Improve Hamming character difference based-on derivative lexical similarity and right space padding

DOI : 10.36909/jer.ICETET.14979

Samah Ali Al-azani*, C. Namrata Mahender

1.2 Department C.S. and I.T ,Dr. Babasaheb Ambedkar Marathawada University, Aurangabad, Maharashtra, India
*[1]alazani183@gmail.com; [2]nam.mah@gmail.com

ABSTRACT

Hamming character difference represents one of the most common problems that can be occurred when students try to answer questions of fill in the gaps that need mostly to one word as the answer. To improve the evaluation of the student answer using Hamming distance, our proposed Hamming model tried to solve the drawbacks of the standard Hamming model by applying a stemming approach to achieve derivative lexical similarity and applying right space padding to deal with unequal lengths of the texts.

Key words: hamming, lexical similarity, questions Answering system, derivatives.

INTRODUCTION

A question answering system is an associate stage in the engineering discipline within the fields of (IR) and the language process that focuses on building systems that mechanically answer queries exhibit by humans in very linguistic communication. a computer understanding of linguistic communication consists of the aptitude of a program system to translate sentences into an indoor illustration so this technique generates valid answers to queries asked by a Valid answers mean answers relevant to the queries exhibit by the user. Because the mental object of linguistic communication, sentences should adequately map the linguistics of this statement, the foremost natural approach is within the simulation of facts contained within the sentences employing a description of real objects likewise as actions and events connected with these objects.

- **What is a question?**

Question is an auditory communication that generally functions as the letter of invitation for information that is anticipated to be provided within the type of a solution. Queries will therefore be understood as a form of an illocutionary act within the field of linguistics or as special varieties of propositions in frameworks of formal linguistics. Queries area unit

typically conflated with interrogatives, the grammatical forms usually want to accomplish the said purpose.

- **What is the answer?**

The answerisa spoken or written reply or response to a question, request, letter, etc., and it is a correct response to a question asked to test one's knowledge.

1. Component of question answering system

A Question answering system contains three main parts: question classification (Question processing), information retrieval (Document processing), and answer extraction (Answer processing). The user prescribes a question using the user question interface. Then this query is used to extract all the possible answers for the input question. The architecture of the Question Answering system is as shown in Fig. (1)

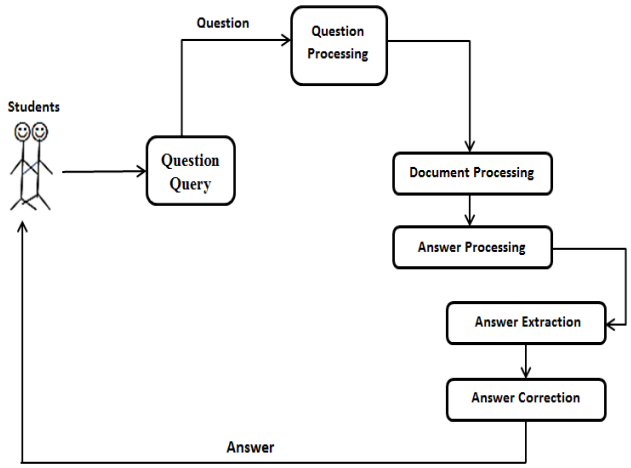

Figure 1 Component of Question Answering system

2. Challenges in Question Answering

The main challenges of a Question Answering System are described as following [2]:

1. **Lexical Gap:** In NLP, the identical which means may be expressed in unique ways. Because an problem can normally best be replied if each referred idea is identified, bridging this gap
2. **Ambiguity:** It is the phenomenon of the identical word having unique meanings; this could be structural and syntactic, lexical and semantic. The identical collection mistakenly denotes unique principles and polysemy, because the identical collection denotes unique however associated principles.
3. **Multilingualism**: at the Web is indicating in unique languages. While Resource Description Framework (RDF).assets may be defined in a couple of languages without delay the use of language tags, there isn't always a unmarried language this is constantly utilized in Web documents. Additionally, customers have unique local languages.

4. Classification Questions based on Domain

There are two types of classification questions based on domain, are as follows

4.1 Open-domain question answering

Open-domain question answering offers with questions on almost anything, and may best depend on fashionable ontologies and global knowledge. Alternatively, those systems generally have plenty of extra data to be had from which to extract the answer.

4.2 Closed -domain question answering

The close-domain question offers questions below a particular area (for example, treatment, Tourist, Economy), and may take advantage of domain-unique understanding frequently formalized in ontologies.

5. Classification QAS based on types

The types of question answering system are classified into several types distributed as follows: 1. Factoid type questions, 2. List type questions 3. Confirmation Questions, 4.Casual Questions 5. Hypothetical Questions, 6. Complex questions.

1. Factoid type questions (what, which, when, who, how)

The factoid questions frequently inception with the wh-word. These inquiries are easy to reply to and actuality based that need answers in a solitary sentence or short expression. For example, the factoid type question "What is the capital of Yemen? requests a city name and it is anything but difficult to answer this kind of factoid question and diminishes the scan space for potential answers. The appropriate response types for factoid type questions are by and large named elements [3]. Factoid type addresses give an acceptable execution in replying. By and large factoid-type questions are an enormous vault of inquiries. Factoid type questions needn't bother with complex normal language handling to get answers. Distinguishing proof of factoid type questions and their sub arrangement is one of the examination issues in the Question Answering framework. Factoid type questions can be replied to by short expressions, for example, associations, people, dates, and areas [4].

2. List type questions

The relief type addresses the need for relief of realities or substances as answers for example decrease names of films in 2017. For the decrease type questions, the appropriate response types are named substances. Consequently, the appropriate responses to rundown questions can give great precision. Question answering frameworks needn't bother with profound regular language preparing to recover answers of rundown type questions. The methods which are applied in factoid type questions can function admirably for list type questions [5]. One of the issues asked in list type inquiry is fixing the limit and incentive for the amount of the substance or the number.

3. Confirmation Questions (yes or no)

Confirmation addresses need answers as yes or no. For example, the Confirmation type question "Is ail a good boy?“appeal the relevant feedback yes or no. To address Confirmation addresses world information, induction component, and good judgment thinking fundamental. One of the upsides of Confirmation type questions asked in QA frameworks is some master clients may jump at the chance to examine for data.

4. Causal Questions [why or how]

The appropriate responses of causal inquiries are not named elements as factoid type questions. Causal inquiries need answers portrayals about an element. Causal inquiries are posed by clients the individuals who want answers as reasons, clarifications, elaborations and so forth identified with specific items or occasions.

5. Hypothetical Questions

Hypothetical questions demand data related to any theoretical occasion and no particular answers to these inquiries. Speculative inquiries ordinarily start with 'what might occur if'. The dependability and precision of these inquiries are low and relies on clients and setting. The normal answer type is spread for speculative sort questions. Thus, the exactness of speculative inquiry noting is low [6].

6. Complex Questions

Is a question that has a expectation this is compund. The presupposition is a proposition that is presumed to apply to the respondent when the question is requested. The respondent will become dedicated to this proposition while he gives any direct answer. The presupposition is referred to as "complicated" because it is a conjunctive proposition, a disjunctive proposition.

6. Similarity measures

Text Similarity Measures are metrics that measure the similarity or distance between two text strings. This depends on the following two groups (lexical similarity) of the text strings or meaning closeness (semantic similarity).

6.1 lexical similarity

It’s a measure of the stage to which the word sets of two given languages are similar. And is only one sign of the combined clarity of the two languages, ago the last also concern on the stage of phonetically, morphological, and syntactical similarity [7][8].

6.2 Semantic similarity

Semantic comparability is a measurement characterized over a bunch of archives or terms, where the distance between things depends on the resemblance of their significance or semantic substance rather than lexicographical closeness. These are numerical apparatuses used to appraise the strength of the semantic connection between units of language, ideas, or examples, through a mathematical depiction got by the examination of data supporting their importance or portraying their nature[9][10].

7. Proposed Method

7.1 Hamming Distance

Hamming Distance metric considers the similarity between any two texts of the same length, where the Hamming distance between any two texts of the same length is the number of positions at which the corresponding characters are different. To understand the concept behind hamming distance, let us assume any two texts. “ABCDEF” and “ABCDSQ. We see the character A in the first location of the text “ABCDEF” is the same character A in the first location of the text “ABCDSQ” , so the distance is 0. Similarly, the characters “BCD” are the same in the second, third, and fourth locations in both texts respectively, also the Hamming distances are 0 in these locations. But, the characters E and F of the first text are different from the characters S and Q of the second text in the same fifth location; therefore, the Hamming distances are 1. Hamming distance is also used in binary strings, where it calculates the distance between the binary vectors. The general formula is:

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

Hamming distance is used for error correction or error detection in network data transmissions. Also, it is used in coding theory, where this paper introduces a proposed model for comparing equal text answers for one-word gap questions. The main requirement for the hamming distance algorithm to work is the lengths in both texts must be the same, so if the lengths are different between the two texts, the distance appears wrongly. So, the standard Hamming model evaluates the correctness of the answer only when both the answers (student answer and model answer) have the same number of character lengths, and there is no difference (missing, mistake, and added characters) in the student answer. Standard Hamming model also not considers the answer is correct if both the student answer and model answer have lexical similarity and they are derivatives of the same root such as play, played, player, Playground, players, and playing. The proposed Hamming model tried to solve these issues by applying the right space padding pre-process to the answer that is smaller than the other answer. Right padding space makes the text lengths of both answers are equal. Also, the proposed Hamming model tried to solve the issue of the lexical similarity for derivatives of the word by applying another pre-process called a stemming process for both the student and model answers that are related to the same root.

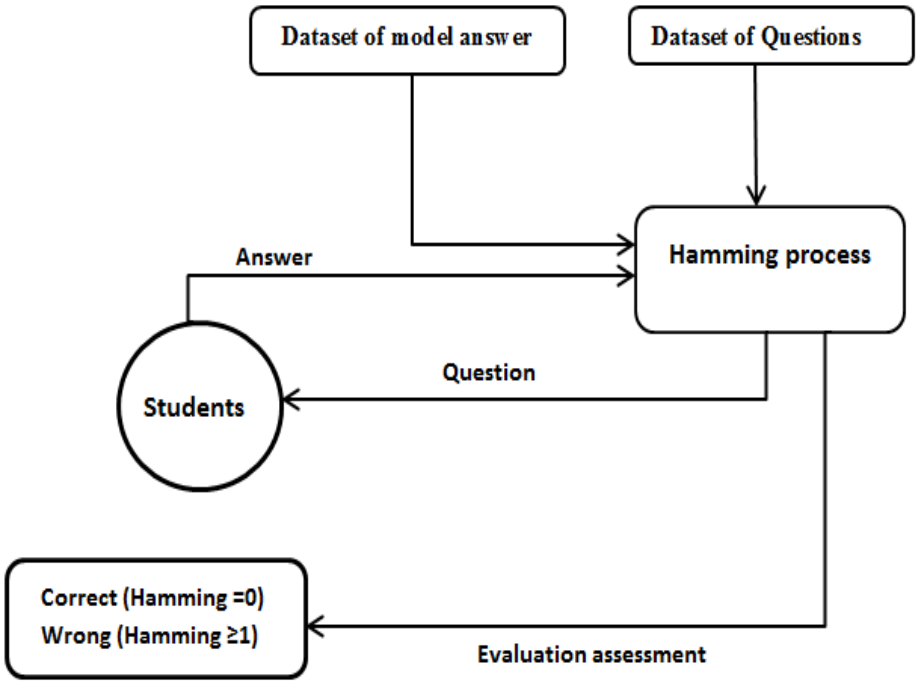

Figure 2. Standard Hamming Distance

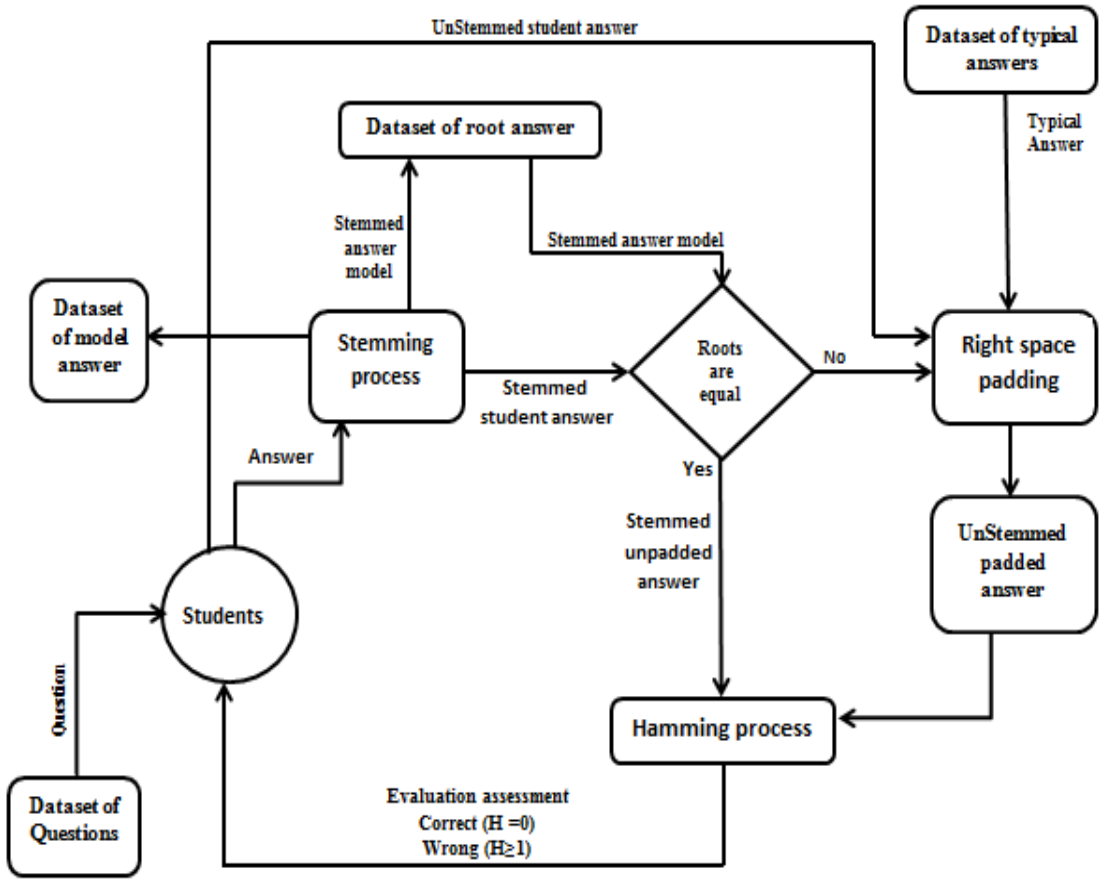

Figure 3.Proposed Hamming Distance

7.2 Data collection

To collect the required data for the proposed Hamming model, we designed some questions and stored them in a dictionary of questions. For example, the question: **It is easy to fix_____ in a formal organization?**We supposed the typical answer (model answer) is **Responsibility**, and we had collected answer from 60 students, where 40 students answered was derived and compared with the model answer: 20 students answered **response**, 10 students answered **responsive**, 7 students answered **responsiveness**, 3 students answered **responsible**, 8 students answered wrong completely (all characters are missing) such as **sport**, and 7 students answered wrong semi-completely (far from or close to the correct answer) such as **resp, respon, responsib, responsibilit** (the word has missing letters), and 5 students answered typical answer (**Responsibility),** and so on for other questions. Also, we used two dictionaries to deal with the model answer, one dictionary to store the roots of the answers (for example respons) and it is used when both the answers have the same root, and other dictionary to store the typical answer (for example Responsibility) and it is used when both the answers have different roots. Therefore, three dictionaries are used in the proposed model, one for questions and two dictionaries for the answer.

7.3 Pre-processing stage

The proposed Hamming model used two pre-processes, the first pre-process is called the ***stemming process*** that takes both the answers for the student and the model. Stemming returns the roots of both the student and model answers. The derivatives of the answer have lexical similarity with the same meaning. The ***model stemmed answer*** is produced and stored in the ***dictionary of the root answers***. If both the student and model answers have the same root, the proposed Hamming model will use the root of the answers to make the exactly lexical similarity between the student answer and model answer. But, if both the answers have different roots, then another pre-process called ***right space padding*** will be used between the ***student unstemmed answer*** and the ***typical answer*** that is extracted from other dictionary called ***dictionary of typical answers***, and it is applied to the small length of the answer to becoming equal to the other answer in length. The right space padding pre-process is not performed if both the answers have exactly the same root characters, but it is performed if both the answers have different root characters.

7.4 Processing Stage

The main process of the proposed process is the ***Hamming process*** that takes both the answers of the student and the model as ***stemmed unpadded answers*** if both the roots of the student and model are the same. But, it takes both the answers as ***unstemmed padded answers*** if both the roots of the student and model are different, and then perform the Hamming operation for strings. Therefore, there is no possibility that the two answers are different in length, Also, in most cases; the proposed model returns 0 values as Hamming distance for derived answers. The proposed Hamming model gives a value greater than 0 as Hamming distance if there is missing or/and wrong in characters.

7.5 Algorithm of the proposed Hamming model

Step 1: Give the question to the student from the questions dictionary.

Step 2: Take the student answer and perform Stemming process

ps = PorterStemmer()

stdent_answer_nostim=input("Answer: ")

stdent_answer = ps.stem(stdent_answer_nostim)

and take model answer to Perform Stemming operation manually and stored it in the root answer dictionary.

Step 3: perform the right space padding on the unstemmed answers when the roots are not the same

```
if len(stdent_answer_nostim)>len(value):
    m=len(stdent_answer_nostim)-len(value)
    value=value+ (" "*m)
if len(stdent_answer_nostim)<len(value):
    m=len(value)-len(stdent_answer_nostim)
stdent_answer_nostim=stdent_answer_nostim+(" "*m)
```

Step 4: Perform Hamming operation on the stemmed unpadded answers or **unstemmed padded answers**

r=hammingDist(stdent_answer,value)

or

r=hammingDist(stdent_answer_nostim,value)

Step 5: Hamming distance returns 0 value for the correct answer and > 0 for the wrong answer.

8. Result and Discussion

The result of the proposed Hamming model that is appliedto the questions with 60 students is shown in the table below. The standard Hamming model gives defined hamming distance only when the lengths of both student and model answers are the same, where it achieves 0 value for the correct answer (typical answer), and achieves greater than 0 value for the wrong answer. But, it gives undefined Hamming distance when the lengths of both the answers are different, and it can't consider the lexical similarity with the same meaning for derivatives of the model answer. The proposed Hamming model gives defined hamming distance when the

answer lengths are equal or unequal, and it achieves 0 value for a correct answer when the student answer is typical or derivative. With undefined standard hamming distance, 40 students (66.7 %) get wrong with derivatives in their answers. With defined standard hamming distance 5 students (8.3 %) get correct with their typical answers. With defined proposed hamming distance, 45 students (75%) get correct with derivatives and typical answers. But with both undefined standard hamming distance and defined standard and proposed hamming distance, 4 students (6.7 %) get wrong with mistake missing character in their answer, 7 students (11.7 %) get wrong with only missing characters in the answers, and 4 students (6.7%) get wrong with a mistake and no missing characters in the answers.

Table 1 Comparison between standard Hamming method and proposed Hamming method.

It is easy to fix_____ in a formal organization? (question) model answer: Responsibility					
Number of Students	Student answer	Hamming distance		Evaluation	
		standard Hamming model	proposed Hamming model	standard Hamming model	proposed Hamming model
20	response	undefined	0	wrong	correct
10	responsive	undefined	0	wrong	correct
7	responsiveness	undefined	0	wrong	correct
3	responsible	undefined	0	wrong	correct
4	sport	undefined	14	wrong	wrong
7	resp	undefined	10	wrong	wrong
5	Responsibility	0	0	correct	correct
4	Responsibixxxx	4	4	wrong	wrong

```
1. It is easy to fix_______ in a formal organization?

Answer: responsive
Hamming distance is  0
```

```
1. It is easy to fix_______ in a formal organization?

Answer: responsiveness
Hamming distance is  0
```

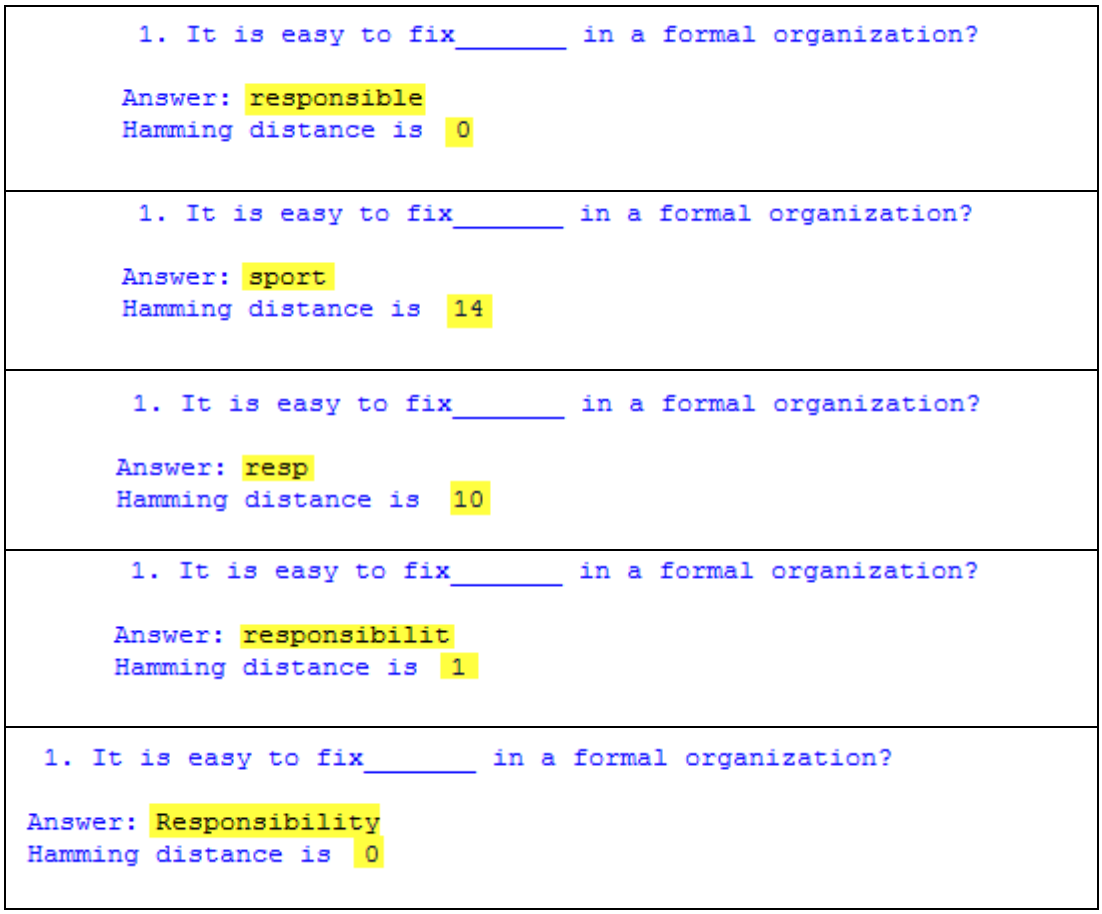

Figure 4. Hamming distance of the proposed Hamming model

9. Conclusion and future work

The proposed Hamming model performs better than the standard Hamming model where it gives a zero-character difference in most derivative answers. Most students get the correct answer because their answers and model answers achieved the derivative lexical similarity. With undefined standard hamming distance, 40 students (66.7 %) get wrong with derivatives in their answers. With defined standard hamming distance 5 students (8.3 %) get correct with their typical answers. With defined proposed hamming distance, 45 students (75%) get correct with derivatives and typical answers. But with both undefined standard hamming distance and defined standard and proposed hamming distance, 4 students (6.7 %) get wrong with mistake missing character in their answer, 7 students (11.7 %) get wrong with only missing characters in the answers, and 4 students (6.7%) get wrong with a mistake and no missing characters in the answers. The future work will investigate how to integrate the semantic approach to achieve semantic similarity.

10. Reference

1. A.Clementeena, Dr.P.Sripiya," A literature survey on question answering system in natural language processing, "international journal of engineering and Technology, 7,2.33, 452-455, 2018.

2. Hoffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J. and NgongaNgomo, A.C., "Survey on challenges of question answering in the semantic web. Semantic Web, 8(6), pp.895–920. 2017.

3. . Youzheng, Hori, Hisashi, Leveraging social Q&A collections for improving complex question answering, Computer Speech, and Language, 29, 1–19, 2015.

4. Amit Mishra, Sanjay Kumar Jain, A survey on question answering systems with classification, Journal of King Saud University – Computer and Information Sciences, 28, 345–361, 2016.

5. Kumar, S. G., and Zayaraz, G. 2014. Concept relation extraction using Naive Bayes classifier for ontology-based question answering systems. J. King Saud Univ.

6. SetioBasuki, AyuPurwarianti, Statistical-based Approach for Indonesian Complex Factoid Question Decomposition, International Journal on Electrical Engineering and Informatics,8, 2,356-373, June 2016.

7. Vijaymeena MK, Kavitha K. A survey on similarity measures in text mining, Machine Learning and Applications:
An International Journal. 2016; 3(1):19−28. https://doi.
org/10.5121/mlaij.2016.3103.

8. Aliguyev RM. A New Sentence Similarity Measure and
Sentence Based Extractive Technique for Automatic Text Summarization, Expert Systems with Applications. 2009;36:7764−72.https://doi.org/10.1016/j.eswa.2008.11.022.

9. Rada, R., Mili, H., Bicknell, E., and Blettner, M. 1989. Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1):17-30, January/February.

10. Li Y., Bandar Z.A., and McLean D. 2003. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Trans. on Knowledge and Data Engineering, 15(4), 871-882.