

A Hybrid Intelligent System for Abstractive Summarization

Dania Sagheer

PhD Student

Department of Artificial Intelligence
Faculty of Informatics Engineering, Aleppo
University
P.O.Bx G502

Fadel Sukkar

Professor

Department of Artificial Intelligence
Faculty of Informatics Engineering, Aleppo
University

ABSTRACT

In this paper, we present a new technique for abstractive summarization of Arabic texts. A system of knowledge base and fuzzy logic has been designed and implemented to simulate human ability of understanding the content of Arabic text, and to create abstractive summary for this text. The knowledge base has been designed for financial and economic field. It consists of facts and if-then rules. The sentences have been parsed by previous stage. The sentences of summary have been obtained using knowledge based system, then Fuzzy system has been designed for selecting appropriate summary of sentences. General membership function has been designed to obtain all the mathematical shapes of membership functions. The peak of the membership function has been designed for hierarchy relations of concepts, and for the destination of semantic relations. The edges of the function has been designed for semantic relations of concepts, and for the domain of semantic relations. The system has been tested on texts for different subjects. The texts have been taken from EASC University Corpus (Essex Arabic Summaries Corpus). The results of this research have shown the effectiveness of the novel hybrid system in terms of semantic, meaning and right composition.

Keywords

Abstractive Summarization, Knowledge Base, Fuzzy Logic

1. INTRODUCTION

In recent years, the amount of information in World Wide Web has been increased, so the need of auto summarization has become necessary to quickly understand large volume of information, and to help readers make a decision to read the full text of document. Reading summary instead of full text can save the time and effort [1].

Text summarization is an important branch of NLP, we can define it as a process of extracting important information from original text and present that information in the form of summary.

Text summarization has two techniques extractive summarization and abstractive summarization, extractive technique extracts the important sentences from original text as they built, where abstractive summarization presents a new sentences that have same meaning of source text.

We propose a novel system to present an abstractive summarization of Arabic text document.

Most researches have focused on extractive summarization and statistical approaches [2]. Abstractive summarization is considered a challenge for most researchers because it needs a semantic understanding of language [3].

Researches on Arabic language have been seen less than other languages [4]. Arabic language has rich morphological system, and other NLP analyzers.

Some research works have presented valuable reviews on abstractive summarization methods [5].

summarization features and techniques used for text summarization have shown in [6].

Abstractive summarization can be classified into two categories i.e. structured based approach and semantic based approach [7].

Some approaches have used clustering methods to get abstractive summary [8], whether other approaches have depended on user's query, expanded once by using the Arabic WordNet thesaurus and domain specific knowledge base [9].

In this paper, knowledge base has been created to get text summary. The knowledge base consists of if-then rules. Facts represent the concepts of commercial and financial field. Rules represent the semantic relationships between concepts and hierarchal relationships control others.

In the Fuzzy logic system, the predicates of created sentences have been the linguistic input variables, but the linguistic output variables have been formed to get summary of sentences. General membership function has been designed to obtain all mathematical shapes of membership functions. The peak of the membership function has been designed for hierarchy relations of concepts in rule base, and for the destination of semantic relations. The edges of the function has been designed for semantic relations of concepts in rule base, and for the domain of semantic relations.

Knowledge based system has established the sentences from the relations in knowledge base, then fuzzy logic system is ready to get the appropriate sentences.

The novel system has been tested on Arabic texts from different subjects. The texts have been taken from EASC University Corpus (Essex Arabic Summaries Corpus) [10].

The paper is organized as follows: Second section illustrates knowledge base, third section illustrates fuzzy logic system, fourth section describes the results, and the fifth section concludes the paper.

2. KNOWLEDGE BASED SYSTEM

knowledge base has been created to describe commercial and financial field. The concepts in this field as bank, money, client, and consumer have been represented as facts with two attributes, the first attribute has expressed concept, and the second attribute has expressed the importance of the concept. The hierarchy relationships between concepts in this field

have been represented as facts with two attributes, the first attribute has expressed of sub concept. The semantic relationships also have been represented as facts with two attributes, Predicate of fact has expressed semantic relation between concepts, the first attribute has expressed of domain relation, and the second attribute has expressed destination relation.

2.1 Preprocessing of Text

The text has been entered and segmented into sentences using stop words and punctuations as comma and point. The sentence has been split into words.

2.2 Selecting of Text Concepts

The concepts of the text have been selected. The rule has been created to find whatever the word of text is a concept. The search engine has worked to find matching between the concepts represented by facts in knowledge base, and words extracted from text.

2.3 Creating of Sentences

The sentences have been created by selecting the text concepts, then concepts of greater importance is chosen. The created sentence has had three predicates, the subject, verb, and objective.

The rules have been created using recursive technique to find the concepts that have greater importance. The attribute has expressed the importance of fact, concludes to the output of previous parsing processing.

Backtracking has been used to find the facts which represented semantic relationships. The selected concept has represented the subject of sentence. semantic relations have represented verb of sentence, and the other concept has represented the objective of sentence.

3. FUZZY LOGIC SYSTEM

Fuzzy logic system has been created to select better sentences to form the summary. The designed system has given a value for every created sentences. This value has assigned to the importance of sentence. The summary consists of sentences that have great fuzzy value. The fuzzy logic system consists of linguistics variables, membership functions, fuzzy rules and fuzzy Inference.

3.1 Linguistics Variables

The words of created sentence have formed the input of fuzzy system. The input and output of system have been determined by linguistics variables.

The first linguistic variable has represented of the first word of created input sentence.

The second linguistic variable has represented the second word of created input sentence.

The third linguistic variable has represented the third word of created input sentence.

The created sentence has been assumed to contain three words.

The output linguistic variable has represented the sentence of summary. So the fuzzy system has three input linguistics variables, and one output linguistic variable.

3.2 Membership Functions

The linguistics values of linguistics variables have been determined for selecting the membership functions. The membership function has been selected for all variables.

The membership function consists of fuzzy sets. This function has described the hierarchy and semantic relations between concepts contained in knowledge base. So, the fuzzy system can prefer sentence from other sentence through the importance of concepts contained in the sentence.

The importance of concepts has appeared through the relations between other concepts. General membership function has been designed to obtain all mathematical shapes of membership functions. General membership function has represented the input and output linguistics variables. The difference between variables values for general membership function has resulted in all membership functions.

The forms of membership functions have illustrated(see Figure 1).

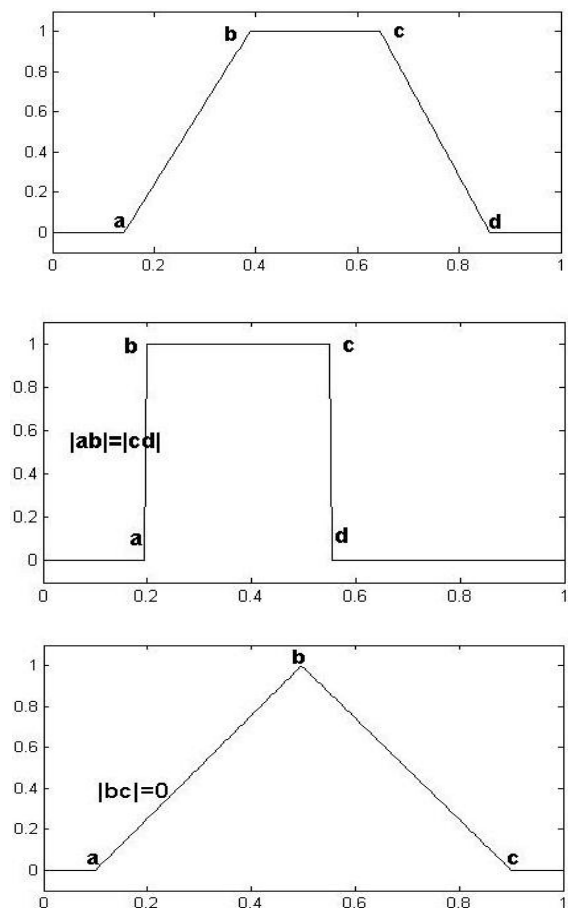


Fig 1: membership functions forms

For the first and third linguistic input variables, the peak of the membership function has been designed for hierarchy relations of concepts in knowledge base. and the edges of the function has been designed for semantic relations of concepts in knowledge base.

For the second linguistic input variable that has expressed verb of created sentence and the semantic relation in knowledge base, the peak of the membership function has been designed for the destination of semantic relations, and

the edges of function has been designed for the domain of semantic relations.

For the linguistic output variable that has represented the sentence of summary, the membership function has designed as same as membership functions of input variables.

Output linguistic variable includes three fuzzy sets that equals to the length of sentence. the fuzzy set has included all the words probability as a part of the sentences in this field.

The words in first position will be selected from The first fuzzy set contained all the words probability will be the first position in the sentence.

The words in second position will be selected from The second fuzzy set contained all the words probability will be the second position in the sentence.

The words in third position will be selected from The third fuzzy set contained all the words probability will be the third position in the sentence.

3.3 Fuzzy Rules

The fuzzy rules have been created to include all sentences. Every sentence has three fuzzy rules, so the fuzzy system has $3 \times \text{number of created sentences}$ fuzzy rules. For each sentence, the fuzzy value has been obtained. The fuzzy value has been calculated through aggregation of real time sets, the sets have been the output of each rule, and the sets are aggregated into one fuzzy set.

For example:

If firstWord is "bank" then first word in Sentence is "Masref.

If secondWord is "present" then second word in Sentence is "present".

If thirdWord is "financial operations" then third word in Sentence is "financial operations".

3.4 Fuzzy Inference

The following steps have been applied for each input sentence:

Step 1: Fuzzification of The Input Variables

For the first variable, the membership degree of first word from sentence has been found from fuzzy sets of first variable. The fuzzy sets of first and third variable have been created appropriately with the selected importance concepts concluded from knowledge based system. For the second variable, the membership degree of second word from sentence has been found from fuzzy sets of second variable. The fuzzy sets of second variable have been created appropriately with the number of created sentences concluded from knowledge based system and entered for fuzzy logic system.

Step 2: Rule Evaluation

The second step is to take the fuzzified inputs, and apply them to the antecedents of the fuzzy rules. A single number is obtained that represents the result of the evaluation.

Step 3: Aggregation of The Output Rule

Aggregation is the process of unification of the outputs of all rules, the membership functions are taken of all rules, and combine them into a single fuzzy set.

The input of the aggregation process is the list of membership functions, and the output is one fuzzy set for the output variable. considered method of aggregation is max-out.

Step 4: Defuzzification

There have been several defuzzification methods, but probably the most popular one has been the centroid technique. It finds the point where a vertical line would slice the aggregate set into two equal masses. Mathematically this center of gravity (COG).

Input the next sentence and apply the same steps until the created sentences finish.

The system has arranged descending the fuzzy values for each sentence, and has detected the sentences of summary.

4. RESULTS

The system has been tested on texts on different subjects. The texts have been taken from EASC University Corpus (Essex Arabic Summaries Corpus).

4.1 First Text

The text in Arabic:

"يعتني التمويل بتحديد احتياجات الأفراد والمنظمات والشركات من موارد النقود، وتحديد سبل جمعها واستخدامها. إن مصطلح التمويل يجمع بين دراسة النقود وغيره من الأصول، وإدارة هذه الأصول ورقابته، وعلم إدارة المال. يستخدم التمويل من قبل الأفراد أو الحكومات أو منظمات الأعمال مثل الشركات أو المدارس والمنظمات غير الربحية. الهدف من النشاط التمويلي يتحقق عبر الاستخدام الأمثل للأدوات المالية."

The text in English:

"Funding takes care of identifying the needs of individuals, organizations, and companies from money resources, and identifying ways to collect and use them. The term fund combines the study of money and other assets, the management and observation of these assets, and the science of money management. Funding is used by individuals, governments or business organizations such as companies, schools, and non-profit organizations. The objective of the financial activity is achieved through the optimal use of financial instruments."

Stage1: Processing of Text

The preprocessing has been done on text, the text has been split on sentences, the sentences has been split into words, the splitting ratios are stop words as conjunctions words and punctuation's letters. the prefixes and suffixes are removed from some words, and the singular forms of word has been obtained from plural forms.

Stage2: Knowledge Based System

The text words extracted from preprocessing stage has been entered to knowledge based system, the knowledge based system has found the text concepts.

Text Concepts:

"تمويل/funding" - "فرد/individual" - "منظمة/Organization"
"شركة/Company" - "حكومة/Government" - "مدرسة/School"
"نقد/cash" - "أصل/Asset" - "مال/Money" - "دراسة/Study"
"إدارة/Management" - "نشاط/activity".

The knowledge based system has given the semantic relations:

"تأخذ(الأفراد، التمويل)/take(individual, funding)" - "يدرس(التمويل، الأصول)/study(funding, assets)"
"استخدم(التمويل، الأموال)/use(funding, money)" - "تأخذ(المنظمات، الأموال)/take(organizations, money)".

The Created Sentences:

“الأفراد تأخذ الأموال/ Organizations take money/ finance ، التمويل يستخدم الأموال/ individuals take funding/ funds study assets/ التمويل يدرس الأصول/ ، uses funds ”.

Stage3: Fuzzy Logic System

Three input linguistic variables have been created as follows:

The first linguistic variable is illustrated (see Figure 2).

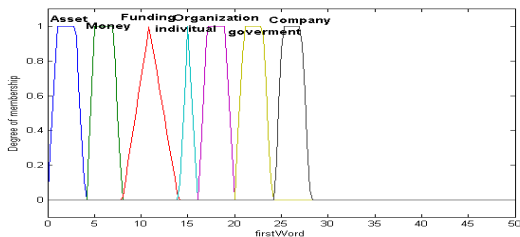


Fig2: first linguistic variable of first text

The second linguistic variable is illustrated (see Figure 3).

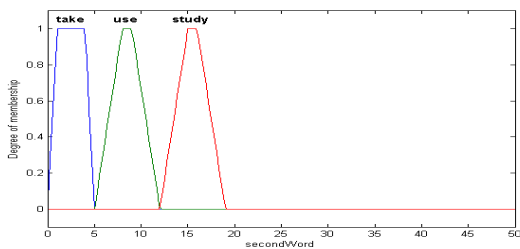


Fig 3: second linguistic variable of first text

The third linguistic variable is illustrated (see Figure 4).

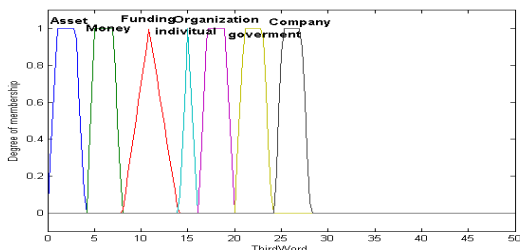


Fig 4: third linguistic variable of first text

The linguistic output variable is illustrated (see Figure 5)

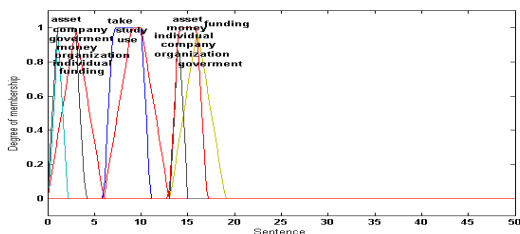


Fig 5: output variable of first text

The fuzzy values are assigned for words of each sentence

The fuzzy values are illustrated (see Table1).

Table1. Fuzzy values of first text

Num	First word	Second word	Third word
1	17.4	2.53	5.98
2	15	2.53	11
3	11	8.38	5.98
4	11	15.8	1.99

The aggregation and defuzzification are done so, the fuzzy values for each sentence are illustrated (see Table 2).

Table 2. Defuzzification of first text

#	Defuzzification form	Fuzzy value
1		8.49
2		10.3
3		9.2
4		9.22

The obtained summary as follows:

“الأفراد تأخذ التمويل/ Individuals take funding/ Finance study assets ، التمويل يستخدم الأموال/ Organizations take money/ ”.

4.2 Second Text

The Text in Arabic:

“المصارف عادة تقرض الناس المبالغ التي يحتاجون لها لقاء فائدة للمصرف، هذا الأسلوب محرم في الإسلام، ويستعاض عنه في المصارف الإسلامية بسد حاجة العميل المقرض عن طريق شراء المصرف حاجة العميل المقرض منه مثل عقار أو واسطة نقل نقداً وبيعها إلى العميل المقرض بالأجل بسعر أعلى يقوم فيما بعد بسداد القرض دفعة واحدة في وقت لاحق يتفق علي تحديده أو بتقسيم المبلغ على دفعات”.

The Text in English:

Banks generally lend people the amounts they need for interest to the bank. This method is illegal in Islam and is replaced in Islamic banks by meeting the need of the borrower by buying the bank the borrower's needs such as a real estate or cash transfer mode and selling it to the borrower at a higher rate After payment of the loan in one lump sum at a later date agreed to be determined or by installments in installments.

Text Concepts:

-Islamic bank / مصرف إسلامي / -interest/فائدة -bank/مصرف / إسلام / Islam قرض/loan-مقرض/borrower- نقدي / cash.

The Semantic Relations:

"البيع بالأجل(المصرف الإسلامي، العميل) / (Islamic bank, sale on term)
"borrower" - "محرم(الفائدة، الإسلام) / (illegal(interest, Islam)
"يشترى(المصرف، نقداً) / (buy(Islamic bank, cash) - "سداد(المقرض، القرض) / (payment(loan, borrower)

The Created Sentences:

The Islamic Bank sells the / المصرف الإسلامي يبيع المقرض بالأجل /
interest is illegal in / الفائدة محرمة في الإسلام /
Islamic bank buy cash / المصرف الإسلامي يشتري نقداً /
المقرض يسدد القرض/ borrower payment loan/.

The first linguistic variable is illustrated (see Figure 6):

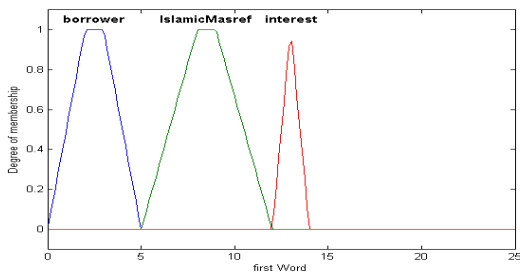


Fig6: first linguistic variable of second text

The second linguistic variable is illustrated (see Figure 7), and the third linguistic variable is illustrated (see Figure 8).

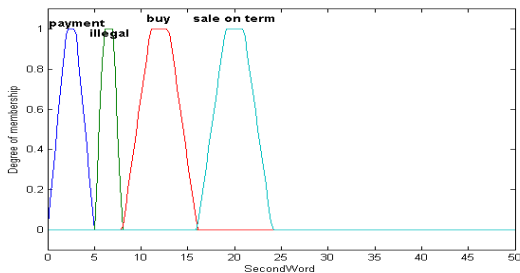


Fig 7: second linguistic variable of second text

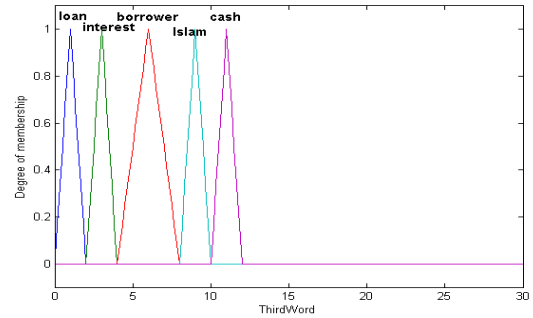


Fig 8: third linguistic variable of second text

The output linguistic variable is illustrated (see Figure 9)

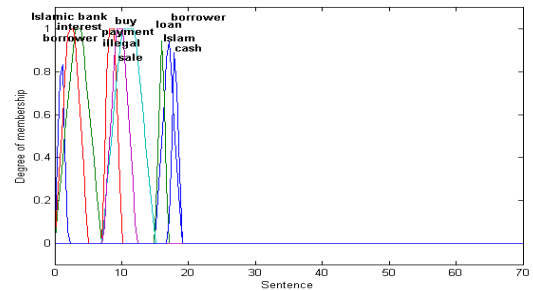


Fig 9: output linguistic variable of second text

The fuzzy values have been assigned for words of each sentence They have been illustrated (see Table 3).

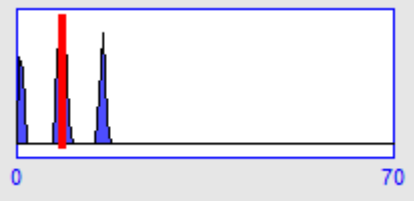
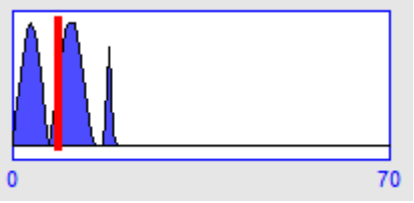
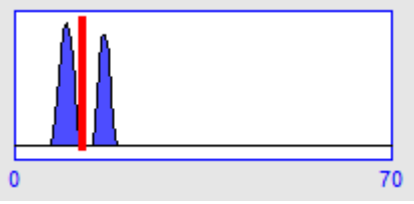
Table 3. Fuzzy values of second text

Num	First word	Second word	Third word
1	8.31	11.6	5.89
2	12.9	6.41	9.03
3	8.31	19.8	10.9
4	2.55	2.49	1.02

The aggregation and defuzzification have been done so, the fuzzy values for each sentence are illustrated (see Table 4).

Table 4. Defuzzification of second text

Num	Defuzzification forms	Fuzzy Value
1		9.38

2		8.8
3		8.71
4		12.9

The summary obtained as follows:

المقترض بسدد القرض/borrower payment loan - المصرف الإسلامي
 يبيع المقترض بالأجل/ The Islamic Bank sells the borrower on term
 -الفائدة محرمة في الإسلام / interest is illegal in Islam -المصرف
 الإسلامي يشتري نقداً/cash Islamic bank buy

4.3 Third Text

Text in Arabic:

"المصرف هو مؤسسة مالية تهدف إلى تسهيل العمليات المالية للعملاء، وحفظ الأموال وتشغيلها، وهو المكان المناسب لحفظ النقود والأموال فيه، والمكان الذي يمكن الاقتراض منه عند الحاجة، مقابل ضمان ما يغطي قيمة القرض، أو مقابل تعهد شخص ضامن يكفل تغطية أو إعادة أصل القرض عند تأخر المدين بالسداد، يدار المصرف من قبل صيرفي و أكثر من موظف".

Text in English:

"The bank is a financial institution that aims to facilitate the financial operations of customers and save money and operate it. This is the place to save money and money in it, where it can be borrowed from when required, in exchange for guaranteeing the value of the loan or against the pledge of a guarantor to cover or repay the loan at The debtor is late to pay. The bank is managed by Sirafi and more than one employee"

Text Concepts:

"مؤسسة/ "institution - "عملية مالية/ "financial operations
 "عميل/ "consumer - "حفظ/ "save - "مال/ "money
 "اقتراض/ "borrowing - "أصل/ "asset - "مدين/ "debit - "مصرف/ "bank
 "صيرفي/ "sirafi - "موظف/ "employee - "قرض/ "loan".

The Created Sentences:

"المصرف يقدم عمليات مالية/ The bank offers financial operations
 المصرف يخرن الأموال/ The bank stores money ، المصرف يتعامل مع
 المدين/ The bank deals with the debtor ، الموظف يعمل في المصرف/
 The employee works in the bank ، المصرف يقرض العملاء/ The
 bank lends customers ، المصرف يمنح القروض/ money ، المصرف يدير
 Al Serifi يدير المصرف/ The bank gives loans ، المصرف يأخذ قرض/ The
 debtor takes a loan ، المصرف يقوم بالحفظ/ "The bank is saving

The first input linguistic variable has been illustrated (see Figure 10)

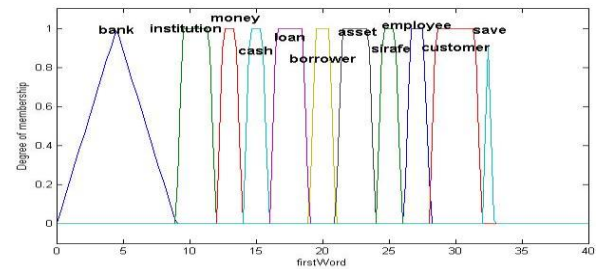


Fig 10: first linguistic variable of third text

The second input linguistic variable has been illustrated (see Figure 11)

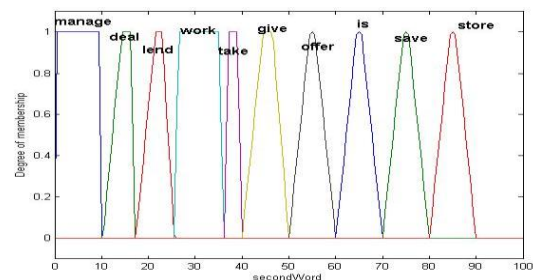


Fig 11: second linguistic variable of third text

The third input linguistic variable has been illustrated (see Figure 12).

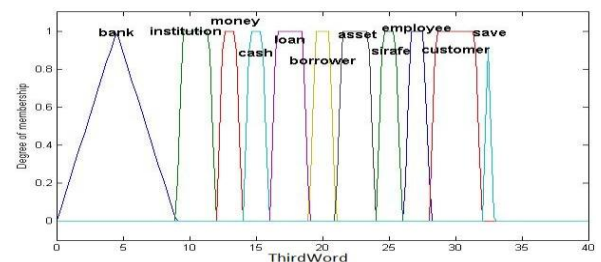


Fig 12: third linguistic variable of third text

The output linguistic variable has been illustrated (see Figure 13).

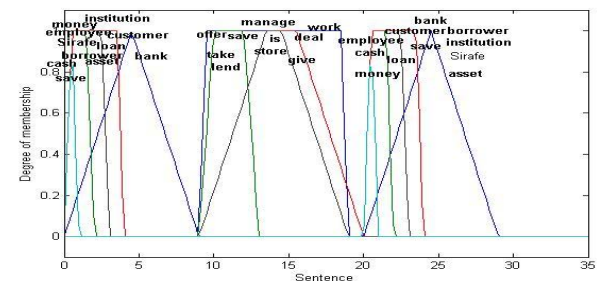


Fig 13: output linguistic variable of third text

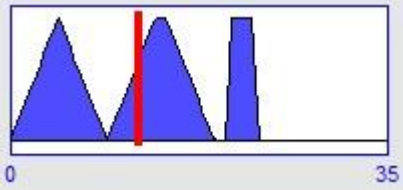
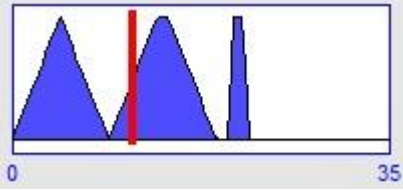
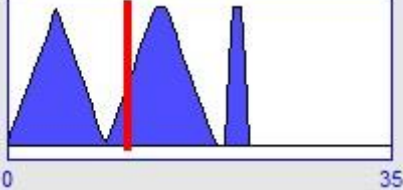
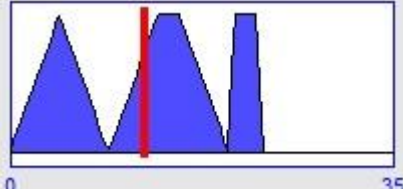
The fuzzy values have been assigned for words of each sentence They have been illustrated (see Table 5).

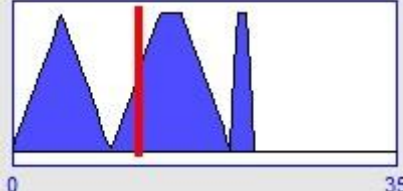
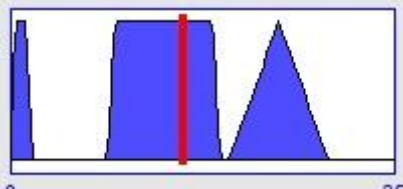
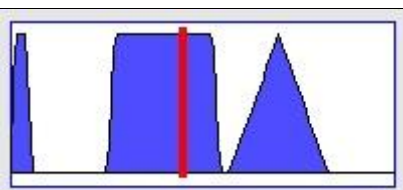
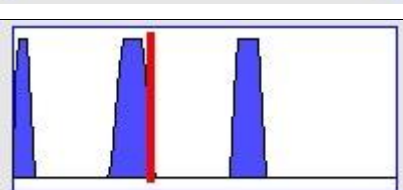
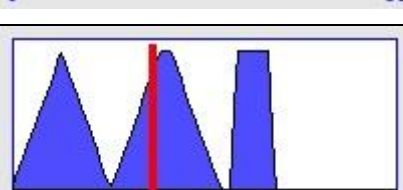
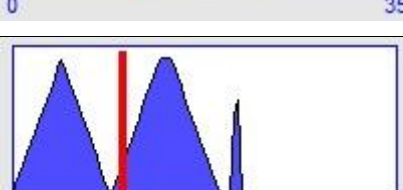
The aggregation and defuzzification have been done, so the fuzzy values for each sentence have been illustrated (see Table 6).

Table 5. Fuzzy values of third text

Num	First word	Second word	Third word
1	4.5	55	10
2	4.5	75	13
3	4.5	85	15
4	4.5	45	17
5	4.5	15	20.5
6	26	5	4.5
7	27.5	30	4.5
8	20.5	38	17
9	4.5	22	30
10	4.5	65	33.32

Table 6. Defuzzification of third text

#	Defuzzification Forms	Fuzzy value
1		12.2
2		11.2
3		11.2
4		12.6

5		11.7
6		16
7		16
8		13
9		12.9
10		10.2

The obtained summary as follows:

"الصيرفي يدير المصرف/ Al Sirafi manages the bank، الموظف يعمل في المصرف./ The employee works in the bank.، المدين يأخذ القروض/ The debtor takes loans، المصرف يقرض العملاء و يمنح القروض. ويقدم العمليات المالية/ The bank lends customers, Loans are granted, And provides financial operations."

5. CONCLUSION

Hybrid system of knowledge base and fuzzy logic has been designed and implemented to present abstractive summary of Arabic text. The text has been processed. The text concepts have been extracted from knowledge based system execution. The search engine of knowledge based system has obtained semantic relations between concepts, and has created the

sentences. The created sentences have been the input to fuzzy logic system that has given a fuzzy value for each sentence. The sentence of greater fuzzy value has most importance of sentence to be in summary. The system has been tested on texts of EASC University Corpus (Essex Arabic Summaries Corpus). The results of this research have shown the effectiveness of the novel system in terms of semantic, expression, abbreviation, and consequently in summary.

6. REFERENCES

- [1] Binwahlan, S. M., "Extractive Summarization Method for Arabic Text - ESMAT", *International Journal of Computer Trends and Technology (IJCTT)* V21(2):103-109, March 2015.
- [2] Nenkova, A. and McKeown, K., "Automatic Summarization", *Foundations and Trends in Information Retrieval* Vol. 5, Nos. 2–3 (2011) 103–233.
- [3] Genest, E. P. and Lapalme, G. 2011. Framework for Abstractive Summarization using Text-to-Text Generation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 64–73.
- [4] Kwaik, A. K. 2011 Automatic Arabic Text Summarization System (AATSS) Based on Semantic Feature Extraction. Master Thesis. Information Technology, University of Gaza.
- [5] Kasture, N. R., Yargal, N., Singh, N.N., Kulkarni, N. and Mathur, V., "A Survey on Methods of Abstractive Text Summarization", *International Journal for Research in Emerging Science and Technology*, Volume-1, Issue-6, November-2014.
- [6] Gaikwad, K. D. and Mahender, N. C., "A Review Paper on Text Summarization", *International Journal of Advanced Research in Computer and Communication Engineering* Vol. 5, Issue 3, March 2016.
- [7] Khan, A. and Salim, N., "A Review on Abstractive Summarization Methods", *Journal of Theoretical and Applied Information Technology*, 2014.
- [8] Banerjee, S., Mitra, P. and Sugiyama, K. 2015. Multi-Document Abstractive Summarization Using ILP Based Multi-Sentence Compression. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*.
- [9] Imam, I., Nounou, N., Hamouda, A. and Abdul Khalek, A. H. 2013, Query based Arabic Text Summarization, *International Journal of Computer Science & Technology*, vol /4/, Issue /2/.
- [10] Arabic Corpora. Essex Arabic Summaries Corpus: http://www.essex.ac.uk/linguistics/research/arabic/arabic_corpora/easc.aspx