2009 Ninth IEEE International Conference on Advanced Learning Technologies

Concept Extraction from student essays, towards Concept Map Mining

Jorge Villalon, Member IEEE, and Rafael A. Calvo, Senior Member IEEE
School of Electrical & Information Engineering, University of Sydney
{villalon,rafa}@ee.usyd.edu.au

Abstract

This paper presents a new approach for automatic concept extraction, using grammatical parsers and Latent Semantic Analysis. The methodology is described, also the tool used to build the benchmarking corpus. The results obtained on student essays shows good inter-rater agreement and promising machine extraction performance. Concept extraction is the first step to automatically extract concept maps from student's essays or Concept Map Mining.

1. Introduction

Essays are an excellent reflection of students' understanding [1], and therefore widely used in secondary and tertiary education. Educational researchers have stated that writing is a task where higher cognitive functions, such as analysis and synthesis, are fully developed [1]. For this reason they represent substantial component of undergraduate and graduate education [2].

Thanks to the increased use of Learning Management Systems to submit and manage the essay assignments, a number of new educational applications are being envisioned. They address the high workload that essay assessment requires [3], plagiarism detection [4], etc. This project explores the creation of automatic concept maps [5], that allow students and teachers to visualize the essays in new ways.

Concept Map Mining (CMM), the automatic extraction of Concept Maps (CMs) from essays can surface student's understanding about a topic as structured information [6]. The CMM process can be broken into three steps: Concept Extraction, Relationship Extraction and Topology Extraction.

An algorithm for the automatic extraction of concepts from essays using grammatical parsers and Latent Semantic Analysis is presented. A benchmarking corpus is developed with an ad-hoc tool used to annotate a corpus of students' essays. Finally the inter-rater agreement is discussed and compared to the accuracy of the automatic concept extraction algorithm.

This paper is structured as follows: Section 1 explains in more detail CMM and CE, section 2 discusses previous work on CE and explains our algorithm, section 3 reports on the annotation of the corpus, section 4 present the analysis of the results and section 5 concludes.

2. Concept Map Mining and Concept Extraction

2.1. Concept Map Mining

CMM refers to the task of automatically generating a CM from an essay for educational purposes. Its aim is to extract a representation of the semantic information contained in the essay, that is to surface the understanding a student has about a topic [6].

CMM is contextualized in a learning scenario, in which the CMs are obtained from essays with different quality, and consumed by humans (teachers, tutors or the students themselves). This context demands particular requirements for the CMs, which are discussed in more detail in [6], two of these requirements affect directly the Concept Extraction task:

Simplicity: CMs must be the best possible summary of the complete essay. Therefore, information must not be redundant (synonyms and redundant propositions must be avoided), and information loss must be minimized.

Subjectivity: CMs represent the author's understanding about a topic and writing skills. Terminology used by the student is also important for assessing the outcome, so the CMs should be represented in the same way the author did, this is, using the same terms. Therefore, the words for the concepts and relations must be extracted literally from the document, and the hierarchy of concepts must

978-0-7695-3711-5/09 $25.00 © 2009 IEEE
DOI 10.1109/ICALT.2009.215

reflect the importance of the concepts relative to what was written in the particular document.

2.2. Concept Extraction

Concept extraction (CE) can be broken into identification of all possible concepts and selection of the most important ones (summarization). The essay (document) D contains all potential words (or phrases) that could become part of the CM, which formally comprises: Concepts (C), Relationships (R), and a Topology (G). We can formalize this by saying that the document contains a triplet $D \subset \{C_d, R_d, G_d\}$ where C_d corresponds to all the concepts, R_d corresponds to all the propositions, and G_d corresponds to the levels of generalization expressed in the essay.

According to this formalization Concept Identification corresponds to identify C_d from D, and Concept Summarization corresponds to filter C_d to form C. Figure 1 summarizes all the steps in the CMM process.

Mathematically Concept Identification is represented by equation 1, where C_d is the set of all n-grams (ng_i) identified in D, for which $C_{id}(ng_i)$ (the function that discriminates if an n-gram is a concept) is 1.

$$C_d = \{ng_i \in D / C_{id}(ng_i) = 1\}$$

Analogously, Concept summarization is expressed by equation 2, where C is the set of all concepts c_i in C_d, for which C_{rk}[8] (the function that ranks the concepts) is above a threshold α_c.

$$C = \{c_i \in C_d / C_{rk}(c_i) > \alpha_c\}$$

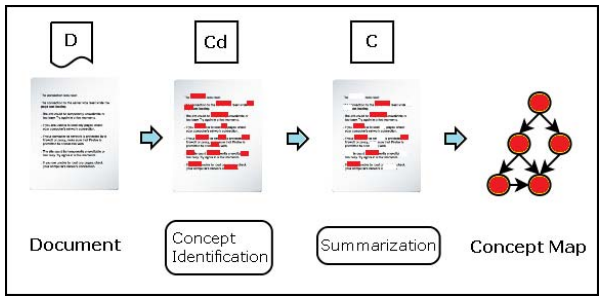

Figure 1: CE in the CMM process

3. Previous work on automatic concept extraction

3.1. Concept identification

Concept identification is a common task to several applications, however they use different definitions for concept, hence different methods.

Previous methods start from the idea that concepts can be found as word or phrases contained in sentences, which are then divided in smaller phrases in one of two ways: Using grammatical or syntactical information. The former can be found in ontology learning [9], glossary extraction [10] and information retrieval systems [11]. Sentences are divided in smaller phrases using a shallow grammar parser, these sub-phrases can be noun or verb phrases. The latter uses syntactical information like punctuation, conjunctions or list of non-semantic words to separate phrases within a sentence. This approach can be found in keyword extraction systems [12].

3.2. Summarization

According to Gong and Liu, "A generic summary provides an overall sense of the document's contents. A good generic summary should contain the main topics of the document, while keeping redundancy to a minimum" [13]. Generic summarization has two approaches to the identification of the main topics, and to avoid redundancy: Statistical and graph based [14]. The former uses the frequency of the phrases, sentences or paragraphs used for the summary as a way to decide which are the most relevant. The latter creates a graph of text passages, and then distances between them are calculated to create the connections between nodes. Finally, the weighted graph is used to identify the most central nodes and create the summary.

A more sophisticated version of the statistical idea, is the use of Latent Semantic Analysis (LSA) for the analysis of the topics in a document [13]. It is argued that a document consists of several topics, some of which are described intensively in several sentences, and hence form the major content of the document. By using LSA, salient topics or concept in the document can be identified in singular vectors, one sentence per vector was then retrieved. The performance was tested using a manually annotated corpus, achieving results around 60% of precision. The same approach was further explored by Steinberger et al. [15], modifying the algorithm by adding anaphora resolution previous to the analysis, improving the performance.

4. Our implementation

4.1. Concept Identification using grammar trees

As a first attempt to create an automatic concept extraction algorithm, we based our approach purely on our definition of CMM and CE. In this way, we took Novak's definition of concepts [5] as labels identifying events or objects, for concept identification, and the idea of a summary were information is maximized while redundancy is kept to a minimum.

As mentioned earlier, the concepts found in concept maps are associated to objects or events, therefore are nouns. We selected a grammatically based division for our first version of the algorithm, where noun phrases were identified and nouns and compound nouns were selected. To identify compound nouns, the Stanford parser was used to obtain a grammar tree, which is the grammatical analysis of a sentence. The actual nouns were identified using *tree regular expressions*.

4.2. Summarization using Latent Semantic Analysis

LSA is a statistical technique developed by Deerwester et al. [16] to improve document retrieval performance by overcoming two linguistic problems: Synonymy and polysemy. The former occurs when two words share the same meaning and the latter when a single word has more than one meaning. Using LSA, these meaningful relations can be learned from a corpus of background knowledge.

LSA follows a classic Text Mining process, documents are pre-processed, features are extracted, a model is created and then used for a particular application.

The pre-processing corresponds to the extraction of the terms from text passages. It starts with the tokenization which is the recognition of terms, symbols, urls, etc. Then a predefined set of terms that don't provide any meaning are removed, these are known as stop words.

Feature extraction identifies features that represent the objects in study, in this case text passages. In LSA terms, appearances are counted for each text passage and for the whole corpus. These values, known as term frequency (TF) and document frequency (DF) are used to create a more complex feature value called weight.

The creation of the model in LSA comprises three basic steps: First, it creates a term by text passage matrix A with values a_{ij} indicating the weight of the i^{th} term for the j^{th} text passage. Second, the matrix is decomposed using Singular Value Decomposition in three other matrices:

$$A = U * \sum * V^t$$

The columns from matrices U and V are orthogonal, and form a new base for the space, these are the eigenvectors. Matrix Σ contains the eigenvalues of the decomposition, sorted so the first value is the highest. Each eigenvalue represents the explained variance of its corresponding eigenvectors, in other words, if we only use the first eigenvector to represent all documents and terms, the separation of the vectors is maximal.

Reducing the dimensions of the space is the key step in LSA, the lower values in matrix Σ are set to 0, then multiplying back, an approximation of the original matrix is created. If we keep the k higher values in Σ we obtain:

$$A_k = U_k * \sum_k * V_k^t$$

The resultant model is a matrix representing terms and text passages. Using this vector representation of text passages and terms, distances can be calculated using the angle between the vectors. New text passages that are not contained in the corpus can be projected on the semantic in the same way.

To implement the summarization step for CE, a weight was assigned to each compound noun found in the concept identification step. According to previous LSA approaches to summarization, each eigenvector represents a topic in the document. Given that the eigenvectors are sorted by their explained variance, we selected those terms with the higher loads, that were also identified as nouns. Concepts were selected moving from each eigenvector until a maximum of 25 concepts was reached.

5. Benchmarking corpus creation

To create a benchmarking corpus for the Concept Extraction task within CMM, three aspects are important: Selection of the corpus, a methodology for the annotation, and an annotation tool.

5.1. Selection of the corpus

Our aim at this stage in our research is to identify the inherent complexities that CMM will present. We wanted to control factors affecting the quality of essays as much as possible. Therefore we restricted the corpus to a single genre, single topic, and similar length for all essays, and similarity among students.

We collected a corpus of 43 essays (with a total of 411 paragraphs and 18,388 words) written by

University of Sydney students as part of a diagnostic activity and marked by two tutors. The activity included the reading of three academic articles about English as a global language.

5.2. Methodology for the annotation

Novak proposed a method to construct good generic CMs [5], the basic procedure consist of, firstly, stating a good focus question, that will guide the next steps. Then, a list of the most relevant concepts must be created, this list must be ranked from the more general inclusive concepts to the more specific. Novak refers to this list as "the parking lot", from where concepts are taken and put in the concept map one by one, starting from the most general ones, and link each one to the previous concepts when needed to form good propositions. He explains that some concepts could be left in the parking lot if they don't provide new information to the map [5].

The methodology for the annotation of the corpus followed the same steps that Novak proposed, with two extra limitations: Concepts must use words or phrases that can be found literally in the essay, and propositions (a triple concept - linking word - concept) must be contained within a single paragraph.

5.3. Annotation tool

Figure 2 shows the interface for the CE, it has three parts: An essay display, a concepts list, and a CM box. The user is asked to select a list of concepts that express the knowledge elicited in the essay.

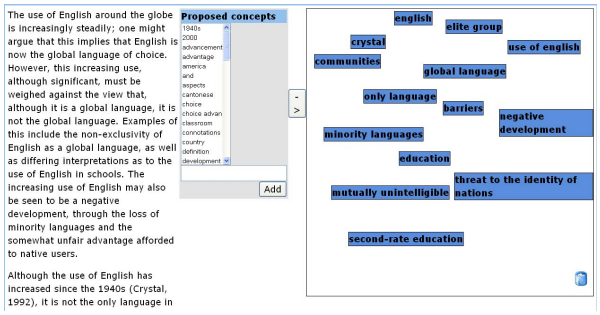

Figure 2: Concept Extraction interface of the annotation tool.

Annotators add concepts to the CM box from the list, until they decide that the knowledge elicited in the essay is covered. The essay display shows the document to ease the annotator's job. The concepts list, suggests concepts to the user, these were obtained using the concept identification method. The user can select multiple concepts from the list, and add them to the CM by clicking on the arrow pointing towards the CM box. In case the user wants to add new concepts that are not in the list, a text box with an "Add" button allows doing so. The tool will make sure that any new concept appears literally in the document, and will not add any concept that is not found in it. To delete a concept, the user has to drag its box and drop it in the garbage bin at the bottom right of the CM box.

6. Results

The markers identified an average of 12 concepts per essay, and the agreement between them was affected by two phenomena: Compound nouns and synonymia. The former corresponds to concepts that are contained lexically within another, an example is "economic and social inequalities" which was annotated by one marker as "inequalities" and as "social inequalities" by the other. The latter corresponds to the use of different words to refer to the same concept, where concepts like "world" were selected by one marker and "globe" by the other.

To overcome these problems we added three options to our comparison algorithm: Use of substrings, use of synonyms, and stemming. The use of substring was implemented using regular expressions and the synonyms were implemented using WordNet [17]. Stemming was implemented using Apache's Lucene Snowball analyzer, which uses the well known Porter's stemmer algorithm. Problems with substrings where found with phrases containing articles and/or preposition, e.g. "communities of nations", because they are common words to many phrases. To overcome this problem, articles and prepositions were not considered in the comparison for substrings.

Previous summarization studies used the standard precision performance measure from information retrieval, comparing inter-human and human to automatic agreement [13]. With S_{man} and S_{aut} the manual and automatic selection (set of concepts) respectively, precision is calculated as follows:

$$P = \frac{|S_{man} \cap S_{aut}|}{\min(|S_{aut}|, |S_{man}|)}$$

The average inter-human agreement for all 43 essays was 56.84%, which is low compared to agreement achieved on assessment tasks which is around 80% [3], but higher than more subjective tasks such as sentence selection for summarization, which is 40% [13]. Further analysis show that the biggest differences are caused by the selection of topics from the essay. Different topics are covered using up to three

concepts by the markers, hence choosing one topic over another, causes a strong difference between the annotations.

The agreement between our algorithm and the human markers was calculated separately, using the same method for the comparison between humans. A simple t test showed that there was no significant difference between the distances between the algorithm and each marker. Table 1 shows the averaged results.

Table 1: Results from the automatic CE task

	Inter-human	Compound nouns & LSA
Average	56.84%	39.58%

Although the automatic algorithm doesn't perform as well as humans, results are promising because highly subjective tasks such as summarization don't present high correlations, and this is our first implementation. A more detailed analysis showed that the minimum agreement was of 16%, this tells us that most essays show an obvious central topic, that the algorithm was able to select. However, as the coverage of knowledge is extended, the agreement between humans starts decreasing. This is also reflected in the automatic algorithm, because the number of concepts chosen by a human marker to cover a topic are not clear, and given that our algorithm was choosing only one concept per eigenvector, it will cover more topics, but in less detail.

7. Conclusion

We are working towards automatically extracting concept maps from essays. Work on the creation of a benchmarking corpus set was reported, the devised tool allowed human markers to reliably annotate the essays.

We also reported on the results of an algorithm for the automatic extraction of concepts, based on state of the art algorithms. The results showed that even though its performance is less than the inter-human agreement, it is performs as good as previous work on summarization. Further analysis showed that the performance is related to the way concepts are chosen by humans. We believe that understanding this phenomenon and using it for the automatic selection of concepts could lead to big improvements.

Finally, as future work, we plan to expand our corpus with new essays, implement a second version of the algorithm following the obtained ideas, and move forward to relationship extraction.

8. References

[1] J. Emig, "Writing as a Mode of Learning," *College Composition and Communication,* vol. 28, pp. 122--128, 1977.

[2] T. Moore and J. Morton, "Authenticity in the IELTS academic module writing text," *IELTS research reports,* 1999.

[3] T. Miller, "Essay Assessment with Latent Semantic Analysis," *Journal of Educational Computing Research,* vol. 29, pp. 495--512, 2003.

[4] H. Maurer, F. Kappe, and B. Zaka, "Plagiarism - A Survey," *Journal of Universal Computer Science,* vol. 12, pp. 1050--1084, 2006.

[5] J. D. Novak and D. B. Gowin, *Learning How To Learn*: Cambridge University Press, 1984.

[6] J. Villalon and R. Calvo, "Concept Map Mining: A definition and a framework for its evaluation," in *International Conference on Web Intelligence and Intelligent Agent Technology*, 2008.

[7] N.-S. Chen, P. Kinshuk, C.-W. Wei, and H.-J. Chen, "Mining e-Learning Domain Concept Map from Academic Articles," *Computers & Education,* vol. 50, pp. 694--698, 2008.

[8] J. J. G. Adeva and R. A. Calvo, "Mining Text with Pimiento," *IEEE Internet Computing,* vol. 10, pp. 27-35, 2006.

[9] R. Navigli and P. Velardi, "Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites," *Computational Linguistics,* vol. 30, pp. 151--179, 2004.

[10] D. Bourigault and C. Jacquemin, "Term extraction + term clustering: An integrated platform for computer-aided terminology," in *EACL*, 1999.

[11] I. Bichindaritz and S. Akkineni, "Concept Mining for Indexing Medical Literature," *Lecture Notes in Computer Science,* vol. 3587, pp. 682--692, 2005.

[12] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "KEA: practical automatic keyphrase extraction," in *Fourth ACM conference on Digital libraries*, 1999.

[13] Y. Gong and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," in *International conference on Research and development in information retrieval*, 2001, pp. 19--25.

[14] S. Afantenos, V. Karkaletsis, and P. Stamatopoulos, "Summarization from medical documents: a survey," *Artificial Intelligence In Medicine,* vol. 33, pp. 157--177, 2005.

[15] J. Steinberger, M. Poesio, M. A. Kabadjov, and K. Jezek, "Two uses of anaphora resolution in summarization," *Information Processing and Management,* vol. 43, pp. 1663-1680, 2007.

[16] S. Deerwester, S. Dumais, G. Furnas, and T. Landauer, "Indexing by latent semantic analysis," *Journal of the American Society For Information Science,* vol. 41, pp. 391-407, 1990.

[17] G. A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM,* vol. 38, pp. 39--41, 1995.