ADAPTING THE AUTOMATIC ASSESSMENT OF FREE-TEXT ANSWERS TO THE STUDENTS

Diana Pérez and Enrique Alfonseca

Adapting the Automatic Assessment of Free-Text Answers to the Students

Diana Pérez and Enrique Alfonseca
Computer Science Department, EPS, UAM
Calle Francisco Tomás y Valiente, 11
28049, Madrid, Spain
Diana.Perez,Enrique.Alfonseca@uam.es

Abstract

In this paper, we present the first approach in the field of Computer Assisted Assessment (CAA) of students' free-text answers to model the student profiles. This approach has been implemented in a new version of Atenea, a system able to automatically assess students' short answers. The system has been improved so that it is now able to take into account the students' preferences and personal features to adapt not only the assessment process but also to personalize the appearance of the interface. In particular, it is now able to accept students' answers written in Spanish or in English indistinctly, by means of Machine Translation. Moreover, we have observed that Atenea's performance does not decrease drastically when combined with automatic translation, provided that the translation does not reduce greatly the variability in the vocabulary.

1. Introduction

Most of the existing distance education courses rely on objective testing exercises, such as Multiple Choice Question (MCQ) or fill-in-the-blank items. However, in the opinion of many researchers (Whittington & Hunt, 1999), in order to fully assess the students' learning progress, these should be complemented with open-ended questions. Therefore, the field called Computer-Assisted Assessment (CAA) of open-ended questions has been created to study how the computer can be used to automatically assess students' free-text answers. This field has received a great deal of attention. Nowadays there are more than fifteen different systems that face it (Valenti et al., 2003).

Concerning user modelling, up to date, there have been several attempts to make adaptive CAA systems: by adapting the problem selection (Mitrovic and Martin, 2004), the navigation through the problems (Gutiérrez et al., 2004; Sosnovsky, 2004) or the feedback provided to the students (Lutticke, 2004). Besides, Computer Adaptive Testing systems such as SIETTE (Guzmán and Conejo, 2002) are able to modify the order in which the test items are presented according to the students' performance during the test.

In previous work, we have developed a non-adaptive CAA system for evaluation of free-text answers called Atenea, whose main aim is to provide students with more practical training before their exams, and to help teachers as a double-checker of their scores (Alfonseca and Pérez, 2004).

This system has been extended with capabilities for gathering the students' profiles. They can be used to adapt the assessment process (for example, by being more or less strict with novice or advanced students in a topic) and to personalize the interface (for example, by being more or less childish).

In particular, we focus in this paper on the adaptation to the students' language, so they can write their answer in the language that they choose (Spanish or English). Using automatic Machine Translation techniques, the texts are translated to the language in which the teachers' references are written. We hypothesise that the automatic translation does not decrease the performance and it may even improve the results, provided that the variability of the vocabulary in the student answers is not greatly affected by the automatic translation. In order to test it, we have performed several experiments that confirm our hypothesis.

The paper is structured as follows: in Section 2 we explain the main features of the non-adaptive version of Atenea; next, in Section 3, we describe the adaptation to the user. Section 4 focuses on the Atenea's multilingual capacity, and, finally, Section 5 ends with final remarks and perspectives of future work.

2. Atenea

Atenea (Alfonseca and Pérez, 2004) is a Computer-Assisted Assessment system for automatically scoring students' short answers (see Figure 1). It is a web-based application, but it can also be used locally.

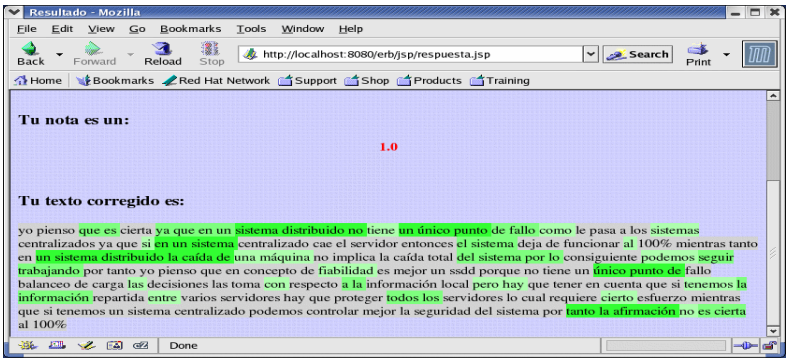

Figure 1. Snapshot of Atenea

The system has been tested with English and Spanish texts, but the procedures used could be easily ported to other languages. It is based on the combination of several Natural Language Processing (NLP) modules whose core idea is that a student's answer is better when it is closer to the answers written by the teachers (the reference answers). There should be at least

three different references per question to capture several possible paraphrasings.

The basic procedure that Atenea follows is to randomly choose a question from its database and to retrieve its references. The student has to write the answer to that question. Then, depending on Atenea's configuration, several NLP techniques can be performed: stemming, removal of closed-class words, and naïve Word Sense Disambiguation (Alfonseca and Pérez, 2004). After they have been processed, the student's answer and references enter the statistical module called ERB (Alfonseca and Pérez, 2004) that compares them using a modified version of the n-gram co-occurrence scoring algorithm called Bleu (Papineni et al., 2001). The output of this module is double: the student receives a final score, and the answer, in which the words are marked up so that the regions which are more similar to the references appear annotated with a colour code (see Figure 1).

For the evaluation of Atenea we have created a corpus of Spanish students' answers from real exams in our home university, described in Table 1. The number of students ranged from 14 to 295 depending on the question.

SET	NC	MC	NR	MR	Type	RS
1	79	51	3	42	Def.	[0,0.5]
2	96	44	4	30	Def.	[0,0.5]
3	11	81	4	64	Def.	[0,1]
4	143	48	7	27	Def.	[0,1]
5	295	56	8	55	A/D	[0,0.5]
6	117	127	5	71	Y/N	[0,1.5]
7	117	166	3	186	A/D	[0,1]
8	14	118	3	108	Y/N	[0,1]
9	14	116	3	105	Def.	[0,1]

Table 1. Evaluation answer sets. Columns indicate: number of candidate student answers (NC), their mean length (MC), number of references (NR), their mean length (MR), question type (Def., definitions; A/D, advantages and disadvantages; Y/N, yes-no with justification), and range of scores (RS)

3. Adaptation to the user

It can be seen that the aforementioned version of Atenea does not take into account any information about the students. We believe that the assessment must be adapted to their particular features.

Concerning the personalisation of the interface, it has been included in order to make the assessment process more engaging. The students have full control over the interface, although default values are set for those who do not want to have their environment personalized. This is also important not only from an aesthetic point of view, but also functionally, because it can help students with some disability.

Concerning the adaptation of the assessment to the user profiles, it can also result in a better scoring process, for instance, by providing more challenging questions as the student passes the easier ones.

The features of the students that have been modelled in Atenea have been chosen because of their relevancy (Barrutieta et al., 2003) and they are:

- **Language:** The statement of the question should be presented in the language of the student. Equally, the answer should be evaluated with the right NLP tools and resources for that language.
- **Experience:** Advanced students would not be correctly evaluated with the questions that they already have passed. Thus, more complex questions will be chosen for them.
- **Stop condition:** The student is given the option of choosing the number of questions to answer in one session, and specifying the amount of time to dedicate to the session. The system would present questions as long as none of the two conditions is fulfilled.
- **Feedback:** According to the student's aim, the system gives him or her the possibility of receiving just the score (summative assessment). On the other hand, together with the score, as additional feedback, the answer can be returned annotated with a colour code, as mentioned in Section 2 (formative assessment).
- **Interface:** The appearance of the interface can change according to the students' preferences.
- **Age:** The statement of the question should be easier for children than for adults. Besides, the references need to be different since children's language and vocabulary are both expected to be simpler.

For each of these characteristics, the teacher can specify the values they can take, e.g. *English* and *Spanish* for the language, *Novice* and *Advanced* for the experience. All of them configure several adaptive paths that the new version of Atenea follows for each student. These are all implemented as stereotypes, as there is a finite set of values for each characteristic. Table 2 shows a possible set of features and values that may have been set by a teacher.

Language	Spanish	English
Experience	Novice	Advanced
Time	Chronometed	Unlimited
Feedback	Basic	Detailed
Interface	Default	Personalised
Age	Child	Adult
History of use	Questions already answered and scores got	

Table 2. Features of the students which are modelled by Atenea, and some possible example values

Apart from these stereotypes, Atenea records the mark obtained by the students for each question. This will be used to decide which will be the next

question asked. If the mark obtained by the student in a question exceeds half of the maximum score for that question, it will be considered correct, and the student will not be asked that question again.

To exemplify the adaptation process, let us suppose that Antonio is a Spanish 34-year-old engineer who is attending a course to deepen in the study of Operating Systems, and he is using Atenea to get more practical training. When he logs into the system for the first time, Antonio is asked to fill several forms so that Atenea can store his information in its database. Next, Atenea chooses a question in Spanish technical formal style and presents it to him. It is important to highlight how the system is going to be stricter with Antonio by choosing the references written in formal technical language; the references written in a simpler language would be used with novice students. Furthermore, only the questions that were not answered correctly will be repeated in the future.

4. Adaptation to the user's language

Traditionally, it was necessary to ask the teachers to write each reference answer several times (one per language) so as to assess students' answers in different languages. However, we have observed how teachers were troubled by this task. This, combined with the fact that the quality of the references is crucial if we want to achieve a good assessment procedure, led us to look for another way to approach the adaptation to the student's preferred language.

Therefore, we tried by using an automatic Machine Translation (MT) system to translate the student's answer to the language in which the teacher's references are written. By doing so, the teacher is only asked to write the references in his or her mother tongue while the students can continue writing in their preferred languages. The MT engine used is Altavista Babelfish (available at http://babelfish.altavista.com/).

On the other hand, it was uncertain whether using translated versions of the students' answers decreases the accuracy of the assessment. In a previous experiment, we asked the teachers to write by hand reference answers both in English and Spanish, for a set of questions. Next, we have tried the following two configurations:

- Firstly, we evaluated answers written by students in Spanish, by comparing them with the Spanish references. This would be traditional use of Atenea.
- Secondly, we translated each answer automatically into English, using Babelfish, and we evaluated it by comparing it with the English references. This shows the use of Atenea combined with an MT engine, for the situation in which a student is writing in a different language than that of the reference answers.

In both cases, the performance of Atenea is calculated in the following way: we asked a group of teachers to manually assess all students' answers, and

then we calculated the Pearson correlation between these manual scores and Atenea's scores for three different configurations, which are the following:

- ERB, in which the n-grams from the student's answer and the references are compared to each other.
- CC, in which the closed-class words (prepositions, determiners, conjunctions, etc.) are first removed before the use of ERB.
- WSD+CC, in which, after removing the closed-class words, a naïve Word Sense Disambiguation procedure is executed on all the nouns, verbs, adjectives and adverbs, before ERB is applied to the text.

The results for all these configurations are presented in Table 3.

	ERB			**CC**			**WSD+CC**		
SET/L	Spanish	English	DifC	Spanish	English	DifC	Spanish	English	DifC
1	0.5244	**0.5330**	-0.01	0.5754	0.5479	0.03	0.4655	**0.4841**	-0.02
2	0.3210	0.1660	0.16	0.3234	0.2892	0.03	0.2844	**0.3264**	-0.04
3	0.7490	0.2594	0.49	0.7774	0.5760	0.2	0.7988	0.7125	0.09
4	0.6608	0.5937	0.07	0.6811	0.6066	0.07	0.6933	**0.7655**	-0.07
5	0.1979	**0.2449**	-0.05	0.2437	**0.3213**	-0.08	0.3040	**0.3282**	-0.02
6	0.4027	0.3649	0.04	0.4159	0.3450	0.07	0.3838	0.3586	0.03
7	0.3970	**0.4583**	-0.06	0.4326	**0.4515**	-0.02	0.5261	0.4699	0.06
8	0.7495	**0.8691**	-0.12	0.6942	**0.7026**	-0.01	0.7716	0.6803	0.09
9	0.8113	**0.8171**	-0.01	0.4832	**0.6759**	-0.19	0.5053	**0.6826**	-0.18
MEAN	0.5348	0.4785	0.06	0.5141	0.5018	0.01	0.5259	**0.5342**	-0.01

Table 3. Correlation between the teacher's and Atenea's scores in Spanish and English for different Atenea's configurations (in bold when the correlation for the translated texts is higher than for the original texts)

As can be seen, the correlations decrease, but not dramatically, and they even improve for some datasets. Particularly, in the last configuration, WSD+CC, the average of the correlations with the automatic translation reaches 53%, which is equivalent to the best correlation obtained without translation (in the ERB configuration).

Our hypothesis is that this decrease is partly due to *the reduction of the vocabulary* that is usually produced in an automatic translation. This reduction is probably introducing noise in the evaluation process. If this were the case, we should be able to prove the following two results:

- There is actually a reduction of the vocabulary due to the automatic translation of the answers.
- This reduction is positively correlated to the decrease of Atenea's performance.

4.1. Vocabulary reduction

Table 3 shows the number of words found in the student answers collected for each dataset, in Spanish, and the number of words found in the English translations. It is clear that, for all datasets, there are less distinct words in the

English datasets than in the original Spanish answers. It can be seen that, indeed, the automatic translation has decreased the variability of the vocabulary.

SET	1	2	3	4	5	6	7	8	9	MEAN	All Sets
Spanish	818	716	332	919	1474	1847	1607	415	342	905	4558
English	674	631	284	781	1174	1541	1337	408	326	770.3	3419
DifV	0.18	0.12	0.14	0.15	0.2	0.17	0.17	0.02	0.05	0.15	0.25

Table 4. Number of different words in the Spanish and English datasets. Row DifV shows the percentage of vocabulary reduction due to the translation

The same result can also be achieved if we measure the percentage of words from the candidate texts that appear in the references, and vice versa. When this is done for the Spanish and the translated texts, we can see, in Table 5, that the percentages are higher for the English texts, as there is a higher number of repetitions of a more limited vocabulary.

SET	% cans. in refs.		% refs. in cans.	
	Spanish	English	Spanish	English
1	49.17	47,23	90	95,68
2	47.42	48,56	92,8	88,37
3	53.97	63,33	77,87	83,73
4	62.73	64,69	95,81	97,95
5	69,79	73,65	98,65	99,56
6	61,57	64,02	96,69	97,81
7	70,44	72,23	96,03	97,7
8	68,4	68,2	86,1	90,03
9	66,67	58,25	87,75	94,33
MEAN	61,13	62,24	91,3	93,91

Table 5. Percentage of overlapping between candidates (cans.) and references (refs.) in the Spanish and English datasets

4.2. Correlation between vocabulary reduction and Atenea's performance

Finally, we should check that the reduction of the vocabulary is correlated to Atenea's performance. In other words that a bigger reduction of the number of different words in the translated texts implies that Atenea's scores are more similar to the scores given by the teachers.

In order to prove this fact we have calculated: (a) the difference between the Spanish and English performance of Atenea, as shown in Table 3 in column DifC, (b) the percentage of vocabulary reduction, as shown in Table 4 in row DifV, and (c) the correlation between the paired values of (a) and (b). Table 6 shows the results of this experiment.

SET	DIFV.	DIFC.		
		ERB	CC	WSD+CC
1	0.18	-0.01	0.03	-0.02
2	0.12	0.16	0.03	-0.04
3	0.14	0.49	0.2	0.09
4	0.15	0.07	0.07	-0.07
5	0.20	-0.05	-0.08	-0.02
6	0.17	0.04	0.07	0.03
7	0.17	-0.06	-0.02	0.06
8	0.02	-0.12	-0.01	0.09
9	0.05	-0.01	-0.19	-0.18
MEAN	0.15	0.06	0.01	-0.01
CORR.		0.14	0.32	0.14

Table 6. Correlation between the percentage of reduction of the variability of the vocabulary and Atenea's performance for different configurations

As can be seen, in the three cases there is a positive correlation between the reduction in vocabulary (column DifV) and the decrease of Atenea's performance (column DifC). Please note that column DifC is positive when Atenea's performance has decreased after the translation, and negative otherwise.

5. Conclusions and future work

The field of Computer Assisted Assessment (CAA) of free-text answers can and should benefit from incorporating user modeling techniques that let adapt the assessment process and the systems' interfaces in order to achieve a fairer assessment and to provide a more enjoyable and engaging environment for the students who need or want more training.

However, until now, no system has been developed to apply adaptation to the assessment of open-ended questions. Thus, we have developed a new version of Atenea that provides different adaptive paths to the students according to their goal (formative or summative), level of experience, language, age, stop condition and interface's preferences.

Special interest has been given to the possibility of choosing the language in which Atenea is going to work (Spanish or English). As we had observed the unwillingness of the teachers to write references in several languages for each question, we have studied whether it is possible to integrate Atenea with a Machine Translation engine to incorporate multilinguality. In this way, it is possible that teachers only have to write the references in their mother tongue, while the students continue writing their answers in their preferred language. In this paper, we support the idea that this is feasible, and that there should not be a large loss in vocabulary variability due to the automatic translation if we want to keep the level of correlation.

The combination of techniques from the fields of User Modeling, Machine Translation and free-text CAA opens many interesting future lines of research.

Some of them are: to study further uses of the automatic translation to support the Atenea's evaluation of answers written by students, and to apply more strategies of adaptation for free-text CAA systems, such as learning styles (Paredes and Rodriguez, 2002), that could be regarded in this field as assessing styles.

References

Alfonseca, E. & Pérez, D. (2004). Automatic Assessment of Short Questions with a Bleu-inspired Algorithm and shallow NLP. In LNCS 3230, Springer-Verlag, *Advances in Natural Language Processing*, 25-35.

Alfonseca, E. & Carro, R.M. & Freire, M. & Ortigosa, A. & Pérez, D. & Rodríguez, P. (2004). Educational adaptive hypermedia meets computer assisted assessment. In Proceedings of the International *Workshop of Educational Adaptive Hypermedia* in the Adaptive Hypermedia Conference.

Barrutieta, G. & Abaitua, J. & Díaz, J. (2003). User modelling and content selection for multilingual document generation. In Proceedings of the *Eighth International Symposium of Social Communication*.

Gutiérrez, S. & Pardo, A. & Delgado, C. (2004). An Adaptive Tutoring System Based on Hierarchical Graphs. In LNCS 3137, Springer-Verlag, *Adaptive Hypermedia and Adaptive Web-Based Systems,* 401-404.

Guzmán, E. & Conejo, R. (2002). An adaptive assessment tool integrable into Internet-based learning systems. In Proceedings of the *Educational Technology: International Conference on TIC's in Education*, 1, 139-143.

Lutticke, R. (2004). Problem solving with Adaptive Feedback. In LNCS 3137, Springer-Verlag, Adaptive Hypermedia and Adaptive Web-Based Systems, 417-420.

Mitrovic, A. & Martin, B. (2004). Evaluating Adaptive Problem Selection. In LNCS 3137, Springer-Verlag, *Adaptive Hypermedia and Adaptive Web-Based Systems,* 185-194.

Papineni, K. & Roukos, S. & Ward, T. & Zhu, W. (2001). BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

Paredes, P. & Rodriguez, P. (2002). Considering Learning Styles in Adaptive Web-based Education. Proceedings of the Sixth Word MultiConference on Systemics, Cybernetics and Informatics, 481-485.

Sosnovsky, S. (2004). Adaptive Navigation for Self-assessment quizzes. In LNCS 3137, Springer-Verlag, Adaptive Hypermedia and Adaptive Web-Based Systems, 365-371.

Valenti, S. & Neri, F. & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. In Journal of Information Technology Education, 2, 319–330.

Whittington, D. & Hunt, H. (1999). Approaches to the computerized assessment of free text responses. In Danson, M. (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough, UK.