



[Revista signos](#)

versión On-line ISSN 0718-0934

Rev. signos v.38 n.59 Valparaíso 2005

<http://dx.doi.org/10.4067/S0718-09342005000300004>

Servicios Personalizados

Artículo

- Artículo en XML
- Referencias del artículo
- Como citar este artículo
- Traducción automática
- Enviar artículo por email

Indicadores

- Citado por SciELO

Links relacionados

- Similares en SciELO

- Permalink

Revista Signos 2005, 38(59), 325-343

ARTÍCULOS

About the effects of combining Latent Semantic Analysis with natural language processing techniques for free-text assessment*

Sobre los efectos de combinar Análisis Semántico Latente con otras técnicas de procesamiento de lenguaje natural para la evaluación de preguntas abiertas

Diana Pérez

Enrique Alfonseca

Pilar Rodríguez

Universidad Autónoma de Madrid,

España

Alfio Gliozzo

Carlo Strapparava

Bernardo Magnini

Instituto para la Investigación Científica y Tecnológica,

Italia

[Dirección para correspondencia](#)

ABSTRACT: This article presents the combination of Latent Semantic Analysis (LSA) with other natural language processing techniques (stemming, removal of closed-class words and word sense disambiguation) to improve the automatic assessment of students' free-text answers. The combinational schema has been tested in the experimental framework provided by the free-text Computer Assisted Assessment (CAA) system called Atenea (Alfonseca & Pérez, 2004). This system is able to ask randomly or according to the students' profile an open-ended question to the student and then, assign a score to it. The results prove that for all datasets, when the NLP techniques are combined with LSA, the Pearson correlation between the scores given by Atenea and the scores given by the teachers for the same dataset of questions improves. We believe that this is due to the complementarity between LSA, which works more at a shallow semantic level, and the rest of the NLP techniques used in Atenea, which are more focused on the lexical and syntactical levels.

Key Words: LSA, free-text assessment, computer assisted assessment, e-learning.

RESUMEN: Este artículo presenta la combinación de Análisis Semántico Latente (LSA) con otras técnicas de procesamiento del lenguaje natural (lematización, eliminación de palabras funcionales y desambiguación de sentidos) para mejorar la evaluación automática de respuestas en texto libre. El sistema de evaluación de respuestas en texto libre llamado Atenea (Alfonseca & Pérez, 2004) ha servido de marco experimental para probar el esquema combinacional. Atenea es un sistema capaz de realizar preguntas, escogidas aleatoriamente o bien conforme al perfil del estudiante, y asignarles una calificación numérica. Los resultados de los experimentos demuestran que para todos los conjuntos de datos en los que las técnicas de PLN se han combinado con LSA la correlación de Pearson entre las notas dadas por Atenea y las notas dadas por los profesores para el mismo conjunto de preguntas mejora. La causa puede encontrarse en la complementariedad entre LSA, que trabaja a un nivel semántico superficial, y el resto de las técnicas NLP usadas en Atenea, que están más centradas en los niveles léxico y sintáctico.

Palabras Clave: LSA, preguntas abiertas, evaluación asistida por ordenador, e-learning.

INTRODUCTION

Computer Assisted Assessment can be defined as the field that studies how to use effectively computers in the assessment of student learning. It is a very general definition, since CAA is a broad field that covers from systems as simple as HTML forms to sophisticated systems with a complex assessment and feedback process. In this work we focus on the assessment of open-ended questions because according to the general opinion of the field, it is also necessary to address this kind of assessment in order to fully assess the student learning process.

On the other hand, automatically assessing free-text answers is a difficult task. Although it has been studied since the 60s (Page, 1966), it has not been able to show real progress until the late 90s, when Natural Language Processing (NLP) techniques were applied to assess students' free-text answers for the first time.

Nowadays, there are more than fifteen different systems that have appeared to tackle this task (Valenti, Neri, & Cucchiarelli, 2003). Although they are based on different techniques, all of them share a common core idea: a student's answer should receive a higher score when it is closer to a reference or to a group of references written by an expert in the topic or collected from other sources such as textbooks or Internet. It is important to notice that students can write the same idea in hundreds of different ways. Due to this paraphrasing problem, the automatic scores highly depend on the quality of these references.

In particular, our approach relies on the combination of statistical NLP techniques (stemming, removal of closed-class words and Word Sense Disambiguation) and Latent Semantic Analysis (LSA). All of these modules are integrated into Atenea (Alfonseca & Pérez, 2004), the system that we have developed for evaluating students' free-text answers written both in Spanish and in English.

LSA is a statistical method for inferring meaning from a text. It has already been used for evaluating free-text students' answers with good results (Foltz, Laham and Landauer, 1999; Dessus, Lemaire & Vernier, 2000). In this work we study different possibilities for combining our CAA system Atenea with LSA. Atenea has, currently, several modules, including one that calculates n-gram statistics between the student answer and the references, and several NLP components. In previous work, we have shown that

combining Atenea's statistical module with LSA achieves better results than just using each of them independently (Pérez, Gliozzo, Strapparava, Alfonseca, Rodríguez & Magnini, 2005). In this paper, our motivation is to study how the use of the NLP components in the combination with LSA affects the results. Hence, we present different configurations of Atenea and how the results vary when the LSA module is used.

The paper is organised as follows: Section 1 presents the state-of-the-art of Computer Assisted Assessment of free-text students' answers and the use of LSA for this task; Section 2 describes the Atenea system without using LSA; Section 3 details the LSA configuration used; and, Section 4 focuses on its integration with other NLP techniques within Atenea. Finally, the article ends with the main conclusions and some lines of future work.

1. Review of the state-of-the-art

1.1. Free-text CAA

Table 1 presents several of the existing free-text CAA systems with the technique that underpins each of them and the results provided by their authors. It is important to highlight that the results are not fully comparable because these systems are using different corpora and metrics.

System	Reference	Technique	Results	Domain
PEG	(Page, 1966)	Statistical	Corr: .87	Non factual disciplines
E-rater	(Burstein, Leacock, & Swartz, 1998)	Statistical and NLP	Agr: .97	Non-native English writing
Larkey's system	(Larkey, 1998)	Text Categorization	EAgr: .55	Social and opinion
IEA Landauer, 1999)	(Foltz, Laham & Landauer, 1999)	LSA	Agr: .85 military	Psychology and
SEAR	(Christie, 1999)	Information Extraction	Corr: .45	History
Apex Assessor	(Dessus, Lemaire & Vernier, 2000)	LSA	Corr: .59	Sociology of education
IEMS	(Ming, Mikhailov & Kuan, 2000)	Indextron	Corr: .8	Non mathematical
IntelliMetric	(Vantage, 2000)	NLP	Agr: .98	k-12 and creative writing
ATM	(Callear, Jerrams-Smith & Soh, 2001)	Information Extraction	Not Available	Factual disciplines
C-rater	(Burstein, Leacock, & Swartz, 2001)	NLP	Agr: .83	Comprehension & algebra
Automark	(Mitchell, Russell, Broomhead, & Aldridge, 2002)	Information Extraction	Corr: .95	Science
BETSY	(Rudner & Liang, 2002)	Bayesian networks	CAcc: .77	Any text classification
PS-ME	(Mason & Grove-Stephenson, 2002)	NLP	Not Available	NCA or GCSE exam
CarmelTC	(Rosé, Roque & VanLehn, 2003)	Machine Learning and Bayes classif.	f-S: .85	Physic
Auto-marking	(Sukkariéh, Pulman & Raikes, 2003)	NLP & Pattern Match.	EAgr: .88	Biology
Atenea	(Alfonseca & Pérez, 2004)	Statistical, NLP and LSA	Corr: .56	Computer Science

Table 1. Overview of the main features of several free-text CAA systems¹.

Nonetheless, it can be seen that the values achieved show the great amount of international research that has been dedicated to the field in the last years, a fact that has even given rise to commercial systems. Institutions such as the Scottish Qualifications Authority (SQA), the Computer-Assisted Assessment (CAA) Centre in the U.K. or the U.S.A. Department of Education and Educational Testing Services (ETS) in the United States support the research in this field.

In fact, CAA has many possibilities of application. Some of them are: assigning problems to students, scoring them (summative assessment), returning feedback (formative assessment), and evaluating the assessment effectiveness (Blayney & Freeman, 2003). In particular, free-text CAA systems open the following capabilities: creation of links to theoretical explanations to clarify the weak points exposed by

the assessment of the students' free-text answers, support to teachers who cannot assist a large number of students, help to students showing them where their mistakes are, and instantaneous feedback.

1.2 Application of LSA for free-text CAA

Though LSA was not originally created for assessing free-text answers, its ability to give an idea of the semantic similarity between texts and to provide content-based feedback makes it particularly suited to e-assessment (Miller, 2003; Haley, Pete, Nuseibeh, Taylor & Lefrere, 2003).

Therefore, this technique has been used for several of the aforementioned free-text CAA systems. In particular, the Intelligent Essay Assessor (IEA, Foltz et al., 1999) and the Apex Assessor (Dessus et al., 2000) rely on LSA as the core technique of their system with good results.

2. Atenea

Atenea (Alfonseca & Pérez, 2004) is a CAA system for automatically scoring students short answers written in Spanish or in English. It relies on the combination of shallow NLP techniques and statistically based evaluation procedures. Figure 1 shows a snapshot of the interface, in Spanish, of the on-line version of Atenea.



Figure 1. Interface of Atenea.

After a student logs into the system, Atenea chooses a question according to his or her profile (Pérez & Alfonseca, 2005). This takes into account the previous performance of the student on other questions in a test set, and on other test sets. It is also possible for the teacher to define alternative reference texts depending on stereotypes (e.g. age, language, experience, etc.), which will also adapt the assessment process.

Once the question and the reference answers have been chosen by the system, it compares the student's answer with the ideal answers, to see how similar they are. The student receives a numerical score, and the answer marked up with a colour background indicating which portions have more coincidences with the reference texts. From this output, students can discern which ones are their weak points. Figure 2 shows an example of feedback page.

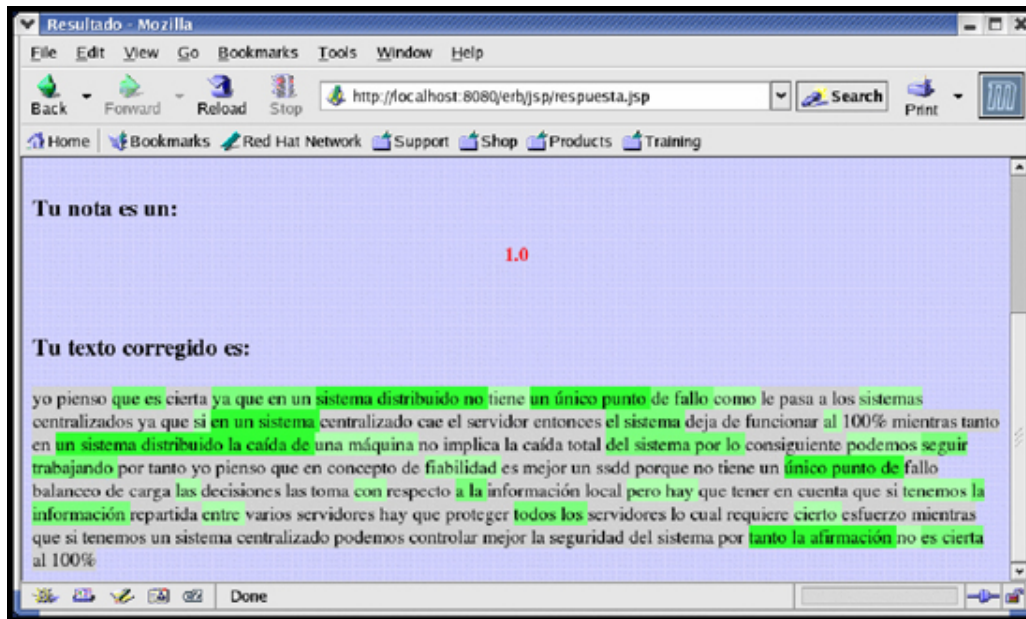


Figure 2. Feedback for student answer (the Spanish statement "Tu nota es un:" can be translated as "Your score is:" and "Tu texto corregido es:" as "Your processed answer is:").

A web-based wizard (see Figure 3) has been developed to facilitate the task of introducing new datasets of questions. It allows augmenting an existing dataset, or creating a new one; modifying existing questions or adding new ones; and, modifying existing question statements, maximum scores or references.

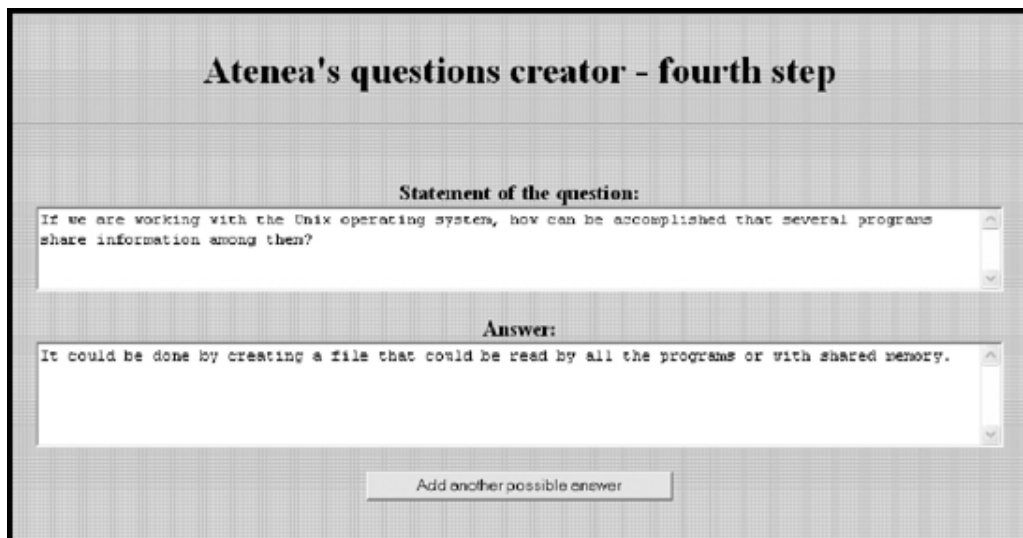


Figure 3. Atenea question authoring tool.

The system can also be retrained in such a way that the references for each question can be chosen from those written by the teacher, and from the best answers written by other students, in a way that maximises the accuracy of the assessment.

The internal architecture of Atenea is composed of a statistical module, called ERB, and several Natural Language Processing (NLP) modules based on the wraetlic tools (Alfonseca, 2003).

The statistical module of Atenea relies on the BiLingual Evaluation Understudy (BLEU) algorithm (Papineni, Roukos, Ward & Zhu, 2001). Basically, it looks for n-gram coincidences between the student's answer and the references. Its pseudocode is as follows:

1. For several values of n (usually from 1 to 3), calculate the Modified Unified Precision (MUP) of the student answer, i.e. the percentage of n-grams from the student's answer which appears in any of the references:

$$MUP(n) = \left(\sum_{i=0}^{num} \min(count_i, \max RC_i) \right) / ICand$$

where $count_i$ is the frequency of the i -th n-gram in the student's answer, $\max RC_i$ is the maximum

frequency of that n-gram in a reference text, and $|Cand|$ is the length of the student's answer.

2. Combine the MUPs obtained for each value of n as:

$$combMUP = \sum_{n=1}^3 \log MUP_n / 3$$

3. Apply a brevity factor to penalise the texts shorter than all the references:

$$BP = e^{(1 - |Ref|/|Cand|)} \text{ if } |Cand| < |Refs|$$

4. The final BLEU score is:

$$BLEU = BP \times e^{combMUP}$$

BLEU is basically a precision metric: it measures which n-grams in the candidate answer appear in the references. In the case of scoring student's answers, we both want the answer to be correct and complete. Therefore, we have modified this metric to calculate as well the percentage of the references that is covered by the student's answer. To do that, the Brevity Factor is substituted by a Modified Brevity Penalty (MBP) calculated in the following way: for each reference, calculate the percentage of n-grams that is covered by the candidate text, and next, we add up all those percentages. Figure 4 shows an example in which 5% of the first reference, 10% of the second one and 20% of the third one appears in the student's answer. Therefore, we can assume that 35% of a complete answer is covered by the candidate text. The results using this MBP clearly outperform those obtained using the original algorithm.

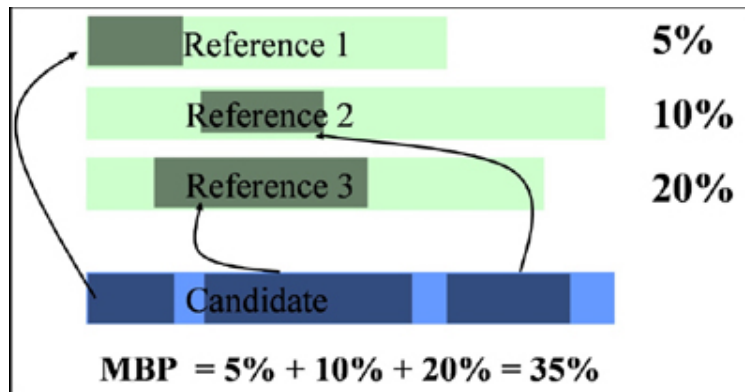


Figure 4. Procedure for calculating the MBP factor.

Concerning the NLP modules, it is possible to configure Atenea to indicate which modules to use. The possibilities that they add to the system are:

- *Stemming*: matching nouns and verbs inflected in different ways.
- *Removal of closed-class words*: ignoring functional words that are irrelevant to extract the students' answer meaning.
- *Word-Sense Disambiguation*: identifying the sense intended by both the teacher and the student and, thus looking if it is the same.

3. LSA

The simplest technique to estimate the similarity between the topics of two texts is the Vector Space Model (VSM) that is based on the calculation of the cosine between the vectors that represent each text. Let $\Gamma = \{t_1, t_2, \dots, t_n\}$ be a corpus, $V = \{w_1, w_2, \dots, w_k\}$ its vocabulary, \mathbf{T} the $k \times n$ term-by-document matrix representing Γ , such that $t_{i,j}$ is the frequency of word w_i into the text t_j . The VSM is a k -dimensional space \mathbf{R}^k , in which the text $t_j \in \Gamma$ is represented by means of the vector \vec{t}_j such that the i^{th} component of \vec{t}_j is $t_{i,j}$.

However, this approach does not deal well with lexical variability and ambiguity. For example, the two sentences "he is affected by AIDS" and "HIV is a virus" do not have any words in common and thus, using VSM their similarity is zero because they have orthogonal vectors, although the concepts they express are very closely related. On the other hand, the similarity between the two sentences "the laptop has been infected by a virus" and "HIV is a virus" would turn out very high, due to the ambiguity of the word **virus**. To overcome this problem, the notion of *Domain Model* (DM) was introduced. It is

composed by soft clusters of terms. Each cluster represents a semantic domain (Gliozzo, Magnini & Strapparava, 2004), i.e. a set of terms that often co-occur in texts having similar topics. A DM is represented by a $k \times k'$ rectangular matrix \mathbf{D} , containing the degree of association among terms and domains, as illustrated in Table 2.

Table 2. Example of Domain Matrix.

Word	Medicine	Computer Science
HIV	1	0
AIDS	1	0
Virus	0.5	0.5
laptop	0	1

Domain Models can be used to describe lexical ambiguity and variability. Lexical ambiguity is represented by associating one term to more than one domain, while variability is represented by associating different terms to the same domain. For example, the term **virus** is associated to both the domain Computer Science and the domain Medicine (ambiguity) while the domain Medicine is associated to both the terms **AIDS** and **HIV** (variability). More formally, let $\Delta = \{D_1, D_2, \dots, D_{k'}\}$ be a set of domains, such that $k' \ll k$. A Domain Model is fully defined by a $k \times k'$ domain matrix \mathbf{D} representing in each cell $d_{i,z}$ the domain relevance of term w_i with respect to the domain D_z . The domain matrix \mathbf{D} is used to define a function $\Delta: \mathbf{R}^k \rightarrow \mathbf{R}^{k'}$, that maps the vectors \vec{t}_j expressed into the classical VSM, into the vectors \vec{t}'_j in the domain VSM. Δ is defined by²:

$$\Delta(\vec{t}_j) = \vec{t}'_j (\mathbf{I}^{IDF} \mathbf{D}) = \vec{t}_j \quad (1)$$

where \mathbf{I}^{IDF} is a diagonal matrix such that $i^{IDF}_{i,i} = IDF(w_i)$, \vec{t}_j is represented as a row vector, and $IDF(w_i)$ is the Inverse Document Frequency of w_i .

Vectors in the domain VSM are called Domain Vectors. Domain Vectors for texts are estimated by exploiting Formula 1, while the Domain Vector \vec{w}'_i , corresponding to the word $w_i \in V$, is the i^{th} row of the domain matrix \mathbf{D} . To be a valid domain matrix such vectors should be normalized (i.e. $\langle \vec{w}'_i, \vec{w}'_i \rangle = 1$).

In the Domain VSM the similarity among Domain Vectors is estimated by taking into account second order relations among terms. For example the similarity of the two sentences "*He is affected by AIDS*" and "*HIV is a virus*" is very high, because the terms AIDS, HIV and virus are highly associated to the domain Medicine.

LSA (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990) is an unsupervised technique for estimating the similarity among texts & terms in a corpus. It has been proposed to induce Domain Models from corpora (Gliozzo, Giuliano & Strapparava, 2005a; Gliozzo & Strapparava, 2005b). It is performed by means of a Singular Value Decomposition (SVD) of the term-by-document matrix \mathbf{T} describing the corpus. The SVD algorithm can be exploited to acquire a domain matrix \mathbf{D} from a large corpus Γ in a totally unsupervised way. SVD decomposes the term-by-document matrix \mathbf{T} into three matrixes $\mathbf{T} \approx \mathbf{V} \Sigma_{k'} \mathbf{U}^T$ where $\Sigma_{k'}$ is the diagonal $k \times k$ matrix containing the highest $k' \ll k$ eigenvalues of \mathbf{T} , and all the remaining elements set to 0. The parameter k' is the dimensionality of the Domain VSM and can be fixed in advance³. Under this setting we define the domain matrix D_{LSA} ⁴ as $D_{LSA} = \mathbf{I}^N \mathbf{V} \sqrt{\Sigma_{k'}}$ (2) where \mathbf{I}^N is a diagonal matrix such that:

$$i^N_{ii} = \frac{1}{\sqrt{\langle \vec{w}'_i, \vec{w}'_i \rangle}}, \quad \vec{w}'_i \text{ is the } i^{th} \text{ row of the matrix } \mathbf{V} \sqrt{\Sigma_{k'}}.$$

The Domain Kernel, denoted by K_D , can be exploited to estimate the topic similarity among two texts while taking into account the external knowledge provided by a Domain Model. It is defined by:

$$K_D(ti, tj) = \frac{\langle \Delta(ti), \Delta(tj) \rangle}{\sqrt{\langle \Delta(tj), \Delta(tj) \rangle \langle \Delta(ti), \Delta(ti) \rangle}} \quad (3)$$

where Δ is the Domain Mapping defined in Formula 1. To be fully defined, the Domain Kernel requires a Domain Matrix \mathbf{D} . In principle, \mathbf{D} can be acquired from any corpora by exploiting any (soft) term clustering algorithm. Anyway, we believe that adequate Domain Models for particular tasks can be

better acquired from collections of documents from the same source. The matrix D_{LSA} , defined by Formula 2 is acquired using the whole unlabeled training corpora available for each task, so tuning the Domain Model on the particular task in which it will be applied.

4. Experimental settings

We have modified Atenea's architecture so that after the NLP modules chosen have done the processing to the student's answer and its references; both the ERB and the LSA module are in charge to compare them. With the introduction of the LSA module we expect to add a notion of semantic similarity between the student's answer and the references. In this way, not only the style of the answer but also the content is addressed.

The LSA algorithm applied follows the pseudo-document methodology described by (Berry, 1992). We have defined the LSA score as the mean of the pseudo-document similarities between the student's answer and each vector representing a reference. Because of the results obtained in previous experiments (Pérez et al., 2005), we have chosen for training the Ziff-Davis part of the North American Collection with 142.580 articles from different journals and magazines in Computer Science. The LSA score is combined with the ERB one using the following linear combination:

$$\text{Atenea' score} = \alpha \times \text{ERB} + (1 - \alpha) \times \text{LSA} \quad (4)$$

The best value for α is obtained, for each test set, empirically, to maximize Atenea's assessing accuracy, measured as the Pearson correlation between Atenea's and the teacher's scores for the same questions data set.

The test corpus used contains ten different questions about Operating Systems and Object Oriented Programming, nine of them obtained from exams in our home university, and the last one consisting of a set of definitions of "Operating System" obtained from the Internet. In total, there are 924 student answers and 44 references written by teachers. Table 3 describes the datasets.

In previous work we have observed the robustness of the n-gram-based algorithm to cope with automatic translations of the answers and the references. The correlation between the teachers' scores and the system's does not vary in a statistically significant way if we work with the original student's answers or with automatic translations (Alfonseca & Pérez, 2004; Pérez & Alfonseca, 2005). Given that the LSA system had been trained on an English corpus, and we did not have a large corpus on Computer Science available in Spanish, we have chosen to work with a translation of our test corpus into English, performed by Altavista Babelfish⁵.

Table 3. Evaluation sets⁶.

SET	NS	MS	NR	MR	Type	ERB	LSA	P(E+L)
1	38	67	4	130	Def.	0.61	0.49	0.61
2	79	51	3	42	Def.	0.54	0.20	0.38
3	96	44	4	30	Def.	0.20	-0.01	0.10
4	11	81	4	64	Def.	0.29	0.52	0.48
5	143	48	7	27	Def.	0.61	0.50	0.64
6	295	56	8	55	A/D	0.19	0.24	0.23
7	117	127	5	71	Y/NJ	0.33	0.29	0.38
8	117	166	3	186	A/D	0.39	0.39	0.46
9	14	118	3	108	Y/NJ	0.75	0.78	0.81
10	14	116	3	105	Def.	0.78	0.87	0.90
MEAN	92.4	87.4	4.4	81.8	—	0.47	0.43	0.50

5. Experiment and results

We have tested five different configurations of Atenea: stemming (A1); removal of closed-class words (A2); stemming and removal of closed-class words (A3); Word Sense Disambiguation (WSD, A4); and WSD and removal of closed-class words (A5). In each case, the scoring is done by calculating the n-gram score described above on the processed text. In the case of WSD, nouns, verbs, adjectives and adverbs are substituted by sense identifiers from WordNet before the score is calculated.

Table 4 shows the results obtained for these combinations without using the LSA module. Table 5 evidences how they improved when the LSA module is used, and we give the same weight to Atenea

and LSA ($\alpha = 0.5$). Values highlighted in bold indicate that the combination with LSA has produced a better result. Finally, Table 6 shows the results obtained after optimising the value of empirically.

Table 4. Results of Atenea for five different configurations.

SET	A1	A2	A3	A4	A5
1	0.618881	0.540396	0.582110	0.632213	0.595169
2	0.468772	0.582403	0.550087	0.440452	0.495139
3	0.232850	0.450573	0.405473	0.234355	0.418775
4	0.346364	0.462735	0.536814	0.311094	0.494748
5	0.639496	0.665886	0.708401	0.655028	0.718707
6	0.236440	0.306757	0.337241	0.225270	0.322892
7	0.291880	0.369568	0.411027	0.282297	0.408539
8	0.374665	0.412187	0.440874	0.368593	0.434522
9	0.765208	0.740003	0.677579	0.761126	0.691731
10	0.850650	0.709258	0.718523	0.829618	0.706292
MEAN	0.482521	0.523977	0.536813	0.474005	0.528651

Table 5. Results of the combination (Atenea+LSA) for the five different configurations.

SET	M(A1+L)	M(A2+L)	M(A3+L)	M(A4+L)	M(A5+L)
1	0.630219	0.564336	0.599315	0.642924	0.611331
2	0.451979	0.573117	0.551557	0.428057	0.506119
3	0.212284	0.419053	0.377003	0.211843	0.385953
4	0.388911	0.499870	0.555359	0.355465	0.518276
5	0.645359	0.674847	0.714907	0.659699	0.724796
6	0.244271	0.312712	0.340969	0.233106	0.326627
7	0.302367	0.380963	0.417945	0.293451	0.416635
8	0.391221	0.430773	0.454752	0.385953	0.449158
9	0.784240	0.764052	0.712132	0.780152	0.722886
10	0.884809	0.780588	0.783078	0.868666	0.771207
MEAN	0.493566	0.540031	0.550702	0.485932	0.543299

Table 6. Results of the combination
($\alpha = 0.174, 0.346, 0.323, 0.151$ and 0.298 , respectively).

SET	P(A1+L)	P(A2+L)	P(A3+L)	P(A4+L)	P(A5+L)
1	0.638036	0.577326	0.610560	0.645999	0.622724
2	0.344846	0.534323	0.509780	0.316856	0.463320
3	0.138333	0.367555	0.334028	0.123294	0.329082
4	0.486621	0.521720	0.566815	0.473688	0.537167
5	0.649906	0.680228	0.719279	0.659279	0.729381
6	0.265019	0.316759	0.343870	0.257485	0.330220
7	0.329238	0.388714	0.423370	0.326698	0.424366
8	0.433712	0.444291	0.466936	0.436597	0.464349
9	0.806440	0.777007	0.737520	0.805299	0.749767
10	0.926590	0.821233	0.828300	0.924139	0.827158
MEAN	0.501874	0.542916	0.554046	0.496933	0.547754

Before these experiments were performed, the best correlation obtained using the statistical module of Atenea combined with LSA was 50% (Pérez et al., 2005). As can be seen in Table 6, when the optimisation procedure is applied, together with some NLP modules, the average correlation increases up to 55.40%.

CONCLUSIONS

Automatically assessing free-text students' answers is a long-standing problem that is still far from being completely solved. In this paper, we describe a new approach that consists in combining statistical, Natural Language Processing and Latent Semantic Analysis techniques.

This approach has been implemented in the Atenea system, an adaptive free-text CAA system, able to assess students' answers. The core idea of the system is that a student's answer is better and thus, it should receive a higher score, when it is more similar to the teachers' answers (or references) for the same question.

In order to be able to handle some of the many different expressions that convey the same meaning, Atenea can be configured to process both the student's answer and the references with stemming, removal of closed-class words, and/or Word Sense Disambiguation (WSD) techniques. Due to the stemming step, morphological differences are not taken into account in the comparison. The removal of closed-class words is useful to ignore coincidences in those words that are less relevant to the general meaning of the answers. Finally, WSD attempts to identify the sense in which polysemous words are used to find out if the senses intended by the teacher and the student are the same.

After that NLP processing, the n-gram co-occurrence scoring procedure ERB produces a value that is combined with the LSA one to finally give the students their score.

The goodness of the procedure is measured with the Pearson correlation between the automatic scores given by the system and the manual scores given by the teachers for the same datasets of questions. In particular, the mean correlation has improved up to the state-of-the-art 56% value, when both the students' answers and the references were first stemmed, then the closed-class words were removed and the ERB and LSA techniques were combined. In this way, it has been seen that LSA and the rest of NLP techniques combined can produce better results than any of them separately.

This result is also interesting since it determines the optimum combination of techniques to use Atenea with students. That is, the on-line version of Atenea will be configured to stem, remove closed-class words, use ERB and LSA and combine their scores to finally produce the student's score.

Atenea benefits from a Machine Translation engine to be able to process answers written in Spanish and in English. In particular, the system is able to automatically translate the students' answers to the language in which the references (teachers' answers) are written, without a significant decrease in accuracy. In fact, in a few cases, the Pearson correlation even improves slightly, as noted in (Pérez & Alfonseca, 2005).

Given that the LSA subsystem had been trained just for the English language, it is not able to process Spanish answers yet. Therefore, we have taken advantage of this feature in Atenea in the evaluation performed, so we have been able to score as well Spanish answers, by automatically translating them into English, using LSA.

This paper opens many promising prospective lines. Firstly, to try a more complex combinational between Atenea and LSA. Secondly, in the general architecture of Atenea it is easy to integrate other NLP tools such as a rhetorical analyzer. It would be very interesting to incorporate it because it would identify both in the student's answer and in the teacher's answer the fragments of the text in which a new idea is introduced, advantages or disadvantages are given, a concluding remark is provided, etc., achieving a better comparison process that could lead to a more accurate score to combine with the LSA score.

Other interesting possibilities are to try this combination in other systems. For instance, it should be possible to add an LSA module to other free-text CAA systems such as SEAR (Christie, 1999), which is based on Information Extraction (IE). Conversely, systems such as IEA (Foltz et al., 1999), that only relies on LSA, might be combined with shallow NLP techniques.

NOTES

¹ The results are presented according to the metrics indicated by their authors (Corr: correlation; Agr: Agreement; EAgr: Exact Agreement; CAcc: Classification accuracy; f-S: f-Score). When the authors have presented several values for the results, the mean value has been taken.

² In (Wong, Ziarko & Wong, 1985) a similar schema is adopted to define a Generalized Vector Space

Model, of which the Domain VSM is a particular instance.

³ It is not clear how to choose the right dimensionality. In our experiments we used 400 dimensions.

⁴ When D_{LSA} is substituted in Formula 1 the Domain VSM is equivalent to a Latent Semantic Space (Deerwester et al., 1990). The only difference in our formulation is that the vectors representing the terms in the Domain VSM are normalized by the matrix IN , and then rescaled, according to their IDF value, by matrix I^{IDF} . Note the analogy with the tf-idf term weighting schema (Salton & McGill, 1983), widely adopted in Information Retrieval.

⁵ Available at <http://world.altavista.com>

⁶ Columns indicate: number of student answers (NS), their mean length (MS), number of references (NR), their mean length (MR), question type (Def., definitions; A/D, advantages and disadvantages; Y/NJ, yes-no with justification), ERB, LSA and their combination results.

REFERENCES

Alfonseca, E. (2003). *Wraetlic user guide version 1.0*. [on line]. Retrieved from: <http://www.ii.uam.es/~ealfon/eng/download.html> [Links]

Alfonseca, E. & Pérez, D. (2004). *Automatic assessment of short questions with a Bleu-inspired algorithm and shallow NLP*. Proceedings of the 4th International Conference, ESTAL 2004, Alicante, Spain. [Links]

Berry, M. (1992). Large-scale sparse singular value computations. *International Journal of Supercomputer Applications*, 6(1), 13-49. [Links]

Blayney, P. & Freeman, M. (2003). *Automated marking of individualised spreadsheet assignments: The impact of different formative self-assessment options*. Proceedings of the 7th Computer Assisted Assessment International Conference, Loughborough, U.K. [Links]

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Bradenharder, L. & Harris, M. (1998). *Automated scoring using a hybrid feature identification technique*. Proceedings of the Annual Meeting of the Association of Computational Linguistics, Montreal, Canada. [Links]

Burstein, J., Leacock, C. & Swartz, R. (2001). *Automated evaluation of essays and short answers*. Proceedings of the 5th Computer Assisted Assessment International Conference, Loughborough, U.K. [Links]

Callar, D., Jerrams-Smith, J. & Soh, V. (2001). *CAA of short non-MCQ answers*. Proceedings of the 5th Computer Assisted Assessment International Conference, Loughborough, U.K. [Links]

Christie, J.R. (1999). *Automated essay marking - for both style and content*. Proceedings of the 3rd Computer Assisted Assessment International Conference, Loughborough, U.K. [Links]

Deerwester, S., Dumais, S., Furnas, G., Landauer, T. & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407. [Links]

Dessus, P., Lemaire, B. & Vernier, A. (2000). *Free text assessment in a virtual campus*. Proceedings of the 3rd International Conference on Human System Learning, Paris, France. [Links]

Foltz, P., Laham, D. & Landauer, T. (1999). The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2). [on line]. Retrieved from: <http://imej.wfu.edu/articles/1999/2/04/index.asp> [Links]

Gliozzo, A., Magnini, B. & Strapparava, C. (2004). *Unsupervised domain relevance estimation for word sense disambiguation*. Proceedings of the Empirical Methods in Natural Language Processing Conference, Barcelona, Spain. [Links]

Gliozzo, A., Giuliano, C. & Strapparava, C. (2005a). *Domain kernels for word sense disambiguation*. Proceedings of ACL, Michigan, U.S.A. [Links]

Gliozzo, A. & Strapparava, C. (2005b). *Domain kernels for text categorization*. Proceedings of (CONLL), Michigan, U.S.A. [Links]

Haley, D., Pete T., Nuseibeh, B., Taylor, J. & Lefrere, P. (2003). *E-Assessment using Latent Semantic Analysis*. Proceedings of the 3rd International LeGE-WG Workshop: Towards a european learning GRID

infrastructure to support future technology enhanced learning, Berlin, Germany. [[Links](#)]

Larkey, L. (1998). *Automatic essay grading using text categorization techniques*. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, U.S.A. [[Links](#)]

Mason, O. & Grove-Stephenson, I. (2002). *Automated free text marking with paperless school*. Proceedings of the 6th International Computer Assisted Assessment Conference, Loughborough, U.K. [[Links](#)]

Miller, T. (2003). Essay Assessment with Latent Semantic Analysis. *Journal of Educational Computing Research*, 29(4), 495-512. [[Links](#)]

Ming, Y., Mikhailov, A. & Kuan, T. (2000). Intelligent essay marking system [on line]. Retrieved from: http://ipdweb.np.edu.sg/lt/feb00/intelligent_essay_marking.pdf [[Links](#)]

Mitchell, T., Russell, T., Broomhead, P. & Aldridge, N. (2002). *Towards robust computerised marking of free-text responses*. Proceedings of the 6th International Computer Assisted Assessment Conference, Loughborough, U.K. [[Links](#)]

Page, E. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, 47(1), 238-243. [[Links](#)]

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, Yorktown Heights, New York, U.S.A. [[Links](#)]

Pérez, D., Gliozzo, A., Strapparava, C., Alfonseca, E., Rodríguez, P. & Magnini, B. (2005). *Automatic assessment of students' free-text answers underpinned by the combination of a BLEU-inspired algorithm and Latent Semantic Analysis*. American Association for Artificial Intelligence (AAAI) Press. Proceedings of the 18th FLAIRS International Conference, Florida, U.S.A. [[Links](#)]

Pérez, D. & Alfonseca, E. (2005). *Adapting the automatic assessment of free-text answers to the students*. Proceedings of the 9th Computer Assisted Assessment (CAA) international conference, Loughborough, U.K. [[Links](#)]

Rosé, C., Roque, A., Bhembe, D. & VanLehn, K. (2003). *A hybrid text classification approach for analysis of student essays*. Proceedings of the HLT-NAACL workshop Building Educational Applications Using Natural Language Processing, Edmonton, Canada. [[Links](#)]

Rudner, L. & Liang, T. (2002). *Automated essay scoring using bayes' theorem*. *Journal of Technology, Learning, and Assessment*, 1(2). [on line]. Retrieved from: <http://www.bc.edu/research/intasc/jtla/journal/v1n2.shtml> [[Links](#)]


Salton, G. & McGill, M. (1983). *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill. [[Links](#)]

Sukkarieh, J., Pulman, S. & Raikes, N. (2003). *Auto-marking: Using computational linguistics to score short, free text responses*. Proceedings of the 29th Annual Conference of the International Association for Educational Assessment, Manchester, U.K. [[Links](#)]

Valenti, S., Neri, F. & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2, 319-330. [[Links](#)]

Vantage Learning Tech. (2000). *A study of expert scoring and intellimetric scoring accuracy for dimensional scoring of grade 11 student writing responses*. Technical Report RB-397, Vantage Learning Technology, Newtown, Philadelphia, U.S.A. [[Links](#)]

Wong, S., Ziarko, W. & Wong, P. (1985). *Generalized vector space model in information retrieval*. Proceedings of the 8th International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, U.S.A. [[Links](#)]

 **Dirección para correspondencia:** Diana Pérez (diana.perez@uam.es). Tel.: (34-91) 4972267 Fax: 4972235. Departamento de Ingeniería Informática, Escuela Politécnica Superior, Universidad Autónoma de Madrid. Ciudad Universitaria de Cantoblanco, Calle Francisco Tomás y Valiente 11, Madrid, España.

Recibido: 30-VI-2005 **Aceptado:** 7-X-2005

*Project TIN2004-03140. Spanish Ministry of Science and Technology.

Todo el contenido de la revista, excepto dónde está identificado, está bajo una Licencia Creative Commons

Pontificia Universidad Católica de Valparaíso

Av. El Bosque 1290, 5º piso, Sausalito

Viña del Mar - Chile

Tel.: (56) (32) 2274000



revista.signos@ucv.cl