Natural Language Processing meets User Modeling for automatic and adaptive free-text scoring

Combinando técnicas de Procesamiento de Lenguaje Natural y Modelado de Usuario para la evaluación automática y adaptativa en texto libre

Diana Pérez Marín
EPS, UAM
Madrid, Spain
diana.perez@uam.es

Ismael Pascual Nieto
EPS, UAM
Madrid, Spain
ismael.pascual@uam.es

Pilar Rodríguez
EPS, UAM
Madrid, Spain
pilar.rodriguez@uam.es

Resumen: Tradicionalmente, los sistemas de evaluación automática de respuestas en texto libre se han basado únicamente en el uso de técnicas de Procesamiento de Lenguaje Natural. De esta forma se ha ido mejorando el rendimiento de estos sistemas, pero no se ha podido ofrecer a los estudiantes la posibilidad de una evaluación formativa adaptada a sus necesidades y a su nivel real de conocimiento en función de las respuestas en texto libre proporcionadas al sistema usado. En este artículo, se describe un procedimiento en el que técnicas de Procesamiento de Lenguaje Natural y Modelado de Usuario se combinan para generar y mantener un modelo de estudiante en sistemas de evaluación automática de respuestas en texto libre. De esta forma, la evaluación de las respuestas en texto libre, no es sólo automática sino también adaptada a las características específicas de cada estudiante en cada momento.
Palabras clave: Evaluación de respuestas en texto libre, Modelado de Usuario, Hipermedia Adaptativa, Blended Learning

Abstract: Free-text Computer Assisted Assessment (CAA) systems are able to automatically score free-text students' answers using Natural Language Processing techniques. Traditionally, free-text CAA systems have not included any possibility of adaptation, or kept a student model. In this paper, a procedure in which Natural Language Processing and User Modeling techniques are used together to generate and keep a student model in free-text CAA systems is described. That way, it is possible to offer the students not only an automatic assessment of their free-text answers, but also adaptation to their specific formative needs and their real level of knowledge. The student model is extracted from the students' free-text answers to the questions asked by the system and, the model is used by the system to choose the next question to ask the student. That way, not only the model is derived from the students' answers, but the students' answers keep the model updated.
Keywords: Free-text Computer Assisted Assessment, User Modeling, Adaptive Hypermedia, Blended Learning

1 Motivation

The benefits of a convergence between Natural Language Processing (NLP) and User Modeling (UM) techniques have already been discussed by several researchers.

For instance, Zukerman and Litman (2001) claimed that obtaining the model from free-text introduced by the user into the system would increase its accuracy, Reiter et al. (2003) described how it is possible to acquire and use limited user models for Natural Language Understanding, and Johansson (2002) gathered several benefits of UM in dialogue systems: providing the user with tailored help and advice, eliciting information, and helping resolving ambiguity (Kass and Finin, 1988); avoiding redundancy and incomprehensibility in answers and explanations, taking into account goals and plans for which the user really needs some requested information, and detecting misconceptions of the user and informing the user about them (Kobsa, 1990); and, enhancing effectiveness (i.e. to reach the

ISSN: 1135-5948

correct decision for a specific user), efficiency (i.e. to reach the correct decision in an economical way) and, acceptability (i.e. to support the decision-making process in a comprehensible and agreeable and enjoyable way for a specific user) (Sparck-Jones, 1989).

However, many of these authors have usually only focused on fields such as Natural Language Generation (NLG), or Natural Language Understanding (NLU), mainly in relation to pragmatics. For instance, NLG systems that consult the user model to do content planning, or NLU systems that build the user model to represent user's plans and goals.

Little has been said about automatically exploiting the user models to other domains such as, for instance, education. User models, or student models for the particular case of the academic context, contain static and dynamic information that could be used to choose which questions should be asked to the students.

For instance, if an estimation of how well each concept is known by each student was kept in each student model, it would be possible to identify which concepts are less understood. That way, it would be possible to directly ask the students about those misconceptions, instead of repeating questions about concepts already known.

In our opinion, the combination of the benefits of UM in Intelligent Tutoring Systems (ITS) (Kay, 2001) with the new possibilities that NLP could also bring to e-learning (i.e. using electronic media to teach), Blended Learning (i.e. using traditional and electronic media to teach), and Computer Assisted Assessment (i.e. using computers to assess students' knowledge) are many.

In particular, in this paper, we describe the combination of Natural Language Processing and User Modeling techniques to automatically generate each student's model from his or her free-text answers to a free-text CAA system.

In this context, a student model can be defined as the set of static and dynamic information about each student. The static component consists of the student's personal data (i.e. name, age, language, etc.) which do not usually change during the course. The dynamic component consists of the information about how the student has used the concepts in the answers which changes during the course.

The dynamic component of the model is the focus of this paper. We have called it the student's conceptual model. It can be defined as a network of interrelated concepts in which each concept is associated a confidence-value (CV). This CV indicates how confident an automatic free-text scoring system is that the student knows each concept according to a set of metrics.

The paper is organized as follows: Section 2 briefly reviews some concepts about student models and educational systems able to keep student models; Section 3 gives an overview of the procedure to generate the conceptual model from the students' free-text answers; Section 4 reviews the Natural Language Processing and User Modeling techniques used in the procedure; Section 5 discusses the results achieved; Section 6 exposes the limitation of the current approach and how it can be extended to other domains; and, Section 7 ends with the main conclusions drawn about the benefits of using UM+NLP for free-text scoring.

2 Related work

Student models are one of the main components of Intelligent Tutoring Systems (Kay, 2001). There is not limits or restrictions regarding the information that can be kept in one student model, or how the information has to be organized.

Therefore, there are many different types of student models, and multiple criteria according to which they can be classified. One of them is concerning the relationship between domain and student knowledge. That is, depending on how the student knowledge represents the domain knowledge, student models can be classified as (Labidi and Sergio, 2000; Mitrovic, 2001):

- Overlay: The student model is a projection of the domain model, i.e. the student knowledge is considered as a subset of the domain knowledge.
- Bug: The bug model is based on a library of possible mistakes that could be made up by the student in its pedagogical activities.
- Perturbation: The perturbation model is an hybrid model that involves the concepts of the overlay and bug together.
- Constraint-based: Opposite to the previous models, it does not compare the student's knowledge to the domain knowledge. It rather focuses on correct knowledge by checking if all the constraints of a certain domain are satisfied by the student.

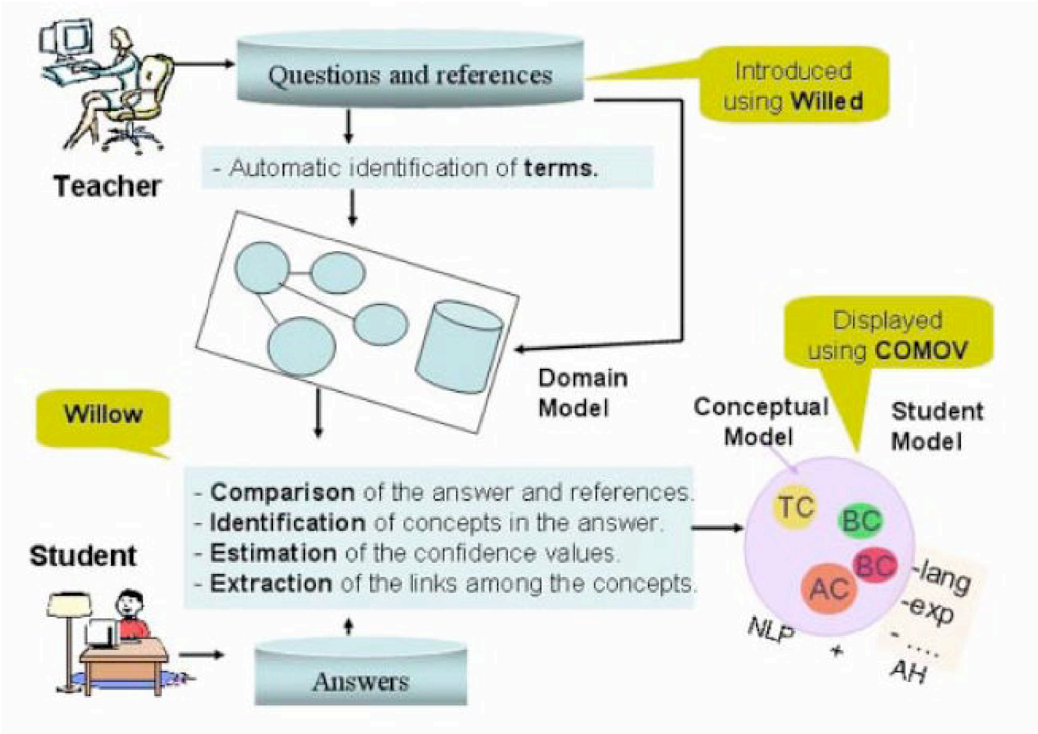


Figure 1: Overview of the procedure to generate students' conceptual models

The student model that we propose in this paper can be classified as a perturbation model. It is because it is not considered that the student model is an exact projection of the domain model, or a set of possible mistakes made by the students. On the other hand, it is considered that the model reflects several possible misconceptions made by the students regarding a certain domain. Furthermore, a constructivist view is follow in the paper according to which each student builds his or her knowledge as s/he interacts with the world.

Several educational applications keep a student model to support personalized learning. For instance, ALE (Kravcik and Specht, 2004) keeps a model of the students with information about their learning style to adjust the navigation possibilities in the course to them; ConceptLab (Zapata-Rivera and Greer, 2001) supports knowledge construction and visualization using concept maps to represent the student's view of the domain; and, E-tester (Guetl et al., 2005) diagnoses student's knowledge with a conceptual frequency histogram student model.

However, none of these systems use any kind of NLP technique to automatically generate and update the model. In fact, the most related system to our approach is E-tester.

E-tester keeps a model of the frequency in which each term is used by each student to be compared with the frequency in which it is used by the teachers in a set of model answers. However, the focus is only on individual concepts without taking into account the relationships among the concepts.

3 Overview of the procedure

Figure 1 shows an overview of the procedure as implemented in the Will Tools (Pérez-Marín, 2007). The Will Tools consist of the following systems: Willow, a free-text CAA system; Willed, an authoring tool; Willoc, a configuration tool; and, COMOV, a conceptual model viewer.

First of all, the teacher is asked to use Willed to introduce the questions and its correct answers (references) in the database. The references are automatically processed to generate the domain model. Next, whenever a student answers one of the questions proposed by Willow, not only s/he gets instant feedback but, his or her use of the concepts of the domain model is analyzed to automatically generate his or her student model. The student model consists of personal data gathered from the student and the generated conceptual model.

Finally, the conceptual model can be shown to teachers and students with COMOV to identify which concepts of the lessons should be reviewed and, which ones have already been assimilated. The model is also used by Willow to choose the next question to ask, and the content of the model is updated with the new answers provided by the students.

This procedure can be used just from the answers of one student to generate an individual student's conceptual model. Or, the procedure can be used from the answers of a group of students to generate a group conceptual model.

In any case, by using this procedure, the student becomes immediately aware, by looking at his or her generated model, of which concepts s/he should still review more; and, the teacher gains access to a monitoring tool that reports not only information such as how many questions each student and the whole group have answered but, by looking at each student's generated model, which concepts each student seems to have understood or misunderstood, and by looking at the group conceptual model, the average results of the whole group.

The Will Tools were used for the first time in the 2005 year by a Spanish group of the Operating Systems subject of an Engineering degree. A year later, they were used during a whole semester in the same domain, and in the first semester of 2007-2008 academic year they have been also used in a non-technical domain. The results gathered in all these experiments support the feasibility of the procedure. However, given that this paper is focused on the combination of the Natural Language Processing and User Modeling techniques for the procedure to work, the experiments are mentioned here just to let the reader know that the procedure has been implemented and has already been used with students during several years.

4 NLP+UM techniques for automatic and adaptive free-text scoring

As can be seen in Figure 1, one of the inputs to Willow is a student's answer in plain text. Other inputs are the knowledge contained in the domain model as introduced by the teacher in Willed (Willow's authoring tool) and, as external lexical resources: WordNet 1.7 for English and the Spanish EuroWordNet for Spanish (Vossen, 1998). Both WordNet and the Spanish EuroWordNet have been used as processed by Alfonseca (2003).

Regarding the NLP techniques, in Willed, it is implemented the automatic identification of the concepts of the course with the ***Term Identification*** module that can be implemented using the C4.5 algorithm (Quinlan, 1993), considering that a term is the label of a concept.

The features considered as attributes for the training should be at least:

- The relative frequency (freqRel.) of appearance of the term in the domain-specific corpus (i.e. a corpus of free-text students' answers) with respect to its frequency in the generic corpus (i.e. journal news on Computer Science). This is because terms tend to be specific to a certain knowledge field and thus, to appear more frequently in the specific corpus and consequently, have a relative frequency higher than one.
- The sequence of part-of-speech (POS) tags of the words composing the sample (e.g. determiner+noun+adjective). This is because terms tend to contain certain POS tags such as nouns, adjectives, etc. but not others such as verbs or adverbs.

The resulting list of terms can be shown to the teacher, so that they can modify it as they consider more adequate to finally produce the list of concepts of the course and, the CV of all these concepts is set to zero.

In Willow, the NLP techniques used are the following:

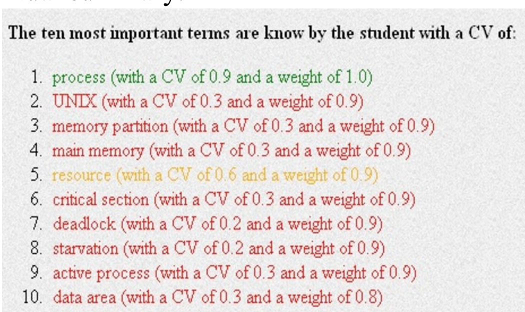
- **Stemming**, **removal of closed class words**, **Word Sense Disambiguation**, and/or **Multiword Identification** for processing the free-text answers provided by the students and the correct answers (references) provided by the teachers to make their comparison easier (following the core idea of Willow and many other free-text scoring systems that the more similar the student's answers to the teachers' answers are, the better they are). Additionally, a simple **pattern extraction** mechanism is implemented to find out relationships *"BC linking word BC"* in the students' answers.
- **Statistical n-gram comparison** using the Evaluating Responses with Bleu **(ERB)** algorithm (Pérez-Marín, 2007), which is more focused on the style of the answer; and, **Latent Semantic Analysis (LSA),** which is more focused on the content of the answer to compare the processed student's answers and teachers' answers. Each of them gives as output a numerical score: ERB between 0 (bad answer) and 1 (good

answer) and LSA between -1 (bad answer) and 1 (good answer). They can be used together or independently. In the case, that ERB and LSA scores are used together, it could be done as a linear combination of their normalized values in a common scale (e.g. from 0 up to 100).

Regarding User Modeling, the **question planner** chooses the next question to ask according to the level of difficulty of the questions that the student is able to pass; and, the **clarification questions** technique generates new questions (not typed by the teacher) in the form *"What is X?"* each time a concept is wrongly used by the student (i.e. the ERB and LSA metrics indicate that the student's answer is very different to the teachers' answers).

Finally, the student is given as feedback the numerical score (result of the combination of ERB and/or LSA techniques), the processed answer (with the concepts marked) and the correct answers (as provided by the teacher). Furthermore, the student and the teacher can look at the generated student's model for each particular student and the whole class in the **Conceptual Model Viewer**.

The student conceptual model is represented in five different formats so that each student can choose the format s/he considers more illustrative or look at all of them. These formats are: concept map, conceptual diagram, bar chart, table and textual summary. Figure 2 shows a sample concept model represented as textual summary.


The ten most important terms are know by the student with a CV of:

1. process (with a CV of 0.9 and a weight of 1.0)
2. UNIX (with a CV of 0.3 and a weight of 0.9)
3. memory partition (with a CV of 0.3 and a weight of 0.9)
4. main memory (with a CV of 0.3 and a weight of 0.9)
5. resource (with a CV of 0.6 and a weight of 0.9)
6. critical section (with a CV of 0.3 and a weight of 0.9)
7. deadlock (with a CV of 0.2 and a weight of 0.9)
8. starvation (with a CV of 0.2 and a weight of 0.9)
9. active process (with a CV of 0.3 and a weight of 0.9)
10. data area (with a CV of 0.3 and a weight of 0.8)

Figure 2: Sample generated student's conceptual model presented as textual summary

Parameter	Value
Language	Spanish / English
Area-of-knowledge	Operating Systems
Metric of goodness	Pearson correlation between Willow's and teachers' scores

Table 1: Parameters used

5 Results

The results of using these NLP techniques combined differ according to the language and area-of-knowledge in which the techniques are applied. Furthermore, it depends on the metric of goodness of the procedure chosen.

Nevertheless, in order to give a general idea of the values that can be reached, some results are presented using the parameters indicated in Table 1 (see Pérez-Marín, 2007 for the experimental details):

- For Spanish, the optimum combination is to choose **Term Identification** to extract the BCs using the C4.5 algorithm (Quinlan, 1993) achieving 74% F-score, **stemming** and **ERB** reaching up to 54% Pearson correlation (an average value in the field, Valenti et al., 2003) Furthermore, if a **Genetic Algorithms Module** is included to automatically choose some of the best students' answers of one year as correct answers of the following year, the Pearson correlation is increased up to **63%** (Pérez-Marín, 2007).

- For English, the optimum combination is to choose **Term Identification** to extract the BCs using the C4.5 algorithm (Quinlan, 1993) also achieving 74% F-score, **stemming, removal of closed-class words**, and **ERB+LSA** reaching up to **56% Pearson correlation**. Given that the procedure has not been applied to English students yet, no results are available to decide how this percentage can be increased by applying the **Genetic Algorithms Module.**

Finally, when the metric used is to measure the Pearson correlation between the scores achieved by the students in the final exam, and the scores achieved by the students in the generated conceptual model as the average of the CVs of all the concepts of the model, a **50%** statistically significant positive correlation is found ($p=0.0063$) (measured over the results of 31 Spanish Engineering students using the optimum combination of NLP techniques, for more experimental details see Pérez-Marín, 2007).

Furthermore, it is possible to validate the automatically generated students' conceptual models at the concept level of granularity. In order to that, first of all, a human teacher has to be asked to estimate how well a group of students know a set of concepts. Secondly, Willow is used to estimate the same concepts for the same group of students. Finally, the mean quadratic error between the estimation made by the human and the estimation calculated by the system per each concept per each student is calculated. In particular, 0.08 mean quadratic error was attained measured over 65 concepts of nine students (for more experimental details, see Pérez-Marín, 2007).

Qualitatively, it has also been observed that the higher the scores achieved by the students in the final exam, the more complex that their conceptual models are.

6 Applications to other domains and/or languages

The results previously indicated have been focused on an Engineering area-of-knowledge, specifically Operating Systems because we work as teachers in that area. However, the proposed combination of techniques is not limited to Engineering degrees.

On the contrary, it can be applied to other areas-of-knowledge provided that it is not necessary to assess creative thinking or do mathematical calculations, which are completely out of scope of this work.

Moreover, the procedure can also be applied to other languages different than Spanish or English, just by taking the requirements explained in Table 2 into account.

NLP technique	Sp.	En.	Ot.
Term Identification	Yes	Yes	No
Stemming	Yes	Yes	No
Removal of closed-class words	Yes	Yes	No
ERB	Yes	Yes	Yes
LSA	No	Yes	No

Table 2: NLP techniques already implemented for the procedure to work in different languages (Sp.: Spanish; En.: English; Ot.: Other)

The steps to apply the procedure to generate the students' conceptual models from free-text answers typed into a free-text ACAA system such as Willow to other domain and/or language are the following ones:

1. Provided that the language in which the procedure is to be applied is English or Spanish, the teachers and students can use Willow or a free-text CAA with an English/Spanish interface without further modification. For other languages, the systems' interfaces should be translated into the target language.
2. Teachers should create a new area-of-knowledge with the authoring tool and afterwards, fill in the forms about its name, description, features and topics that the area-of-knowledge comprises.
3. Next, teachers should introduce the questions. In particular, it is necessary, per each question: its statement and several correct answers (e.g. three) to capture as much lexical variability as possible, maximum score, topic and level of difficulty.
4. For English and Spanish courses, it is not necessary to acquire any other Natural Language Processing techniques or resources than the already implemented in Willow, and can go directly to the next step. On the other hand, for courses written in a different language, it would be necessary to have a stemmer, a Part-of-Speech (POS) tagger, and a specific and generic corpora for the Term Identification Module. It is necessary to classify the candidate n-grams in the references as terms. The specific corpus is given by the correct answers provided by the teachers and the generic corpus can be automatically retrieved from the web.
5. The administrator of the authoring tool should apply the Term Identification module to the correct answers provided by the teachers to generate the first list of terms. This list can be reviewed by the teachers, and the resulting list of terms is stored as the concepts of the conceptual model together with their frequency in the teachers' correct answers.
6. Teachers ask their students to register in the free-text CAA system.
7. Students answer the questions asked by the free-text CAA system, which is keeping track of each student's use of the concepts found in his or her free-text answers. The system is also looking for patterns *"BC+linking words+BC"* and, in general, processing the answer with the NLP techniques chosen to be compared with the teachers' answers.

8. During the course, teachers and students can see the generated conceptual models using a conceptual model viewer.
9. For the next course, teachers who are willing to use the procedure again can ask the administrator of the tools to tune some internal parameters to improve the accuracy of free-text scoring. In particular, from the information gathered this year: genetic algorithms can be run on the teachers' and students' answers to choose the best references among this set of texts; and, teachers can be asked to manually score a set of answers to calculate the Pearson correlation between the automatic and the teachers' scores for these answers. This is because, as it has been seen in the previous section, the optimum combination of NLP techniques for a different area-of-knowledge and/or language may change and should be optimized to each particular case.

7 *Conclusions*

The benefits of the convergence between User Modeling (UM) and Natural Language Processing (NLP) techniques have been claimed by several researchers (Zukerman and Litman, 2001; Johansson, 2002; Reiter et al., 2003). However, these authors have usually focused the application of these techniques to the Natural Language Understanding and Natural Language Generation fields.

In this paper, it has been studied that UM and NLP techniques can be combined to permit the automatic and adaptive assessment of students' free-text answers.

Traditionally, many of the existing free-text CAA systems have formative assessment purposes: to serve as double-checker of the scores given the teachers, or to provide more training to the students before their final exams (Valenti et al., 2003). However, up to date, none of them has kept a student model that guides the student towards the correct answer, or provide him or her with more detailed feedback than a numerical score, correct answer or link to the theoretical explanation.

However, the benefits of generating a student model from his or her free-text answers are several, both for teachers and for students.

For teachers, there are two main benefits that can be highlighted:

- To receive more feedback to know how well the students have understood the concepts taught in the lessons. The teachers can look at the generated students' conceptual models in conceptual model viewers such as COMOV. Moreover, teachers can identify several types of misunderstandings that have been classified in the taxonomy of detectable errors detailed below,

* For concepts:

– Ignorance: Whenever a students does not use a certain concept, the concept is associated a CV of zero, and it may indicate that the student ignores that concept.

– Misconceptions: Some concepts may seem to be known by students as sometimes they use those concepts. However, the students might have wrongly used the concepts in their answers. Thus, these concepts are associated a CV below 0.5 in a 0 (no knowledge) to 1 (perfect knowledge) scale of estimation.

* For links:

– Ignorance: Whenever a student does not relate two concepts, the teacher can notice the lack of links between these two concepts, and it may indicate that the student ignores that the concepts are related.

– Erroneous links: Sometimes students erroneously relate two concepts in their answers. This evidences an error in the cognitive structure of the student as s/he believes that the concepts are related in a wrong way. It is fundamental to correct this situation to allow the student to continue learning meaningfully and linking correctly new concepts to the existing ones (Ausubel, 1963).

- To keep track of the students' learning progress by looking at the representations of the student model several times during the course, i.e. from a concept map representation, the teachers can easily see the conceptual evolution of the students by observing how the new concepts modify the previous ones, and the new links that are being created.

For students, there are the following benefits:

- To be able to get more personalized and efficient training before their final exams as the system finds out which concepts are worst known. The system starts asking each particular student about concepts with low CV, instead of going over concepts already known.

Furthermore, the system chooses the questions related to those concepts that are in the level of difficulty that the student is able to handle. The reason for that is to increase the students' motivation to keep answering new questions so that the new questions are not too difficult (i.e. too complicated to give any answer) or too easy (without any real interest for the student).

- To be guided towards the correct answer (promoting reflective thinking instead of memorizing the answer) with a set of clarification questions automatically generated by the system.

- To have access to always updated feedback as the model has not to be created by the teacher or the student. On the contrary, the model is automatically generated from the students' answers in a transparent process for the student who only has to answer the questions of the free-text CAA system.

References

Alfonseca, E. 2003. 'An Approach for Automatic Generation of on-line Information Systems based on the Integration of Natural Language Processing and Adaptive Hypermedia techniques', PhD dissertation, Computer Science Department, Universidad Autónoma de Madrid, available on-line at http://alfonseca.org/eng/research/thesis.html

Guetl, C., H. Dreher, and R. Williams. 2005. 'E-tester: A computer-based tool for auto-generated question and answer assessment', E-Learn, Canada.

Johansson, P. 2002. *'NLP Techniques for Adaptive Dialogue Systems'*, available on-line at (link reviewed on 20th April 2008): http://www.ida.liu.se/~ponjo/downloads/nlp1paper/johansson_nlp1_paper.pdf

Kass, R. and T. Finin. 1988. 'Modeling the User in Natural Language Systems', *Computational Linguistics,* 14: 5–22.

Kay, J. 2001. 'Learner Control', *User Modeling and User-Adapted Interaction,* 11: 111-127.

Kobsa, A. 1990. 'User Modeling in Dialog Systems: Potentials and Hazards', *AI & Society,* 4(3): 214-231.

Kravcik, M. and M. Specht. 2004. 'Flexible Navigation Support in the WINDS Learning Environment for Architecture and Design'. Proceedings of the Adaptive Hypermedia International Conference, The Netherlands.

Labidi, S. and N. Sergio. 2000. 'Student modeling and semi-automatic domain ontology construction for shiecc'. In Proceedings of the 30th ASEE/IEEE Frontiers in Education Conference.

Mitrovic, A. 2001. 'Cosc420 lecture notes: Cognitive modeling and intelligent tutoring systems'.

Pérez-Marín, D. 2007. '*Adaptive Computer Assisted Assessment of free-text students' answers: an approach to automatically generate students' conceptual models*', Ph.D. dissertation, Escuela Politécnica Superior, Universidad Autónoma de Madrid, available on-line at (link reviewed on 20th April 2008): http://www.eps.uam.es/~dperez/index1.html

Quinlan, J.R. 1993. 'C4.5: Programs for Machine Learning', Morgan Kaufmann Publishers.

Reiter, E., S. Sripada, and S. Williams. 2003. 'Acquiring and Using Limited User Models in NLG', *European Workshop on Natural Language Generation*, Hungary, pp. 87-94.

Sparck Jones, K. 1989. 'Realism about User Modeling'. In Kobsa, A., Wahlster, W., eds.: *User Models in Dialog Systems*. Springer-Verlag, Symbolic Computation Series, pp. 341–363.

Valenti, S., F. Neri and A. Cucchiarelli. 2003. 'An Overview of Current Research on Automated Essay Grading', Journal of Information Technology Education 2, pp. 319–330.

Zapata-Rivera, J.D. and J.E. Greer. 2001. 'Externalising learner modelling representations'. Proceedings of Workshop on External Representations of AIED: Multiple Forms and Multiple Roles, pp. 71–76.

Zukerman, I. and D. Litman. 2001. 'Natural Language Processing and User Modeling: Synergies and Limitations', *User Modeling and User-Adapted Interaction,* 11: 129-158.