Computational Linguistics and Chinese Language Processing
Vol. 23, No. 2, December 2018, pp. 47-66

Supporting Evidence Retrieval for Answering Yes/No Questions

Meng-Tse Wu*, Yi-Chung Lin* and Keh-Yih Su*

Abstract

This paper proposes a new n-gram matching approach for retrieving the supporting evidence, which is a question related text passage in the given document, for answering Yes/No questions. It locates the desired passage according to the question text with an efficient and simple n-gram matching algorithm. In comparison with those previous approaches, this model is more efficient and easy to implement. The proposed approach was tested on a task of answering Yes/No questions of Taiwan elementary school Social Studies lessons. Experimental results showed that the performance of our proposed approach is 5% higher than the well-known Apache Lucene search engine.

Keywords: Supporting Evidence Retrieval, Q&A for Yes/No Questions.

1. Introduction

Supporting evidence retrieval is a key step in the question-answering task. It locates the related text passage from the given documents according to the question content so that the system can efficiently answer the question only based on the retrieved passage. The goal of supporting evidence retrieval is to merely keep necessary information (but filter out the irrelevant content as much as possible) to reduce the associated inference time.

Previous supporting evidence retrieval approaches can be classified into three categories: (1) Term matching approaches (Chen, Fisch, Weston & Bordes, 2017), (2) Syntactic/Semantic scoring approaches (Murdock, Fan, Lally, Shima & Boguraev, 2012; Jansen, Sharp, Surdeanu & Clark, 2017), and (3) Translation model based approaches (Berger, Caruana, Cohn, Freitag & Mittal, 2000; Jeon, Croft & Lee, 2005; Xue, Jeon & Croft, 2008; Zhou, Cai, Zhao & Liu, 2011). Term matching approaches, such as Lucene search[1], used the vector space model and some language models adopted in *Information Retrieval* (Manning, Raghavan & Schütze,

* Institute of Information Science, Academia Sinica
E-mail: {moju, lyc, kysu}@iis.sinica.edu.tw

[1] http://lucene.apache.org/

2008). On the other hand, syntactic/semantic scoring approaches (Murdock *et al*., 2012; Jansen *et al*., 2017) retrieved the supporting evidence by conducting the syntactic/semantic analysis of each document sentence. They detected certain terms or structures in the question and then weighted the candidates differently by the appearance of those terms or structures. Finally, approaches that utilize a translation model were widely adopted in the *Community QA* systems (Berger *et al*., 2000; Jeon *et al*., 2005; Xue *et al*., 2008; Zhou *et al*., 2011). They used phrase-based or word-based translation models to find the similar historical questions from the new queried question. In the task of supporting evidence retrieval, we could let the question play the role of new queried question and the supporting evidence play the role of historical questions, and then adopt the translation model to find the supporting evidence.

Term matching approaches are widely adopted in the search engine due to its efficiency. However, they do not consider the local context of each term, not even mentioning the associated syntactic/semantic information. Therefore, they usually result in low accuracy. On the other hand, syntactic/semantic scoring approaches utilize syntactic/semantic meaning of each document sentence. They can understand the questions more in the syntactic/semantic level. However, those approaches are not only time consuming but also task orientated. Finally, translation model based approaches are widely adopted in the Community QA systems. However, they need large training data to train the translation models, and are thus not suitable for the tasks with only small amount of training data.

To overcome the problems mentioned above, we aimed at the approach that is efficient, general and accurate enough. Therefore, the approach of term (most of them are unigrams) matching is still adopted in this paper for computation efficiency and generalization. However, to further consider the phrase and local context, it is extended into n-gram for considering the local dependency. It thus avoids the drawbacks of previous approaches.

Given a question, our goal is to find a related passage, from the given corpus, that contains minimum but sufficient information to answer the question. In other words, good supporting evidence should include sufficient related information and less irrelevant and redundant information for the given question. On the other hand, supporting evidence can be extracted in different granularity. For instance, they are specified as top 5 articles in (Chen *et al*., 2017). The smaller the granularity is, the harder the approach is to find the appropriate supporting evidence (since we need to locate it more accurately). In our task, we define the supporting evidence as a text passage with consecutive sentences in the same paragraph, which will be explained in Section 4.3. We propose two scoring functions for finding the supporting evidence: QE-BLUE and modified F-measure. QE-BLUE is converted from the CR-BLEU score (Papineni, Roukos, Ward & Zhu, 2002) which only considers n-gram precision and is used in evaluating the performance of a machine translation system. In contrast, the modified F-measure takes both recall and precision of n-grams into consideration.

Therefore, the modified F-measure is able to evaluate the portion of the matched terms in the question. In comparison with those term matching approaches, the proposed method provided better performance. On the other hand, in comparison with those semantic scoring approaches, the proposed method is more efficient, easy to implement and task independent. In summary, we make the following contributions in this paper:

- We studied the desired characteristics of extracted supporting evidence.
- We proposed a novel scoring function for retrieving the supporting evidence by jointly considering precision and recall of n-grams.
- We adopted and tested several techniques for improving the supporting evidence retrieval.
- We conducted the experiments to show the superiority of the proposed approach.

The remainder of this paper is organized as follows. Section 2 illustrates the desired characteristics that an effective supporting evidence retrieval algorithm should possess. The proposed approach is introduced in Section 3. Section 4 shows the experimental result. The error analysis of the proposed approach is then given in Section 5. The related work is introduced in Section 6. Finally, Section 7 concludes this paper.

2. Desired Characteristics

Question:我們|應該|完全|聽從|父母|的|建議|，|選擇|加入|學校|的|團隊|。

"**We** should fully follow the **advice** of **parents** for **choosing** which school **group to join**."

Evidence:我們|可以|依照|自己|的|興趣|，|參考|老師|和|父母|的|建議|，|選擇|加入|不同|的|團隊|學習|。

"**We** can consider our own interest and refer to the **advice** from the teachers and **parents** for **choosing** which learning **group to join**."

Figure 1. A question and its corresponding supporting evidence

From the question and its supporting evidence shown in Figure 1, we can see that they share many words (which are marked in bold and underlined). This is because the questions usually use the same words or sentences to describe the same thing.

Let s_i stand for the i-th matched word, w_j stand for the j-th unmatched word, w_j^* stand for the j-th string which purely consists of an arbitrary number of unmatched words, and $|w_j^*|$ denote the number of words contained in w_j^*. The desired characteristics of an effective supporting evidence retrieval algorithm are listed as follows.

Characteristic-1: Prefer more matching occurrences

Candidate1: $\underline{\mathbf{s_1}}\ w_1^*$

Candidate2: $\underline{\mathbf{s_1}}\ w_2^*\ \underline{\mathbf{s_1}}\ w_3^*$

In the above pattern, we prefer Candidate-2 as the supporting evidence since the same matched term appears more times. Consider the following Example-1:

Example 1

Question: 安平古堡|是|位於|台灣|南部|的|古蹟|。

"**Fort Zeelandia** is a monument located in south Taiwan."

Candidate-1: 安平古堡|位於|台灣|南部|，

"**Fort Zeelandia** is located in south Taiwan."

Candidate-2: 安平古堡|位於|台灣|南部|，|*安平古堡*|*有約*|*400*|*年*|*歷史*|。

"**Fort Zeelandia** is located in south Taiwan. ***Fort Zeelandia*** *has a history of about 400 years.*"

This preference is illustrated with the above Example-1. We prefer Candidate-2 here since it additionally mentions that 安平古堡 ("Fort Zeelandia") has a long history which entails that it is a monument. As a result, we prefer more occurrences of a matching term because it may contain more information we need.

Characteristic-2: Prefer less unmatched terms

Candidate1: $\underline{\mathbf{s_1}}\ w_1^*$

Candidate2: $\underline{\mathbf{s_1}}\ w_2^*$

Suppose $|w_1^*|$ is larger than $|w_2^*|$, then we prefer Candidate-2 in this case because it contains less number of unmatched terms whic are assumed to be the irrelevant information. Consider the following Example-2:

Example 2

Question: 小丑魚|是|一|種|熱帶|海水魚|。

"**Clownfish** is a tropical sea fish."

Candidate-1:小丑魚|原|生於|印度洋|和|太平洋|較|溫暖|的|水中|，|*包括*|*大堡礁*|*和*|*紅海*|。

"**Clownfish** are native to the warmer waters of the Indian Ocean and Pacific Ocean, *including the Great Barrier Reef and the Red Sea.*"

Candidate-2:**小丑魚**|原|生於|印度洋|和|太平洋|較|溫暖|的|水中|，
"**Clownfish** are native to the warmer waters of the Indian Ocean and Pacific Ocean."

Candidate-1 in Example-2 contains the extra information "*包括大堡礁和紅海*" ("including the Great Barrier Reef and the Red Sea") which is irrelevant to our question. Therefore, we prefer Candidate-2 which contains less unmatched terms.

Characteristic-3: Prefer more different term-types

Candidate1: $\mathbf{s_1}$ w_1^* $\mathbf{s_1}$ w_2^*

Candidate2: $\mathbf{s_1}$ w_3^* $\mathbf{s_2}$ w_4^*

Suppose $|w_1^*| = |w_3^*|$ and $|w_2^*| = |w_4^*|$, and both Candidate-1 and Candidate-2 match two terms in the above pattern. However, Candidate-1 has the same two terms s_1 but Candidate-2 has two different terms s_1 and s_2. In this case we prefer Candidate-2 as the supporting evidence because it recalls more terms from the question. Consider the following Example-3:

Example 3

Question: **電腦**|和|**手機**|已|成為|現代人|生活|的|必需品|。
"**Computers** and **mobile phones** have become necessities for modern life."

Candidate-1: **電腦**|在|許多|工作|中|廣泛|使用|，|***電腦***|也|讓|我們|生活|更加|進步|。
"**Computers** are widely used in many jobs. **Computers** also make our lives more advanced."

Candidate-2: **電腦**|在|許多|工作|中|廣泛|使用|，|***手機***|也|讓|我們|生活|更加|進步|。
"**Computers** are widely used in many jobs and **mobile phones** make our lives more advanced."

Candidate-1 only mentions the information about *computer* twice; however, Candidate-2 contains the information about both *computer* and *mobile phone* (which provide more question-related information). As the result, we prefer the candidate-2 that matches more term-types.

According to the desired characteristics of the supporting evidence mentioned above, "*Prefer more matching occurrences*" and "*Prefer less unmatched terms*" could be reflected through the *precision-rate*; and "*Prefer more different term-types*" could be reflected through the *recall-rate*. Following two cases (Table 1 and Table 2) illustrate the effect of precision and recall in retrieving the supporting evidence candidates.

Table 1. Precision and Recall for question-case-1: Question: w_1 w_2 w_3 $\underline{s_1}$ w_4

Candidate	Terms	Precision	Recall
1	w_5 $\underline{s_1}$ $\underline{s_1}$	2/3	1/5
2	w_6 w_7 w_8 $\underline{s_1}$ $\underline{s_1}$ $\underline{s_1}$	3/6	1/5

Table 1 shows that the precision-rate could truly reflect the desired Characteristic-1 and Characteristic-2. Therefore, with the precision-rate, we can successfully select the human desired Candidate-1 as the supporting evidence.

Table 2. Precision and Recall for question-case-2: Question: w_1 w_2 w_3 $\underline{s_1}$ w_4

Candidate	Terms	Precision	Recall
1	w_5 w_6 w_7 $\underline{s_1}$ $\underline{s_1}$	2/5	1/5
2	w_8 w_9 w_{10} $\underline{s_1}$ $\underline{s_2}$	2/5	2/5

However, the precision-rate alone is not enough to meet the desired Characteristic-3. For example, the precision-rate cannot tell the difference between two candidates in Case-2, since both the candidates match two terms. However, by measuring the recall-rate we can choose the better candidate that matches more terms of the question.

According to the above two cases, it clearly shows that both precision-rate and recall-rate should be involved in the scoring function for obtaining the best supporting evidence

3. Proposed Method

3.1 QE-BLEU Scoring Function

Intuitively, BLEU score (Papineni *et al*., 2002), which is a widely used metric in evaluating machine translation quality via comparing the machine-translation output with human-translation references, could be adopted for this task as it can check the similarity between the question content and the passage of the supporting evidence. BLEU score (also called CR-BLEU score, where C stands for candidate and R stands for reference) is originally defined as:

$$BLEU = BP * \prod_{n=1}^{4} p_n^{w_n} \tag{1}$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases} \tag{2}$$

where p_n is the modified n-gram precision between machine translation candidate and a set of human translation references, w_n is the n-gram weight, r and c are the reference and candidate

lengths, respectively. BP is the brevity penalty which penalizes the candidate that is shorter than the reference. BLEU score combines each n-gram precision by multiplication.

As shown in Equation (1), CR-BLEU score only cares about the precision-rate of a candidate. However, we actually more care about the recall-rate in retrieving supporting evidence. We thus adapt the original CR-BLEU metric by letting the given question plays the role of translation candidate and each possible supporting evidence as the translation reference. Therefore, we propose an alternative QE-BLEU score which is defined as follows:

$$QE\text{-}BLEU = BP * \prod_{n=1}^{4} p_n^{w_n} \tag{3}$$

$$BP = \begin{cases} 1 & if\ \ len_E \leq len_Q \\ exp\left(1 - len_E/len_Q\right) & if\ \ len_E > len_Q \end{cases} \tag{4}$$

where Q and E denote the question and the evidence passage, respectively. Question and evidence thus correspond to the candidate and the reference, respectively, in the original function of CR-BLEU.

3.2 Modified F-measure Scoring Function

On the other hand, F-measure is a widely used evaluation metric in information retrieval which considers both precision and recall of the information retrieved (Chinchor, 1992; Sasaki, 2007). We thus prefer the F-measure, instead of BLEU score, for this task as both precision and recall are required to meet the desired characteristics listed in Section 2.

Inspired by BLEU score metric, we also apply n-gram model to consider the word order information. Therefore, we proposed a new Modified F-measure:

$$Modified\ F\text{-}measure = \sum_{n=1}^{4} \left(\frac{1}{\frac{\alpha}{p_n} + \frac{1-\alpha}{r_n}} \right)^{w_n} \tag{5}$$

where p_n and r_n denote the n-gram precision and recall of the question passage, respectively; and w_n is the corresponding n-gram weight as that in BLEU score; α is an adjustable parameter ranging from 0 to 1. If α is close to 0, Modified F-measure becomes more recall-oriented; on contrary, it becomes more precision-oriented when α is close to 1. The adopted precision and recall are defined as follows:

$$Precision = \frac{\#matched\ words\ in\ evidence}{\#words\ in\ evidence} \tag{6}$$

$$Recall = \frac{\#matched\ words\ in\ question}{\#words\ in\ question} \tag{7}$$

4. Experiments

4.1 Data Sets Adopted

We evaluate various approaches on a Taiwan elementary school social studies Yes/No question supporting evidence benchmark data set, which was created by two part-time workers and decided by the third person when there is a conflict. The original corpus consists of 178 lessons, and each lesson is composed of several paragraphs and then followed with its associated questions. We randomly divide those lessons into a development-set (124 lessons) and a test-set (54 lessons). Afterwards, we arbitrarily selected 202 and 414 questions from the development-set and the test-set, respectively. Afterwards, each question is annotated with its supporting evidence benchmark. The statistics of the benchmark is showed in Table 3.

Table 3. The statistics of the benchmark data set.

Data-Set	Development-Set	Test-Set
#Lesson	124	54
#Question	202	414
Averaged #paragraphs per lesson	26.8	30.6
Averaged #sentences per paragraph	3.7	3.6
Averaged #words per sentence	5.0	5.0
Averaged #characters per sentence	9.1	9.0

4.2 Procedure

Step 0: Preprocessing:

The raw texts of lessons and questions are segmented into words via HanLP[2] package. The punctuations are then eliminated after the segmentation (as the punctuations are only used for segmenting sentences). We had tested the case of eliminating *stop words*, but the result seems not much different. Therefore, we keep all the words in the following experiments.

After the preprocessing process, we retrieve the supporting evidence via following four steps:

Step 1: Paragraph-based search

Given a question and its corresponding lesson, we first locate the top-1 paragraph with *Apache Lucene* search engine. This step is used to cut down the search space of locating the supporting evidence.

[2] https://github.com/hankcs/HanLP

Step 2: Sentence-level candidate generation

After the above paragraph-based search, we generate various supporting evidence candidates by increasingly concatenating the consecutive sentences (up to the whole paragraph). For example, if we have a paragraph with three consecutive sentences A, B and C in order, then we will generate the following six different candidates: A, B, C, AB, BC, and ABC.

Step 3: Candidate scoring

This step is the focus of our approach. We use either QE-BLEU or Modified F-measure to score each candidate according to the given question passage.

Step 4: Select the top-1 candidate

After scoring the candidates with a specific scoring function, we then choose the candidate with the highest score as the supporting evidence.

4.3 Experiments Settings

Smoothing: We adopt the package jbleu[3], which uses the smoothing method-3[4] adopted in (Chen & Cherry, 2014) to smooth both QE-BLEU and Modified F-measure. After smoothing, they will get a small non-zero value (instead of zero) when there is no match for a given n-gram.

Weight optimization: Last, there are four n-gram weights in QE-BLEU; however, there are four n-gram weights and one additional parameter α in Modified F-measure. These parameters affect the performance of the proposed scoring functions significantly. We adopt *Particle Swarm Optimization*[5], which is known for being able to escape from the local maximum points, to automatically search for their optimal values on the development-set. We then use the obtained optimal parameters to evaluate the performance on the test-set. There are two α values tested in the Modified F-measure approach. α=0.5 is the situation to weight precision and recall equally; α=0.13 is obtained by optimizing the Modified F-measure with equal n-gram weights. And finally, α=0.12 (without smoothing) and α=0.21 (with smoothing) are the optimal values obtained by jointly optimizing the n-gram weight and α value.

4.4 Experiment Results

For various reasons, there are some benchmarks that cannot be generated by our candidate generation procedure (Step-2). Table 4 briefly lists different reasons and their associated percentages. As shown in Table 4, 16.2% of the questions are originally marked as the case

[3] GitHub repository, https://github.com/jhclark/multeval/tree/master/src/jbleu

[4] It basically assigns a geometric sequence to the n-gram that has 0 matches.

[5] https://pythonhosted.org/pyswarm/

that no appropriate evidence can be found in the text. 12.8% of the selected top-1 paragraph is different with the desired paragraph. 13.8% of the benchmarks are not a consecutive passage within a paragraph. In order to focus on comparing the effectiveness of various scoring functions, we eliminate those types of questions that the desired benchmark cannot be included in the candidate-set, and only evaluate the performance on the remaining questions (total 237 questions remained) in the following tests.

The performances of various approaches are shown in Table 5. Apache Lucene Core 5.5.0 is regarded as our baseline which uses the vector space model and a pre-specified scoring function for ranking. We adopted two widely used scoring functions, TF-IDF and BM25, as our baselines. The performances of equally weighting the n-gram are listed in the table "Equal N-gram Weight". The "+Smoothing" column shows the experiments that involve smoothing technique. The table "Optimal Weight" shows the experiments that adopt the optimized parameters which include various n-gram weights and the α value (for Modified F-measure). Again, the columns labeled with "+Smoothing" are the experiments that adopt smoothing technique with optimal weights. Table 5 shows that the overall performance of both QE-BLEU and Modified F-measure with optimal weight and smoothing technique outperform the baseline Apache Lucene (TF-IDF) about 5%.

Table 4. The statistics of the benchmark evidences that are not covered by the generated candidate-set (measured on the test set).

No evidence in the text	16.2% (67/414)
Non-Top-1 paragraph	12.8% (53/414)
Non-consecutive passage	13.8% (57/414)
Total	42.8% (177/414)

Table 5. The performances of various approaches.

Baseline:

Apache Lucene(TF-IDF)	54.43%
Apache Lucene (BM25)	46.84%

Equal N-gram Weight:

	Equal N-gram Weight	+Smoothing
QE-BLEU	37.13%	52.32%
Modified F-measure (α=0.5)	37.55%	42.19%
Modified F-measure(α=0.13)	58.23%	50.63%

Optimal Weight: (α=0.12, 0.21)

	Optimal N-gram Weight	+Smoothing
QE-BLEU	40.93%	**59.49%**
Modified F-measure	**59.92%**	**59.49%**

5. Error Analysis and Discussion

Apache Lucene: We find that Apache Lucene makes errors (for selecting the desired candidate) in the cases that the top-1 paragraph contains more sentences. This is mainly due to that IDF weight is adopted in both BM25 and TF-IDF, and IDF weight is based on the diversification of the documents to give the term weights. The term which appears in many documents is thus given a lower weight. However, various supporting evidence candidates are actually from the same paragraph (due to the way that they are created). Therefore, the term which appears in many candidates may actually be the key word (in the question) that we should pay attenuation to. As a result, Apache Lucene is not a preferable method for supporting evidence retrieval because it is related to the term distribution in the supporting evidence candidates. As shown in Table 5, the performance of BM25 is lower than that of TF-IDF. The reason is that the IDF matrix in BM25 is more sensitive, which deteriorates the performance in this task.

QE-BLEU: We find that most errors resulted from the QE-BLEU approach is due to the brevity penalty factor, as it penalizes the length of evidence candidates when the length of a candidate is longer than that of the question. In principle, the brevity penalty factor is mainly introduced to avoid involving unnecessary sentences in the evidence. However, as we mentioned in Section 1, the supporting evidence selection is only affected by the relevant and irrelevant information but not the question length. If we punish the evidence of which the length is larger than the question length, we tend to get the supporting evidence that is shorter, and might lose some relevant information.

Modified F-measure: As shown in Table 5, we test two α values: 0.5 (i.e., equally weighting precision and recall) and 0.13 (which is the optimal value obtained from the developoment-set). The performances are found about 8%~20% better when we adopt the optimal α value. However, both QE-BLEU and Modified F-measure get the same performance in the "+Smoothing" column in "Optimal Weight" (Modified F-measure improves 14 cases against QE-BLEU, but it also deteriorates the same number of cases). Furthermore, the optimal α value (α = 0.13) shows that recall is more important than precision since α= 0.13 < 0.5.

However, this model is found that it tends to find the evidence which is the longest among the candidates if we only consider recall. To avoid involving unnecessary sentences in

the evidence, the proposed approaches actually adopt two different stategies: QE-BLEU relies on *Brevity Penalty* (which penalizes the longer passage regardless of its content) and Modified F-measure relies on *Precision* (which penalizes the passage with more irrelevant content). However, utilizing *Precision* is better than adopting *Brevity Penalty* since *Brevity Penalty* only penalizes the passage with the length being longer than that of the quesiton without considering its content. To show the effect of this issue, we further extend the experiments to test Top-N (intead of Top-1 only) accuracy-rates to demonstrate the superiority of Modified F-measure. Table 6 shows that the performances of Modified F-measure are better under the columns of Top-2 and Top-3. Where "+Both" means that the experiments are under the setting of the optimal N-gram weight and smoothing technique.

Table 6. Top-N accuracy rates of QE-BLEU (+Both) and Modified F-measure (+Both)

	Top-1	Top-2	Top-3
QE-BLEU (+Both)	59.49%	72.15%	79.32%
Modified F-measure (+Both)	59.49%	73.84%	79.75%

Last, we further check 30 wrong cases from the Modified F-measure with optimal parameters along with smoothing technique. It is observed that those associated errors are mainly due to six different types as shown in Figure 2. They will be further illustrated as follows.

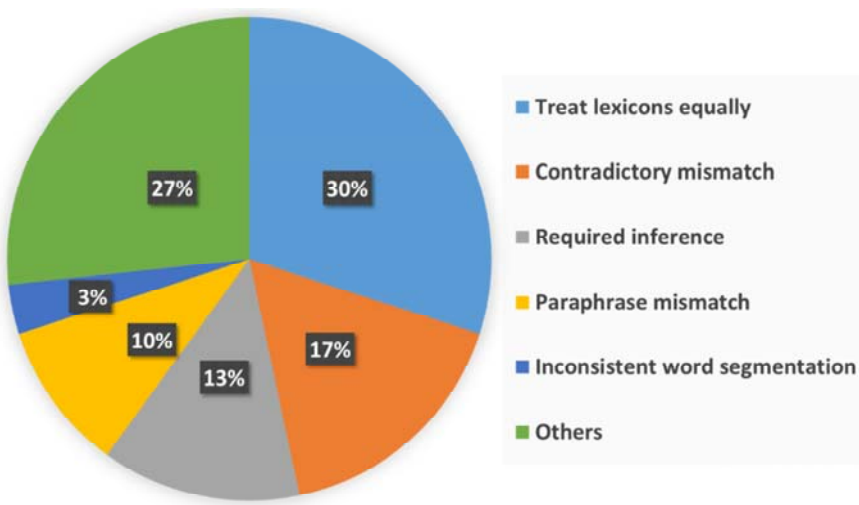

Figure 2. Error types of Modified F-measure

(1) Treat lexicons equally (30%): Since we match the terms without considering which terms are more important in the sentence, some error occurs due to the *focusing-words* are not weighted more. For example:

Question: 班級|幹部|要以|公平|，|公正|的|態度|，|引|導|同學|遵守|團體|秩序|。

"**Class leaders** should guide the classmates to abide by group order with a fair and just attitude."

Top-1 candidate: 並以|公平|，|公正|的|態度|，|引|導|同學|遵守|團體|秩序|，

"and guide them to abide by group order with a fair and just attitude."

Benchmark: 擔任|班級|幹部|，|要能|作為|同學|的|榜樣|，|並以|公平|，|公正|的|態度|，|引|導|同學|遵守|團體|秩序|，

"As a **class leader**, you should be able to serve as a role model for classmates and guide them to abide by group order with a fair and just attitude.

The terms "班級" and "幹部" ("Class leaders") are the important topic words in this Yes/No question. However, they are interleaved with other unmatched words in the first half of the benchmark. The Top-1 candidate, instead of the benchmark, is thus selected because it possesses a higher precision-rate. This kind of error need a specific technique to find the focusing-words in the sentence and give different term weights according to the degree of importance of the terms in the sentence.

(2) Contradictory mismatch (17%): Some Yes/No questions are designed to describe the wrong fact. Therefore, the sentence which describes the wrong fact would not match the evidence sentence in the lesson, but this unmatched evidence sentence still should be regarded as a part of the supporting evidence. For example:

Question: 在|從前|，|農民|參與|民俗|藝|陣|的|目的|，|是|為了|反抗|政府|而|集結|組成|的|。

"In the past, **the purpose** that farmers participate in folk art array was to assemble against the government."

Top-1 candidate: 臺灣|的|民俗|藝|陣|從前|多|是|業|餘|的|組織|，|村民|利用|農閒|時|參與|藝|陣|，

"Taiwanese folk art array used to be an amateur organization, and villagers used leisure time to participate in the folk art array."

Benchmark: 臺灣|的|民俗|藝陣|從前|多|是|業|餘|的|組織|，|村民|利用|農閒|時|參與|藝|陣|，|**既**|**可**|休閒|娛樂|、|練武|強身|，|也|間接|連絡|情誼|，|凝聚|地方|的|向心力|。

"Taiwanese folk art array used to be an amateur organization, and villagers used leisure time to participate in the folk art array. **The purpose** of it is for leisure, martial arts and also connect with friendship, condense the centripetal force of the place."

The sentence "是為了反抗政府而集結組成的" ("was assembled against the government") is the wrong fact in the question which describe the incorrect purpose of forming "民俗藝陣" ("folk art array"). Although the sentences "既可休閒娛樂、練武強身，也間接連絡情誼，凝聚地方的向心力" are not matched, they in fact provide the supporting evidence to conclude that the associated statement in the given question is incorrect. Therefore, they should be included in the supporting evidence. This kind of error also need to identify the focusing-words in the sentence, and emphasize them with larger weights.

(3) Require real-world knowledge (13%): This kind of errors is caused by the shortage of real-world knowledge. For example:

Question: 開|漳|聖王|陳元光|因|開發|漳州|有|功|而|被|當地|人們|所|**信仰**|。

"Chen Yuanguang, the Kaizhang Shengwang, was **believed** by the local people for his contribution in developing Zhangzhou."

Top-1 candidate: 宜蘭縣|壯圍鄉|開|漳|聖王|廟|祭祀|開|漳|聖王|。|因|唐朝|武進士|陳元光|開發|漳州|有|功|，

"In the Zhuangwei Township of Yilan County, the Kaizhang Shengwang Temple worship the Kaizhang Shengwang. Because Chen Yuanguang had contributed in developing Zhangzhou,"

Benchmark: 宜蘭縣|壯圍鄉|開|漳|聖王|廟|祭祀|開|漳|聖王|。|因|唐朝|武進士|陳元光|開發|漳州|有|功|，|**當地人**|**建**|**廟**|**祭祀**|，|**是**|**漳州人**|**的**|**保護神**|。

"In the Zhuangwei Township of Yilan County, the Kaizhang Shengwang Temple worship the Kaizhang Shengwang. Because Chen Yuanguang had contributed in developing Zhangzhou, **the local people built temples to honor him, he is the protecting god of the people of Zhangzhou.**"

To deal with the errors of this category, we need to know that the sentences "當地人建廟祭祀，是漳州人的保護神" ("the local people built temples to honor him, he is the protecting god of the people of Zhangzhou") implies "信仰" ("believe").

(4) Paraphrase mismatch (10%): Since we only count those "exactly matched" words, we

cannot match two terms that describe similar concepts but use different word-types. For example:

Question: **參觀**|名勝古蹟|時|要|維護|環境|的|整潔|。
"We should maintain a clean environment when **visiting** famous places and monuments."

Top-1 candidate: 古蹟|時|，|應|遵守|規定|並|維護|環境|整潔|；
"monuments, you should abide by the regulations and maintain a clean environment."

Benchmark: **拜訪**|名勝|，|古蹟|時|，|應|遵守|規定|並|維護|環境|整潔|；
"When **traveling to** famous places and monuments, you should abide by the regulations and maintain a clean environment."

The terms "參觀" ("visiting") and "拜訪" ("traveling to") have similar meaning but are not matched in string. Therefore, the capability of detecting paraphrasing is needed to deal with this kind of problems.

(5) Inconsistent word segmentation (3%): This type of errors is caused by the inconsistent word segmentation between the word in questions and lessons. For example:

Question: **名勝古蹟**|的|環境|維護|是|政府|的|責任|，|與|參訪|民眾|無關|。
"The environmental maintenance of **historical sites** is the responsibility of the government and has nothing to do with the visitors."

Top-1 candidate: 維護|家鄉|的|**名勝**|，|**古蹟**|，|需要|政府|機關|與|民間|機構|積極合作|，
"The maintenance of **historical sites** in the hometown requires active cooperation between government agencies and private institutions."

Benchmark: 維護|家鄉|的|**名勝**|，|**古蹟**|需要|政府|機關|與|民間|機構|積極合作|，|加強|對|**名勝**|，|**古蹟**|的|管理|與|修復|，|也|需要|居民|共同|關心|與|愛護|。
"The maintenance of **historical sites** in the hometown requires active cooperation between government agencies and private institutions in order to strengthen the management and restoration of **historical sites**. It also requires the resident's care and protection."

Because the same string "名勝古蹟" ("historical sites") is segmented differently in the question (as one word: "名勝古蹟") and in the candidates (as two words: "名勝" and "古蹟"), the system thus regards the second sentence in the benchmark as a purely irrelevant string.

(6) Others (27%): The errors in this category are either the cases that are caused by multiple error types mentioned above or the errors that only occupy a small portion. In the following example:

Question: 位|在|山地|丘陵|的|地方|適合|發展|林業|，|畜牧業|。

"It is suitable for the development of forestry and animal husbandry in the hilly areas."

Top-1 candidate: 山地|丘陵|等|地方|，|發展|出|林業|，|畜牧業|等|活動|；|而|居住|在|平原|地區|的|居民|，

"In hilly areas where forestry, animal husbandry and other activities are developed; for those residents living in the plains,"

Benchmark: 山地|丘陵|等|地方|，|發展|出|林業|，|畜牧業|等|活動|；

"In hilly areas where forestry, animal husbandry and other activities are developed;"

The error is caused by multiple reasons. First, because we treat lexicons equally, the last sentence in the Top-1 candidate matches the stop words which are not important. Second, the last sentence in the Top-1 candidate cannot express a meaning completely by its own. We need to detect the coherent of the sentence to deal with this kind of problem. An another example:

Question: 近年來|各縣市|親水|步道|，|河濱公園|的|設立|都|是|河川|整治|的|成果|，|不但|改善|了|河流|的|水質|，|也|提高|了|居民|的|生活品質|。

"In recent years, **some city's** hydrophilic trails and the establishment of the riverside park are the result of river remediation, which not only improves the water quality of the river, but also improves the quality of life of the residents."

Top-1 candidate: 在|整治|過程|後|，|改善|了|河流|的|水質|，|也|提高|居民|的|生活品質|。

"After the remediation process, the water quality of the river has been improved and the quality of life of the residents has also been improved."

Benchmark: **高雄市|的|愛河|曾|遭受|嚴重|污染|**，|在|整治|過程|後|，|改善|了|河流|的|水質|，|也|提高|居民|的|生活品質|。

"**The love river in Kaohsiung has been seriously polluted.** After the remediation process, the water quality of the river has been improved and the quality of life of the residents has also been improved."

In this case, we need an extra module to link "高雄市" (Kaohsiung) to "各縣市" (some city) because "Kaohsiung" is an instance of "some city".

6. Related Work

As mentioned in Section 1, the previous studies of retrieving supporting evidence can be grouped into three categories: matching terms, conducting syntactic/semantic analysis, and scoring with a translation model. Term matching approaches focus on retrieving the related query from a large scale of documents by using similarity functions and word weight functions. For example, DrQA system (Chen *et al*., 2017) was developed for large scale applications such as retrieving the relevant documents from Wikipedia. In their document retriever model, they evaluated the similarity of the articles and questions by the score of TF-IDF weighted bag-of-word vectors. They also improved the model by taking bi-gram counts. However, those approaches usually do not consider word order and local context.

Syntactic/semantic scoring approaches are specially developed to deal with certain QA datasets. The DeepQA pipeline in IBM Watson system (Murdock *et al*., 2012), which is used in the task Jeopardy![6], presented four passage-scoring algorithms to retrieve the supporting evidence by scoring the passages. The scoring algorithms operate on the syntactic-semantic graphs constructed from analyzing the syntactic and semantic information of the documents. The QA system in (Jansen *et al*., 2017) was developed for standardized science exams. They extracted the focus words according to their scores of the concrete concepts. The words are scored by the psycholinguistic concreteness and rated from 1 (highly abstract) to 5 (highly concrete) by human. Nonetheless, this kind of approaches is more complex and their operations are usually more time-consuming.

Translation model based approaches are widely adopted in community Q&A tasks. They mainly check the similarity between the queried question and those historical questions kept in the archive with a translation model (in which a higher translation score implies that they are more similar). In our case, this approach translates the given question into the specified supporting evidence candidate via a translation model, and then assigns the obtained translation score as the associated score of that candidate. These approaches can be further categorized into word-based approaches and phrase-based approaches. Word-based approaches (Berger *et al*., 2000; Jeon *et al*., 2005; Xue *et al*., 2008) adopt word translation probabilities in a language model to rank the similarity. Zhou *et al*. (2011) further extended this model into a phrase-based one and obtained better performances. This kind of approaches clearly needs large benchmark data which is expensive to construct in our task.

In comparison with previous term matching approaches, our proposed n-gram matching

[6] Jeopardy! is an American television game show.

approaches further consider word order and local context, and thus improve the retrieval accuracy. On the other hand, for those syntactic/semantic scoring approaches, the proposed approaches can operate more efficiently due to the use of simple string matching. Last, comparing with those translation model based approaches, our approaches do not need large training data.

7. Conclusion

Two different models are proposed in this paper to retrieve supporting evidence for the given Yes/No question: *QE-BLEU* and *Modified F-measure*. In comparison with previous approaches, the proposed approaches provide better accuracy and efficiency. Both of them adopt n-gram to incorporate phrases and local context; however, the Modified F-measure takes care of both precision and recall, while QE-BLEU only handles recall of the question. Experiment results have shown that both of them outperform Lucene Apache search engine by 5%.

Our main contributions mainly are: (1) We proposed and tested two novel approaches to retrieve the supporting evidence, and have obtained better performances. (2) We list the desired characteristics of the supporting evidence retrieved. (3) We implement and compare various refinement techniques, including smoothing and optimization, for the proposed approaches.

References

Berger, A., Caruana, R., Cohn, D., Freitag, D. & Mittal, V. (2000). Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 192-199. doi: 10.1145/345508.345576

Chen, B. & Cherry, C. (2014). A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. doi: 10.3115/v1/W14-3346

Chen, D., Fisch, A., Weston, J. & Bordes, A. (2017). Reading Wikipedia to Answer Open-Domain Questions. Retrived from *arXiv preprint arXiv:1704.00051*

Chinchor, N. (1992). The statistical significance of the MUC-4 results. In *Proceedings of the 4th conference on Message understanding*, 30-50. doi: 10.3115/1072064.1072068

Jansen, P., Sharp, R., Surdeanu, M. & Clark, P. (2017). Framing QA as Building and Ranking Intersentence Answer Justifications. *Computational Linguistics*, *43*(2), 407-449. doi: 10.1162/COLI_a_00287

Jeon, J., Croft, W. B. & Lee, J. H. (2005). Finding similar questions in large question and answer archives. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, 84-90. doi: 10.1145/1099554.1099572

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval*. New York, NY: Cambridge university press.

Murdock, J. W., Fan, J., Lally, A., Shima, H. & Boguraev, B. K. (2012). Textual evidence gathering and analysis. *IBM Journal of Research and Development*, *56*(3-4), 8:1-8:14. doi: 10.1147/JRD.2012.2187249

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, 311-318. doi: 10.3115/1073083.1073135

Sasaki, Y. (2007). The truth of the F-measure. Teach Tutor mater. Retrived from *http://www.flowdx.com/F-measure-YS-26Oct07.pdf*

Xue, X., Jeon, J. & Croft, W. B. (2008). Retrieval models for question and answer archives. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 475-482. doi: 10.1145/1390334.1390416

Zhou, G., Cai, L., Zhao, J. & Liu, K. (2011). Phrase-based translation model for question retrieval in community question answer archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, *1*, 653-662.