Chen, W. et al. (Eds.) (2016). Proceedings of the 24thInternational Conference on Computers in Education. India: Asia-Pacific Society for Computers in Education

Graph-Alignment Approach towards Identifying Gaps in Student Answer

Archana SAHU[a*], Plaban Kumar BHOWMICK[b]
[a]*Centre for Educational Technology, Indian Institute of Technology Kharagpur, West Bengal, India*
[b]*Centre for Educational Technology, Indian Institute of Technology Kharagpur, West Bengal, India*
*sahuarchana7@gmail.com

Abstract: Formative assessment aims to provide hints regarding quality of student answers. The hints are in the form of gaps or irrelevant entities in the student answers on comparison with the corresponding model answers. The main objective of the present work is to extract gaps in short answers provided by students. In this work, each of the student and model answers for a given question are transformed into graphs by extracting different relations present in the answers. The nodes of the resultant answer graphs are aligned by considering the similarity of neighborhood topologies of a pair of nodes in addition to the node-node similarity. This leads to the effective extraction of gaps in the form of meaningful phrases in student answers, rather than mere words. Evaluation metrics for reporting performance of the proposed task have been formally defined in this paper. We have compared our proposed methodology with a word-word alignment based baseline system. The proposed methodology outperforms the baseline system in a significant margin.

Keywords: Short answer grading, Formative Feedback, word-to-word alignment, Relation Extraction, Graph Alignment.

1. Introduction

Automatic evaluation of student generated discourses such as essays, short answers has attracted considerable attention of natural language processing and educational technology research communities. These systems work with the aim to ease the burden of human graders, who are generally entrusted with responsibilities to grade innumerable writings in a short time. Evaluation of student responses can be classified into holistic evaluation and formative evaluation.

In the present work, our aim is to focus on providing formative feedback to the student answers driven by the following requirements pertaining to the educational scenario:

- Apart from the grade, students also desire to know the areas where they are lagging so that they could work upon them to improve their learning.
- At the same time, instructors wish to recognize where students are struggling so that they could adapt their instructional strategy.

This is where the concept of formative assessment comes into focus where the objective is to provide feedback to the students regarding gaps that are there in the student provided answers as well as irrelevant entities in them. The task of generating formative assessment can be divided into three sub-tasks: a) Identifying gaps, b) Corrections to remediate the identified gaps and c) Entities irrelevant to the question. The task of formative assessment can be defined as follows:

Definition 1: Formative Assessment
The problem of formative assessment can be described as follows:
Input: A pair of short answers **(S, M)** to the same question **Q** where:
S: Student answer
M: Model answer
Output: Gaps and redundant entities in student answer

Generation of formative feedback may vary with the type of answers that are being evaluated. For example, feedback that a student receives in concept completion type answer should contain concepts that are missing in the student answer; whereas, formative assessment of argumentative answers should point out the unsupported claims. Various cases involving following kinds of answers present the central challenges to the system of formative assessment (FA):

- **Concept-completion:** A question from these kind of answers has a prompt and each prompt is like a slot which has to be filled up with appropriate concept(s). Both M and S would have such slots that are filled with some concepts. The FA system is put to test in such circumstances when it should be capable of detecting synonymy and paraphrasing to match the concepts in the pair of answers. This would also lead to extraction of correct gaps as well as the redundant entities in S, too.
- **Definition-type of answers:** This kind of answer seeks the definition involving a concept. The challenge for FA here lies in effective extraction of salient points that defines the 'concept' matching of the same so that the crucial terms missing in S can be highlighted in addition to the redundant entities in it.

With reference to the challenges faced by the formative assessment system, following are our contributions for the same.

- **Identification of gaps in student answer:** Meaningful gaps in the form of phrases are extracted in student answer
- **Evaluation metric:** New evaluation measures have been formally defined to evaluate the task of formative assessment.
- **Relation extraction and Graph alignment approach:** Extraction of relations in the sentences of answers uncover the key elements. This is then followed by alignment of answer graphs and appropriate thresholding of similarity between the aligned nodes, which helps in pinpointing the gaps in student answers.

In this paper, we focus on the extraction of gaps in short answers. The central idea in the proposed methodology is to compare a student answer and the corresponding model answer by transforming the answers into their corresponding abstract representations. The abstract representation is realized through answer graphs generated by extracting the relations present in the respective answers.

The next section gives an overview of the approaches discussed so far in the research area of formative assessment. This is followed by the description of baseline system and the proposed system based on Graph-alignment. The evaluation metrics for FA have been defined in Section 4, followed by the experimental results and analysis.

2. Related Work

The approaches proposed for formative assessment of student answers have been discussed in this section.

C-rater (Sukkarieh & Bolge, 2008; Sukkarieh & Blackmore, 2009) has adopted concept-based scoring to enable individualized formative feedback for each student answer (concept-completion answers have been tested). At first, a sample of students' answers (manually annotated with evidence for each of the concepts in the test question, such that concept C entails evidence E) and the corresponding test question are available. Then a set of model sentences are manually written by referring to the manually annotated students' answers. The pairs of students' answers and corresponding model answers are linguistically processed using a deeper parser followed by extraction of Linguistic features based on hand-written rules. The result is a flat structure representation consisting of phrases, predicates and relations between them. Then, pronoun resolution and morphological analysis of the entities extracted as part of the linguistic features is done. The above process is repeated for all the pairs of answers selected as training data and then the features extracted are used to train a Goldmap matching model. The answer pairs are manually labeled as 0 or 1 for no-match, match respectively.

Unseen student answers and the corresponding model answers are processed as follows:

- Spelling correction
- Part-of-speech tagging and deeper parsing (OpenNLP parser outputs a deep constituent parse tree)
- Feature extraction from parse tree (same as done for training data)
- Pronoun resolution and morphological analysis

The features extracted for a pair of unseen student answer and model answer, is applied to the trained Goldmap matching model (based on maximum entropy modeling). A probability on the match between the unseen student answer and model answer is obtained. A specific threshold is decided to determine a match. C-rater gives quality feedback to students with details such as, which concepts they get right in their answers and which concepts they get wrong.

No comprehensive evaluation for concept-based scoring and linguistically-driven feedback has been carried out for C-rater. There are no such notable works till date, regarding formative assessment of Definition answers and Explanation answers.

Some recent works such as Auto-marking (Sukkarieh, Pulman, & Raikes, 2003) aim to provide formative feedback for each student answer in the following manner:

- A panel of experienced teachers may be employed to look at samples of student answers and sort each of the answers into a feedback category depending on its semantic content.
- The teachers write formative feedback for each category, leading to a sample of student answers matched to appropriate formative feedback. When an unseen student answer was submitted, it is initially compared with all the sample answers. The formative feedback of the category of the sample student answers to which the unseen student answer is close enough, is assigned to the student answer.

Systems such as OpenMark (Butcher, 2008) are being used for formative assessment. OpenMark has been designed by Open University such that feedback at multiple levels of learning could be included.

WriteToLearn (Landauer, Lochbaum, & Dooley, 2009) is an iterative writing tool in which students write essays and receive feedback about their writing. Feedback regarding various aspects of writing such as ideas, organization, word choice, sentence fluency is given to the students.

Review of the existing works in the current problem area resulted in the following research gaps:

- **Limitation of approaches:** Although some efforts have been put forward in identifying gaps in student answers, no formal approach has been adopted till date.
- **No evaluation of formative assessment (FA) systems:** There is a clear absence of appropriate literature regarding a systematic study for evaluation of FA systems. The evaluation measures based on which a particular automated FA system could be declared as close to human assessment have not been discussed anywhere.

3. Research Design

The main objective of the present work is to generate formative feedback for the student answers. As existing literature does not contain a comprehensive study in the current problem domain, we intend to compare our proposed method with a baseline system.

3.1 Baseline System

The baseline system adopts a naive approach in the form of simple word-to-word alignment between a pair of answers (Sultan, Bethard, & Sumner, 2014). The related words in the two answers are aligned by exploiting semantic and contextual similarities of the words. The words in model answer not participating in the resulting word-to-word alignment are designated as gaps in the student answer. This is shown with an example shown in Figure 1.

It is observed that the gaps in student answer are extracted as words rather than complete phrases, for example, ‘function’ and ‘members’ indicate independent entities or different gaps but actually refer to one gap i.e., ‘function members’. The gaps may also contain parts of phrases that are actually not gaps, for example ‘members’ (occurred twice, the first one refers to ‘function members’ and the second one refers to ‘data members’). Such occurrences of partial phrases lead to confusion regarding the actual gaps in student answer.

Example: What are the elements typically included in a class definition?
Model answer: The elements typically included in a class definition are function members and data members.
Student answer: An object and data are included in a class definition.
Word-to-word alignment:
included ⟶ included
in ⟶ in
a ⟶ a
class ⟶ class
definition →
definition and
⟶ and
data ⟶ data

Figure 1. Example trace of the baseline system

3.2 Proposed System

In this work, we propose a graph alignment based approach towards formative assessment. The proposed approach is decomposed into several subtasks.

3.2.1 Relation extraction

A set of triples of the form <Subject, Predicate, Object> is extracted from each of the answers in a <student answer, model answer> pair (Del Corro & Gemulla, 2013). Answer graphs are constructed for each of the answers in the <student answer, model answer> pair using the triples corresponding to each of the answers. As discussed earlier, an answer graph is an unweighted and undirected graph where nodes represent different phrases and edges represent different predicates. The construction of an answer graph is illustrated in Figure 2.

Question: What are the elements typically included in a class definition?
Answer: The elements typically included in a class definition are function members and data members.

S-P-O triples:

“The elements included in a class definition” “are” “function members”
“The elements included in a class definition” “are” “data members typically”

Answer graph constructed from the S-P-O triples

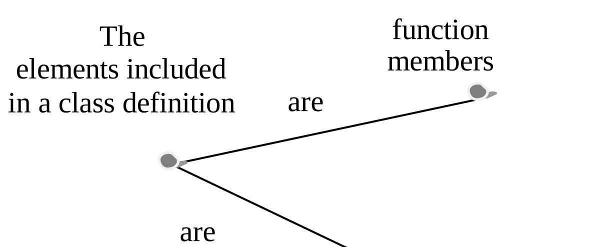

Figure 2. Construction of answer graph for an answer

3.2.2 Alignment of Answer graphs

The goal of alignment of answer graphs is to obtain one or more mapping(s) between the nodes of the input answer graph pair and for each mapping, the set of common edges.

The IsoRank algorithm (Singh, Xu, & Berger, 2007; 2008) has been originally used to obtain an alignment between multiple protein-protein interaction (PPI) networks. It is based on the idea that a protein in one network is matched well to a protein in another network if the protein sequences as well as their neighborhood topologies match well. It is used in the proposed system for matching of a pair of answer graphs. The three stages of IsoRank algorithm are explained as follows:

- **Stage 1: Structural alignment between a pair of nodes in answer graphs**
  This is done by the iterative computation of similarity between their neighborhood topologies using *Power Method* (Golub & Van Loan, 1996). An eigenvalue approach is followed for the score computation. Let the similarity score between all pairs of nodes in answer graphs be denoted as �

  $$� = A� \quad (1)$$

  �represents the principal eigenvector of A. A indicates the support provided to each node-pair, due to the matching between their respective neighboring nodes.
- **Stage 2: Combination of Structural alignment and Content alignment**
  Only structural similarity will not give a sense of similarity between a pair of nodes. So, it is necessary to look into the similarity between content in each of the nodes. Hence, the eigenvalue equation in equation (1) should contain a term representing the content-based similarity. The content-based similarity between a pair of nodes is denoted by �. It is computed using the word- vector similarity measures (Mikolov, Chen, Corrado, & Dean, 2013).

  The combination of Structural alignment and Content alignment between the nodes of the answer graphs is represented as:

  $$� = \alpha � + (1 - \alpha)� \qquad (2)$$

  α acts as a tuning parameter to control the weight of similarity score involving neighborhood topologies of each pair of nodes, relative to that of node-to-node semantic similarity measures between the pair of nodes.
- **Stage 3: Extraction of node-node mapping between node pairs from input answer graphs**
  Now �in equation represents a bipartite graph connecting two sets of nodes V_S and V_M. Finding global alignment between Student answer graph and Model answer graph is now mapped to a bipartite matching problem which can be solved with a greedy approach (Singh, Xu, & Berger, 2008).

In other words, the IsoRank algorithm obtains suitable mapping between a pair of answer graphs provided as an input to it, for example as shown below in Figure 3.
Recent advancements in PPI network alignment namely SPINAL (Scalable Protein Interaction Network Alignment) (Aladağ & Erten, 2013) claim to obtain improved alignment between PPI networks as compared to that obtained using IsoRank algorithm. SPINAL involves an additional concept of *contributors* (C). These are pairs of vertices with higher chances of existence in the optimum one-to-one alignment. In the process of computation of neighborhood topology similarity scores for a pair of vertices in a pair of PPI networks, only the *contributors* in the immediate neighborhood contribute to the neighborhood similarity score inverse proportional to its degree product. This is in contrast to IsoRank algorithm in which each and every pair of vertices in the immediate neighborhood contribute to the neighborhood similarity score of a pair of vertices corresponding to a pair of PPI networks.

3.2.3 Identification of common subgraphs between the input answer graphs

Let $�_M$ and $�_S$ represent the graphs of a pair of model answer and student answer. Common subgraphs (which may be disconnected from each other) are then identified from the resulting global alignment between $�_M$and $�_S$ in the following manner:

- Let node �1 in $�_M$be aligned to node �2 in $�_S$ and node �1 in $�_M$ be aligned to node �2 in $�_S$.

  �1-����1-�1 and �2-����2-�2 are edges in the answer graphs, which are called as the supporting edges for the corresponding edge to be created in the common subgraph between the

  pair of answer graphs, where ����1 and ����2 represent the predicates for the corresponding edges. Hence the output subgraph would contain an edge between the nodes �1/�2 and �1/�2 with predicate having weight equal to the similarity between the predicates ����1 and ����2, as shown below:

����ℎ� = ���(����1, ����2) (3)

- The nodes �1/�2 indicate that �1, �2 refer to the same node in the common subgraph.

Similarly, �1/�2 indicate that �1 refer to the same node in the common subgraph.

An illustration for identification of common subgraph is depicted in Figure 4.

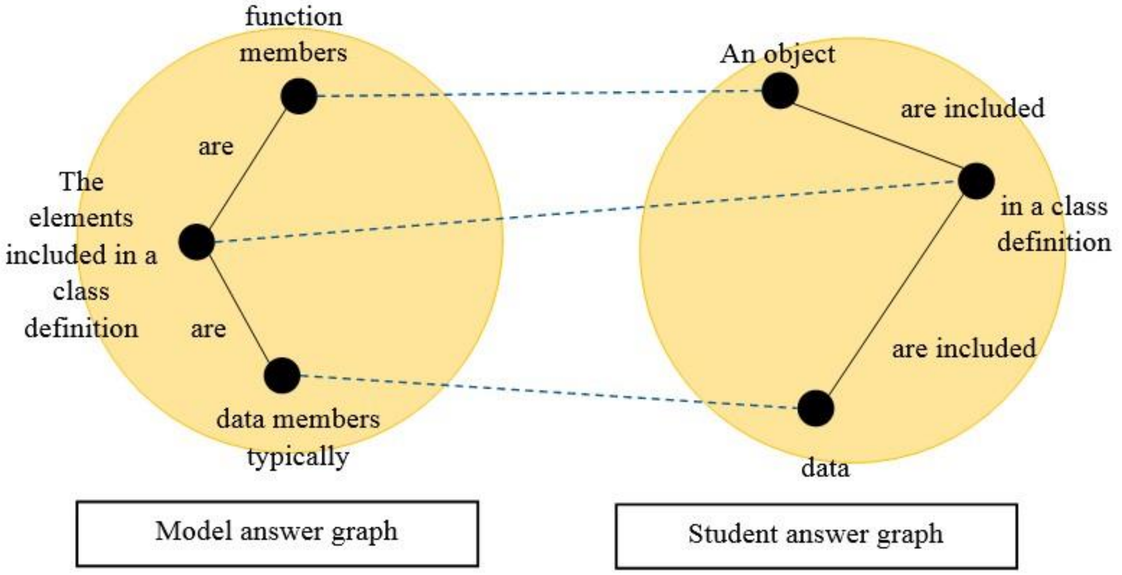

Figure 3. Alignment between Model answer graph and Student answer graph

“The elements included in a class definition” ⟶ “in a class definition”
“Function members” ⟶ “An object”
“Data members typically” ⟶ “Data”

3.2.4 *Extraction of formative feedback candidates*

The structure of the induced common subgraph is analyzed in order to extract formative feedback candidates. It is assumed that the length of student answer and model answer are the same, leading to the corresponding answer graphs having same number of nodes.

Let $|V_M|$ be the number of nodes in model answer M.

Let $|V_S|$ be the number of nodes in student answer S.

Now, $|V_M| \neq |V_S|$.

Let $V_M = \{V_{M1}, V_{M2}, V_{M3}, V_{M4}\}$ and $V_S = \{V_{S1}, V_{S2}, V_{S3}, V_{S4}\}$ indicate sets of vertices in �$_M$ and

�$_S$,

respectively, which are aligned optimally as follows:

$$V_{M1} \longrightarrow V_{S3}$$
$$V_{M2} \longrightarrow V_{S2}$$
$$V_{M3} \longrightarrow V_{S4}$$
$$V_{M4} \longrightarrow V_{S1}$$

If there exist edges e$_M$, e$_S$ in �$_M$and �$_S$, respectively, such that:

$e_M = (V_{M1}, V_{M2})$ and $e_S = (V_{S3}, V_{S2})$ then e$_{MS}$ is an edge in the common subgraph such that:

$e_{MS} = (V_{M1}/V_{S3}, V_{M2}/V_{S2})$

where V_{M1}/V_{S3} and V_{M2}/V_{S2} are single nodes in the induced common subgraph connected by a single edge e_{MS} having weight equal to the predicate similarity of the edges e_M and e_S (in the individual student and model answer graphs �$_M$ and �$_S$).

Similarity scores involving different components with the common subgraph are computed.

- **Node-pair similarity:** Similarity between a pair of nodes residing at each end of a common edge.

$$sim_1 = Similarity(V_{M1}, V_{S3})$$

$$sim_2 = Similarity(V_{M2}, V_{S2})$$

- **Predicate similarity:** Similarity between a pair of predicates representing a common edge.

$$sim_3 = Similarity(e_M, e_S)$$

- **Node-pair similarity:** There do not exist any common edge containing the nodes V_{M3} and V_{S4} or the nodes V_{M4} and V_{S1}. Hence these node-pairs form a disconnected common subgraph. Similarity between such pair of nodes is shown as follows:

$$sim_4 = Similarity(V_{M3}, V_{S4})$$

$$sim_5 = Similarity(V_{M4}, V_{S1})$$

The gaps in S are determined as follows:

$$\begin{aligned} \text{Gaps in S} &= V_{M1}, && sim_1 < \rho_1 \\ &= V_{M2}, && sim_2 < \rho_2 \\ &= e_M, && sim_3 < \rho_3 \\ &= V_{M3}, && sim_4 < \rho_4 \\ &= V_{M4}, && sim_5 < \rho_5 \\ &= \text{none}, && \text{otherwise} \end{aligned}$$

Where, ρ_1, ρ_2, ρ_3, ρ_4, ρ_5 are the values of threshold (limiting factors) in each case, values below which indicate potential chances of Gaps in S.

If there do not exist such edges e_M, e_S in �$_M$ and �$_S$, respectively, then single-node disconnected common subgraphs where each of which consist of common nodes represented by the pair of aligned nodes such as V_{M1}/V_{S3}, V_{M2}/V_{S2}, V_{M3}/V_{S4}, V_{M4}/V_{S1} are constructed.

Gaps in S = V_{M1}, V_{M2}, V_{M3}, V_{M4} (except e_M)

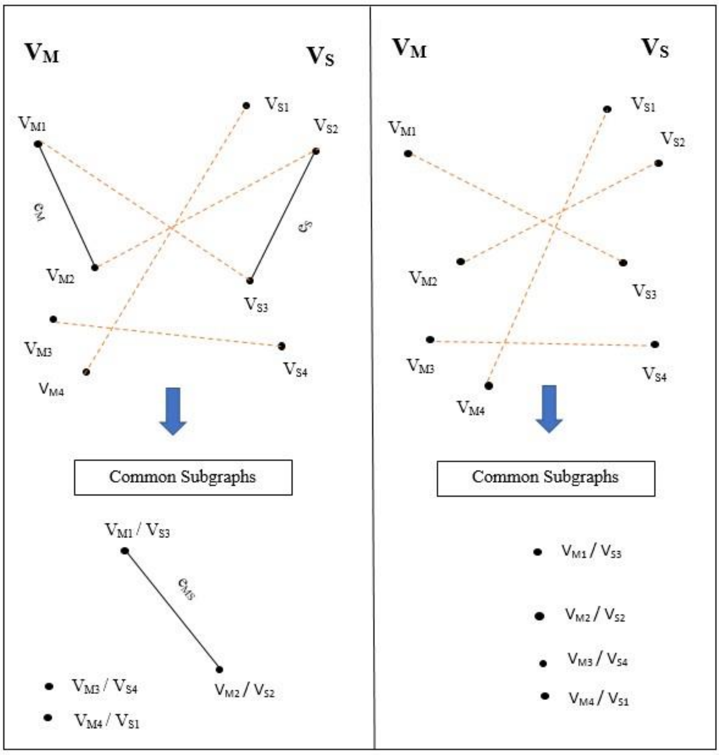

Figure 4. Identification of common subgraphs

An example for extraction of gaps in student answer on alignment of a pair of answer graphs is shown in Figure 5.

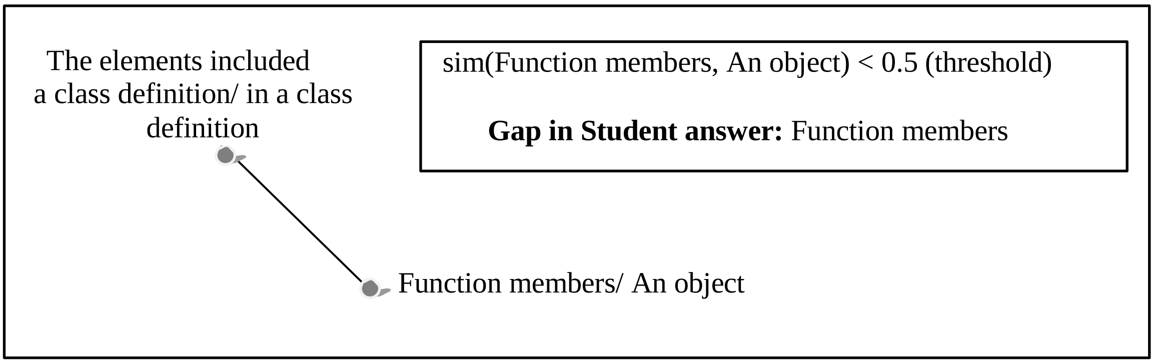
The elements included a class definition/ in a class definition
sim(Function members, An object) < 0.5 (threshold)
Gap in Student answer: Function members
Function members/ An object

Figure 5. Extraction of gaps in Student answer

4. Evaluation

4.1 Test Bed

The data set comprising of questions and answers, as part of assignments of a Data Structures course at the University of North Texas[1]is used as test bed for the purpose of experimentation in formative assessment. There are student answers provided by a class of undergraduate students and corresponding model answers to around 4 questions spread across an assignment and an examination. The questions involve concept-completion type of answers. A total of 40 student answers have been considered. Gold standard data for the student answers with regard to detection of gaps has been manually prepared.

4.2 Evaluation of formative assessment system

There are no research works so far that have clearly described the evaluation metrics for formative assessment system. We define Micro-average precision and Macro-average precision to measure performance of a formative assessment system.
Let the true labels or gold standard annotations for a pair of answers <S, M> be denoted as *G* and FA system prediction labels be denoted as *P*, where:

$$G = \{g_1, g_2, \dots, g_n\}$$

$$P = \{p_1, p_2, \dots, p_n\}$$

Here:

$g_i = \{phrase1, phrase2, \dots\}$ where $phrase1, phrase2, \dots$ indicate missing keywords/ phrases of M in S, known as gaps in S in gold standard data.

$p_i = \{phrase1, phrase2, \dots\}$ where $phrase1, phrase2, \dots$ indicate missing keywords/ phrases of M in S, known as gaps in S predicted by the FA system.

The following evaluation metrics are defined for gap identification problem. For a pair of answers (S_i, M_i) in the dataset:

TP_i = The number of system predicted gaps p_i that belong to the set of gold standard gaps g_i for the i^{th} question.

$$TP_i = \sum_{k=1}^{|p_i|} (p_{ik} \in g_i) \tag{4}$$

[1] Expanded version of the data set used by (Mohler & Mihalcea, 2009)

FP_i = Number of non-gaps (keywords/phrases of M actually not gaps in S) wrongly identified as gaps.

$$FP_i = \sum_{j=1}^{|M_i|} k_j(M_i) \notin g_i \qquad (5)$$

Precision is the fraction of gaps detected by the FA system that are relevant in S.
For a pair of answers (S$_i$, M$_i$) in the dataset, the precision is defined as:

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \qquad (6)$$

$$Macro\text{-}Average\ Precision = \frac{1}{|P|} \sum_{i=1}^{|P|} Precision_i \qquad (7)$$

$$Micro\text{-}Average\ Precision = \frac{\sum_{i=1}^{|P|} TP_i}{\sum_{i=1}^{|P|} TP_i + \sum_{i=1}^{|P|} FP_i} \qquad (8)$$

The following experiment is conducted to measure the efficiency of the Baseline system as well as the proposed system with regard to extraction of gaps in student answers.

Experiment: Computation of the above defined evaluation metrics between the True gaps and System detected gaps in Student answer

As explained previously, the number of **TP**, **FP** are calculated accordingly for each pair of answers. **Precision** is computed for each pair of answers. The **Micro-average precision** and **Macro-average precision** for the entire testing data is then computed accordingly.

Table 1: Comparison of evaluation metrics between that computed using Baseline system and Proposed System for FA

	Baseline System for FA	Proposed System (based on Graph alignment of answer graphs)
Micro-average Precision (Gaps)	0.6077	**0.6375**
Macro-average Precision (Gaps)	0.6094	**0.6508**

It is observed from Table 1, that the Proposed System for FA performs significantly better (Micro-average precision = 0.6375 and Macro-average precision = 0.6508) as compared to the Baseline system (Micro-average precision = 0.6077 and Macro-average precision = 0.6094) with regard to extraction of gaps in student answers. The explanation for the above results can be given as follows:

- It is to be noted that the phrases occurring as gaps in student answer may be mere repetitions of some portions of the respective questions. This has been remedied by suitable matching and detection of potential gaps in questions and then figuring out the actual ones in the proposed system.
- Nevertheless, the detection of gaps by the Proposed System for FA shows quite better results than the baseline system, which is evident clearly from above.

5. Conclusion and Future Scope

In this work, the task of providing formative feedback to the students is carried out that is desirable for their future enhanced performance in exams. A baseline system based on word-to-word alignment is prepared as well a system based on Graph-alignment is proposed for FA. Gaps extracted using the baseline system are words rather than phrases. Due to this, there is a false notion of more gaps rather than few of them. The proposed system based on Graph-alignment approach manages to extract more meaningful gaps in the form of phrases in student answers. This is clearly observed from the significantly high values of evaluation metrics computed using proposed system as that compared to Baseline system. The concept of using neighborhood similarity of pairs of nodes in addition to the traditional node-node similarity scores in Graph alignment have resulted in a huge improvement in the precision measures.

This system has been successful in providing formative feedback. In addition, to the best of our knowledge, this is the first time that supporting evaluation has been provided for evidence. Evaluation measures have been newly defined in the field of formative assessment.

We plan to perform the experiment for a larger span of answers, with a wide variety at the same time. Presently, we have worked upon a limited set of concept-completion type of answers. It is our aim to extend the experimental work and analysis towards the definition and explanation-type of answers too. It is perceived that the experimental results could be improved by considering better extraction of relations from the answers as well as better graph alignment approaches in alignment of answer graphs. We plan to work on the same in future.

References

Aladağ, A. E., & Erten, C. (2013). SPINAL: scalable protein interaction network alignment. *Bioinformatics*, 917-924.

Butcher, P. (2008). Online assessment at the Open University using open source software: Moodle, OpenMark and more.

Del Corro, L., & Gemulla, R. (2013). Clausie: clause-based open information extraction. *22nd international conference on World Wide Web* (pp. 355-366). International World Wide Web Conferences Steering Committee.

Golub, G. H., & Van Loan, C. F. (1996). Matrix computations. *3*.

Landauer, T. K., Lochbaum, K. E., & Dooley, S. (2009). A new formative assessment technology for reading and writing. Theory into Practic. *48*(1), 44-52.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector spac. *arXiv preprint arXiv*, 1301-3781.

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. *12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 567-575). Association for Computational Linguistics.

Singh, R., Xu, J., & Berger, B. (2007). Pairwise global alignment of protein interaction networks by matching neighborhood topology. *Annual International Conference on Research in Computational Molecular Biology* (pp. 16-31). Springer Berlin Heidelberg.

Singh, R., Xu, J., & Berger, B. (2008). Global alignment of multiple protein interaction networks with application to functional orthologydetection. *National Academy of Sciences, 105*(35), 12763-12768.

Sukkarieh, J. Z., & Blackmore, J. (2009). c-rater: Automatic Content Scoring for Short Constructed Responses. *FLAIRS Conference.*

Sukkarieh, J. Z., Pulman, S. G., & Raikes, N. (2003). Auto-marking: using computational linguistics to score short, free text responses. *International Association for Educational Assessment (IAEA), Manchester, UK.*

Sukkarieh, J., & Bolge, E. (2008). Leveraging C-rater’s automated scoring capability for providing instructional feedback for short constructed responses. *Intelligent Tutoring Systems, Springer Berlin Heidelberg*, (pp. 779-783).

Sultan, M. A., Bethard, S., & Sumner, T. (2014). Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*, 219-230.