Indexing Student Essays Paragraphs using LSA over an Integrated Ontological Space

Gaston G. Burek   **Maria Vargas-Vera**   **Emanuela Moreale**

Knowledge Media Institute,
The Open University
Milton Keynes , UK, MK7 6AA
{g.g.burek,m.vargas-vera,e.moreale}@open.ac.uk

Abstract

A full understanding of text is out of reach of current human language technology. However, a shallow Natural Language Processing (NLP) approach can be used to provide automated help in the evaluation of essays. The main idea of this paper is that Latent Semantic Indexing (LSA) can be used in conjunction with ontologies and First order Logic (FOL) to locate segments relevant to a question in a student essay. Our test bed, in a first instance, is a set of ontologies such the AKT reference ontology (describing academic life), Newspaper and a Koala ontology (concerning koalas' habitat).

1   Introduction

This paper describes a novel methodology aiming to support evaluators during the essay marking process. This approach allows measuring semantic similarity between structured (i.e. ontology and binary relations derived from the essay question) and unstructured (i.e. text processed as a bag of words) information by means of Latent Semantic Analysis (LSA) and the cosine similarity measure (Deewerster *et al.*, 1990).

Previous studies (Foltz *et al.*, 1998; Wiemer-Hastings and Graesser, 1999) have used LSA to measure text coherence and comprehension by comparing units of text (i.e. sentences, terms or paragraphs) to determine how semantically related they are. The work presented in this paper is based on the use of "pseudo" documents: these are temporary documents containing a description of knowledge entities extracted from available domain ontologies (i.e. ontological relations). Both pseudo documents and paragraphs in student essays are represented as vectors. Essay paragraphs are indexed according to a measure of semantic similarity (called cosine similarity). The ontological space acts as a mediated schema, a set of virtual relations among knowledge entities related by their degree of similarity. A new knowledge entity can be added in this space and automatically a similarity measure is calculated for all the entities within the space.

1.1   Motivation and Context

The main motivation for this work derives from a need for semantics in essay evaluation, whether by a tutor or by the student author in the process of writing. Page (Page, 1968) makes a useful distinction between marking for syntax (i.e. linguistic style) and for content (subject matter) which we will use in our outline. Based on this distinction, four main approaches to essay assessment have been reported (Williams, 2001). Early systems such as PEG (Page, 1966) relied mainly on syntactic and linguistic features and required a sample of the essays to be marked by a number of human judges. E-rater (Burstein *et al.*, 1998) uses a combination of statistical and natural language processing techniques for the purpose of extracting linguistic features of the essays to be graded. Again, the essays are evaluated against a set of human–graded essays acting as a benchmark. In the LSA method of essay grading, an LSA space is constructed based on domain specific material and the student essays. LSA grading performance is about as reliable as human graders (Foltz, 1996). Text categorisation (Larkey, 1998) also requires a database of graded essays, so that new essays can be categorised in relation to them.

In short, the approaches seen so far [1] have either concentrated on syntactic and linguistic features or used domain knowledge in the form of keywords and documents about the domain. What we are proposing in this paper is that a further distinction should be made between using implicit (keywords, documents) and explicit content representations (see Fig.1, our contribution is marked in bold). We

---

[1] Kukich presents in her article Beyond Automated Essay Scoring a time line of research developments in the field of writing evaluation (Kukich, 2000).

then argue the case for adding explicit domain knowledge in the form of domain ontologies. In particular, we merge ontologies, LSA and FOL. An advantage of this approach is that it does not require a corpus of graded essays, except for validation. This feature enables tutors (or students in need of feedback) to evaluate essays on particular topics even when there are no pre-scored essay examples available. Effectively, this capability may reduce the overall time required to prepare a reliable evaluation scheme for a new essay question.
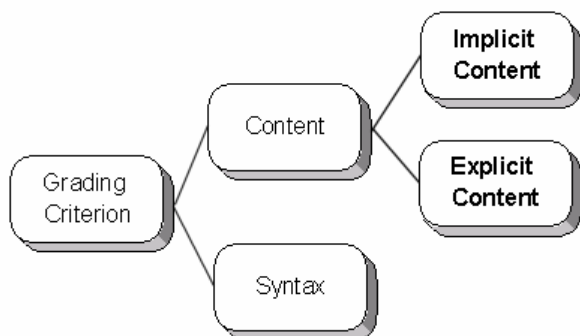

Figure 1 - Grading Criteria for Student Essays

2   LSA and the Cosine Similarity

In the vector space model, a term-to-document matrix is built in which the entries are weighted frequencies of pre-processed terms occurring in a collection of documents. Dimension reduction methods (such as LSA), when applied to the semantic vector space model, improve information retrieval, information filtering and word sense disambiguation. The reduction in dimensions reduces the noise in text categorisation, reduces the computational complexity of cluster creation, and produces the best statistical approximation to the original vector space model. Likelihood curves characterise with a quantity the level of significance of the reduced model dimensions. Also, the significance of each dimension follows a Zipf distribution (Li, 1992) indicating that the reduced model dimensions represent latent concepts (Ding, 1999). The dimensions in the reduced vector space model can be compared measuring semantic similarity between each of them by means of the cosine similarity. The cosine of the angle between two vectors is defined as the inner product between the vectors **v** and **w** divided by the product of the length of the two vectors.

$$Cos\theta = \frac{v.w}{\|v\|.\|w\|}$$

3   Indexing Essays Paragraphs

An index of relations within the ontologies related to the semantic space is obtained for each binary relation derived from the essay question. Then a subset containing the higher ranked relations is selected and the similarity between each of the relations in the subset and all the documents containing essays paragraphs is also calculated by applying LSA. Finally, an average similarity value is obtained for the paragraph over the number of relations in the subset.

3.1   An Ontology Integration Method to Build the Semantic Space

A collection of "pseudo" documents is created for each of the classes within the ontologies describing the domains tackled in the essay. The ontologies are described quantitatively using probabilistic knowledge (Florescu *et al.*, 1997).

Each of these documents contains information (name, properties and relations) about a class. The documents are represented by a vector space model (Salton *et al.*, 1971) where each column in the term-to-document matrix represents the ontological classes and the rows represent terms occurring in the pseudo documents describing those knowledge entities.

Relations within the available ontologies are also represented by a vector space model where the columns in the term–to–document matrix are a combination of two or more vectors from the term–to–document matrix representing classes. Each column represents the relation held between the combined classes. A new column representing the binary relation derived from the essay question is added to the new matrix: this new column contains the weighted frequencies of terms appearing as arguments within the relation. For each essay question, one or more binary relations are derived through parsing. For instance: given the query "Do koalas live in the jungle?" the binary relation is live_in (koala, jungle). In the case of this example, the vector representing the question contains a frequency of one in the rows corresponding to the terms koala and jungle.

LSA is applied to the term–to–document matrix representing the ontological relations, the vector space model is reduced and the cosine similarity is calculated to obtain the semantic similarity between the vectors of the reduced space model. For each column, a ranking of similarity with the rest of the columns will be obtained.

	Researcher (APO)	StandartAd (NO)	Salesperson (NO)	Student (KO)	Parent (KO)	Koala (KO)	Newspaper (APO)	Newspaper (NO)
Researcher (APO)	1.0000	0.5160	0.5215	0.8811	0.8524	0.8536	0.5036	0.5905
Standart Ad. (NO)	0.5160	1.0000	0.9999	0.0496	-0.0078	-0.0057	0.9998	0.9959
Salesperson (NO)	0.5215	0.9999	1.0000	0.0561	-0.0013	0.0007	0.9997	0.9965
Student (KO)	0.8811	0.0496	0.0562	1.0000	0.9983	0.9984	0.0353	0.1388
Parent (KO)	0.8524	-0.0078	-0.0014	0.9983	1.0000	0.9999	-0.0222	0.0815
Koala (KO)	0.8536	-0.0057	0.0008	0.9985	0.9999	1.0000	-0.0201	0.0837
Newspaper (APO)	0.5036	0.9998	0.9997	0.0353	-0.0222	-0.0201	1.0000	0.9945
Newspaper (NO)	0.5905	0.9959	0.9965	0.1388	0.0815	0.0837	0.9945	1.0000

Table 1 – Cosine similarity for classes belonging to different ontologies.

3.1.1 Weighting Scheme

Given the term–to–document matrix containing a frequency f_{ij}, the occurrence of a term in all the pseudo documents j is weighted to obtain matrix. The entries of the matrix are defined as,

$$a_{ij} = l_{ij} g_{ij} d_j ,$$

where, l_{ij} is the local weight for term i in the pseudo document j, g_j is the global weight for term i in the collection and d_{ij} is a normalisation factor. Then, as defined by Guo and Berry (Guo and Berry, 2003),

$$a_{ij} = \log_2 \left(f_{ij} + 1 \right) \left(1 + \frac{\sum_j p_{ij} \log_2 \left(p_{ij} \right)}{\log_2 \left(n \right)} \right) ,$$

where,

$$p_{ij} = \frac{f_{ij}}{\sum_j f_{ij}} .$$

4 Experiments on semantic similarity

In order to evaluate how well LSA captures similarity, this section will describe three preliminary experiments for measuring semantic similarity between knowledge entities (i.e. binary relations and classes) of three different ontologies, the Aktive Portal Ontology (APO), the Koala Ontology (KO) and the Newspaper Ontology (NO).

4.1 Experiment 1

The aim of this experiment is to evaluate how well LSA captures similarity between classes that belong to different ontologies. Eight classes have been selected randomly from within three ontologies and described in "pseudo" documents. The words included in each of the documents correspond to the names of the classes and slots related to the class described. The terms have been stemmed and stop words deleted before applying LSA to the term–to–document matrix built using the weighted frequencies of the term occurring within the eight documents describing the classes. Terms have been weighted according to the weighting scheme presented in section 2.1.1 with d_j=1, the only difference being that terms corresponding to classes names have been multiplied by two. The similarity measures for the eight classes are obtained (See Table 1) after applying LSA with a rank of two and the cosine similarity measure to the term–to–document matrix.

The results from this experiment show that, in terms of the cosine similarity measure, the class "Researcher" appears to be very similar to the class "Student" in a different ontology. The same results also show that the two classes "Newspaper" belonging to two different ontologies are very similar to each other.

4.2 Experiment 2

The aim of this experiment is to evaluate the ability of LSA to measure similarity between a predicate argument and different classes. The query is represented as an added column in the term–to–document matrix which already contains as columns the same documents representing the eight classes used in the first experiment. The column representing the query argument contains only one term corresponding to the name of one of the classes within the ontologies used in this experiment. The frequency of this term is the entry in the added column with a frequency of one multiplied by two as all the other terms representing names of classes. The results for the cosine similarity measure between the eight classes plus the query containing the term "student", "newspaper" and "animal" after applying LSA with a rank of four (see Table 2) indicate that the most similar classes for the query containing the

term "student" are the following classes: "Student" from KO, "Researcher" from APO, and "Parent" from KO.

Argument (student)	Argument (animal)	Argument (newspaper)	Classes
0.0018	0.0000	0.0099	Researcher (APO)
0.0000	0.0000	0.1403	Standart Ad. (NO)
-0.0001	0.0000	-0.0042	Salesperson (NO)
0.4473	0.0000	-0.0080	Student (KO)
-0.0013	0.5563	-0.0084	Parent (KO)
-0.0006	0.4374	-0.0085	Koala (KO)
0.0000	0.0000	0.5127	Newspaper (APO)
-0.0001	0.0000	0.8112	Newspaper (NO)
1.0000	1.0000	1.0000	**Queries**

Table 2 – Semantic similarity between arguments and classes belonging to different ontologies

For the query containing the term "newspaper" the results shows that the most similar classes are "Newspaper" from APO, "Newspaper" from NO and "Standard Advertising" also from NO. Finally, for the query containing the term "animal", the most similar classes in order of similarity closeness are "Parent" from KO and "Koala" also from KO.

The results of this experiment indicate that LSI may be accurately used as a measure of similarity between a keyword representing a query predicate argument and a set of documents representing classes that belong to a set of different available ontologies.

4.3 Experiment 3

The aim of this experiment is to evaluate the cosine similarity measure as a measure of semantic similarity between binary relations derived from a question or query and relations held between two classes. This measure is based on the same methodology and procedures applied to both experiments described above. For this experiment, eighteen classes have been selected arbitrarily from the three available ontologies (see Table 3).

The binary relations held among the selected classes are represented as documents in a term–to–document matrix that is the union of the two pseudo documents describing the related classes. Following the same procedure as in the previous experiment, a new column representing the binary relation derived from a question is added to the matrix, but in this case it contains the terms describing the two arguments of the binary relation.

Newspapers Ontology (NO)			
ID Relation	**Relation name**	**Class1**	**Class2**
OBR1	Sales Person	Advertisement	Salesperson
OBR2	Purchaser	Advertisement	Person
OBR3	Published in	Content	Newspaper
OBR4	Content	Newspaper	Content
OBR5	Employees	Organisation	Employee
OBR6	Prototype	Newspaper	Prot. Newspaper

Aktive Portal Ontology (APO)			
ID Relation	**Relation name**	**Class1**	**Class2**
OBR7	Has gender	Researcher	Gender
OBR8	Has appellation	Researcher	Appellation
OBR9	Owned by	Newspaper	Legal Agent
OBR10	Has Size	Organisation	Organisationsize
OBR11	Headed by	Organisation	Afiliated Person
OBR12	Organisation part of	Organisation	Organisation

Koala Ontology (KO)			
ID Relation	**Relation Name**	**Class 1**	**Class 2**
OBR13	Has gender	Animal	Gender
OBR14	Has habitat	Animal	Appellation
OBR15	Has children	Animal	Animal

Table 3 – Ontological Binary Relations (OBR) used in Experiment 3

The cosine similarity between fifteen predicates and the available relations after applying LSA with a rank of four (see Table 4) show that, in eight of the fifteen cases, the similarity value is higher for the relations held between classes than between predicate arguments. In the rest of the cases, the similarity values are very close for two or more relations including the one held between classes that are the same as the predicate arguments. Another interesting observation is that, Question Binary Relation 3 (QBR3) has a cosine value more similar to Ontological Binary Relation 9 (OBR9), OBR3 and OBR4. In the case of QBR5, the cosine value is higher when measuring similarity with OBR11 and OBR12 than, for example, the cosine value when measuring similarity with OBR3 and OBR4. Similar results were obtained for QBR6 where, apart from OBR6, OBR9 has the cosine value closest to one. Other similar results are obtained for QBR11 and QBR12 where OBR5 is closer to a value of one than OBR7, OBR8 and OBR9.

	QBR1	QBR2	QBR3	QBR4	QBR5	QBR6	QBR7	QBR8	QBR9	QBR10	QBR11	QBR12	QBR13	QBR14	QBR15
OBR1	0.3520	0.3033	0.1993	0.1993	0.2588	0.1713	-0.0007	0.0000	0.0487	-0.0030	0.0051	-0.0048	0.0000	0.0000	0.0000
OBR2	0.3628	0.3286	0.2170	0.2170	0.1896	0.1864	0.0006	0.0000	0.0528	-0.0033	0.0053	-0.0053	0.0000	0.0000	0.0000
OBR3	0.0900	0.0023	0.2864	0.2864	0.0002	0.2631	-0.0005	0.0000	0.0771	0.0017	0.0223	0.0027	0.0000	0.0000	0.0000
OBR4	0.0900	0.0023	0.2864	0.2864	0.0002	0.2631	-0.0005	0.0000	0.0771	0.0017	0.0223	0.0027	0.0000	0.0000	0.0000
OBR5	-0.0007	0.0039	-0.0013	-0.0013	0.3925	0.0000	0.0001	0.0000	-0.0006	0.0304	0.0566	0.0468	0.0000	0.0000	0.0000
OBR6	-0.0003	0.0024	0.2730	0.2730	0.0003	0.3284	0.0010	0.0000	0.0880	0.0011	0.0184	0.0017	0.0000	0.0000	0.0000
OBR7	0.0000	0.0032	-0.0004	-0.0004	0.0001	-0.0004	0.9572	0.3621	-0.0013	-0.0016	0.0143	-0.0012	0.0130	0.0130	0.0000
OBR8	0.0000	0.0032	-0.0004	-0.0004	0.0001	-0.0004	0.9567	0.3666	-0.0014	-0.0016	0.0143	-0.0012	0.0109	0.0109	0.0000
OBR9	0.0002	0.0002	0.2971	0.2971	-0.0029	0.2738	0.1477	-0.0028	0.9300	0.0147	0.0264	0.0082	-0.0002	-0.0002	0.0000
OBR10	-0.0002	0.0115	0.0014	0.0014	0.0633	0.0014	0.0999	0.0458	0.3012	0.4894	0.5181	0.3599	0.0190	0.0190	0.0000
OBR11	-0.0002	0.0113	0.0012	0.0012	0.0545	0.0012	0.1153	0.0454	0.2882	0.4304	0.4759	0.3161	0.0196	0.0196	0.0000
OBR12	-0.0002	0.0119	0.0015	0.0015	0.0522	0.0014	0.1019	0.0478	0.3146	0.4189	0.4623	0.3061	0.0221	0.0221	0.0000
OBR13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7312	0.0000	0.0000	0.0000	0.0000	0.0000	0.5397	0.5397	0.4910
OBR14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7490	0.0000	0.0000	0.0000	0.0000	0.0000	0.5202	0.5202	0.4767
OBR15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7550	0.0000	0.0000	0.0000	0.0001	0.0000	0.5594	0.5594	0.5261
QBR	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 4 –Cosine similarity between the Question Binary Relations (QBR) and the Ontological Binary Relations (OBR).

The results of this experiment indicate that the presented methodology is able to detect similarity between compact representations (binary relation arguments) and more expanded representations such as the pseudo documents representing the binary relations within the three available ontologies.

4.4 Experiments discussion

We expect that using LSA together with the cosine similarity measure, we will be able to pick up semantic similarity between the compacted and expanded representations of the binary relation and paragraphs from student essays. The main difference between our approach and other essay scoring approaches (e.g. The Intelligent Essay Assessor; Laundauer *et al*., 2000) where the scores are calibrated using LSA with pre-scored essay examples, is that our approach scores paragraphs using LSA and the cosine similarity with ontologies describing the essay domain. The experiment results in the previous sections validate our view showing that the cosine similarity may be used as a reliable score for semantic similarity between knowledge entities belonging to different data sources (i.e. terms, classes and binary relations).

5 Conclusion and Future Work

This paper introduces the idea of "explicit content" and its use in essay evaluation. The main contribution of the paper is then the idea that ontologies and First Order Logic (FOL) can be used together with LSA to locate segments relevant to a question in a student essay.

Our main interest is to provide help to tutors in grading and to students for feedback purposes. In fact, even outside the realms of grading, we believe that it will help annotate and rank paragraphs more relevant to queries. In our proposal, we went about doing this supplementing the widely-used LSA method with added semantics (ontologies) and First Order Logic (FOL). Our approach therefore attempts to bridge the gap between statistical and semantic approaches.

There is clearly a lot more work needed to make this technology work well enough for large-scale deployment. Further work may include a visualisation service that also provides a visualisation of annotation of segments relevant to the current question types around the lines of the work described in (Moreale and Vargas-Vera, 2003; Moreale and Vargas-Vera, 2004).

6 Acknowledgements

References

J. Burstein, K. Kukich, S. Wolff, C. Lu and M. Chodorow. 1998. Enriching Automated Essay Scoring Using Discourse Marking. *Proceedings of the Workshop on Discourse Relations and Discourse Markers,* Annual Meeting of the Association of Computational Linguistics, August, Montreal, Canada.

C. H. Q. Ding. 1999. A Similarity-Based Probability Model for Latent Semantic Indexing. *Proc. 22nd ACM SIGIR Conference*, p. 59–65.

S.C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, R.A. Harshman. 1990. Indexing by Latent Semantic Analysis. *JASIS* 41(6): 391–407.

D. Florescu, D. Koller, A. Levy. 1997. Using Probabilistic Information in Data Integration. *Proceedings of the 23rd VLDB Conference, Athens,* Greece.

P.W. Foltz,W. Kintsch, and T.K. Landauer. 1998. The Measurement of Textual Coherence with Latent Semantic Analysis, *Discourse Processes*, Vol. 25, Nos. 2–3, 1998, p. 285–308.

P.W. Foltz. 1996. Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers,* 28, 197-202.

D. Guo and M. W. Berry. Knowledge –Enhanced Latent Semantic Indexing. *Information Retrieval,* 6 (2): 225-250, 2003.

K. Kukich. 2000. The Debate on automated essay grading, Beyond Automated Essay Scoring. *IEEE Transactions on Intelligent Systems.* September/October 15 (5):27–31.

L.S. Larkey. 1998. Automatic Essay Grading Using Text Categorization Techniques. *Proceedings of the Twenty First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Melbourne, Australia, p. 90–95.

T. K. Laundauer , D. Laham and P.W.Foltz. 2000. The Debate on automated essay grading, The Intelligent Essay Assessor. *IEEE Transactions on Intelligent Systems.* September/October 15 (5):27–31.

W. Li. 1992. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6):1842–1845.

E. Moreale and M. Vargas-Vera. 2004. A Question-Answering System Using Argumentation. Third International Mexican Conference on Artificial Intelligence (MICAI-2004), *Lecture Notes in Computer Science (LNCS 2972)*, Springer-Verlag, p. 26–30. ISBN 3-540-21459-3.

E. Moreale and M. Vargas-Vera. 2003. Genre Analysis and the Automated Extraction of Arguments from Student Essays. *The Seventh International Computer Assisted Assessment Conference (CAA-2003).* Loughborough University, 8-9.

E.B. Page. 1968. Analyzing Student Essays by Computer. *International Review of Education*, 14, 210–225.

E.B. Page. 1966. The Imminence of Grading Essays by Computer. *Phi Delta Kappan*, p. 238–243.

G. Salton, A. Wong, and C. Yang. 1971. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620, 1971.

P. Wiemer-Hastings and A.C Graesser. 2000. Select-a-Kibitzer: A Computer Tool that Gives Meaningful Feedback on Student Compositions. *Interactive Learning Environments 2000.*Vol.8, No.2, p. 49–169. Curtin University of Technology

R. Williams. 2001. Automated essay grading: An evaluation of four conceptual models. In A. Herrmann and M. M. Kulski (Eds), Expanding Horizons in Teaching and Learning. *Proceedings of the 10th Annual Teaching Learning Forum, 7-9 February 2001.* Perth: Curtin University of Technology.