Abstract

Some algorithms and techniques for automated grading of the conceptual content of written tests are described. The algorithms are based on a statistical approach - the Latent Semantic Analysis (LSA) [6]. Other applications of LSA are also described, such as the cross-language information retrieval. An LSA-based automated free-form written test assessment application is under development at the Department of Computer Science, Technical University of Cluj-Napoca. It will be part of a Web-based virtual university.

1. Introduction

The automatic grading of written tests for student assessment has two purposes: (a) getting feedback in different stages of learning a course or module by taking an automatically graded written test in a regular fashion, say weekly; (b) taking even the final examination for a course/module in the same way. It is generally difficult to incorporate large numbers of written assignments due to the effort required to evaluate "manually" large numbers of written works. Students’ ability to convey information verbally is a very important educational goal, which is not sufficiently well assessed by multiple choice tests, very popular because of their amenability to automation. As a consequence, grading and criticizing written works is important for students to better learn not only content but also the skills of thinking and writing. This paper describes a couple of algorithms for automated grading of the conceptual content of written tests.

The statistical approach on which the algorithms rely is the Latent Semantic Analysis (LSA) [6], an augmented vector space model. It is augmented by introducing a mathematical method of statistical machine learning. Students’ written works and the textbooks are latent semantically analyzed. In LSA, the semantic similarity between words and/or text passages is evaluated as the cosine of the angle of the vectors corresponding to the words and/or the passages. The vectors are represented in an about 300-dimensional semantic space. The similarity comparison made by LSA is the basis for the implementation of the automated grading of written tests.
2. The latent semantic analysis and its applications

LSA is a method for the extraction and representation of the meaning of words as related to the meanings of text passages containing these words and, symmetrically, the meaning of passages as a function of the meanings of the words in a passage. Thus LSA induces the contextual-usage meaning of words. This induction is done by statistical computations applied to large corpora of text [6]. This mathematical processing is relevant for language corpora similar in volume to the language exposure of people (different texts read, all the sentences used in verbal communication, in letters etc.).

The lexical-semantic explanation of LSA is that the set of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that captures the similarity of meaning of words and passages to each other. The measures of word-word, word-passage and passage-passage relations produced by LSA are well correlated with several cognitive phenomena involving semantic similarity and association [6]. Consequently, the latent semantic analysis is also a new theory of knowledge induction and representation. After processing a large sample of machine-readable text, LSA represents the words and the passages (sentences, paragraphs, written works) from the original corpus or new ones, as vectors in a very high dimensional semantic space (50-1500). LSA is based on singular value decomposition (SVD), an algebraic matrix decomposition method.

The semantic similarity estimates derived by LSA are computed via a powerful mathematical processing that is capable of correctly inferring much deeper relations than the surface level co-occurrence frequencies or correlations in usage. Generally, these similarity estimates are much better predictors of human meaning-based reasoning than the surface level contingencies in probabilistic methods, e.g. naïve Bayes [8].

LSA takes as its input – the training data - raw text parsed into words defined as unique character strings and grouped into meaningful passages such as sentences or paragraphs. LSA begins its processing by representing the text as a matrix in which each row corresponds to a unique word and each column corresponds to a text passage or other context. Each cell contains the number of occurrences of the word of its row in the passage denoted by its column. Then each cell frequency is weighted by a function that expresses both the word’s importance in the particular passage and the degree to which the word type carries information in the domain of discourse in general.

In the next step, LSA computes the SVD (singular value decomposition) of the matrix. SVD decomposes a rectangular matrix into the product of three other matrices. The first one represents the row entities as vectors of derived orthogonal factor values. The second one is a diagonal matrix containing scaling values (coefficients, factors) and the third one represents the original column entities in the same way as the first matrix. The result of the multiplication of the three matrices by keeping the order first, second, third, will be exactly identical to the original matrix. The fact that any matrix can be subject to SVD, using no more diagonal coefficients than the smallest dimension of the original matrix, can be proved mathematically.

If the diagonal of the second matrix is truncated to only the greatest $k$ coefficients, the other ones being set to zero, the multiplication performed as above will lead to a least-squares best fit of the original matrix. This dimensionality reduction reflected in the final matrix representation combines the surface information in the corpus into a deeper abstraction, that captures the mutual implications of words and passages. This is the machine learning done by LSA. Hypothetically, the optimally reduced space for the representation of the acquired knowledge has the same dimensionality as the source that generated the discourse (the human
speaker’s or writer’s semantic space). The reduction of the dimensionality to this value captures much indirect information contained in the huge number of constraints in the input corpus. This way it captures semantic structural relations and mutual entailments latent in the local observations available to experience, i.e. in the training data for LSA. For computational reasons, for very large corpora, only word/passage count matrices of limited dimension (about a few thousands) can be kept.

Besides the automatic grading of written test answers [6, 2] which is the object of the present paper, LSA has already been tested in a number of other application domains. The first one is information retrieval. Latent Semantic Indexing (LSI) works better than systems that depend on literal matches between terms in queries and documents. Its superiority can be explained by its ability to correctly match queries to and only to documents of similar topical meaning when the query and the document use different words. Compared to the standard vector method (essentially LSI without dimension reduction), e.g. naïve Bayes [8], LSI showed a 16% improvement [1]. Another advantage of LSI is its direct application in cross-language information retrieval. The query and the document could be expressed in different languages. Text and/or term similarity comparison breaks the language frontier. The contextual usage meaning of a word and its translation in different languages tend to be the same in all the natural languages.

From the multitude of application areas in which LSA has also been tested, it is worth mentioning the comprehensibility and coherence checking of discourses [3] and matching students with textbooks at the optimal level of conceptual complexity for learning [10]. This latter application area pertains also to the area of educational technology. The idea of recommending textbooks of appropriate level to students has the aim of easing student learning. This starts from the observation that people learn most efficiently when the text on a subject is neither too hard, with too many new concepts for the student, nor too easy, involving too little new knowledge construction. The method semantically compares a written work of the student on the selected topic with a set of possible recommended readings on that topic. The sophistication of these recommended readings varies in a reasonable range. Finally, the most appropriate textbook in terms of conceptual complexity will be indicated by the maximum similarity with the conceptual complexity of the knowledge owned by the student as reflected in his/her own written work.

3. Automated grading of written tests

LSA-based grading methods concentrate on the conceptual content, the knowledge conveyed by a written work, rather than its style, or even its syntax or argument structure. The holistic approach for automated grading leads to a holistic grade which measures the overall similarity of content. The results show that the reliability of the LSA-based holistic approach in grading is in fact equivalent to that of teachers [2, 6].

There are a couple of different variants of the method [6]. The first step for all of the variants consists in the construction of an LSA space, by training on domain representative text (the recommended textbooks read by students or a number of chapters therein, similar texts from other sources, and also the written works to be graded). The result of this step is a representation of the information contained in the course. In the second step, the written works of the students are assigned LSA vectors derived as the sum of the vectors of all the words in the work.
The first variant presupposes that a sample of written works is already graded by teachers. The first step of this variant computes the vector cosine between each written work not yet graded and each pre-graded one. The computed angle represents the degree to which the two works discuss information of a similar quality of knowledge content. If the angle between two works is small, then these works should be similar in content. It is important to observe that two written works can be considered to have almost identical content if they convey the same meaning, even if they contain few or none of the same words. In the final step, each work that is not yet graded is assigned the average grade of a small number of pre-graded works that are the most similar with the one not yet graded.

The second variant for LSA-based holistic grading presupposes the existence of an exemplary text on the topic of the written test. This text could be written by a teacher or another expert author and is used as a standard. The grade of a written work is computed as its LSA cosine with the standard.

The first step of the third variant of holistic grading computes the cosine between each sentence from standard recommended textbooks and each sentence of a written work. The second step chooses the maximum of the cosines between one textbook sentence and each sentence of the written work. The second step is repeated for each textbook sentence. The third step computes the semantic similarity measure between the written work and the recommended reading by cumulating the values computed in step 2 for all the textbook sentences or for a set of sentences selected by the teacher as being very important from the recommended reading.

The first step of variant number 4 constructs the matrix of distances (1-cosine) between all the written works. The second step computes the SVD of this matrix and constructs a least square approximation of it by reducing the dimensionality of the diagonal matrix of the SVD. This dimensionality reduction is taken to an extreme by keeping a single dimension that best reconstructs all the distances. In the third step, the point of a written work along this dimension is taken as the measure of its quality. The presupposition of this variant is that the most important dimension of difference among a set of written works on a given subject is their overall quality.

An automated computer-based or Web-based written test grader must be able to determine if it cannot grade a written work reliably. There are a couple of situations that cause unreliability for the computed grade. For instance, a written work is pretty different from the works for which it has been trained; a written work is too creative, off topic, or violates the standard formats or structures for written tests; a written work is too similar to other works or to the textbooks etc. The automated grader must flag all these situations and forward the “anomalous” works to the teacher for an additional, “classical” assessment. There are a number of techniques that can detect such situations; they must be added to the test grader. As a particular instance of these additional functionalities, it is worth mentioning the ability to detect different levels of plagiarism.

4. Getting feedback for learning

Our LSA-based automated written test grader will be Web-based and will allow students to submit their written works on a given subject from their Web browsers. After a reasonable amount of time, the students will receive the answer containing the automatically computed grade for their written work; the answer will also contain a set of questions and statements of additional subtopics that are missing from their work. The student will have the
possibility to revise and update his/her work immediately and then submit it again for a new
evaluation. During the different stages of learning a course or a module, such an exercise will
have the role of getting feedback for learning.

Such a system will allow students to practice the writing without the need that all their
toy or draft works to be assessed by the teachers. Thus students will receive feedback and
make multiple revisions during a session that could last one or two hours.

In order for the answer returned by the system to contain the appropriate comments
pertaining to a written work, individual sentences in the work will be compared against
textbook sentences that correspond to the different subtopics of the test subject. If no sentence
in the written work will be found to match a given subtopic, then the student will receive
feedback about the fact that his/her work doesn’t properly cover that subtopic. The automatic
grader will also have the ability to guide the student to the corresponding chapters and
passages in the Web-based on-line textbooks for the course.

5. Implementation

At this stage, our LSA-based automated free-form written test assessment application
inspired from systems like the Intelligent Essay Assessor (IEA) [2] is under development. The
first thing to experiment was to verify the power of the latent semantic analysis when used for
grading written tests. For this purpose, I have already implemented the second variant for
automatic grading, which is the simplest one. For LSA training I have chosen the first three
pages of the text taught in one of the lectures of the Databases course for students in the
second year at Computer Science and Engineering at Technical University of Cluj-Napoca. I
have then applied LSA to evaluate the semantic similarity of the fourth page in the same
lecture with the previous three pages LSA was trained on. The first four pages of the lecture
contain only text and no equations, pseudo-code or graphical diagrams. The grade returned by
this experiment was 9. This is a good image of the reality, because the fourth page of any
lecture is indeed sufficiently similar with the first three in the conceptual content. The fourth
page indeed uses concepts introduced in the first three pages with the aim of defining new
concepts.

The implementation obviously requires the use of some statistical techniques (the
vector space model, SVD, dimensionality reduction etc.). The package of statistical routines R
is used for this purpose. R comes with its own programming language [9]. Besides some
linguistic preprocessing implemented in C, the main part of the implementation thus far is
written in R. In order to be able to correctly populate the word/passage matrices of LSA, some
linguistic preprocessing is required: a morphological analysis in order to keep only the root
form of each word and a part-of-speech tagging that indicates what words are worth keeping
in the matrix representation (prepositions, conjunctions etc. are ignored as irrelevant for the
(latent) semantic analysis). The computational linguistic tools in VIE (Vanilla Information
Extraction) [5] in the computational framework of the GATE (General Architecture for Text
Engineering) package [4] has been used for all of this preprocessing. VIE comes equipped
with a morphological analyzer for English implemented in LEX [7]. With the aim of also
applying the implementation to English-Romanian cross-language latent semantic indexing, I
have already implemented a morphological analyzer for the Romanian language in LEX in the
VIE/GATE framework. The implementation so far covers a considerable part of the
Romanian nouns. Nouns are the most important for information retrieval.
The Department of Computer Science of the Technical University of Cluj-Napoca will integrate this application as a module in a Web-based virtual university. All the courses prepared for the automatic written test grading are content-based courses not based on algorithms, equations, formalisms and graphical diagrams; consequently, the written works will only contain free-form natural language.

6. Conclusions

The LSA-based free-form written test assessment represents a new approach for the automatic evaluation of knowledge expressed in written works. Nevertheless, the technique is only able to analyze natural language, i.e. it is not able to assess whether an algorithm expressed in pseudo-code or in a programming language is correct or not, nor whether an equation or a graphical diagram in Mathematics, Physics, Chemistry etc. is correct or not. Consequently, the humanistic area is the most suitable for the approach. The question to explore in the near future is to evaluate the effectiveness of such an approach even in exact sciences, by trying to automatically grade written works describing algorithms or even equations (where possible) in natural language, i.e. in words.

Future work will indeed include the further development of the implementation of all the variants of the automatic grading of written tests, including the growing towards a realistic size of the text material given to LSA as training data, that is, portions of textbooks or whole textbooks. Another future plan is to extend the morphological analyzer for Romanian to also cover the verbs, which are equally important as the nouns in the semantic analysis.

References