Current Developments in Technology-Assisted Education: e-Learning standards - Learning Objects - Digital Games - National, Regional and Local experiences - Virtual presentations, Vol. III. Ed. A. Méndez-Vilas, A. Solano Martín, J.A. Mesa González and J. Mesa González. Badajoz: FORMATEX, 2006.
2007-2011

Automatic Identification of Terms for the Generation of Students' Concept Maps

D. Perez-Marin[*,1], **I. Pascual-Nieto**[1], **E. Alfonseca**[1,2] and **P. Rodriguez**[1]

[1] Universidad Autonoma de Madrid, C/ Francisco Tomás y Valiente 11, 20849, Madrid, Spain.
[2] Tokyo Institute of Technology, 4259 Nagatsuta Midori-ku Yokohama 226-8503 Japan.

Willow [1], an adaptive multilingual free-text Computer-Assisted Assessment system, automatically evaluates students' free-text answers given a set of correct ones. This paper presents an extension of the system in order to generate the students' concept maps while they are being assessed. To that aim, a new module for the automatic identification of the terms of a particular knowledge field has been created. It identifies and keeps track of the terms that are being used in the students' answers, and calculates a confidence score of the student's knowledge about each term. An empyrical evaluation using the students' real answers show that it is robust enough to generate a good set of terms from a very small set of answers.

Keywords automatic term identification; concept maps; e-assessment; e-learning

1. Introduction

Concept maps can be defined as visual illustrations displaying the organization of concepts and outlining the relationship among or between these concepts. Traditionally, teachers ask their students to draw their concept maps about a certain knowledge field. In this way, they can review how well the students understand these concepts. Moreover, they can find possible misconceptions by looking at how students have related the concepts [2]. Despite their seeming usefulness, concept maps are not yet a common representational media, and they are not used extensively in the classrooms. This could be due to the fact that it is time consuming to learn how to create them, and they are difficult to manage in paper [3-4].

Therefore, it would be interesting to automate the generation of the students' concept maps. As we show later, this can be done from the students' answers to a free-text Computer Assisted Assessment (CAA) system [5] such as Willow [6]. In order to build this concept map, the identification of the most important concepts in the students' answers is a necessary first step.

A term is usually defined as a word or a multi-word expression that is used in specific domains with a specific meaning. Term extraction is an important problem in the Natural Language Processing (NLP) area [7]. Proposed solutions to term extraction usually analyse large collection of domain-specific texts and compare them to general-purpose text, in order to find domain-specific regularities that indicate that a particular word or multi-word expression is a relevant term in that domain. Term candidates are usually returned ranked according to some specific metric or weight that indicates its relevancy. In this work we focus on nominal terms (nouns or multi-word noun phrases), and do not consider domain-specific verbs. Therefore, throughout this paper, the word "term" is used to refer to nominal terms only.

Several techniques have been devised to identify and extract the terms of a text:

1. Statistical corpus-based approaches such as in [8,9].
2. Linguistic processing techniques such as part-of-speech patterns, or the use of parsers [10,11].
3. Hybrid approaches which combine statistical techniques and linguistic knowledge [12,13].

Concepts are usually labeled by terms [14] and a traditional procedure to choose them was by consulting a group of experts or assessors [15]. However, there are some critics to this approach, as leaving the decision to humans make it subjective [16] and two humans tend not to agree completely.

Up to our knowledge, no previous attempt before this article has been done to use NLP techniques to automatically extract the terms for generating concept maps for educational purposes. This would be

* Corresponding author: e-mail: diana.perez@uam.es, Phone: +34 91 497 22 67, Fax: +34 91 497 22 35

interesting to make the procedure more objective and to free the teachers of the additional task of having to identify these terms by themselves. Our approach has been to build a module that uses a statistical corpus-based technique to identify the terms (main concepts) of a particular knowledge field. This module is inside a multilingual on-line system called Willow [1], whose purpose is the automatic evaluation of students' free-text answers given a set of correct ones (references). The references are introduced by the teacher the course, using an authoring tool [17].

This paper is organised as follows: Section 2 presents a general introduction to Willow; Section 3 focuses on the new module for automatic identification of terms and the evaluation carried out to analyze its performance; finally, Section 4 ends with the main conclusions and lines of future work.

2. Willow

Willow is an on-line application that automatically evaluates students' answers written in natural language. It is **multilingual** in the sense that it is able to process English and Spanish texts, and even evaluate answers written in one language by comparing them to references written in the other language. The necessary linguistic processing is performed using the wraetlic toolkit [18]. It is also **adaptive**, because the questions shown to the students depend on their varying user models. The main aim of Willow is to engage the students in an interactive set of questions so that they can get more training before their final exams. The goal is not to substitute the teacher and the traditional exam but to complement it as a double-scorer and a supplier of extra exercises with instant feedback.

The core idea of Willow is to compare the student's answer with the references so that the more similar they are, the higher the score the student achieves. In order to compare the texts and taking into account the problem of paraphrasing (many different ways to express the same information) several NLP techniques have been implemented. Moreover, the student's answer is not only compared with one reference but with several written by different teachers.

The user profile kept by Willow in order to adapt the assessment includes a static and a dynamic component. The static features, which are based on stereotypes, include information about the students' age, mother tongue and level of experience [19]. The dynamic adaptation procedure keeps track of how well the students are answering the questions. Its purpose is to adjust the level of difficulty of the questions to the students' level of knowledge [1].

As can be expected, the system provides different feedback to the students and to the teachers. For students, the feedback includes a numeric score, the student's processed answer where the fragments that most contributed to the score are highlighted, and the teachers' reference answers to which the student answer has been contrasted (see Figure 1 left). This feedback is provided in a gradual manner, with the aim that the student has to think properly the answers before seeing the correct references. Following this line, a set of questions have been implemented to guide the students to the correct answer without directly presenting it to them the first time they fail in answering a question [20].

On the other hand, the teacher receives as feedback a graphical representation of the students' conceptual model as a concept map (see Figure 1 right) [6]. Willow keeps track of the terms used by the teachers in the references to see if they are being used correctly by the students. Iniatially, a zero confidence value is assigned to each student's term because Willow ignores how well the students know those concepts. As the students answer more questions, the frequency of use and how they are related among themselves is registered and the confidence values are updated.

Finally, the concepts are presented with a two-colour schema in which the background colour indicates the type of concept (basic or more complex as it groups several basic concepts) and the foreground colour represents graphically the confidence-value (red for values lower than 0.4, yellow for values between 0.4 and 0.6 and green for values higher than 0.6). Moreover, they are linked according to membership relationships and other type of links found in the students' answers. In this way, the teacher can identify where students' misconceptions are and which concepts should be reviewed.

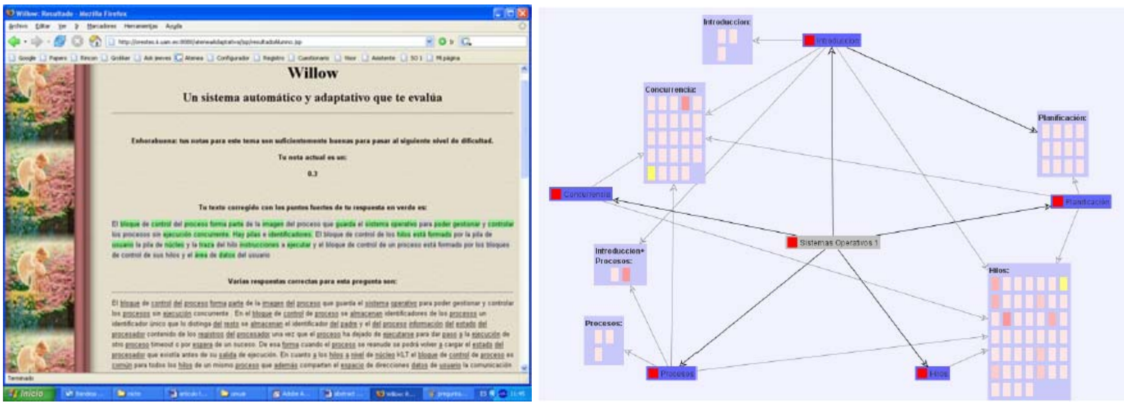

Fig. 1 Left: snapshot of Willow’s feedback page. Right: example of a generated student’s concept map

3. Willow’s Term Identification module

Our approach treats the problem of extracting terms as a classification task. We define that a term can be a single word (unigram), a sequence of two words (bigram) or a sequence of three words (trigram). Thus, each n-gram found in the text, with n varying from 1 to 3, can be classified as either being a term (class 1) or not (class -1). The input of the module is a text (usually a reference or several) and the output is the list of terms automatically found.

The approach followed is inially based on [21]. As in that work, the C4.5 algorithm is used to learn a decision tree [22]. Due to its statistical nature, this algorithm has the advantage of being equally applicable to the two languages Willow currently processes, Spanish and English.

In our experiments, the decision tree has been trained using the set of references of the questions about Operating Systems stored in Willow’s database. This contains 4617 words in Spanish, and 4636 in the English version. In order to find the domain-specific terms, this domain-specific corpus is compared to a generic corpus. In the case of Willow, there was the choice of using a general-purpose corpus (e.g. the British National Corpus). However, we finally opted for collecting a corpus of news on Computer Science. This is because the answer sets used in the experiment belong to very specific domains inside Computer Science (e.g. Operating Systems), so we would like the system to skip general computing terms and to extract only terms belong to those particular sub-fields. The collected corpora contain 50.823 words for Spanish and 157.340 words for English.

Three different human annotators reviewed the specific corpora (references) by hand to build a gold standard. The criteria agreed to determine that a certain n-gram was a term was that it was specific to the domain and that it was a noun or a noun phrase. Afterward, the terms that appeared in the three lists were automatically incorporated in the gold standard, and the three annotators discussed about the discrepancies until an agreement was reached in each case. In this way, a final agreed list was created as gold standard with 76 terms for Spanish and 89 for English.

The metric chosen to measure the goodness of the procedure was the F-score:

$$\text{F-score} = \frac{(\beta^2 + 1)\ \text{x}\ p\ \text{x}\ r}{\beta^2\ \text{x}\ p + r} \quad \text{where} \quad p = \frac{\text{no. of correct terms found}}{\text{number of terms found}} \quad \text{and} \quad r = \frac{\text{no. of correct terms found}}{\text{total number of terms}}$$

where p indicates the precision and r, the recall.

For the learning phase, the samples were chosen so that the distribution of classes were balanced (50% terms and 50% non-terms). Regarding the features considered as attributes, they were the relative frequency of appearance of the term in a corpus of students’ answers with respect to its frequency in the generic corpus and the sequence of part-of-speech tags of the words (e.g. noun, verb, adjective, etc.). The reason of choosing these features was that they are related to the nature of which a term is:

- The relative frequency is important because terms tend to be specific to a certain knowledge field. Thus, words with a relative frequency (frequency in the specific corpus / frequency in the generic corpus) lower than 1 should be discarded as they are too common.
- The part-of-speech (pos) is relevant because a term is usually a noun or a simple multi-word noun phrase. In most of the cases, the syntactic structure of noun phrases is not as complex as that of a clause or a sentence, so it should be possible to characterise using with regular expressions on the pos tags. For instance, the sequence of tags "determiner+noun+adjective" covers noun phrases such as "the operating system". Thus, if a word is a finite-tense verb it will probably not be part of a nominal term.
- Moreover, by examining the list of terms in the gold standard, we observed that every single term in the corpus can be represented by the following regular expression: NC* NP* ADJ* PREP* NC2* NP2* (zero or more common names, proper names, adjectives, prepositions, more common names and more proper names). Thus, for each n-gram extracted from the corpus, it is matched to the previous regular expression, and each of the pos tags receives a weight equal to the number of words belonging to that class. The weights are later normalised so that they all add up 1. For instance, if the extracted n-gram is "algorithm of Dekker", Table 1 shows how it is matched to the regular expression, and the weight assigned to each of the six pos tags.

POS	NC	NP	ADJ	PREP	NC2	NP2
word	algorithm	--	--	of	--	Dekker
value	0.33	0	0	0.33	0	0.33

Table 1 Example of configuration of the pos attributes for the learning phase of C4.5

In this way, the results achieved after performing a 10-fold stratified cross-validation are as follows:

Spanish		English	
precision	59.74%	precision	66.00%
recall	98.26%	recall	86.01%
F-score	74.3%	F-score	74.69%

Table 2 Results of using C4.5 (70% learning phase or training, 30% test) to identify terms

It can be seen that, even using small corpora, we have been able to reach results that are higher than the results attained by other related systems such as [8] with 67.81% F-score for English.

4. Conclusions and future work

This paper shows a promising line of work that combines techniques from NLP and e-Learning to be able to extract terms for creating automatically a concept map representing the students knowledge. We believe it will be very fruitful in filling the gap between what is being taught and what actually students learn, so the teachers have simple representations of the knowledge their students are acquiring.

A fundamental first step in the automatic creation of the concept map is the identification of the most relevant terms in the area of knowledge being taught. To that aim, the C4.5 algorithm has been used to automatically identify terms in a very small set of reference answers written by the teachers for a free-text CAA system. If has been able to attain F-score of around 74% both for English and Spanish. In our experiments, recall is well above precision, which is appropriate given that the list of extracted terms is later reviewed manually by the teacher. Therefore, it is important that most of the relevant terms are identified, as the noise can be removed by the teacher during the manual review phase.

A relevant result of this work is that it has proved to be able to extract high-quality term candidates even though the domain-specific corpora are very small, of around four thousand words each. We believe that it is due to the fact that the reference answers as written by the teachers are very high-quality and focused texts, so a small amount of them provides a good amount of data for the identification.

Concerning future work, this experiment has to be integrated with automatic procedures to learn relationships between the extracted terms, and to infer the amount of knowledge each student has from his or her answers to the free-text CAA system, so that more complete concept models can be generated for each of the students in the class.

Acknowledgements This work has been sponsored by Spanish Ministry of Science and Technology, project number TIN2004-0314.

References

[1] D. Perez-Marin, E. Alfonseca and P. Rodríguez. On the Dynamic Adaptation of Computer Assisted Assessment of Free-Text Answers. In Adaptive Hypermedia and Adaptive Web-Based Systems: 4th International Conference (AH 2006), Dublin, Ireland. Editors: Vincent Wade, Helen Ashman and Barry Smith, Lecture Notes in Computer Science, volume 4018, Springer Berlin / Heidelberg, 2006.
[2] J. Novak and D. Gowin. Learning How to Learn. Cambridge, U.K.: Cambridge University Press, 1984.
[3] R. Kremer. Concept Mapping: Informal to Formal. In: Proceedings of the International Conference on Conceptual Structures (ICCS). Maryland, U.S.A., 1994.
[4] L. Hsu, R. Edd, S. Hsieh, and R. Msn. Concept Maps as an Assessment Tool in a Nursing Course. Journal of Professional Nursing 21(3), 141–149, 2005.
[5] S. Valenti, F. Neri, and A. Cucchiarelli. An Overview of Current Research on Automated Essay Grading. Journal of Information Technology Education 2, 319–330, 2003.
[6] D. Perez-Marin, E. Alfonseca, M. Freire, P. Rodriguez, J.M. Guirao and A. Moreno-Sandoval. Automatic Generation of Students' Conceptual Models underpinned by Free-Text Adaptive Computer Assisted Assessment. In the Proceedings of the International Conference on Advanced Learning Technologies, ICALT, Kerkrade, The Netherlands, July 2006.
[7] M.T. Cabré, R. Estopá and J. Vivaldi. Automatic term detection: a review of current systems. In Recent advances in computational terminology, vol. 2 of NLP, pp. 53-87. John Benjamins, 2001.
[8] P. Pantel and L. Dekang. A Statistical Corpus-Based Term Extractor. In Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence. London, UK, pp. 36-46, Springer-Verlag, 2001.
[9] P. Drouin. Term extraction using non-technical corpora as a point of leverage. In Terminology 9(1), 2003.
[10] D. Bourigault. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In Proceedings of the Fourteenth International Conference on Computational Linguistics-COLING 92, pp. 977-981. Nantes, 1992.
[11] A. Voutilainen. Nptool, a detector of English noun phrases. In Proceedings of the Workshop on Very Large Corpora, pp. 48-57. Columbus, 1993.
[12] J. Justeson and S. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering 3(2), 259–289, 1996.
[13] A. Maynard and S. Ananiadou. Identifying terms by their family and friends. In: Proceedings of COLING. Saarbrucken, Germany, pp. 530–536, 2000.
[14] C. Rovira. El editor de mapas conceptuales DigiDocMap y la norma Topic Maps. [on-line] http://www.hipertext.net/web/pag261.htm, 2005.
[15] M. Ruiz-Primo. Examining concept maps as an assessment tool. In: Concept Maps: Theory, Methodology, Technology. Proc. of the 1st International Conference on Concept Mapping. Spain, 2004.
[16] D. Leake, A. Maguitman, T. Reichherzer, A. Caas, M. Carvalho, M. Arguedas and T. Eskridge. Googling from a concept map: towards automatic concept-map-based query formation. In: Concept Maps: Theory, Methodology, Technology. Proceedings of the 1st International Conference on Concept Mapping.Spain, 2004.
[17] E. Alfonseca, R.M. Carro, M. Freire, A. Ortigosa, D. Perez and P. Rodriguez. Authoring of Adaptive Computer Assisted Assessment of Free-text Answers. Journal of Educational Technology and Society, Special Issue on Authoring of Adaptive Hypermedia, International Forum of Educational Technology & Society, ISSN 1176-3647. Volume 8, Issue 3, July 2005.
[18] E. Alfonseca, A. Moreno-Sandoval, J. M. Guirao and M. Ruiz-Casado. The wraetlic NLP suite. In proceedings of the Language Resources and Evaluation Conference, LREC-2006, Genoa, Italy.
[19] D. Perez and E. Alfonseca. Adapting the automatic assessment of free-text answers to the students. In Proceedings of the 9th international conference on CAA, Loughborough, UK, July 2005.
[20] D. Perez-Marin, E. Alfonseca and P. Rodriguez. A free-text scoring system that generates conceptual models of the students' knowledge with the aid of clarifying questions. In Proceedings of the 4th Workshop on Application of Semantic Web Technologies for Adaptive Educational Hypermedia, Dublin, Ireland, June, 2006
[21] A. Ballester, A. Martín Municio, F. Pardos, J. Porta-Zamorano, R. J. Ruiz-Ureña and F. Sánchez-León. Combining statistics on n-grams for automatic term recognition. In Proceedings of the Language Resources and Evaluation Conference, LREC-2002, Las Palmas, Spain.
[22] J.R. Quilan. C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann, 1993.