Semantic Comparison of Texts for Learning Environments

Patricia Gounon[1] and Benoît Lemaire[2]

[1] I.U.T. de Laval,
Département SRC
52, rue des docteurs Calmette et Guérin
BP 2045
53020 Laval Cedex 9
France
Tel: +33 2 43 59 49 23
Patricia.Gounon@univ-lemans.fr

[2] L.S.E.
University of Grenoble 2
BP 47
38040 Grenoble Cedex 9
France
Tel: +33 4 76 82 57 09
Benoit.Lemaire@upmf-grenoble.fr

Abstract

This paper presents a method for comparing a student essay and the text of a course. We first show that the comparison of complex semantic representations is better done with sub-symbolic formalisms than symbolic ones. Then we present a method which rely on Latent Semantic Analysis for representing the meaning of texts. We describe the implementation of an algorithm for partitionning the student essay into coherent segments before comparing it with the text of a course. We show that this pre-processing enhances the semantic comparison. An experiment was performed on 30 student essays. An interesting correlation between the teacher grades and our data was found. This method aims at being included in distance learning environments.

Keywords

AI in Education,
Free-text assessment,
Latent Semantic Analysis,
Semantic Representation

Track

Paper Track

Conference Topics

AI in Education and Intelligent Tutoring Systems,
Natural Language Processing

Semantic Comparison of Texts for Learning Environments

Patricia Gounon[1] and Benoît Lemaire[2]

[1] L.I.U.M., University of Maine, France
Patricia.Gounon@univ-lemans.fr
[2] L.S.E., University of Grenoble 2, France
Benoit.Lemaire@upmf-grenoble.fr

Abstract. This paper presents a method for comparing a student essay and the text of a course. We first show that the comparison of complex semantic representations is better done with sub-symbolic formalisms than symbolic ones. Then we present a method which rely on Latent Semantic Analysis for representing the meaning of texts. We describe the implementation of an algorithm for partitionning the student essay into coherent segments before comparing it with the text of a course. We show that this pre-processing enhances the semantic comparison. An experiment was performed on 30 student essays. An interesting correlation between the teacher grades and our data was found. This method aims at being included in distance learning environments.

1 Introduction

There is a huge demand nowadays for intelligent systems being able to assess student texts produced in distance learning environments. Students at a distance want assessments on the course they just work on. Multiple-choice questions can be found in most of the existing learning environments but their design is very time-consuming and they are quite rigid. Free text assessment is much more precise but require sophisticated AI techniques.

The specificity of the problem is that the student texts need to be assessed with respect to a domain, corresponding to a course. For instance, a student text about "the financial crash of 1929" need to be assessed with respect to a correct representation of that domain in order to detect missing information or, inversely, parts with too much details. One could imagine to assess student texts with respect to an ontology or any symbolic description of the domain but very few domains are represented in such formalisms. Most of the courses being taught are represented as... texts. Therefore, the challenge is to compare a text to another text. Since the phrasing will not be the same in both texts, the comparison need to be performed at the semantic level.

One way to achieve that goal is to automatically transform each text, the student text and the reference text, into a semantic representation and to compare both representations. The question is: which knowledge representation formalisms to rely on? And which comparison method to use?

After describing various approaches of that problem, we briefly present a statistical model of semantic knowledge representation, called Latent Semantic Analysis (LSA). We relied on this model to partition a text into paragraphs. We then used this segmentation to design a method for assessing a student text. For each method, we performed an experiment with real student texts.

2 Knowledge Representation

2.1 Symbolic approaches

Artificial intelligence has produced many expressive languages for representing the meaning of words. Some of them are symbolic approaches, other can be called sub-symbolic. All the symbolic formalisms have basically the same structure: each concept is associated with a node which is linked to other nodes, thus describing the relations between nodes. Semantic networks [14] are a well-known example of such formalism. Description logics [2] rely also on this formalism while being more formal and rigourous. The main advantage of these approaches is that the representation is explicit: it is possible, even for a non-computer scientist, to understand the meaning of a node as well as its relations with other nodes. The drawback, however, is that the main part of the knowledge needs to be coded (or at least verified) by a human, even if a system can assist that task.

2.2 Latent Semantic Analysis

LSA is more a sub-symbolic approach since the knowledge representation is not so explicit. LSA represents the knowledge by assigning high-dimensional vectors to words and pieces of texts. To do so, LSA analyses huge corpus of texts divided into paragraphs. The underlying idea is that:

- two words are semantically similar if they appear in similar paragraphs;
- two paragraphs are semantically similar if they contain similar words.

This kind of mutual recursion is implemented by a singular value decomposition algorithm applied to a word-paragraph matrix. This huge matrix is then reduced to 300 dimensions. The aim of this paper is not to describe this algorithm which is presented in details elsewhere [3].

Once the whole corpus has been analyzed and all vectors created, it is straightforward to compare two words or two pieces of texts or a word and a piece of text at the semantic level. The measure of similarity is just the cosine of the angle between the corresponding vectors. Therefore, the similarity between two words or two set of words is a number between -1 (lowest similarity) and 1 (highest similarity). In spite of a lack of syntactical analysis, this measure of semantic similarity has been proven successful in various experiments [4, 5, 10]. Basically, if the corpus is big enough, LSA has performances on semantic judgement between pairs of words that compare with human ones. It is worth noting that the number of dimensions plays an important role. Experiments show that performances are maximal for dimensions around 300 [4].

We believe that such a representation of the meaning of words can be used for representing knowledge, provided that a knowledge source can be represented by a piece of text, that is a set of words. Therefore, a knowledge source can be represented by a vector in the high-dimensional space defined by LSA.

2.3 Comparison of both approaches

As we mentioned above, symbolic approaches of knowledge representation have the great advantage of being explicit. Therefore, various reasoning methods can rely on these representations. However, knowing whether one representation is semantically similar to another one is not obvious because this formalism is not intended for comparison but rather for description. Therefore, most of the methods used for comparing representations are based on surface features : number of common nodes, subsumption relations, etc. This is often the case in the field of machine learning or case-based reasoning. The representation is rich but the comparison is poor. In the other way, LSA representations are not suited for drawing inferences since they are not explicit. However, they are better at making semantic comparisons between entities. In other words, the symbolic approach make absolute representations while LSA provides relative representations. Our goal being to compare representations and not drawing inferences, LSA seemed to us an interesting model of knowledge representation.

Another advantage of LSA is that any new text can be given quickly a new representation, provided that the words are part of the initial corpus with a sufficient frequency. This is not exactly the case with symbolic representations: building a representation from a novel text might be difficult.

LSA is a fully automatic method: there is no need to code knowledge by hand. As we will show, our system can therefore be used in any domain, provided that there exists texts describing that domain.

3 Assessing a student essay with LSA

Assessing a student essay with LSA implies that a high-dimensional space is computed from a text describing the domain. Usually, this text is a course. Additionnal texts in the same domain might be required to ensure the accurracy. Each word of the domain will then be represented by a vector.

Several environments rely on LSA to assess a student essay. The essay is usually represented as a 300-dimensions vector. It is compared with other vectors representing several reference texts or parts of reference texts. The feedback to the student is usually composed of (1) a global assessment score; (2) an indication of the topic that are (or not) well covered by the essay;

Intelligent Essay Assessor (IEA) [7, 8] is a system which is based on reference texts that are pre-graded essays. Two kinds of assessments are proposed:

- an holistic score corresponding to the score of the closest pre-graded essay;
- a gold standard score which relies on an expert essay.

An experiment with 188 student essays led to a correlation of 0.80 between IEA grades and human grades. However, this system provides no advice on the student essay, which is important for the student to improve the text.

Apex [11] performs a semantic comparison between the essay and the parts of the course previously marked as relevant by the teacher. The whole student essay is successively compared with each of these parts. For instance, if the student has to write a text from the question "What were the consequences of the financial crash of 1929?", the essay will be compared with the following sections of the course: *The political consequences*, *Unemployment and poverty*, *The economical effects*, etc. An experiment with 31 student essays in the domain of Sociology of Education led to a correlation of 0.51 between Apex grades and teacher grades.

Select-a-Kibitzer [15] automatically assesses a student essay and provides feedback on the text. The system is not intended to assess whether the student knows a domain, like in the previous approach. Its goal is rather to assess the task of text composition. Therefore, students are required to write a text on a topic like: "if you could change something about school, what would you change?". Select-A-Kibitzer is based on reference texts that are prototypical sentences of what students usually say about school (food, teachers, school hours, etc.).

Despite interesting correlations with human scoring, these approaches suffer from not taking into account the semantic structure of the essay. Two essays that would have the same sentences but organized differently would get the exact same score. This is not acceptable since it is known that the way the essay is structured is an important predictor of the student comprehension of the domain. Moreover, relying on the essay structure would enhance the feedback to the student by providing a more precise assessment. In particular, the student would be advised if one part of the essay seems to cover several topics.

The goal is therefore to go from a n-to-1 comparison to a n-to-m comparison (cf. Fig 1). We need to segment the essay into coherent parts. One way would be to identify each carriage return indicating the end of a paragraph. However, this approach is not fully reliable since students do not usually segment properly their essays, especially if they have trouble to organize what they know about the domain. Therefore, we need to segment student paragraphs from the content.

4 Partitioning a Text

4.1 Related work

Several methods were designed for partitioning texts into coherent units. A first set of approaches is based on the identification of term repetitions. The idea is that there is a lexical cohesion within a unit. A new unit implies the use of new words. Therefore, term repetition should be an indicator of the lexical cohesion. Hearst [9] implemented such an algorithm and found interesting correlations with human judgements. Reynar [12] relied on a dotplotting algorithm for detecting lexical cohesion. Beefmann [1] also worked at the lexical level but implemented a feature selection algorithm, a method often used in machine learning, to detect

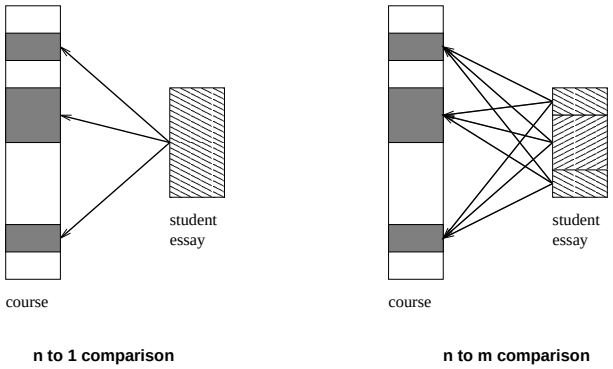

Fig. 1. *Two ways of comparing a course and a student essay.*

topic boundaries. Salton and Allan [13] relied on a statistical method for representing the similarities and the relations between paragraphs on a map. This technique allows the visualization of groups of similar paragraphs.

All of these approaches suffer from not being explicit. In particular, it is hard to explain to the student why topic breaks have been defined at positions X or Y. Moreover, these methods do not really work at the semantic level, which is what is required for student texts.

Another method was proposed by Foltz et al. [6]. It is based on the semantic comparison of adjacent sentences. The idea is that segment boundaries can be defined each time a low similarity is found between adjacent sentences since it should be an indication of a topic shift. Foltz et al. realized an experiment from a psychology textbook to determine whether LSA can detect automatically the ends of chapters. They found that LSA identified half of the ends of chapters.

4.2 Our method

This last method was applied to textbooks but we did not know whether it would work for student essays. So, we decided to implement and test this method. A high-dimensional space was then computed by LSA from the text of the course.

The first step to partition a text is therefore to make a comparison between all pairs of adjacent sentences. Each sentence needs to be represented by a vector. It is worth noting that this vector does not exist beforehand since the sentence did not usually appear in the corpus. LSA computes a new vector for each sentence by just adding the vectors of each word of the sentence. Given the vectors of two sentences, it is therefore possible to compute a semantic similarity between them. This measure returns a coefficient of similarity between -1 and 1. At the end, we have a sequence of coefficients that we use in a second step to identify local minima. Instead of just looking for a low coefficient, we compare each sequence of 3 consecutive coefficients to determine whether the second is lower than the previous and the next ones. A local minimum is an indication of a topic shift in the text. To test this method, we realized an experiment.

4.3 Experiment 1

The goal of this experiment is to compare the topic breaks found by our method with those defined by the students by means of the paragraph structure.

Procedure This first experiment is not concerned with comparing student essays with the course. Inputs are just student essays. However, we need to process the text of the course for computing the high-dimensional semantic space. We submitted LSA with 30 student essays in which all paragraph ends were deleted.

Results The goal of this test was to determine how LSA would partition a text. Although we know that the way students segment essays into paragraphs is far from being optimal, we decided to compare both information sources. We wanted to know whether LSA would find the same breaks as the student. The results concern in particular the number of correct and supplementary cuts. The percentage of error made by LSA was 60%. This means that LSA made several breaks that were not in the initial text. The score of correct breaks was thus 40%. We decided to analyse more precisely these results. Information retrieval researchers make use of several measures to determine the precision of a query: precision and recall. They are based on the following values, for each essay:

1. REL is the number of breaks determined by the student;
2. RET is the number of breaks determined by LSA;
3. RETREL is the number of correct breaks, which is defined here by the number of breaks found by both LSA and the student.

Precision corresponds to the number of common breaks in relation to the number of breaks determined by LSA:

$$precision = \frac{RETREL}{RET}$$

Recall corresponds to the number of common breaks in comparison with the number of breaks determined by the student:

$$recall = \frac{RETREL}{REL}$$

We found a recall of 41% as well as a precision of 34%. These results bring us to put several questions that we develop in the following section.

Comments The goal of this experiment was to determine whether LSA can retrieve the initial structure of a text. Results indicated that, although they are far better than random, they do not correspond well with the student breaks. The main reason is that the student texts were not expert texts: they were written by people who were in the process of learning a domain. We could not expect

them to provide ideal text. Since all student texts were real exam essays, it is also possible that some factors like stress or time pressure might have interfered.

It should be possible to enhance the method by using a local minimum threshold to increase the accurracy of the partitioning. When we compare three consecutive coefficients, the difference between two coefficients can be quite small. In that case, it is probably not a real topic shift. Therefore, another solution would be to indicate a coefficient of certitude for each paragraph after partitioning a text. It can be a coefficient indicating the degree of certitude of each paragraph. More work will be done in that direction in the near future.

5 Comparing Texts

The previous section presented a method for partitioning a text based on the LSA model. Now, the question is: how to take into account the structure of the essay that the previous method identified? In other words, how to assess an essay composed of several paragraphs?

5.1 How to assess a student essay composed of several paragraphs?

To assess a student essay, we relied on the text of the course. This text contained seven chapters and 56 units in the domain of Sociology of Education. We also used an exam question concerning 14 units of this course marked as relevant by the teacher. Contrary to other approaches that we presented above, the assessment is not realized with pre-graded essays.

The first question is: how to evaluate each unit of course? To do that, we compared each paragraph of the essay with each unit of the course. So, we obtained, for each paragraph, several coefficients of semantic similarities between -1 and 1. In the example described in Figure 2, the first paragraph of the essay was compared with each one of the 6 relevant units of the course. This first paragraph was given the following values: 0.02, 0.09, 0.45, 0.08, 0.01, 0.09.

The second question is: how to grade the essays based on these different coefficients? For each unit, we look for the paragraph which covered it the best. In the figure, each best similarity coefficient is marked with a circle. For instance, the unit 1 is best described by the third paragraph. For each paragraph, we have several possibilities:

- only one unit is covered (paragraph 1 in the example). It means that the paragraph is probably coherent. Its grade is the semantic similarity provided by LSA between the paragraph and the course unit (0.45 in the example). In addition, the student is informed that this paragraph is OK.
- several units were developed in the paragraph (paragraphs 3 et 4 in the example). This indicate a paragraph which is not very coherent since it covers several topics. In that case, the grade for that paragraph is the average of the semantic similarity between the essay and the relevant units (0.31 and 0.19 for paragraphs 3 and 4 of the example). The student is warned that these paragraphs need to be revised since they cover several topics.

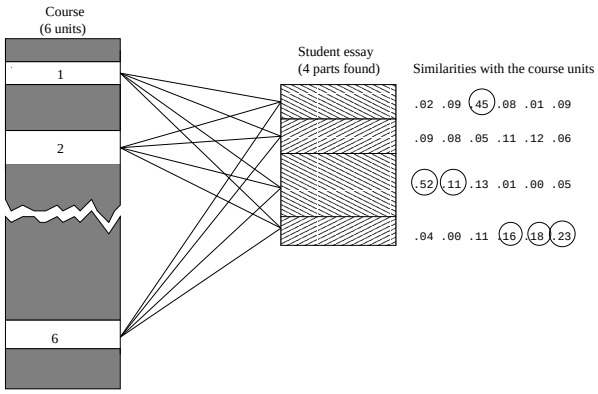

Fig. 2. *Example of comparison between student essay and course*

– no units were found (paragraph 2 of the example). It means that the paragraph does no seem to cover any relevant topic. A message is then provided to the student for requiring a modification of the paragraph.

The grade of the essay is the average of the grades of its paragraphs. This measure between -1 and 1 is converted into a letter from A to E.

Concerning the feedback, more information is given to the student concerning the assessment of the essay :

– the way LSA structured the essay by highlighting the breaks in the text;
– for each of these paragraphs, the course units that seem to have been covered;
– for each of these paragraphs, a grade between A and E;
– a global grade between A and E;
– an indication of the rank of this global grade with respect to the grade of other students. For instance: "You are in the first 25%". This measure is probably more reliable than the absolute global grade. In fact, it is possible that the absolute global grade does not correspond well with the grades given by teachers.

These methods were implemented in an interface written in PHP. The LSA procedures, written in C, were kindly provided by Telcordia Technologies. To test this method, we realized an experiment only concerned with the global grade.

5.2 Experiment 2

The goal of this experiment was to compare the grades given by our system with those given earlier by the teacher.

Procedure We used the text of the course mentioned before, composed of seven chapters and 56 units and the same exam question, corresponding to 14 units of the course. The same 30 student essays were involved.

Results and comments A correlation of 0.62 ($p < .001$) between the system grades and the teacher grades was found. This result is coherent with other researches comparing human and computer grades or even grades produced by two human judges on literary domains.

We compared this result with a test performed with the Apex system described earlier. This test concerned the same exam question and the same student essays. A correlation of 0.51 ($p < .001$) between Apex grades and teacher grades was found [11]. It means that it was useful to structure the student essay before computing the similarity with the text of the course.

It is worth noting that we are not interested in the grade per se. It is an objective value that can be easily compared with human data, but the student could be much more interested in the overall feedback which indicates the parts of the essay where the student should work on to enhance the text. In the previous systems, the student was informed about the topics that were not well covered but the student did not know where to make the modification in the text. Our system provides this indication.

6 Conclusion

We presented in this paper a method for partitioning a text into coherent segments which allows the comparison of a student essay and the text of a course. This segmentation gives better results than usual approaches in which the student essay is considered as a whole. This method has been implemented in a Web-based environment. Students at a distance can connect to the system and learn a domain by writing and revising their essay. This method can be used in any literary domain, provided that there exists texts describing that domain.

We relied on the LSA model to represent the knowledge contained in both the text of a course and a student essay. Results are in the same vein as those found in the literature: LSA seems to be an interesting model of semantic representation of textual knowledge.

Improvements can be done at various levels. First, the method for partitioning the essay and comparing it with the course could be improved by using thresholds for rejecting low similarities. It would also be interesting to test the system with different domains and different teachers. An experiment with real students working on-line would also be very informative, especially for the design of the interface. One more improvement would consist in asking the student to validate the segmentation of the essay. The student would agree or not on the breaks defined by the system. The assessment part would then be based on this segmentation. The result of this step would possibly be more accurate.

It is worth noting that our goal is not just to provide grades but rather to help students at a distance to learn a domain by writing and revising free texts. A grade is a general information which can be useful but it cannot help the student to revise the essay. For this reason, the other kinds of feedback are highly valuable. These feedbacks were possible because of the segmentation of the essay. This way of learning a domain by writing and revising a text is intended

to be included in Web-based learning environments, which are currently mainly based on multiple-choice questions, a method which is domain-dependent and, moreover, quite time-consuming.

Acknowledgments

We would like to thank V. Barré, C. Choquet, A. Corbière, P. Dessus and P. Tchounikine for their comments on an earlier version of this paper.

References

1. Beefermann D., Berger A., and Lafferty J.D., 'Statistical models for text segmentation', *Machine Learning*, **34**(1-3), 177–210, (1999).
2. Borgida A., 'On the relative expressive power of description logics and predicate calculus', *Artificial Intelligence*, **82**, 353–367, (1996).
3. Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., and Harshman R., 'Indexing by latent semantic analysis', *Journal of the American Society for Information Science*, **41**(6), 391–407, (1990).
4. Dumais S.T., 'Improving the retrieval of information from external sources', *Behavior Research Methods, Instruments and Computers*, **23**(2), 229–236, (1991).
5. Foltz P., 'Latent semantic analysis for text-based research', *Behavior Research Method, Instruments and Computer*, **23**(2), 229–236, (1996).
6. Foltz P., Kintsch W., and Landauer T.K., 'The measurement of textual coherence with latent semantic analysis', *Discourse Processes*, **25**, 285–307, (1998).
7. Foltz P.W., Laham D., and Landauer T.K., 'Automated essay scoring: Applications to educational technology', in *Proceedings of the ED-MEDIA Conference*, Seattle, (1999).
8. Foltz P.W., Laham D., and Landauer T.K., 'The intelligent essay assessor: Applications to educational technology', *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, **1**(2), (1999).
9. Hearst M., 'Multi-paragraph segmentation of expository text', in *32nd. Annual Meeting of the Association for Computational Linguistics*, pp. 9–16, Las Cruces, (1994).
10. Landauer T.K. and Dumais S.T., 'A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge', *Psychological Review*, **104**, 211–240, (1997).
11. Lemaire B. and Dessus P., 'A system to assess the semantic content of student essays', *Journal of Educational Computing Research*, **24**(3), 305–320, (2001).
12. Reynar J.C., 'An automatic method of finding topic boundaries', in *Meeting of the Association for Computational Linguistics*, (1994).
13. Salton G. and Allan J., 'Automatic text decomposition and structuring', *Information Processing and Management*, **32**(2), 127–138, (1996).
14. Sowa J.F., *Principles of Semantic Networks: Exploration in the Representation of Knowledge*, Morgan Kaufman, 1991.
15. Wiemer-Hastings P. and Graesser A., 'Select-a-kibitzer: A computer tool that gives meaningful feedback on student compositions', *Interactive Learning Environments*, **8**(2), 149–169, (2000).