A new Passage Retrieval Method in Arabic Question Answering Systems

Lana Alsabbagh (lana.alsabbagh@hiast.edu.sy)
Higher Institute for Applied Science and Technology https://orcid.org/0000-0001-6813-3166
Oumayma AlDakkak
Higher Institute for Applied Science and Technology
Nada Ghneim
Higher Institute for Applied Science and Technology

Research

Keywords: Arabic, Question Answering (QA), Passage Retrieval (P R), Word Embedding

Posted Date: December 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-119562/v1

A new Passage Retrieval Method in Arabic Question Answering Systems

Lana Alsabbagh[1], Oumayma AlDakkak[2], Nada Ghneim[3]

Big Data, Higher Institute for Applied Science and Technology (HIAST), Damascus, Syria

E-mail: 1. lana.alsabbagh@hiast.edu.sy
2. oumayma.dakkak@hiast.edu.sy
3. nada.ghneim@hiast.edu.sy

Abstract

In this paper, we present our approach to improve the performance of open-domain Arabic Question Answering systems. We focus on the passage retrieval phase which aims to retrieve the most related passages to the correct answer.

To extract passages that are related to the question, the system passes through three phases: Question Analysis, Document Retrieval and Passage Retrieval. We define the passage as the sentence that ends with a dot ".". In the Question Processing phase, we applied the traditional NLP steps of tokenization, stopwords and unrelated symbols removal, and replacing the question words with their stems. We also applied Query Expansion by adding synonyms to the question words. In the Document Retrieval phase, we used the Vector Space Model (VSM) with TF-IDF vectorizer and cosine similarity. For the Passage Retrieval phase, which is the core of our system, we measured the similarity between passages and the question by a combination of the BM25 ranker and Word Embedding approach.

We tested our system on ACRD dataset, which contains 1395 questions in different domains, and the system was able to achieve correct results with a precision of 92.2% and recall of 79.9% in finding the top-3 related passages for the query.

Keywords: Arabic, Question Answering (QA), Passage Retrieval (PR), Word Embedding

1.Introduction

Nowadays, search engines became very essential in everyday life, as they are continuously queried by people, but they fail to find accurate and short answers to users' questions. The most that search engines can do is to retrieve a set of documents related to the user's query ranked by their degree of similarity to the initial query and leaves us with the task of searching the accurate answers within these documents (Allam et al., 2012), and this is what the technology recently attempted to solve. Question Answering (QA) is a field in Natural Language Processing (NLP) systems that are concerned with extracting short and accurate answers to users' questions rather than just returning the best documents related to these questions (Allam et al., 2012).

The components of Question Answering systems differ according to their purpose. Some systems specialize in answering questions asked within a specific domain (such as Islamic questions) (Abdi et al., 2020), and some are

concerned with answering questions of a specific type ("why", “when”,…) (Azmi et al., 2017). Other systems, called open-domains Question Answering systems, try to answer many types of questions and within different domains, and they are the most general systems (AL-SMADI et al., 2017; Zhou et al., 2020). Such systems consist of the following basic modules (1) *Question Analysis Module,* including question tokenization, stemming its words, diacritics removal (in Arabic), question reformulation and classification, (2) *Document Retrieval Module,* to extract the most related documents to the question, (3) *Passage Retrieval Module,* that is concerned with extracting the short text passage that are mostly related to the question and expected to contain an answer to the asked question, and finally (4) *Answer Extraction Module,* which includes extracting the exact answer from the returned passages from the previous phase, according to the type of the asked question (Sarrouti et al., 2020). The *Passage Retrieval* phase is important for effective Question Answering systems’ results, a. extracting the correct answer depends on its appearance in the passages resulting from passage retrieval phase, Retrieval is often achieved in open-domain Question Answering systems using traditional methods such as TF-IDF or BM25, such methods depend on representing the question and the passage with weighted vectors in dataset space and ranking the passages according to the similarity of passage vector with the query vector (Karpukhin et al., 2020). One of the most popular applications of the Word Embedding representation is the Information Retrieval (IR) systems, where it can be used in the passage retrieval module within the Question Answering systems (Zuccon et al., 2015). Word Embedding is a set of language modelling and feature learning methods that has been used recently in many Natural Language Processing tasks. This model generally depends on mapping textual words into a low-dimensional continuous space, so each word can be represented by a real-valued vector and thus semantic similar words have similar vector representation (Li et al., 2018). The advantage of this approach is that it is concerned with the words semantic similarity instead of being satisfied with the lexical similarity (Mitra et al., 2017).

In this paper, we propose a new model to implement the passage retrieval component in Arabic Question Answering systems. We considered a combination of lexical and semantic similarities, using BM25 word embedding model for similarity measuring. In the rest of this paper, we represent other related works in QAS domain in section 2. In section 3, we present in details our approach with its different steps. In section 4, we introduce our experimental results with a comparison with similar works. At the end, insights for the future and a short summary are presented.

2. Related Works

Abdi et al. proposed in (Abdi et al., 2020) ASHLK, a question answering system in Hadith using linguistic knowledge. The system retrieves the most related hadiths' sentences for the user's query from a large set of hadiths that represent the system's data set. The system relies on finding the answer by measuring the similarity between user's query and hadiths' sentences by using the graph similarity so that each node in the graph represents a sentence and each edge represents the value of the similarity between the two sentences (nodes), the similarity is measured by the combination between the semantic similarity between the words of the two sentences in addition to the syntactic similarity that takes into consideration the order of words in the sentence (on the assumption that the word order in a sentence reflects the meaning of this sentence), the system will return a specific number of sentences that are more similar to each other and more similar to the user's question as an answer, after eliminating the repeated sentences. The system was tested on a data set consisting of 4000 hadiths

and 3825 queries divided as 2678 queries in training set, and 1147 queries in testing set, and had obtained a precision of 83.5% and a recall of 63%.

LEMAZA (Azmi et al., 2017) is an open-domain Arabic Question Answering system concerned with answering questions of the type “why”. After the *Question Analysis*, and the text *Document Preprocessing*, phases, the most related passages to the question are extracted and verified in the *Document|Passage Retrieval* phase using Lemur retrieval engine. The *Answer Extraction* phase uses RST-based algorithm which has been implemented specifically for answering "why" questions. The system was tested on a manually collected dataset (110 "why" questions), with 79.2% precision of and 72.7 recall.

SOQAL (Mozannar et al., 2019), is a Neural Arabic Question Answering system. It aims to extract accurate answers to questions in different domains, where the form of the answer is a span of text from Wikipedia articles. An Arabic Reading Comprehension Dataset (ARCD) was prepared based on Wikipedia articles, and contained 1395 questions in various domains. The system extracts the set of documents most related to the question by using a hierarchical TF-IDF approach and then get the span of text that is most related to the user's query and returns it as an answer. The system is based on a neural reading comprehension model with a pre-trained bi-directional transformer BERT. The sentence match (SM) metric was used to evaluate the performance of the system and it reached 91.4%.

3. Method

Our system consists of three basic components: Question Analysis (QA), Document Retrieval (DR), and Passage Retrieval (PR), (see Figure1). The input of system is a question posed in Modern Standard Arabic MSA to the system, and the output is the most three related passages to that question.

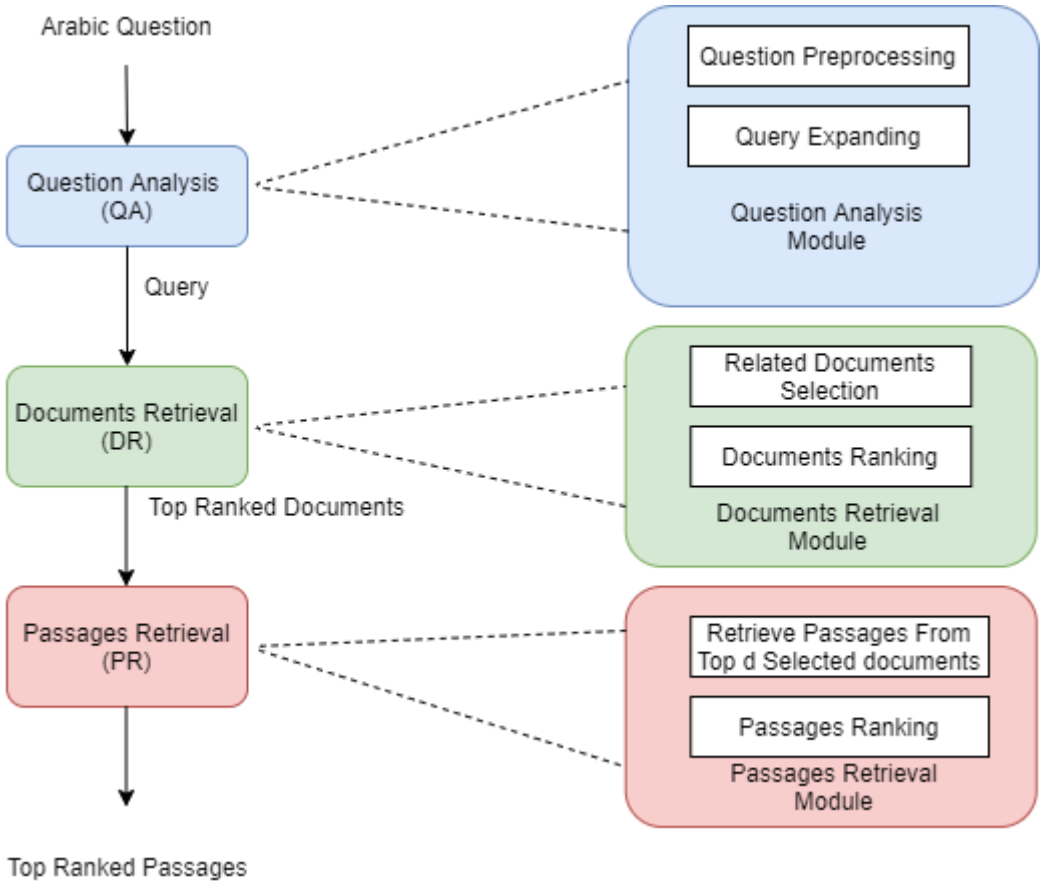

Figure 1 The general structure of the system

3.1 Question Analysis (QA)

3.1.1 Query Preprocessing

At this stage, the question is passed through traditional linguistic processing steps, wich are:

a) Tokenization, by sperating question text into smaller units (words), this done by assuming space as a delimiter.

b) Normalization, by replace some characters that can by written in different ways with its normal form like replace"أ ,إ ,ا ,آ " with "ا".

c) Diacritics removal, as shown in Table 1.

d) Stopwords removal, where stop words include prepositions, conjunctions, etc.

e) Non-relevant Arabic characters and symbols elimination.

f) Extracting the stem of each word.

Table 1 Arabic Diacritics

Name	Shape
Futtha	◌َ
Thummah	◌ُ
Tenween Futtha	◌ً
Tenween Thummah	◌ٌ
Tenween Kusrah	◌ٍ
Kusrah	◌ِ

3.1.2 Query Expansion (QE)

To increase the accuracy of the system, the question was expanded by adding synonyms to each of the question words, Table 2 shows question before and after Query Expansion process. Through experimentation, it was found that the system gives the best results when we take one synonym (if found) for each word of the question, as shown in Table 3. To implement the question expansion phase, the approach of Word Embedding used in the phase of Passage Retrieval (see section 3.3) is also used here, by using a pre-trained word embedding model in Arabic (AraVec) (Soliman et al., 2017).

Table 2 Query Expansion Examples

Question before Query Expansion	Question after Query Expansion
لمن ينتقد خاشقجي في مقالاته الإخبارية؟	لمن من ينتقد نقد خاشقجي في مقالاته كتابته الإخبارية؟
ما هي الكيانات الأولى في المملكة العربية السعودية؟	ما هي الكيانات المنظمات الأولى في المملكة العربية السعودية؟

Table 3 System Accuracy according to Synonymous set size

Synonyms set size	System Accuracy
1	92.2%
2	91.7%
3	91%

3.2 Document Retrieval (DR)

In this stage, the documents are processed by applying the same linguistic processing steps implemented in the previous stage. The Vector Space Model (VSM) method was used to represent each document with a vector using a TF-IDF vectorizer. The question vector is also represented in the same way, so that the most relevant documents are retrieved by applying the cosine similarity between each document vector and question vector. The most five similar documents with the question are selected as input for the next stage (PR).

3.3 Passage Retrieval (PR)

In this phase, we aim to extract the three passages that are most related to the asked question. This is done using the concept of Word Embedding, based on a pre-trained model (AraVec) on an Arabic dataset collected from Arabic Wikipedia articles (Soliman et al., 2017).

To obtain the similarity between the question and one passage, we use the following algorithm, where Sim_{AraVec} represents the similarity between two sentences using AraVec model:

Algorithm 1: AraVec Similarity Measure Algorithm

Input: Two Arabic sentences (S1, S2)
Output: Similarity between sentences using AraVec model
$S1words \leftarrow wordTokenization(S1)$;
$S2words \leftarrow wordTokenization(S2)$;
$sim \leftarrow 0$;
foreach $w1 \in S1words$ **do**
  foreach $w2 \in S2words$ **do**
    $V1 \leftarrow wordVector_{AraVec}(w1)$;
    $V2 \leftarrow wordVector_{AraVec}(w2)$;
    $sim \leftarrow sim + cosineSimilarity(V1, V2)$;
return sim

To extract the most related passages to the question, we first divide the documents resulting from the Document Retrieval Phase by using a sentence segmenter, and we consider each passage to be a sentence that ends with a dot ".". For each passage, we calculate Sim_{AraVec}, between the question and the corresponding passage p $\{v_{p1}, v_{p2}, ..., v_{pn}\}$. We perform a normalization operation for the resulting vector by dividing by the max value. We then sort the values in descending order and choose the top three passages that achieve the largest similarity values.

The previous measurement takes into account the similarity in terms of meaning only. To increase the accuracy of the model, we added another measure, which is BM25 function that takes into account the lexicon similarity. The general form of the model used to achieve the passage retrieval module becomes as follows:

$$sim = \alpha\ Sim_{AraVec} + (1- \alpha\ )\ Sim_{BM25} \qquad (1)$$

where α is a weighting parameter in the range [0, 1] for weighting the significance between semantic and lexical similarity and Sim_{BM25} is the similarity between two sentences using BM25 method.

4. Results and discussion

We tested our system using the Arabic Reading Comprehension Dataset (ARCD), which consists of 1395 questions from Wikipedia articles. We have randomly chosen 100 questions for development and used the rest for testing.

To choose an appropriate value for the parameter α in Equation[1], we calculated the precision and recall for each value in the range [0.1, 0.9] with a step of 0.1. The system achieved the best performance for α = 0.3, as shown in Figure 2.

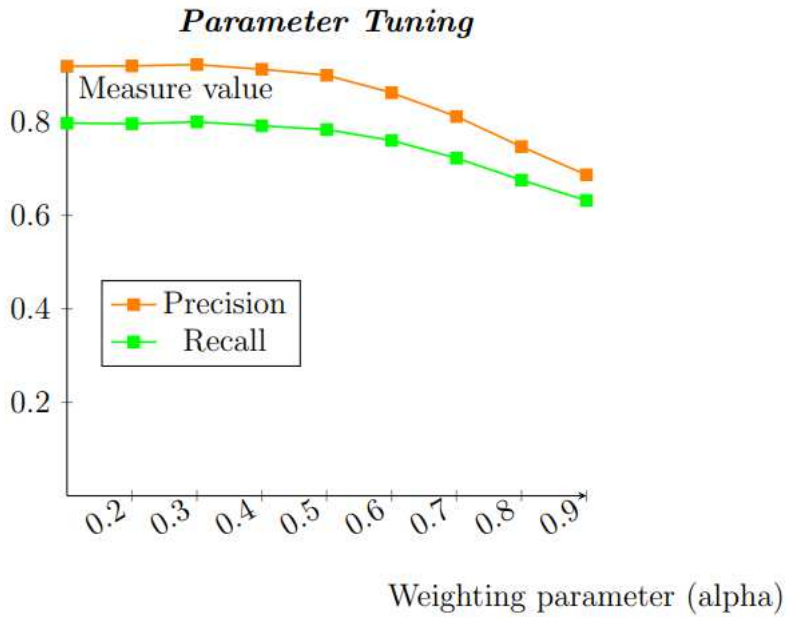

Figure 2 Tuning the weighting parameter (α)

To evaluate the system, we use Precision (P) and Recall (R) metrics, which are calculated using the following equations (Arora et al., 2016):

$$P = \frac{Relevant\ Passages \cap Retrieved\ Passages}{Retrieved\ Passages} \qquad (2)$$

$$R = \frac{Relevant\ Passages \cap Retrieved\ Passages}{Relevant\ Passages\ in\ Dataset} \qquad (3)$$

We compared the performance of our system with other methods performance: ASHLK, LEMAZA, and SQAL (see Table 4). We used *Precision*, *Recall* and *Sentence Match* as metrics for comparison. SQAL is the only system that used ACRD dataset, whereas ASHLK system was tested on Hadith data, and LEMAZA system was tested on their own dataset.

Table 4 Performance comparison with other methods.

System	Data set	Precision	Recall	SM
Our System	ACRD	**92.2%**	**79.9%**	**92.2%**
ASHLK	Hadiths	83.5%	63%	-
LEMAZA	-	79.2%	72.7%	-
SOQAL	ACRD	-	-	91.4%

We notice that based on SM metric, our system performed better than SQAL (+0.8%) on the same dataset. Compared to ASHLK and LEMAZA systems, and though different datasets were used for testing, we can notice that the precision of our system is generally better (with an increment of 8.7%, 13% respectively), and that our system's recall is better with an increment of 16.9%, and 7.2% respectively.

5. Conclusion and future work

In this paper, we presented a new efficient method for passage retrieval in Arabic Question Answering systems. It is based on measuring semantic similarity between words using a pre-trained AraVec model. We considered a combination of lexical and semantic similarities, using BM25 word embedding model for similarity measuring. The method was tested using ACRD dataset. The system was able to achieve a precision of 92.2% and a recall of 79.9%. The system can be improved by using a new method to measure similarity between words that take into account the word context addition to semantic similarity, this can be achieved using a pre-trained context-dependent language models such as BERT.

6. Abbreviations

NLP: Natural Language Processing; QA: Question Analysis; DR: Document Retrieval; PR: Passage Retrieval; VSM: Vector Space Model; QE: Query Expansion; ACRD: Arabic Reading Comprehension Dataset.

7. Declarations

Acknowledgements

There are no acknowledgements.

Availability of data and materials

Dataset from ACRD.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

LA designed and developed the system, interpreted the results and wrote the manuscript under the supervision of OD and NG as an academic supervisor. NG also made contribution to the conception and analysis of the work.

All authors read and approved the fnal manuscript.

Funding

Not applicable.

References

[1] Allam, Ali Mohamed Nabil, and Mohamed Hassan Haggag. "The question answering systems: A survey." International Journal of Research and Reviews in Information Sciences (IJRRIS) 2.3 (2012).

[2] Abdi, Asad, et al. "A question answering system in hadith using linguistic knowledge." Computer Speech & Language 60 (2020): 101023.

[3] Azmi, Aqil M., and Nouf A. Alshenaifi. "Lemaza: An Arabic why-question answering system." *Natural Language Engineering* 23.6 (2017): 877-903.

[4] Al-Smadi, Mohammad, et al. "Leveraging Linked Open Data to Automatically Answer Arabic Questions." *IEEE Access* 7 (2019): 177122-177136.

[5] Zhou, Mantong, et al. "Knowledge-Aided Open-Domain Question Answering." *arXiv preprint arXiv:2006.05244* (2020).

[6] Sarrouti, Mourad, and Said Ouatik El Alaoui. "SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions." *Artificial Intelligence in Medicine* 102 (2020): 101767.

[7] Karpukhin, Vladimir, et al. "Dense Passage Retrieval for Open-Domain Question Answering." *arXiv preprint arXiv:2004.04906* (2020).

[8] Zuccon, Guido, et al. "Integrating and evaluating neural word embeddings in information retrieval." *Proceedings of the 20th Australasian document computing symposium*. 2015.

[9] Mitra, Bhaskar, and Nick Craswell. "Neural text embeddings for information retrieval." *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 2017.

[10] Li, Yang, and Tao Yang. "Word embedding for understanding natural language: a survey." *Guide to Big Data Applications*. Springer, Cham, 2018. 83-104.

[11] Mozannar, Hussein, et al. "Neural arabic question answering." *arXiv preprint arXiv:1906.05394* (2019).

[12] Abdelmgeid Amin, Aly. "Using a query expansion technique to improve document retrieval." (2008).

[13] Soliman, Abu Bakr, Kareem Eissa, and Samhaa R. El-Beltagy. "Aravec: A set of arabic word embedding models for use in arabic nlp." *Procedia Computer Science* 117 (2017): 256-265.

[14] Arora, Monika, Uma Kanjilal, and Dinesh Varshney. "Evaluation of information retrieval: precision and recall." *International Journal of Indian Culture and Business Management* 12.2 (2016): 224-236.

Figures

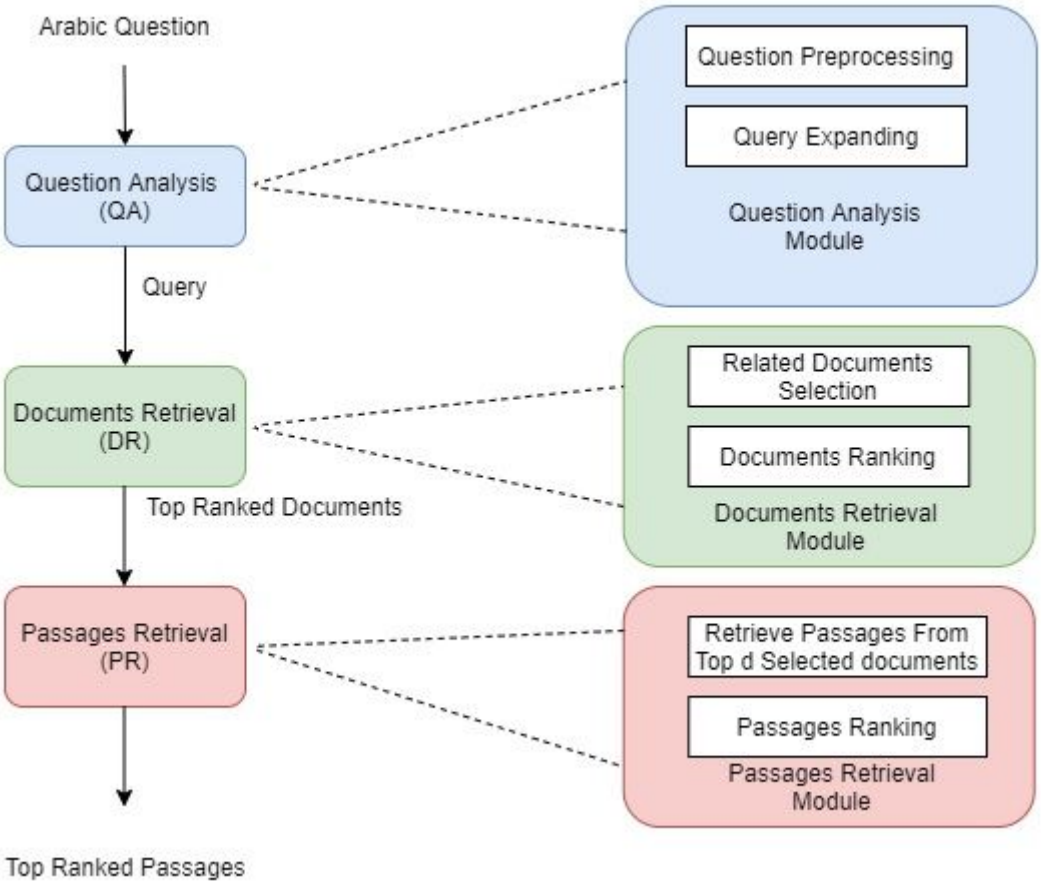

Figure 1

The general structure of the system

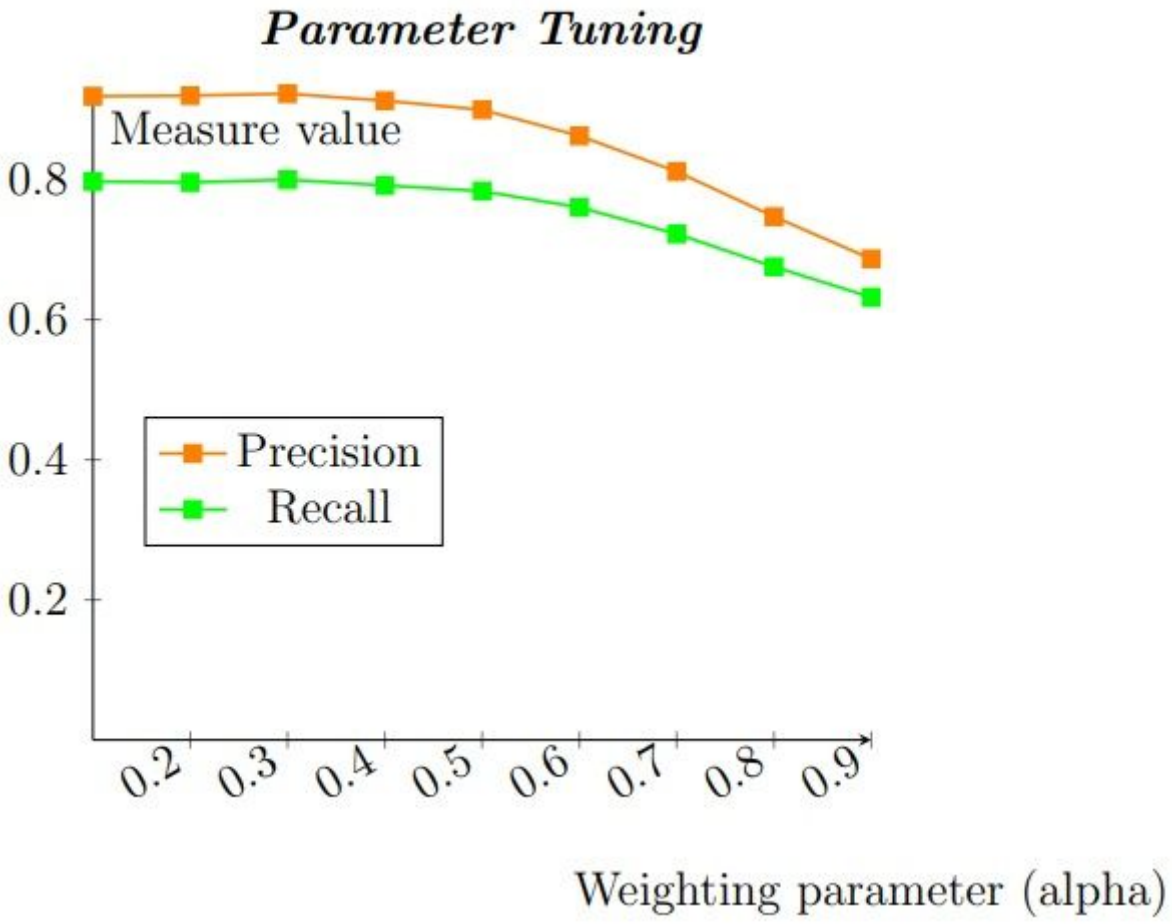

Figure 2

Tuning the weighting parameter (α)