| Noun Position | 67 | 66 | 64.5 | 68.3 | 71.4 | 72.6 | 75.9 |
| Exemplification | 64 | 63.1 | 62.5 | 64.5 | 66.3 | 68.9 | 69.4 |
| Evidence | 68 | 68.6 | 69.2 | 70.5 | 71.4 | 73.4 | 72.3 |
| Attribution | 61 | 63.4 | 71.4 | 73.3 | 75.8 | 76.2 | 77.9 |
| Noun action | 75 | 76.3 | 76.3 | 79.4 | 80.5 | 81.6 | 82.6 |
| Background | 69 | 71.4 | 72.4 | 74.1 | 76.4 | 77.9 | 80.9 |
| Negation | 45 | 53.8 | 73.4 | 75.6 | 77.8 | 79.4 | 86.4 |
| Sample | 63 | 64.5 | 79.5 | 80.5 | 81.4 | 82.5 | 83.5 |
| Connection | 57 | 59.3 | 65.4 | 71.2 | 76.3 | 78.6 | 88.7 |
| Raison | 59 | 64.1 | 61.3 | 63.2 | 65.4 | 66.8 | 75.4 |
| Argumentation | 60 | 62.3 | 64.4 | 65.4 | 67.8 | 68.3 | 73.9 |
| Purpose | 75 | 76.1 | 76.4 | 77.2 | 78.2 | 77.8 | 81.5 |

## 5. Conclusion

The familiarity of the machine learning techniques lead to the development of natural language processing based applications. Saudi vision 2030 enabled opportunities to public and private sectors to automate their processes. There is a lack of unsupervised learning based applications due to the limited number of dataset. The existing dataset could not train the system to produce optimum results. In particular, the existence of Arabic language dataset is very few in number. Thus, in this study, authors proposed a rhetorical relation based corpus that supports machine-learning approaches to improve their accuracy. The proposed study generated more number of documents with the support of crawler and secondary sources such as Arabic WordNet and web repositories. Finally, a Naïve Bayes classifier is used for evaluating the performance of the proposed dataset. The outcome shows the performance of the proposed dataset is superior to the existing dataset. The potential relationship between rhetorical relations and user context is an interesting future research direction: for instance, is there a correlation between the evolution of the information requirement of the user and the rhetorical relations that the retrieval system should increase in a document in a search session involving many query reformulations? Additionally, this research could be extended to include metrics for measuring inter-document graded relevancy, like as in the retrieval of JavaScript Object notation (JSON). To some extent, a document's relevance may be represented in the discourse structure. As a final point, because it uses terms like "rhetorical relation" to refer to an "Inter relation between text fragments," the existing functionalities of the proposed model are based on JSON retrieval models. Researchers will frame a layered approach between text fragments in the future work.

## Acknowledgement

The authors would like to acknowledge the support provided by AlMaarefa University while conducting this research work.

## References

[1] Liu, Y., Li, S., Zhang, X. and Sui, Z. (2016) 'Implicit discourse relation classification via multi-task neural networks', Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16), pp.2750–2756.

[2] Louis, A., Joshi, A. and Nenkova, A. (2010a) 'Discourse indicators for content selection in summarization', in Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, pp.147–156.

[3] K. C. Ryding, A Reference Grammar of Modern Standard Arabic. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[4] F. Aouladomar, ''Towards answering procedural questions,'' in Proc. IJCAI Workshop Knowl. Reasoning Answering Questions, 2005, pp. 1–11.

[5] Walaa Saber Ismail and Masun Nabhan Homsi," DAWQAS: A dataset for Arabic Why Question Answering system, Procedia computer science, vol.142, pp. 123 -131, 2018.

[6] A. Farghaly and K. Shaalan, ''Arabic natural language processing: Challenges and solutions,'' ACM Trans. Asian Lang. Inf. Process., vol. 8, no. 4, pp. 1–22, 2009

[7] Lagrini, S., Redjimi, M. and Azizi, N. (2017) 'Automatic Arabic text summarization approaches', International Journal of Computer Applications, Vol. 164, No. 5, pp.31–37.

[8] Lee, H.Y. and Renganathan, H. (2011) 'Chinese sentiment analysis using maximum entropy', in Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011), pp.89–93.

[9] Li, H., Zhang, J. and Zong, C. (2017) 'Implicit discourse relation recognition for English and Chinese with multiview modeling and effective representation learning', ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), Vol. 16, Nos. 3–19.

[10] Samira lagrini, Nabiha Azizi, Mohammed Regjimi, and Monther Al Dwairi, "Toward an automatic summarisation of Arabic text depending on rhetorical relations", International journal of reasoning - based intelligent systems, Vol.11, No.3, 2019, pp. 203 -214.

[11] Christina Lioma, Birger larsen, Wei Lu, " Rhetorical relations for information retrieval", 35th International ACM SIGIR conference on research and development in information retrieval, USA, August 12-16, 2012.

[12] Regragui, Yassir & Abouenour, Lahsen & Krieche, Fettoum & Bouzoubaa, Karim & Rosso, Paolo. (2016). Arabic WordNet: New Content and New Applications. W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. Text, 8:243–281, 1988.

[13] B. Heerschop, F. Goossen, A. Hogenboom, F. Frasincar, U. Kaymak, and F. de Jong. Polarity analysis of texts using discourse structure. In Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11, pages 1061–1070, New York, NY, USA, 2011.

[14] K. J¨arvelin and J. Kek¨al¨ainen. Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst., 20(4):422–446, 2002.

[15] P. Kingsbury and M. Palmer. From treebank to propbank. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC), pages, 2002.

[16] B. A. Shawar, ''A Chatbot as a natural Web Interface to Arabic Web QA,''Int. J. Emerg. Technol. Learn. (iJET), vol. 6, no. 1, pp. 37–43, 2011.

[17] M. F. Al-Jouie and A. M. Azmi, ''Automated evaluation of school children essays in Arabic,'' Procedia Comput. Sci., vol. 117, pp. 19–22, 2017.

[18] H. Rababah and A. T. Al-Taani, ''An automated scoring approach for Arabic short answers essay questions,'' in Proc. 8th Int. Conf. Inf. Technol. (ICIT), May 2017, pp. 697–702.

[19] W. H. Gomaa and A. A. Fahmy, ''Automatic scoring for answers to Arabic test questions,'' Comput. Speech Lang., vol. 28, no. 4, pp. 833–857,Jul. 2014.

[20] Al-Ayyoub Mahmoud, Nuseir Aya , Alsmearat Khouloud, Jaraweh Yaser and Gupta Brij, 2018. Deep learning for Arabic NLP.Journal of computational science 2018, volume 26.

[21] Mallek Fatma, Belainine Billal and Fatiha Sadat, 2017. Arabic social Media Analysis and Translation.3rd International conference on Arabic Computational Linguistics, ACLing 2017. Dubai,united Arab Emirates.

[22] Karaoui Jihen, Zitoune Benamara Farah and Moriceau Véronique,2017. SOUKHRIYA: Towards an Irony Detection System for Arabic in Social Media. 3rd International conference on Arabic computational Linguistics, ACling 2017. Dubai, Unitd Arab Emirates.

[23] Luqman Hamzah and Mahmoud Sabri, 2018. Automatic Translation of Arabic text-to Arabic sign language. Universal access in the information society.

[24] D. Jurafsky and J. H. Martin, Speech & Language Processing. London, U.K.: Pearson, 2017.

[25] S. K. Ray and K. Shaalan, ''A review and future perspectives of Arabic question answering systems,'' IEEE Trans. Knowl. Data Eng., vol. 28, no. 12, pp. 3169–3190, Dec. 2016.

[26] A. Mishra and S. K. Jain, ''A survey on question answering systems with classification,'' J. King Saud Univ.-Comput. Inf. Sci., vol. 28, no. 3, pp. 345–361, Jul. 2016.

[27] M. Biltawi, A. Awajan, and S. Tedmori, ''Evaluation of question classification,'' in Proc. 2nd Int. Conf. New Trends Comput. Sci. (ICTCS), Oct. 2019, pp. 1–7.

[28] Y. H. Phuong and L. G. T. Nguyen, ''English teachers'questions in a vietnamese high school reading classroom,'' JEELS (J. English Educ. Linguistics Stud.), vol. 4, no. 2, pp. 129–154, 2018.

**Dr. Ashit Kumar** is working as an associate professor in the department of computer science and information systems. His specialization is in the field of artificial intelligence and cyber security. He is a certified ethical hacker from EC Council, U.S.A. He has experience of 18 years in both national and international level of education. He has published more than 60 research papers, and most of them are ISI journals. He has completed five research projects in the local and international levels, respectively. He is the author of 4 books. He is serving as Managing Editor in chief and editorial member in various international journals. He is serving as multiple membership of professional bodies /societies. He is also one of the committee members of Scientific Research, curriculum development and Quality management of educational institutions. He got various meritorious award in both national and international levels.

**Abdul Rahaman Wahab Sait** is an Assistant Professor in Department of Documents and Archive, King Faisal University, Kingdom of Saudi Arabia. His research spans many aspects of data science, machine learning techniques, and soft computing. His most recent projects were the question answering system for Arabic languages and generating discourse relation based datasets. He earned Ph.D. degree in Computer science from Alagappa University, India. He has completed certification courses in machine learning techniques. He has published research papers in reputed journals. In addition, he is serving as a reviewer for ISI and Scopus indexed journals. He has served on numerous advisory committees.

**Ismail Keshta** received his B.Sc. and the M.Sc. degrees in computer engineering and his Ph.D. in computer science and engineering from the King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia, in 2009, 2011, and 2016, respectively. He was a lecturer in the Computer Engineering Department of KFUPM from 2012 to 2016. Prior to that, in 2011, he was a lecturer in Princess Nourah bint Abdulrahman University and Imam Muhammad ibn Saud Islamic University, Riyadh, Saudi Arabia. He is currently an assistant professor in the computer science and information systems department of AlMaarefa University, Riyadh, Saudi Arabia. His research interests include software process improvement, modeling, and intelligent systems.

**Abeer ElHalles** is a multitasking hardworking person, senior student in Health Information System from AlMaarefa University. Her interests are machine learning techniques in health information systems and data science. she achieved excellent grades and Awarded for participating in the work and activities program of the Computer and Information System Department.