Assessing the Student's Knowledge in Informatics Discipline Using the METEOR Metric

Iuliana Dobre
Department of Information Technology, Mathematics and Physics,
Petroleum-Gas University of Ploiesti, Romania
iulianadobre@yahoo.com

DOI:10.5901/mjss.2014.v5n19p84

Abstract

1970s could be considered a frontier since when the computers became accessible to the large mass of users and not only to a restricted segment of the academic society. As a consequence after 1970s, in the next three decades, the computers use has been extended to all domains of activity. In fact, at present can be stated that there is no industry, no activity, no school, no process etc. which has not involved the computer in their development. The higher education organizations didn't take any exceptions and enrolled themselves within the course of events and proceeded to capitalize on the computers use benefits. Many researchers driving different research programs have proved over the time that the computers are the solution to many fundamental and pervasive problems of the educational process. One of these problems is the learning performances measurement. Regardless the educational program proposed or the size of the higher education organization, the performances achieved are still to be measured. The students' knowledge achieved after a teaching and learning cycle is ended have to be assessed in order to establish the course of actions to be taken and to ensure on one hand the proper environment for having a continual improvement of the teaching and learning cycle and, on the other hand, to understand if the teaching and learning objectives of both, teachers and students, have been achieved. One of the greatest achievements obtained during the computers use expansion era was the development of the Intelligent Tutoring Systems (ITSs). Considered at the very beginning of their existence as too complex nowadays, the ITSs are not anymore considered just a fancy tool but are considered the future of many domains. Education is one of these domains which look to have full benefits from the use of the ITSs. Moreover, the Natural Language Processing (NLP) has been brought in the loop of the ITSs development-continual improvement process and integrated successfully within various activities requirements. During the time have been demonstrated that whenever was possible to develop ITSs using Natural Language Processing techniques, than the benefits become even greater then in the case of a classic ITS. This paper, is briefly presenting an ITS that uses NLP techniques developed to teach and learn the course of Computers Programming and C Language, part of the Basics of Informatics curriculum, deemed to provide students with general knowledge about programming of computers and C language. Also, this paper addresses the possibility to use the Metric for Evaluation of Translation with Explicit Ordering (METEOR) as an option for student's knowledge assessment module.

Keywords: *Intelligent Tutoring System; METEOR metric; Natural Language Processing; students' knowledge assessment*

Introduction

Use of ITSs within educational programs proposed by higher education organizations has received a significant attention in the past decades and has virtually blossomed since the Information Technologies and Communications (ITC) developments have invaded the market with really useful gadgets, cheap enough to be affordable to large mass of users, sufficiently feasible to be used for an easier access to large quantities of information from any domain of activity, offering very user friendly interfaces so no need having advance knowledge in computers use domain and also, heavy-duty equipped from manufacturers side than any software could work properly and could communicate at speeds unimaginable some years ago. Nowadays, most of the people afford to purchase at least one such gadget. Either a smartphone, or a tablet, or any of the laptops versions (i.e., notebook, ultrabook etc.) could be easily found in the hands of most of the users. Great benefits were supplied at hand to all users and education of any level should take all possible advantages from such developments.

The author of this paper considers that among all the opportunities offered by hardware developers and consequently, by the software developers, there is still room for improvement and looks to propose a new approach in students' knowledge assessment within the higher education organizations educational programs. ITSs and NLP are not new in education but the use of these in higher education curriculums, separately or together, don't have a very long history. The author believes that disciplines such us Informatics could benefit from the use of NLP techniques within teaching-learning-assessment cycle and started years ago a complex research journey to understand how ITSs and NLP techniques can be brought to work together and how other domains than the ones already using ITSs and NLP could benefit from a such challenging blend.

This paper is presenting briefly, the theoretical part of the final product developed by the author, product which it's proposed for usage within the teaching-learning-assessment cycle of Computers Programming and C Language course, part of the Basics of Informatics curriculum.

ITSs that Uses NLP Techniques

Motivation for ITSs Use within Teaching and Learning

As highlighted by Woolf (2009), for centuries the main forms used for education were the classrooms and the books. Nowadays, the classrooms and the books are still used as main methods to teach and learn but also, the computers have been integrated together with other items provided by ITC. Why the evolution of education has involved the use of computers and ITC gadgets? The answer is offered by figure 1 (Woolf, 2009).

Considered by Woolf (2009) as a "passive method" (p.14), the classroom method it is not considered as a very effective method (Woolf, 2009; Waterman, Matlin, and D'Amore 1993). As stated by Woolf (2009) only "the top fourth of each class, often motivated and gifted students" (p.14) are successful at the end of a teaching and learning cycle. As a support to his conclusions Woolf has adapted the "summative achievement scores vs. students" (p.15) graph from Bloom (1984). The graph has been reported in a study published by Bloom (1984) and is presenting the various ratios of student achievement in different education environments. According to Woolf and Bloom (as cited in Woolf, 2009), in a classroom environment (teacher presenting the course in front of 30 students, teacher/student ratio: 1:30), the student achievement is different from the one-on-one tutoring environment (teacher/student ratio: 1/1). The graph from figure 1 shows by bell curves that the achievement is around a 50% for traditional classroom education and is increasing significantly to a 98% for one-on-one teaching environment.

In conclusion, based on the graph from figure 1, the best performances could be obtained through one-on-one tutoring comparing to traditional education system (the conventional classroom). As a matter of fact, the figure 1 graph has been proved to be realistic by the results obtained through different studies, studies which have evaluated the promises made by innovations and advances of ICT and use of computers in education. Two of these promises, and which could be considered extremely profound and critical for a successful teaching and learning cycle, were:

(a) to engender teaching and learning for all, and, (b) to increase as much as possible the students' performances at individual level.

These promises, as it was highlighted by Woolf (2009) and also by Shute (1991a) were kept and the success was proved after ITSs have been taken onboard and implemented within educational programs. The ITSs were developed based on the resources made available by ITC and Artificial Intelligence. Specialists, researchers and developers acting in the education domain have looked to develop systems capable to provide high quality and effective one-on-one teaching and learning environments.

One of the first meta-studies published in 1991 by Shute (1991b), at the end of the first decade of ITSs use for teaching and learning, shows clear improvement when one-on-one tutoring method has been used. According to Shute (1991b), an evaluation of four ITSs has been performed in order to check the efficiency of those ITSs within the military educational programs. The four ITSs evaluated were: LISP tutor developed by Anderson, Farrell, and Sauers (1984), Smithtown developed by Shute and Glaser (1991a), Sherlock developed by Lesgold, Lajoie, Brunzo, and Eggan (1990) and PASCAL ITS developed by Bonar, Cunningham, Beatty, and Weil (1988), (all cited in Shute, 1991b). The results published by Shute (1991b) show that the performances obtained through conventional teaching and assessment, verbal testing and "paper-and-pencil examination" (p.4), are far lower than the performances obtained through the use of any of the four ITSs, while the comparison between the four ITSs shows similar performances. As an example, the ITS Sherlock, "which provides a coached practice environment for an electronics troubleshooting task" according to authors (Lesgold et al., 1990; as cited in Shute, 1991b, p.4) was used on two groups, one which was receiving on-job training for twenty hours and one group which received same training through the ITS for the same time period. The on-job training group has achieved a rate of success, at the end of the training of 58.9% while the group using the ITS obtained a rate of success of 82.2% (Shute, 1991b).

Another meta-analysis, more recent (Steenbergen-Hu and Cooper, 2014), "synthesizes research on the effectiveness of intelligent tutoring systems (ITS) for college students" (p.331). The analysis has been carried out based on 39 studies performed on 22 different ITSs used within higher education programs (Steenbergen-Hu and Cooper, 2014). Among ITSs taken into consideration by Steenbergen-Hu and Cooper (2014) were enumerated: AutoTutor, Assessment and Learning in Knowledge Spaces and eXtended Tutor-Expert System. The main conclusions of the meta-analysis from Steenbergen-Hu and Cooper (2014) study were:

(a) Can be considered that the ITSs, in general, have a positive moderate effect on students learning, and, (b) The ITSs proved uniformity from the viewpoint of effectiveness regardless the domain/discipline sighted or the ITSs involvement degree in teaching, learning and assessment. The performances obtained by using various ITSs were similar, (c) The ITSs analyzed proved the capability to achieve higher performances than the ones obtained through conventional teaching, learning and assessment, (d) Earlier studies consulted reported that ITSs effectiveness was greater than in the more recent studies consulted.

The above referenced analysis and their conclusions are considered by the author of this paper a sufficient motivation for the use of ITSs within higher education programs. In fact, nowadays, the ITSs could be considered a very important, powerful and effective tool for assuring that the large quantities of data and information which have to be processed by both, teachers and students, during the educational programs, are processed, assimilated and, later, practiced properly.

NLP and ITSs

The NLP as it is understood these days, same like ITSs don't have a very long history. The beginning is considered to be the year 1950 when, according to Wikipedia (Natural Language Processing, n.a., 2014), the first article related to computing machinery and intelligence was published. Afterwards developments and advances have led in 1980s to the beginning of a new revolution in NLP being introduced the "machine learning algorithms for language processing" (Natural Language Processing, n.a., 2014). Since then the machine translations breakthrough many domains but, still not very well present and implemented in some other important domains (i.e., industries training of personnel activating in various industries, higher education in physics, mathematics, and informatics etc.).

Advances in NLP machine learning algorithms have increased significantly the area of research and the number, type and variety of tasks proposed to be researched in this domain. The author considers sufficient to enumerate few of the most important tasks proposed by researchers, and highlighted in the Wikipedia dedicated Web page (Natural Language Processing, n.a., 2014) such as: natural language generation and understanding, machine translations, discourse analysis, questions answering, information extraction, speech and word segmentation etc..

Taking into consideration the advantages offered by NLP the researchers considered NLP as an option to be used within ITSs. Several ITSs have been developed and implemented with a certain rate of success, and from these ITSs could be mentioned: iSTART - Strategy Trainer for Active Reading and Thinking (McNamara, Levinstein, and Boonthum, 2004), ExtrAns – Extracting Answers from Technical Texts Question-Answering System (Molla, Schwitter, Rinaldi, Dowdall, and Hess, 2003), C-Rater – Short-Answer Questions Scoring System (Leacock and Chodorow, 2003), (all cited in Boonthum, Levinstein, McNamara, Magliano, and Millis, 2009), RACAI's Question Answering System (Tufiş, Ştefănescu, Ion, and Ceauşu, 2008b) etc.

According to Boonthum et al. (2009), at present, two are the main challenges in regards of ITSs that uses NLP and implemented in higher education programs, these being as follows:

(a) Students' knowledge assessment based on the NLP techniques, (b) To assure that proper feedback it is provided to students while learning and also, properly guide the students to improve based on the assessments results.

Same Boonthum et al. (2009) considers that the above challenges could be resolved using NLP techniques such as: word-matching, latent semantic analysis (Landauer, Foltz, and Laham, 1998, as cited in Boonthum et al., 2009) and topic models [Steyvers and Griffiths, 2007, as cited in Boonthum et al., 2009).

The author of this paper performed a research of the options to design, develop and implement an ITS that uses NLP to teach and learn the basic in Informatics and concluded that this is possible. Therefore, the author has designed and developed an ITS for teaching and learning the course of Computers Programming and C Language (Dobre, 2013a, 2013b, 2014a, 2014b, 2014c).

METEOR Metric

As stated by Lavie, Sagae, and Jayaraman (2004) and by Agarwal and Lavie (2008), METEOR has been released in the year 2004 by a group of researchers from Language Technology Institute School of Computer Science from Carnegie Mellon University, with the declared purpose defined by Banerjee and Lavie (2005): "to improve correlation with human judgments of MT quality at the segment level" (p.66). In fact, as explained by Wikipedia, METEOR has been designed and developed due to the necessity to solve some of the issues noted after quite popular metrics such as: BLEU (developed by IBM), and NIST (a BLEU derivate) were used (METEOR, n.a., n.d.). Same Web page from Wikipedia defines METEOR as "a metric for the evaluation of machine translation output" (METEOR, n.a., n.d.). Agarwal and Lavie (2008) explained that "Meteor evaluates a translation by computing a score based on explicit word-to-word matches between" (p.115) a candidate text and a given reference text.

The METEOR metric is using same like BLEU and NIST metrics, a reference text to which is compared the candidate text. Also, METEOR metric has the capability to compare the candidate text to more than one reference text in case more than one reference text is available. The comparison is done independently for each reference text and the candidate text it's scored against each reference text. METEOR will select the best scoring obtained.

As explained by developers, METEOR is capable to create a word alignment between the two texts and to realize a map between the texts words, "such that every word in each string maps to at most one word in the other string. This alignment is incrementally produced by a sequence of word-mapping modules. The "exact" module maps two words if they are exactly the same" (Agarwal and Lavie, 2008, p.115, 116).

In order to compare the reference text with the candidate text is necessary to use the so called "n-grams" (Dobre, 2013a; and Callison-Burch, Osborne, and Koehn, 2006; Papineni, Roukos, Ward, and Zhu, 2002 as cited in Dobre, 2013a). The n-grams are defined like "groups of n consecutive words rated as units of measure" (Dobre, 2013a, p.61) and were introduced by the developers of BLEU metric. More detailed various researchers defined similarly the grams as follows: "1-gram means that a word is considered as unit of measure, 2-grams means that two words are considered units of measures and so on" (Callison-Burch et al., 2006; Papineni et al., 2002, as cited in Dobre, 2013a).

METEOR developers, according to Banerjee and Lavie (2005), have designed the metric in a way that was possible to perform an "alignment" (p.117) of the candidate text and reference text. The alignment is defined by Banerjee and Lavie (2005) "as a mapping between unigrams, such that every unigram in each string maps to zero or one unigram in the other string, and to no unigrams in the same string. Thus in a given alignment, a single unigram in one string cannot map to more than one unigram in the other string" (p.117). The alignment is realized in two phases. In the first phase are listed all possible unigram (1-gram) mappings between the candidate text and reference text, this being realized by an external module. As a general example, if a word is identified once in the candidate text and twice in the reference text than the external module will list "two possible unigram mappings" (Banerjee and Lavie, 2005, p.117), one for each appearance of the word. However, the metric will map exactly the word and will not map derivates of that word with the original word. In the second phase, the unigrams mapped are collected in subsets and largest subset it is selected as the resulting alignment set, which means according to Banerjee and Lavie (2005) that "each unigram must map to at most one unigram in the other string" (p.118). It is possible that there will be more than one subset which could be selected as alignment. In such case the metric will select as resulting set the one having "the least number of unigram mapping crosses" (Banerjee and Lavie, 2005, p.118).

Banerjee and Lavie (2005) also considered that "two unigram mappings […] are said to cross if and only if the following formula evaluates to a negative number" (p.118). The formula referenced is presented as equation (1):

$$(pos(t_i)-pos(t_k))*(pos(r_j)-pos(r_l))_{(1)}$$

where: ti and rj are one unigram, tk and rl are another unigram, pos(tx) is the numeric position of the unigram tx in the candidate text and pos(ry) is the numeric position of the unigram ry in the reference text (Banerjee and Lavie, 2005, p.118).

During each stage, according to Banerjee and Lavie (2005), the metric will map only those unigrams "that have not been mapped to any unigram in any of the preceding stages" (p.118). After "all the stages have been run and a final alignment has been produced between" (p.118) the candidate text and the reference text, the metric will score as follows:

(a) The first unigram precision, noted with P, is a ratio calculated between the number of unigrams from the candidate text that are mapped to the unigrams from the reference text, and the total number of unigrams from the candidate text, (b) Same way, the unigram recall, noted with R, is a ratio of the number of unigrams from the candidate text that are mapped to the unigrams from the reference text, and the total number of unigrams from the reference text, (c) As per Banerjee and Lavie (2005), the third step is to calculate the Fmean "by combining the precision and recall via a harmonic-mean (van Rijsbergen, 1979) that places most of the weight on recall" (p.118). Therefore, was developed equation (2) by using the "harmonic mean of P and 9R" (p.118):

$$F_{mean} = \frac{10PR}{R+9P} \quad (2)$$

Unigram precision P, unigram recall R and Fmean "are based on unigram matches. To take into account longer matches, METEOR is computing a Penalty for a given alignment" (Banerjee and Lavie, 2005, p.118). In order to calculate the Penalty, according to Banerjee and Lavie (2005), METEOR metric will group all unigrams from the candidate text which are mapped to unigrams from the reference text, "into the fewest possible number of chunks such that the unigrams in each chunk are in adjacent positions" (p.118) in the candidate text, "and are and are also mapped to unigrams that are in adjacent positions" (p.118) in the reference text. Same researchers stated that "the longer the n-grams, the fewer the chunks, and in the extreme case where the entire system translation string matches the reference translation there is only one chunk" and, vice-versa, "if there are no bigram or longer matches, there are as many chunks as there are unigram matches" (p.118). The Penalty was calculated by Banerjee and Lavie (2005) using equation (3) below:

$$Penalty = 0.5 * \left(\frac{\#chunks}{\#unigrams\_matched} \right)^3 \quad (3)$$

At the end, the METEOR metric will calculate the Score by using equation (4) below (Banerjee and Lavie, 2005, p.119):

$$Score = F_{mean} * (1 - Penalty) \quad (4)$$

The METEOR developers explained that "this has the effect of reducing the Fmean by the maximum of 50% if there are no bigram or longer matches" (Banerjee and Lavie, 2005, p.119). For a single candidate text, METEOR will calculate the Score as per above equations for each reference txt which is available and after will select the best score as the final score. Overall Score for a candidate text it is calculated by METEOR in the same way like other metrics do (i.e., BLEU), "based on aggregate statistics accumulated over the entire test set" (Banerjee and Lavie, 2005, p.119). To calculate the aggregate score, same equations (1), (2), (3), and (4) shall be used in order to obtain the aggregate P, aggregate R, aggregate Penalty (Banerjee and Lavie, 2005).

Another aspect which it's considered very important by the author of this paper is the superiority of METEOR metric comparing to results obtained by using BLEU or NIST metrics. The author considers sufficient to highlight the "human/METEOR correlation" value as presented in (Banerjee and Lavie, 2005, p.119) and which was obtained based on the evaluation performed by the developers of the METEOR metric. The value obtained was 0.964 while "human/BLEU correlation" value obtained was 0.817 (Banerjee and Lavie, 2005, p.119, table 1).

Student's Knowledge Assessment Using METEOR Metric

Proposed ITSs Structure – Brief Overview

The author of this paper has presented more in detail the proposed ITS structure in previous papers (Dobre, 2013a; Dobre, 2014a; Dobre, 2014c), therefore, for the purpose of this article will only briefly mention few important aspects which bridge the proposed specialized ITS with the personal contribution of the author, the students knowledge assessment system, using METEOR metric and which has been incorporated in the Tutoring module. Proposed ITS structure is the same like the classic ITSs structure. The structure used by the author of this paper is presented in figure 2 and was adapted from Nwana (1990).

The conventional four modules: Domain module (contains the definitions, rules, concepts etc.), Student module (collects, selects, makes available various information about students), Tutoring Module (encloses the teaching strategies, the courses to be taught - learnt, the students knowledge assessment system etc.) and Communication Module (provides the communication tools between the parties involved in the use of the ITS), have been designed and developed to cover the teaching, learning & assessment cycle for the course of Computers Programming and C Language (Dobre, 2013a; Dobre, 2014a; Dobre, 2014c).

The proposed ITS was developed using free public license JavaFX technology and for the data bases management has been used MySQL (Dobre, 2013b; Dobre, 2014c). Also, apart these tools were used the METEOR metric together with word-to-word matching NLP technique and the Morphosyntactic tagger Web Service located on the server of Research Institute for Artificial Intelligence "Mihai Drăgănescu" (RACAI), Romanian Academy (Dobre, 2013a; Dobre, 2014a; Dobre, 2014b; Dobre, 2014c, and Tufiş, Ion, Ceauşu, and Ştefănescu, 2008a; Tufiş et al., 2008b; Ion, 2007; Morphosyntactic tagger Web Service from RACAI, n.d., referenced in Dobre, 2014b). For the presentation of the lessons part of the course of Computers Programming and C Language has been used the pdf format which could be easily accessed through Adobe Reader (free public license software). Also, the author considers important to highlight that the ITS it's developed in Romanian language, and all features, documentation etc. are available exclusively in Romanian language.

Proposed Students' Knowledge Assessment System

Initially, the author used the BLUE algorithm to develop the assessment module but after a series of "in-house" tests have identified a series of issues. Therefore, the author decided to review the assessment module and to use the METEOR metric as for students' knowledge assessment (Dobre 2013a; Dobre 2014b; Dobre, 2014c). The metric has been accessed from the address provided by Denkowski and Lavie (2011) and used as provided. No deviations of the METEOR metric have been used. The version used was version 1.4, this version being available in several languages as follows: English, Arabic, Czech, Danish, Dutch, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Romanian, Russian, Spanish, Swedish, and Turkish (Denkowski and Lavie, 2011).

Students' assessment is performed through tests which could be accessed by students after they complete the teaching cycle for each lesson (chapter) made available by Administrator. Each test has a set of questions and a certain time (established by Administrator) for student to answer all questions. The questions are text type and the students answers have to be provided under a text form. The calculations are performed separated, for each question, and at the end through equation (5) it is calculated the final score.

To calculate the score obtained by a student to one question, noted Intr, from any test available in the system, the following algorithm, noted Alg1, steps will be taken by the proposed system:

Step 1 – The student answer text (the candidate text) to the question Intr it is processed using the RACAI Web service for the diacritics and capital letters restoration, if applicable, this service being accessed online, by the proposed ITS, at the RACAI server (RACAI Web service for the diacritics and capital letters restoration, n.d.).

Step 2 – The result obtained through Step 1 it is processed using the Morphosyntactic tagger Web Service located on the RACAI server (Tufiş et al., 2008a; Tufiş et al., 2008b; Ion, 2007; Morphosyntactic tagger Web Service from RACAI, n.d.). Using the RACAI service, the text it is divided firstly at phrase level and after at word level. Afterwards, the words are morpho-syntactic annotated and for each annotated word it is obtained the word lemma.

Step 3 – It is applied the Step 1 to the reference text which is the correct answer uploaded previously by the Administrator to the question Intr, and using the RACAI Web service for the diacritics and capital letters restoration, with the same scope to restore the diacritics and capital letters as applicable (RACAI Web service for the diacritics and capital letters restoration, n.d.).

Step 4 – It is applied Step 2 for the text processed through Step 3, text pertaining to the reference text (correct answer to the question Intr). For the reference text it is obtained each text word lemma.

Step 5 – It is accessed online the Web service METEOR (Denkowski and Lavie, 2011) and the service will consider: e1) the candidate text – the text consisting form the succession of lemmas obtained at Step 2, and e2) the reference text – the text consisting form the succession of lemmas obtained at Step 4.

Step 6 – Using below equation (5) it is calculated the score for question Intr:

$$Scor_{Intr} = 10 * ScorMeteor \quad (5)$$

To calculate the score obtained by a student to a test, noted Test, and which contains a number of questions noted NrIntrTest, has been defined and implemented the below steps of the algorithm noted Alg2:

Step 1 – Attribute value zero to test score, than will have equation (6):

$$Scor_{Test} = 0 \quad (6)$$

Step 2 – For I = 1, NrIntrTest,, +1, execute equation (7):

Attribute $Scor_{Test} = Scor_{Test} + Scor_i$ (7)

where, Scori is the score obtained for the question i applying the algorithm Alg1.

The score obtained through the METEOR metric is a number in the range 0 to 1. Taking into consideration this, the score obtained by the student to one question (result obtained using the algorithm Alg1, at Step 4), is a number in the range 0 to 10. Also, if the number of questions part of test, noted NrIntrTest, is 10, by applying algorithm Alg2 for a test could be obtained a score which is a number in the range 0 to 100. Thus, the student could consider that passed a chapter (lesson) and could go to the next one only if the final score obtained for a chapter is a number in the range 50 to 100. Contrary, the student has to retake the learning cycle of the chapter with the testing phase failed.

Conclusions

The ITSs have proved that could be a reliable tool for improving the performances achieved by all parties involved in the educational process. This paper is presenting a part of the results obtained by the author during the research performed in the domain of ITSs using NLP. The proposed assessment system, a system using the NLP algorithms and applicable for assessing the students' knowledge achieved by them at different stages of teaching-learning and at the end of teaching-learning cycle for a discipline from Informatics domain, it is the personnel contribution of the author of this paper. The author considers that students' knowledge assessment systems using NLP algorithms offer several advantages (i.e., assessment objectivity, teaching-learning cycle centered on the students etc.) against the traditional assessment methods, and which advantages can be used not only for the classic translation machines, dictionaries etc. but also for other educational domains. At present, tutoring system, including the assessment system proposed by the author, it's under in-house evaluation and tests are currently carried out in order to compare the results obtained using METEOR metrics with the ones obtained in a previous research with the BLEU algorithm.

Future work will refer to the capability of the system to offer the services of both NLP algorithms, METEOR and BLEU, within the same ITS, than the Administrator will have the option to choose which one will be used as well as will refer to the possibility to have the proposed system available for use online from any server.

References

Agarwal, A., Lavie, A. (2008). Meteor, m-bleu and m-ter: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output. In Proceedings of the Third Workshop on Statistical Machine Translations, Columbus, Ohio, USA, 115-118.

Anderson, J. R., Farrell, R., Sauers, R. (1984). Learning to program in LISP. Cognitive Science, 8, 87-129.

Banerjee, S., Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan, USA, 65-72.

Bloom, B. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13, 3-16.

Bonar, J., Cunningham, R., Beatty, M., Weil, W. (1988). Bridge: Intelligent tutoring system with intermediate representations. Technical Report, PA: University of Pittsburgh, Learning Research & Development Center, Pittsburgh, USA.

Boonthum, C., Levinstein, I. B., McNamara, D.S., Magliano, J., Millis, K. K. (2009). NLP Techniques in Intelligent Tutoring Systems. Encyclopedia of Artificial Intelligence. IGI Global, Retrieved March, 2014, from http://129.219.222.66:8080/SoletlabWeb/pdf/NLP_Techniques_in_Intelligent_Tutoring_Sy.pdf

Callison-Burch, C., Osborne, M., Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In 11th Conference of the European Chapter of the Association for Computational Linguistic Proceedings of the Conference, EACL 2006, The Association for Computer Linguistic, Trento, Italy, 249-256.

Denkowski, M., Lavie, A. (2011). METEOR – Automatic Machine Translation Evaluation System. On Dr. Alon Lavie Home page, Retrieved March, 2014, from: https://www.cs.cmu.edu/~alavie/METEOR/.

Dobre, I. (2013a). The BLEU Algorithm Utilization for the Students Knowledge Evaluation Under an Intelligent Tutoring System. In Proceedings of the 8th International Conference on Virtual Learning ICVL 2013, Bucharest, Romania, 60-65.

Dobre, I. (2013b). The Design of an Intelligent Tutoring System Using the Natural Language Processing Technologies. In Proceedings of the 8th International Conference on Virtual Learning ICVL 2013, Bucharest, Romania, 73-79.

Dobre, I. (2014a). An Intelligent Tutoring System for Tutoring the Computers Programming and C Language Discipline. In Proceedings of the10th International Scientific Conference eLearning and software for Education, eLSE 2014, Bucharest, Romania, 142-149.

Dobre, I. (2014b). Students' Knowledge Assessment through an Intelligent Tutoring System Using Natural Language Processing based on an Automatic System for Generating Questions. In Proceedings of the10th International Scientific Conference eLearning and software for Education, eLSE 2014, Bucharest, Romania, 150-155.

Dobre, I. (2014c). Sistem de E-learning utilizând tehnologii de prelucrare a limbajului natural. PhD report no.3, Doctoral Advisor Prof. Acad. Dan Tufiş, Institute of Research for Artificial Intelligence. Romanian Academy, Bucharest, Romania.

Ion, R. (2007). Metode de dezambiguizare semantică automată. Aplicaţii pentru limbile engleză şi română. Institute of Research for Artificial Intelligence. Romanian Academy, Bucharest Retrieved March, 2013, from http://www.racai.ro/media/radu-ion-tezadoc.pdf.

Landauer, T. K., Foltz, P.W., Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.

Lavie, A., Sagae, K., Jayaraman, S. (2004). The Significance of Recall in Automatic Metrics for MT Evaluation. In Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004), DC, Whashington, USA, 134-143.

Leacock C., Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. Computers and the Humanities, 37(4), 389-405.

Lesgold, A., Lajoie, S.P., Brunzo, M., Eggan, G. (1990). A coached practice environment for an electronics troubleshooting job. In J. Larkin, R. Chabay & C. Sheftic (Eds.), Computer-assisted instruction and intelligent tutoring systems: Establishing communication and collaboration. Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates.

McNamara, D. S., Levinstein, I. B., Boonthum, C. (2004). iSTART: Interactive Strategy Trainer for Active Reading and Thinking. Behavioral Research Methods, Instruments, and Computers, 36, 222-233.

Molla, D., Schwitter, R., Rinaldi, R., Dowdall, J., Hess, M. (2003). ExtrAns: Extracting Answers from Technical Texts. IEEE Inteligent System, 18(4), 12-17.

Nwana, H. S. (1990). Intelligent Tutoring Systems: an overview. Artificial Intelligence Review, 4, 251-277.

Papineni, K., Roukos, S., Ward, T., Zhu, J. (2002). Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistic (ACL), Philadelphia, USA, 311-318.

Shute, V. J., Glaser, R. (1991a). An intelligent tutoring system for exploring principles of economics. In R. E. Snow & D. Wiley (Eds.), Improving inquiry in social science: A volume in honor of Lee J. Cronbach. Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates.

Shute, V. L. (1991b). Meta-Evaluation of Four Intelligent Tutoring Systems: Promises and Products – Final Technical Paper for Period June 1990 – July 1991. Armstrong Laboratory, Human Resources Directorate, Manpower and Personnel Research Division, Brooks Air Force Base, Texas, USA.

Steenbergen-Hu, S., Cooper, H. (2014). A Meta-Analysis of the Effectiveness of Intelligent Tutoring Systems on College Students' Academic Learning. Journal of Educational Psychology, 106(2), 331-347, doi: 10.1037/a0034752

Steyvers, M., Griffiths, T. (2007). Probabilistic Topic Models. In T. Landauer, D. S. McNamara, S. Dennis & W. Kintsch (Eds.), Handbook of Latent Semantic Analysis, Mahwah, NJ: Erlbaum, New Jersey, USA.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2008a). RACAI's Linguistic Web Services. In Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008. ELRA - European Language Resources Association, Marrakech, Morocco.

Tufiş, D., Ştefănescu, D., Ion, R., Ceauşu, A. (2008b). RACAI's Question Answering System at QA@CLEF 2007. In Carol Peters et al. (Eds.), Advances in Multilingual and Multimodal Information Retrieval. CLEF 2007, Vol. 5152 of Lecture Notes in Computer Science, Springler-Verlag Publisher, 3284-3291.

van Rijsbergen, C. (1979). Information Retrieval. 2nd ed., London, England: Butterworths.

Waterman, M. A., Matlin, K. S., D'Amore, P. A. (1993). Using Cases for Teaching and Learning in the Life Sciences: An Example from Cell Biology. Coalition for Education in the Life Sciences. Woods Hole, Massachusetts, USA.

Woolf, B. P. (2009). Building Intelligent Interactive Tutors – Student centered strategies for revolutionizing the e-learning. Burlington, Massachusetts, USA: Morgan Kaufmann Publishers an imprint of Elsevier Inc.

METEOR. (n.a.; n.d.). In Wikipedia, Retrieved November, 2013, from http://en.wikipedia.org/wiki/METEOR

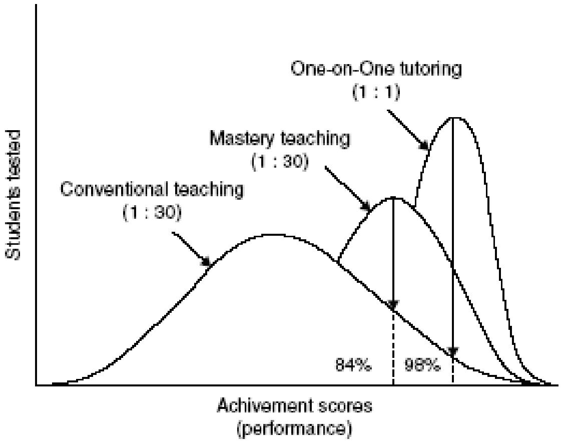
Morphosyntactic tagger Web Service from RACAI. (n.d.). Retrieved October, 2013, from http://www.racai.ro/webservices/TextProcessing.aspx

Natural Language Processing. (n.a.). (2014). In Wikipedia, Retrieved April, 2014, from http://en.wikipedia.org/wiki/Natural_language_processing

RACAI Web service for the diacritics and capital letters restoration. (n.d.). Retrieved October, 2013, from http://dev.racai.ro/

Figures

Figure 1. Advantages of One-on-One Tutoring, as presented by Woolf (2009), adapted by Woolf from Bloom (1984) and reprinted by permission of Sage Publications Inc. in Woolf (2009, p.15)



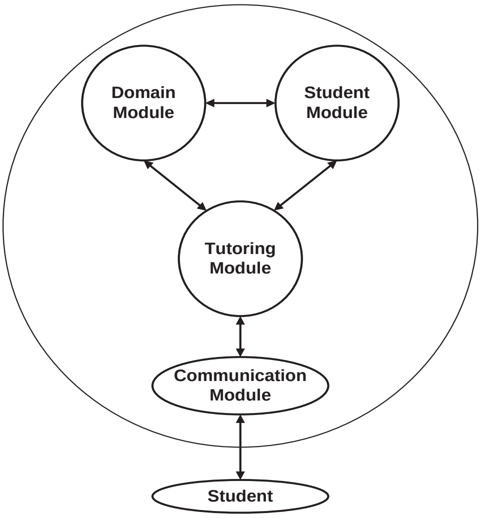


Figure 2. ITS general structure, adapted from Nwana (1990, p.257)

Abbreviations:

n.a. – no author,

n.d. – no date