AUTOMATIC ASSESSMENT OF SHORT FREE TEXT ANSWERS

Fátima Rodrigues and Lília Araújo
GECAD – Knowledge Engineering and Decision-Support Research Center,
Institute of Engineering, Polytechnic of Porto (ISEP/IPP), Rua Dr. António Bernardino de Almeida, Porto, Portugal

Keywords: Computer Assisted Assessment, Part of Speech Tagging, Computerized Grading.

Abstract: Assessment plays a central role in any educational process, because it is a common way to evaluate the students' knowledge regarding the concepts related to learning objectives. Computer assisted assessment is a research branch established to study how computers can be used to automatically evaluate students' answers. Computer assisted assessment systems developed so far, are based on a multitude of different techniques, such as Latent Semantic Analysis, Natural Language Processing and Artificial Intelligence, among others. These approaches require a reasonable corpus to start with, and depending on the domain, the corpus may require regular updates. In this paper we address the assessment of short free text answers by developing a system that captures the way the teacher evaluates the answer. For that, the system first classifies the teacher question by type. Then concerning the type of question, the system permits the teacher define scores associated with subparts of the answer. Finally, the system performs the assessment based on these sub scores. For certain types of questions, paraphrases of answers are also considered in an attempt to obtain a more precise assessment. The system was trained and tested on exams manually graded by a History teacher. Based on the results obtained, we show that there is a good correlation between the evaluation of the instructor and the evaluation performed by our system.

1 INTRODUCTION

The correction of questions in an evaluation process involving a large number of free text answers presents teachers with three major problems. Firstly this procedure is very time expensive, as teachers dedicate approximately 30% of their time to exam correction (Mason and Grove-Stephenson, 2002). In addition to this, it is very difficult to ensure an equitable application of evaluation criteria. This is due not only to the subjective nature of the assessment of free text responses, but also to the lengthy evaluation process. Teachers must be highly concentrated for long periods of time, and therefore assessment is subject to variations of level of concentration and of mood of the human being. This can lead to different evaluation grades for answers with similar quality, thus creating inequities in the assessment process that could be even more pronounced if the evaluation is conducted by different evaluators. Moreover, such task involves human labour that cannot be reused.

In this paper we will describe the various functionalities added to an application in order to perform automatic assessment of short free text answers. Firstly we classify the teacher's questions/answers by type; then a spell checker is applied to the students' answers to correct misspelling errors. After that, both teacher and students' answers are processed through several text pre-processing tasks that reduce them to their canonical form. Such tasks include removing punctuation and words without any semantic associated (stop words); word reduction to its radical (stemming); and a morphological analysis is also performed to tag each word in the sentence with its corresponding part of the speech element, such as, noun, adjective, pronoun, verb, etc. These combined tasks turn text into a canonical form that is more manageable.

Student answers (SAs) should be compared with reference answers (RAs). If there is only one RA for each question, the assessment will be very limited and punitive and may fail. To surpass this limitation we create paraphrases of RAs that will provide different correct variations of RAs, for the same question, with a vocabulary more varied and less restrictive, that will allow a more accurate

Rodrigues F. and Araújo L..
AUTOMATIC ASSESSMENT OF SHORT FREE TEXT ANSWERS.
DOI: 10.5220/0003920800500057
In *Proceedings of the 4th International Conference on Computer Supported Education* (CSEDU-2012), pages 50-57
ISBN: 978-989-8565-07-5

assessment.

With these functionalities, the classroom assessment application will be consistent in the way it scores SAs. Moreover it will provide enormous time savings by reducing the time spent by evaluators and it will allow students to test their knowledge at any time, enabling them to adapt their study according to their progress or individual limitations.

The remainder of this paper is organized as it follows: section 2 provides an insight into some current available systems for automatic assessment. In section 3, an overview of our system is made in terms of its main modules. In the next sections, the classification and pre-processing modules are described in detail. The following section outlines the evaluation process applied and presents the experiments performed and their results. In the last section, conclusions and future work are presented.

2 CURRENT APPROACHES ON AUTOMATIC ESSAY GRADING

Computer Assisted Assessment (CAA) of free text answers is a long-standing problem that has attracted interest from researchers since the 1960s. CAA systems can be distinguished by the way they evaluate essays, either for style or for content, or for both. Another distinguished dimension is the approach adopted for assessing style and/or content. The most important approaches found in existing CAA systems are Statistical, Latent Semantic Analysis (LSA) and Natural Language Processing (NLP). The first CAA systems, which were focused on statistical approaches, captured only the structural similarity of texts. The following systems, which were based on LSA, did more than a simple analysis of co-occurring terms. In fact, they introduced two new approaches to the problem: a comparison based on a corpus, and an algebraic technique which allowed to identify similarities between two texts with different words (Thomas et al. 2004). The latest systems are based on NLP techniques and can do intelligent analyses that capture the semantic meaning of free text documents.

It is beyond the scope of this paper to fully cover the state-of-the-art automated assessment in terms of current implementations. (Valenti et al., 2003) and more recently (Perez Martin et al., 2009) do so in great depth: the former, looks at 10 different systems, while the latter investigates 22 systems. The philosophy of the researchers, the type of assessment, the method of assessing and the format of the programs are described in the two works above mentioned. Some comments from the authors about the efficiency of the systems are supplied as well. It is important to note, however, that the test data and metrics used to analyse all the systems are not consistent, and therefore the results are not necessarily comparable. This is due to the lack of reference materials for testing purposes. This field has no large corpus of essays that can be used as a standard measure of automated grading systems (Valenti et al., 2003). With most research projects marking their own sets of essays and judging according to their own correlation criteria, it is difficult to accurately compare systems. For this reason, it is very difficult to determine which system and methods are the best.

As mentioned above, the distinction is made between grading essays based on content and those based on style. While there are systems that evaluate primarily based on style, Project Essay Grade (PEG) (Page, 1994); or on content, Intelligent Essay Assessor (IEA) (Jerrams-Smith et al., 2001), Educational Testing Service (ETS I) (Whittington and Hunt, 1999), Conceptual Rater (C-Rater) (Burstein et al., 2001), most of the latest systems aim to grade across both dimensions, Bayesian Essay Test Scoring sYstem (BETSY) (Rudner and Liang, 2002), Automark (Mitchell et al., 2002), Paperless School free-text Marking Engine (PS-ME) (Mason and Grove-Stephenson, 2002). Another feature of these systems is that they widely differ in the methods used. For example, IEA is based on LSA, whereas E-rater is based on NLP, BETSY uses Bayesian Networks and PEG works on linguistic features via proxes. None of these systems is adequate for our purposes because they only handle text written in English and require large volumes of text for learning, particularly those based on LSA and Artificial Intelligence techniques. Our system is developed to process short free text answers written in Portuguese.

3 PROPOSED APPROACH

The system was developed under a modular approach. It is mainly composed of four modules:

- A classification module that permits the teacher to classify each question/answer in three main types: enumeration, specific knowledge and essay. The enumeration answer has its own structure that consists of a list of topics separated by commas. Specific knowledge questions are “Wh questions” (Who, What, Where, Which, When or How) and

may also include definitions. The answers to those questions are more limited in terms of vocabulary and should not differ greatly from the RA. Essay answer is really a text free answer, that can be affirmative, negative or merely an opinion;

- A pre-processing module that combines various NLP techniques to convert answers into a canonical form that is more manageable and easily to interpret and compare;
- An evaluation module where we calculate the similarity between SAs and RAs and use various metrics to compare the system with the teacher´s evaluation;
- A feedback module that gives personal comments to students based on missing topics and topics that are not covered deeply enough in their responses. The feedback will consist on information about students' errors and performance, but also of adaptive hints for the improvement of his/her solution. The system also gives feedback to the teacher on the topics less covered by the class or by a student. This module will not be described in detail because it is still under development.

The SAs were collected using the UNI_NET-Classroom application that provides management support to school teachers. The input to our system is a database of answers performed by students that belong to various classes, about an exam conceived by a teacher with its corresponding RAs. Thus the system always has a set of answers concerning the same question to evaluate, which is beneficial for the evaluation process, because it rates terms according to its frequency in the corpus.

3.1 Classification Module

This module is an application with an interface web that allows the teacher to define the questions/answers, its type and scorings for the whole exam. When the teacher inserts a question it begins with the definition of its type according to the three types previously described: enumeration, specific knowledge and essay. Next, the teacher defines the question and its answer. If it is an answer to an enumeration question, the teacher defines each component of the question and the relative scoring for each component on the answer. If it is a specific knowledge answer, the teacher defines the mandatory terms or words that must occur in the SAs without any changes, or variations, which may include proper nouns or not. These mandatory terms are marked by the system with a special tag *<mand>*. Finally each mandatory term has its scoring defined by the teacher, while for an essay question the teacher only determines whether the answer is affirmative or not, or if it is free.

This is not an additional process that the teacher must do in order to this assessment system operates. Indeed, it takes part of any teacher's exam preparation. While preparing an exam, all teachers must perform its correction and assign the adequate scoring to the questions, or part of the questions, according to the importance/difficulty of the issues. This criteria definition is very important because it will allow our system to apply exactly the same assessment criteria as the teacher.

3.2 Pre-processing Module

To increase the performance of the pre-processing module, a dynamic structure was created. So at the beginning of this process all the information concerning an exam, such as, questions/answers teachers and student's answers are carried to this dynamic structure, which turns the information always accessible and avoids a constant reading from disk.

The pre-processing of answers is performed according to the type of question. For enumeration questions these are decomposed in the various sub answers. On the specific questions, we look for the terms that were marked by the teacher as "mandatory" and these ones will not be processed by the pre-processing module.

The pre-processing module is composed of a set of steps which run sequentially in an effort to reduce each sentence to its canonical form. All the steps, except the last one, are applied to both the RA and SAs. The last step, that assigns a list of synonyms to every pair *(word,tag)* is only applied to RAs, since its goal is to make paraphrases of RAs. The order in which each task is executed has a great influence on the final form of sentences. So, each step was implemented in an independent way so that a quick and easy reorganization and combination of the different tasks can be performed and compared. The combination that best fulfils the objectives proposed is the one that loses less information from the answers. For example, if the stop-words are removed, before the spelling checker execution, the words with errors won't be considered in the stop-word process and the sentence will not be placed in its correct canonical form. After some experiences we have reached the following pre-processing task order.

3.2.1 Removal of Punctuation

This first task removes all special characters. Special characters are those which are not integrated in a word, such as punctuation. The accentuation of the words (in this context "accentuation" refers to a mark or symbol used to write words in certain languages to indicate the vocal quality to be given to a particular letter) was maintained so that they would not be regarded as misspellings.

3.2.2 Correction of Spelling Errors

The corrector used for the misspelling verification was Jspell (Almeida and Simões, 2007). All the SAs are verified, and if they have misspelling errors are corrected. Besides detecting misspellings, the Jspell tool also suggests the solution. Thus, the words with misspellings are replaced by the correct ones, so that answers can be compared. In addition to this, errors are counted for further evaluation of the answer.

3.2.3 Removal of Stop-words

A stop-word is a word that is considered irrelevant because it doesn´t affect the semantic meaning of the sentence. The goal of this task is the removal of all stop-words that aren't meaningful for the quotation of the answer. For example, grammatically speaking, words like Yes or No are considered stop-words but they can alter the meaning of an answer, thus changing the scoring given to it, consequently these words aren't removed in this step.

3.2.4 Stemming

In this stage, individual words are reduced to their canonical form or stem. The canonical form of a word is the base or lemma of that word. Stemming will simplify the process of matching words of SAs to RAs and will also help in the process of locating synonyms which will be made in the next steps.

3.2.5 Text Tagging

This task assigns to the words not yet marked as "mandatory" their part of speech tag. This is also performed by using Jspell. This categorization will enable us to compare words with the same part of speech tag.

3.2.6 Synonyms

The list of synonyms of a word depends of their part of speech tag. A word will have one list of synonyms associated with each part of speech tag. The thesaurus is critical to the success of this operation, so we adopted the synonyms of the OpenThesaurusPT project (OpenThesaurus, 2010) that provides synonyms for Portuguese words. It is used in the process of RA paraphrases generation developed.

All the synonyms concerning the words and their part of speech tag, belonging to the RA, are loaded from a text file to our dynamic structure. This way, we provide several paraphrases to the same RA, and the local search is significantly reduced in the moment of the comparison of answers.

3.2.7 Application Example

After the pre-processing tasks, the canonical form of a RA is a list of triples

$(i,word_i<pstag_{i1}>[syn_{i1}..syn_{in}]...<pstag_{in}>[syn_{n1}..syn_{nn}],...,$
$n,word_n,<pstag_{n1}>[syn_{11}..syn_{1n}]...<pstag_{nn}>[syn_{n1}..syn_{nn}])$

which contains the order in sentence of the i-th word, the word, its list of part of speech tags and the list of synonyms associated with each tag. Some word/tag may not have a list of synonyms. The canonical form of SA is simpler than RA, because their words don´t have a list of synonyms. So the student canonical form is

$(i,word_i,<pstag_{i1}>,<pstag_{in}>,..., n, word_n, <pstag_{n1}>, <pstag_{nn}>)$

The following example shows the effects of these operations in a question/answer picked from our database.

Question: *"Caracteriza o Urbanismo Pombalino"*
Question translation: *"Describe the Pombalino urbanism"*

Original teacher answer: *"A Lisboa Pombalina tinha ruas largas e perpendiculares, sistema de esgotos, as casas possuíam uma estrutura em gaiola. O terreiro do Paço passou a chamar-se Praça do Comércio."*
Teacher answer translation: *"The Pombalino Lisbon had wide and perpendicular streets, sewerage system, the houses had a cage structure. The yard of the palace change its name to Praça do Comércio."*

This is an example of an enumeration answer. When the teacher defines the answer he separates it in its subcomponents, so the answer is divided in the following four sub answers:

- "A Lisboa pombalina tinha ruas largas e perpendiculares" (30%)
- "sistema de esgotos"(30%)

- "as casas possuíam uma estrutura em gaiola"(40%)
- "O terreiro do Paço passou a chamar-se <Praça do Comércio<mand>>"(10%)

And each subcomponent has its own relative scoring.

Step 1 - Removal of Punctuation
"A Lisboa pombalina tinha ruas largas e perpendiculares"
"sistema de esgotos"
"as casas possuíam uma estrutura em gaiola"
"O terreiro do Paço passou a chamar-se <Praça do Comércio<mand>>"

In this step, in a RA enumeration is not applicable any action, because in the previous module the answer has been divided in its subcomponents. When processing SAs, this step will look for punctuation signals that will separate the SA in its subcomponents.

Step 2 - Correction of Spelling Errors
"A Lisboa pombalina tinha ruas largas e perpendiculares"
"sistema de esgotos"
"as casas possuíam uma estrutura em gaiola"
*"O terreiro do Paço passou a chamar-se <Praça do Comércio<**mand**>>"*

Step 3 - Removal of Stop-Words
"Lisboa pombalina ruas largas perpendiculares"
"sistema esgotos"
"casas possuíam estrutura gaiola"
"terreiro Paço passou chamar <Praça do Comércio<mand>>"

Step 4 – Stemming
"Lisboa pombalino rua larga perpendicular"
" sistema esgoto"
"casa possuir estrutura gaiola"
"terreiro Paço passar chamar <Praça do Comércio<mand>>"

Step 5 – Text Tagging
*(1,Lisboa,<**np**>) (2,pombalino,<**adj**>)*
*(3,rua,<**nc**>,<**verb**>) (4,larga,<**nc**>,<**adj**>,<**v**>)*
*(5,perpendicular,<**a_nc**>) (6,sistema,<**nc**>)*
*(7,esgoto,<**nc**>,<**v**>)*
*(8,casa,<**nc**>,<**v**>) (9,possuir,<**v**>) (10, estrutura,<**nc**>,<**v**>) (11, gaiola,<**nc**>)*
*(12,terreiro,<**a_nc**>) (13,Paço,<**nc**>)*
*(14,passar,<**v**>) (15,chamar,<**v**>) (16,Praça Comércio,<**mand**>)*

Step 6 – Synonyms
(*1,lisboa,*<**np**>**)** (*2,pombalino,*<**adj**>)
(*3,rua*,<**nc**>,<**verb**>) (*4,larga*,<**nc**>,<**adj**>[amplo espaçoso extenso grande vasto],<**v**>[abandonar ceder deixar desistir]) (*5,perpendicular*,<**a_nc**>**)**
(*6,sistema*,<**nc**>[arrumação maneira método ordem processo]**)** (*7,esgoto*,<**nc**>,<**v**>[ensecar esvaziar exaurir haurir])

(*8,casa*,<**nc**>[lar mansão morada moradia vivenda],<**v**> [agregar associar reunir])
(*9,possuir*,<**v**>[haver ter])
(*10,estrutura*,<**nc**>[arcaboiço arcabouço armação carcaça esqueleto],<**v**>**)** (*11, gaiola*,<**nc**>**)**

(*12,terreiro*,<**a_nc**>[eira eirado terraço terrado praça rossio]) (*13,Paço*,<**nc**>) (*14,passar*,<**v**>[atravessar cruzar galgar transpor correr decorrer percorrer])
(*15,chamar*,<**v**>[invocar]) (*16,Praça* Comércio, <**mand**>)

The tags used are <**np**> for proper noun, <**adj**> for adjective, <**nc**> for common noun, <**v**> for verb, <**a_nc**> for adjective or proper noun, and <**mand**> for mandatory term marked by the teacher. As shown in the example, some words have more than one tag, because the tagger tags each word regardless the context, and each word/tag may have its own list of synonyms. In addition to this, the words are not weighted, because this operation will be performed next, in the evaluation module.

3.3 Evaluation Module

After the pre-processing module, the evaluation of students' responses, based on correct teacher's answers, takes place. Our first approach was the application of Vector Space Model (VSM) technique (Salton et al., 1975). VSM measures how important a word is to a document (answer) in a collection or corpus (the exam). VSM is carried out without any pre-processing task, because this approach does not assign any value to words that appear in all answers. VSM calculates the similarity between texts representing them through vectors. The weight of each word is obtained from the $tf \times idf$ formula:

$$Score_{answ} - \frac{Tot(spellerror, answ)}{Tot(words, answ)} \times \alpha \quad (1)$$

Where, tf_i is the number of times the term appears in the answer, D is the total number of answers and df_i is the number of answers in which the term appears. This method assigns high weights to terms that appear frequently in a small number of answers in the whole exam. Once the term weights are determined, all teacher and students answers are represented by vectors, and the similarity between them is calculated using the the cosine measure. This measure determines the angle between the document vectors when they are represented in a V-

dimensional Euclidean space, where V is the vocabulary size. Given two vectors answers (SA and RA) the similarity is determined by the Eucledian dot product:

$$Sim(RA, SA) = \frac{w_{RA}.w_{SA}}{|VectorSize_{RA}| \times |VectorSize_{SA}|} \quad (2)$$

Where the size of the vectors is based on the weight of its words:

$$VectorSize = \sqrt{(w_1^2 + w_2^2 + ... + w_n^2)} \quad (3)$$

And the vectors product is the cross product between the SA vector and the RA vector.

$$w_{SA}.w_{RA} = \sum_{i=1}^{n} w_{RA,i} \times w_{SA,i} \quad (4)$$

where $w_{RA,i}$ is the weight of term *i* in the RA, and is defined in a similar way as $w_{SA,i}$ (that is, $tf_{RA,i} \times idf_i$).

The denominator in equation 2, called the normalization factor, discards the effect of document lengths on document scores. One can argue whether this is reasonable or not. In fact, when document lengths vary greatly, it makes sense to take them into account. However, in our case, the lengths of the answers don't vary greatly because we are dealing with short answers.

Nevertheless, the results obtained with VSM differed substantially from those given by the teacher, so we had to make some adjustments to adapt to our context.

In the specific case of SAs, a word that appears less frequently may be simply wrong and therefore should not be valued. In order to get around this, we first perform all the pre-processing tasks, removal of punctuation, correction of spelling errors, removal of stop-words, and stemming. Then, the vector space model penalizes answers with different words independently of their meaning. In fact, a student may write answers using different syntaxes, keywords and word counts, but the meaning could be the same as the RA. If there is only one RA for each question, the assessment will certainly be very penalizing. So SAs should be compared with various RAs. In order to do this we compare each SA with various paraphrases of the RA created in the pre-processing module.

3.3.1 Matching of Words

At this moment the RA and SAs are in its canonical form, so the words considered in each answer are meaningful and their frequencies in the answers are low, not only because of the pre-processing tasks applied, but also because we are dealing with short text answers. The canonical form of the SA is compared with the canonical form of the RA used to look for similar words. Each word in the SA is searched in the RA. If the word doesn't exist in the RA we search for its tag in RA. After this if tag is found, the word of SA is searched in the list of synonyms of the word associated with the tag in the RA. In other words, the various SA components of sentences identified by its corresponding part of speech tag, whose word doesn't match, will be compared using the list of synonyms of the corresponding tag in RA in an effort to better match the SAs. If the word exists in the list of synonyms, it is replaced, otherwise it is not. The word-matching algorithm is presented next.

```
// Word-Matching Algorithm

Input: canonical_RA, canonical_SA

//for each words in canonical SA
for s = 1,...,m
//if words doesn't exists in RA
  if !search_word(words,canonic_RA)
   //for each tag words
     for j = 1,...,n
     //if tagsJ exists in RA
       lstsyn=search_tag(tagsJ,canonic_RA)
     //if words is synonym of wordR
      if search_synon(words,lstsyn,synon)
        replace(words,synon)
```

After these changes, we apply the formulas of the VSM algorithm to calculate the similarity between the RA and SA. The training set to assign the weights to words is composed by teacher and students exams. From the similarity value obtained the answer score is calculated.

$$Score_{answer} = AnsSimil \times QuestMaxSc\,ore \quad (5)$$

The score calculated $Score_{answer}$ is combined with the number of spelling errors to calculate the final answer score. Each spelling error has a fix value α defined by the teacher and the penalization is based on the quotient of number of wrong versus total words in sentence. The following formula is used to calculate the final score.

$$Score_{answ} - \frac{Tot(spellingerror, answ)}{Tot(words, answ)} \times \alpha \quad (6)$$

3.3.2 Evaluation and Analysis

We have tested our system with a History exam that has sixteen questions. These questions were marked up to a hundred points. The three different types of questions, which include, enumeration, specific

knowledge and essay, were designed so that the performance of the system can be satisfactorily evaluated with all-types of short-answer questions. Table 1 describes the questions exam in terms of RA length, which gives the number of words in the RA, the average length of SAs, which presents the average number of words in all SAs for that question, as well has, the difference between these last two columns and the maximum answer score.

Table 1: Length of RA and SAs and score answers.

Question	RA length (word)	Avg. SAs length (word)	Diff. between RA length SA length	Max. Answer score
1.1	3	3,81	-0,81	6
1.2	7	11,29	-4,29	7
1.3	11	10,62	0,38	8
1.4	21	16,62	4,38	7
2.1	13	13,48	-0,48	7
4.1	1	1	0	7
4.2	5	1,52	3,48	6
5.1	4	3,48	0,52	7
5.2	13	12,86	0,14	7
6.1	9	1,38	7,62	5
7.1	30	7,76	22,24	6
7.2	22	19,86	2,14	6
7.3	13	6,34	6,66	7

As can be seen by the table, generally SAs are smaller than RA.

Table 2: Summary of assessment results by question.

Quest.	Avg. Teacher scores	Avg. System scores	Scores Diff.	Adjac. Agreem	Avg. Teach-Syst Agreem
1.1	5,35	5,53	-0,18	0,03	0,59
1.2	4,88	3,88	1,00	0,14	0,35
1.3	2,18	1,47	0,71	0,09	0,41
1.4	3,82	2,94	0,88	0,13	0,06
2.1	6	5,94	0,06	0,01	0,06
4.1	2,24	3,18	-0,94	0,13	0,24
4.2	4,59	4,59	0,00	0,00	0,88
5.1	4,47	3,71	0,76	0,11	0,12
5.2	3,53	4,47	-0,94	0,13	0,35
6.1	0,82	2,12	-1,30	0,26	0,41
7.1	1,41	0,53	0,88	0,15	0,71
7.2	3,94	2,24	**1,70**	0,28	0,29
7.3	5,47	4,29	1,18	0,17	0,12

A History teacher took part in the evaluation process. All SAs were assessed by the teacher to determine the level of agreement between teacher's evaluation and our system. The teacher was asked to assign scores to answers using the scorings of the answer, that is, if the answer has a maximum score of 7 points, the teacher may score the question between 0 and 7. Table 2 summarizes the results collected in our experiments.

Average teacher scores and average system scores is the mean of scores given by the teacher and the system by answer, the third column presents the difference between these last two scores. The adjacent agreement is calculated by the following formula:

$$Adjac.\,Agr = \frac{|\,avg.\,teach\ scores - avg.\,syst\ scores\,|}{max.\,answer\ score} \quad (7)$$

The average teacher-system agreement rate for a question is the number of times where teacher and system agree, divided by the total number of SAs.

As illustrated in Table 2, the maximum difference between system scores and the teacher's scores is 1.70 points in a question scored in 6 points, which gives an agreement of 0.28. The average teacher-system agreement rate measure tends to decrease as the complexity of short-answer question type increases. Another important pattern that may be deduced from the data is that, as the average answer length increases, the average teacher-system agreement rate decreases.

In general, in the cases where system and teacher's scores don't match, the teacher usually assigns a higher score than the system. This is due to the fact that the SA is too abstract, or the SA is much smaller than the RA. In these cases, our system's score is lower than expected, because its precision is based on the matching of words, and some SAs are in a format where no matches can be found in the RA, even though teacher evaluation indicates that the student understands the learning concepts. Therefore, experts assess such answers with a higher score, thereby increasing the differences between system and teacher assessments.

To evaluate the results obtained in this study, the Pearson correlation was used (Noorbehbahani and Kardan, 2011). Pearson correlation measures the standard correlation, that is, how much the average teacher scores (X) are associated with the system scores (Y). Equation 8 is applied to calculate this correlation.

$$Correlation(X,Y) = \frac{Covariance\,(X,Y)}{StandDev\,(X) \times StandDev\,(Y)} \quad (8)$$

We have obtained a promising correlation result of 0.78 which shows a reasonable correlation between our system assessment and the teacher's assessment, but this can be improved.

Another analysis of the errors of our system show that the errors fall into two categories: false

negatives (FN) and false positives (FP). A FN occurs when an answer gets lower score than it deserves. A FP occurs when the system assigns more marks to an answer than it deserves. In case of our system's evaluation, the number of FN was much higher than the number of FP. 35% of all errors was FN while only 25% were FP. The relative ratio of FN can be explained based on the difficult of anticipating all the possible paraphrases for an answer. If some correct possibility is missed, then SA will lead to FN. The most relevant scenario that accounts for systems' FP refers to students that don't know the answer to the question, but are fortunate enough to write some words that match with the RA.

4 CONCLUSIONS AND FUTURE WORK

In this study, we proposed a system for free text answer assessment. In the proposed approach, each question has several RAs that are automatically developed by our system, based on the word and its part of speech tag. Answers submitted by students can be compared with several RAs. After the word matching algorithm that searches for similar words of SAs in RA is applied, the similarity score is calculated based on weights of common-words between the SA and the RA. The system was tested in the context of History exams, and some evaluation results were presented. Despite evaluation results showed a good correlation (0.78) between average teacher scores and system scores, we think it is possible to improve system results. We intend to do that by detecting combined words, (occurrences of n-grams), and using as RAs, SAs previously marked by the teacher with the maximum score. This way the system will be improved in a continuous manner, when more and more training examples - RAs are provided, which will permit a more accurate assessment. Also teachers need to obtain feedback on their teaching performance, and students need feedback on their learning performance, these goals will be achieved through the development of the feedback module that we intend to develop next.

REFERENCES

Mason, O., Grove-Stephenson, I., 2002. *Automated free text marking with paperless school. In Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University, UK.*

Thomas P., Haley D., Roeck A., Petre M., 2004. *E–Assessment using Latent Semantic Analysis in the Computer Science Domain: A Pilot Study. In Proceedings of the Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning, pp. 38-44. Association for Computational Linguistics.*

Valenti S., Neri F., Cucchiarelli A., 2003. *An Overview of Current Research on Automated Essay Grading. Journal of Information Technology Education.*

Perez-Marin D., Pascual-Nieto I., Rodriguez P., 2009. *Computer-assisted assessment of free-text answers. The Knowledge Engineering Review, 24(4), pp. 353-374.*

Page, E. B., 1994. *New computer grading of student prose, using modern concepts and software. Journal of Experimental Education, 62(2), pp. 127-142.*

Jerrams-Smith J., Soh V., Callear D., 2001. *Bridging gaps in computerized assessment of texts. In Proceedings of the International Conference on Advanced Learning Technologies, pp.139-140.*

Whittington, D., Hunt, H., 1999. *Approaches to the computerized assessment of free text responses. In Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University, UK.*

Burstein J., Leacock C., Swartz R., 2001. *Automated evaluation of essay and short answers. In Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughborough University, UK.*

Rudner L. M., Liang T., 2002. *Automated essay scoring using Bayes' Theorem. The Journal of Technology, Learning and Assessment, 1(2), pp. 3-21.*

Mitchell T., Russel T., Broomhead P., Aldridge N., 2002. *Towards robust computerized marking of free-text responses. In Proceedings of the Sixth International Computer Assisted Assessment Conference, Loughboroug University, UK.*

Smith, J., 1998. *The book*, The publishing company. London, 2nd edition.

Almeida, J. J., Simões A., 2007. *Jspellando nas morfolimpíadas: Sobre a participação do Jspell nas morfolimpíadas. In Avaliação conjunta: um novo paradigma no processamento computacional da língua portuguesa. IST Press.*

OpenThesaurus, 2010 *http://openthesaurus.caixamagica.pt/, last access, Feb 2011*

Salton G., Wong A., Yang C. S., 1975. *A Vector Space Model for Automatic Indexing. Communications of the ACM, vol. 18, nr. 11, pp. 613–620.*

Noorbehbahani F., Kardan A. A. 2011. *The Automatic assessement of free text answers using a modified BLEU algorithm. Computers & Education 56, pp.337-345.*