Sentence similarity measures for essay coherence

Derrick Higgins
Educational Testing Service

Jill Burstein
Educational Testing Service

Abstract

This paper describes the use of different methods for semantic similarity calculation for predicting a specific type of textual coherence. We show that Random Indexing can be used to locate documents in a semantic space as well as terms, but not by straightforwardly summing term vectors. Using a mathematical translation of the semantic space, we are able to use Random Indexing to assess textual coherence as well as LSA, but with considerably lower computational overhead.

1 Introduction: Text coherence in student essays

We have developed an approach to text coherence for student essays that is comprised of multiple dimensions, so that an instructional application may provide appropriate feedback to student writers, based on the system's prediction of high or low for each dimension. For instance, sentences in the student's thesis statement may have a strong relationship to the essay topic, but may have a number of serious grammatical errors that make it hard to follow. For this student, we may want to point out that on the one hand, the sentences in the thesis address the topic, but the thesis statement as a discourse segment might be more clearly stated if the grammar errors were fixed. By contrast, the sentences that comprise the student's thesis statement may be grammatically correct, but only loosely related to the essay topic. For this student, we would also want the system to provide appropriate feedback, so that the student could revise the thesis statement text appropriately.

Much previous work (such as Foltz et al., 1998; Wiemer-Hastings and Graesser, 2000) looks at the relatedness of adjacent text segments, and does not explore global aspects of text coherence. Hierarchical models of discourse have been applied to the question of coherence (Mann and Thompson, 1986), but so far these have been more useful in language generation than in deter-

This paper appeared in the published proceedings of the 7th International Workshop on Computational Semantics (IWCS), held January 2007, Tilburg, The Netherlands. Republished on www.ets.org with permission of IWCS.

mining how coherent a given text is, or in identifying the specific problem, such as the breakdown of coherence in a document.

Our approach differs in a fundamental way from this earlier work that deals with student writing. First, our approach considers the discourse structure in the text, following Burstein et al. (2003). Our method considers sentences with regard to their discourse segments, and how the sentences relate to other text segments both inside (such as the essay thesis) and outside (such as the essay topic) of a document. This allows us to identify cases in which there may be a breakdown in coherence due to more global aspects of essay-based discourse structure.

In this paper, we concentrate on the subpart of this text coherence project which determines the relationship of an essay sentence to the essay prompt, or question to which the essay is presented as a response.

2 Vector-based semantic representation

The starting point for this work is the observation that related sentences in a text tend to use the same or similar words. While the use of similar terms does not guarantee relatedness, it is almost a precondition, and we believe it should function well as a predictor of relatedness.

A number of different metrics for evaluating the semantic similarity of words have recently been devised. Some of these use human-constructed lexical resources such as WordNet (cf. Resnik, 1995; Hirst and St-Onge, 1997; Leacock et al., 1998), while others are trained by collecting statistics from unannotated or lightly-annotated text corpora (e.g., Lin, 1998).

These methods for assessing similarity between words, however, do not directly induce a method for assessing the similarity between larger units such as sentences and full documents. For this reason, in this project we concentrate on those approaches which develop vector-based semantic representations of words. These approaches represent the meaning of a word as a vector, and compute the similarity between words as the cosine of the angle between these vectors, which suggests that this method of similarity calculation could be generalized to larger semantic units. This extension to document similarity calculation has been shown to be feasible for one method of vector-based semantic similarity calculation (LSA), but has yet to be attempted for others.

2.1 Standard IR vector spaces

The natural first candidate among vector-based approaches to semantics for use in assessing text coherence is the standard model of content vector analysis, used in information retrieval. In this model, each document is associated with a vector of the words in that document, and each word is represented as a vector listing the documents in which the word occurs.

Unfortunately, this is not an effective means of calculating similarity scores for individual sentences. The problem is that the components of a document vector correspond directly to words, and a sentence of 15 words or so is not a large enough sample for the frequencies of terms within the sentence to consistently be found near their expectations. More simply stated, it is hardly surprising to find a sentence about biology which nevertheless does not contain the word 'biology', but it is more remarkable to find a biology article of multiple paragraphs which does not contain the word 'biology'.

If we have two sentences from a treatise on genetics, one may contain the terms 'chromosomal', 'crossover', and 'mutation', while the other contains 'genome', 'DNA', and 'RNA'. In fact, the sentences may have no terms in common, so that their vectors are orthogonal in semantic space, meaning that they are judged by this model to have no semantic connection. This is clearly unacceptable for a model of sentence similarity which we wish to be useful in assessing text coherence, and to address this deficiency, a number of models have been introduced which use dimensionality reduction.

2.2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) is the best known and most widely used of the vector-space methods of semantic similarity using dimensionality reduction. LSA involves the application of Singular Value Decomposition to the document-by-term matrix in order to reduce its rank. Because information is lost in this compression process, some similar vectors will be conflated, or moved closer within the space. Thus, if LSA succeeds, we automatically derive generalizations such as the one we were missing in the problematic example for the standard IR model: there is a class of biology terminology, and when different terms from this class occur in different documents, this is evidence of the documents' relatedness.

Latent Semantic analysis has proved to be a great improvement over the simple vector-space IR model for some domains. The term-term similarity scores it produces are more robust (Landauer et al., 1998), as evidenced by the much-cited results that LSA has been used to achieve a 64% score on a test of 80 synonym items from the Test of English as a Foreign Language. In addition, the document similarity scores can be used in some IR tasks Deerwester et al. (1990). Furthermore, LSA has been shown to be applicable to the task of establishing text coherence, in Foltz et al. (1998), although their application of LSA in this domain is very different from our own.

However, there are some drawbacks to LSA which restrict its applicability. For one thing, Singular Value Decomposition requires a numerical computation which is demanding both in terms of processor time and in terms of memory requirements. LSA is also very dependent on the corpus used to train it. Since term co-occurrence within documents is the basis for the generalizations it derives, it performs best when trained on corpora which are very topically coherent, and which cover a diverse set of topics. An encyclopedia is a good text to use for LSA, but it is hard to obtain high-quality encyclopedic texts at low cost, especially in large quantities. Finally, there are a number of patents which apply to Latent Semantic Analysis, so that it cannot be freely used for all purposes.

2.3 Random Indexing

Another vector-based approach to semantic similarity is Random Indexing (Kanerva et al., 2000; Sahlgren, 2001). Random Indexing differs from LSA in a number of significant ways. First, its primary focus is on term similarity, and the issue of handling document similarities with Random Indexing has yet to be addressed. Second, it does not require a specialized corpus, which is very topically coherent and neatly divided into documents for training. Finally, the algorithm underlying Random Indexing is also very different from that used in Latent Semantic Analysis, and has lower computational requirements.

Random indexing assigns vector representations to words in the following way:

- First, every word in the vocabulary is assigned a sparse, randomly-generated *label vector* of ternary values. The idea is to start with a random sparse vector representation for each word, which can then be used in further computation.

- Second, the *semantic vectors* for each word are initialized to all zeroes.

- Finally, the semantic vectors are trained, using some text corpus. For each word token in the corpus, that word's semantic vector is incremented by the label vectors of each word appearing within a certain distance of it.

In this way, Random Indexing produces a semantic vector for each word in the corpus, which can then be compared to the vectors of other words using the standard cosine similarity metric. Using this method to assess term similarity, Sahlgren (2001) achieves an even better result on the TOEFL synonyms test than the 64% score reported for LSA.

However, as noted earlier, there is no primitive notion of document similarity in Random indexing, since the text it is trained on need not be divided into documents. For this reason, it is not clear exactly how Random Indexing could be applied to the task of assessing sentence similarity for text coherence. Intuitively, it seems promising to somehow represent the meaning of a document as a vector sum of the words which appear within it, but it is not obvious that this approach will work, given that the semantic vectors of Random Indexing were not produced with this application in mind. Later in this paper, however, we will pursue this intuition, and present an extension of Random Indexing which does allow for the comparison of full documents.

3 Experiments

The goal of our project was to produce a measure of semantic similarity which is a good predictor of "relatedness" between sentences, with the ultimate goal of assessing the coherence of an essay. With this aim, we conducted experiments using different methods of constructing semantic vectors for sentences, assessing them on their performance in predicting the relatedness of sentences from a human-annotated development set.

For our preliminary investigations, we used a set of 2330 sentences (from 292 essays) which had been assessed by human annotators as either related or unrelated to the prompt paragraph in response to which the essay was composed. These sentences were taken from essays written by 12th-grade high-school students or college freshmen, and the sentences used were either from the essay thesis, essay conclusion, a main idea sentence of a paragraph, or from an introductory "background" section of the essay. This excludes supporting idea sentences from within a paragraph, which have only a mediated relationship to the essay prompt.

As a way of evaluating each of the approaches to semantic similarity we tried, we measured the overall accuracy of the method for determining relatedness. That is, we assumed some cutoff similarity score, such that sentences above the cutoff in similarity are taken to be related, while those below the cutoff are taken to be unrelated. Then, given this cutoff, we assess what percentage of the 2330 sentences in the data set are correctly classified as related or unrelated.

3.1 LSA

We used the SVDPACKC software package (Berry et al., 1993) to construct an LSA semantic space from a subset of the Wikipedia open-source encyclopedia. While the quality of the articles in this resource is somewhat variable, it was the only encyclopedia available to us for this project. To increase the quality of the training data, we used only encyclopedia articles with a single-word title (such as "Geography"). The primary reason for this choice was to exclude the large number of articles describing small US towns and television shows, and thereby remove a source of bias in the corpus.

The result is an LSA space based on 55,424 terms in 21,885 documents. We used a tf-idf weighting scheme with log weighting of term frequencies and document frequencies to construct the document-by term matrix.

Using this sentence similarity metric to predict sentences' relatedness to the essay prompt, the LSA model achieves a maximum classification accuracy of 70.6%, when the cutoff for separating "related" sentences from "unrelated" ones is set at about 0.08. This is a modest improvement over the baseline accuracy of 67.1%, which we get by simply predicting that all essay sentences are related to the prompt.

3.2 Random Indexing

For our Random Indexing experiments, we used four years' worth of newswire text from the San Jose Mercury News as training data. The Mercury News contains over 30 million word tokens, around five times the size of our encyclopedia corpus, and therefore provided a higher-quality semantic space for Random Indexing (as evinced by performance on the TOEFL benchmark).

As stated above, Random Indexing has not previously been used to calculate similarity scores for units larger than a word, but the similarity of this approach to Latent Semantic Analysis suggests that perhaps an analogous approach to building up document vectors from term vectors could be used. In LSA, a vector for a new document is obtained by first making a sparse vector of the length of the vocabulary, indicating the frequency of each term in the document, and then multiplying this vector by the term matrix T, in order to map the vector to the reduced space. The goal of the experiments of this section was to determine if document vectors could similarly be represented as a vector sum of term vectors under a Random

Indexing approach.

As a first attempt, we tried a very straightforward implementation, in which the vector representation for a document (in this case, a sentence) was computed as the vector sum of the term vectors for each word in the document. (The term vectors had previously been normalized to unit length, and a stoplist was used to prevent the vectors for function words from being included in the sum.)

This straightforward method of applying Random Indexing to sentence similarity calculation yielded a maximum accuracy of 67.12%, with an improvement over the baseline corresponding to just one more correctly classified example. Clearly, this simple implementation of document similarity using Random Indexing is not competitive with the sentence similarity scores provided by our LSA space.

3.2.1 Improved Random Indexing for document similarity

Manual inspection of the sentence similarity scores produced by this simple method of extending Random Indexing term vectors revealed that the scores were quite high for all sentence pairs, regardless of their degree of relatedness. In fact, it was hard to find a pair of sentences which exhibited a similarity score of less than 0.8 using this metric. This observation suggests a specific problem with the current way of aggregating Random Indexing term vectors to produce a document vector.

Suppose we take a set of random normalized term vectors and produce a document vector to represent them, by summing the vectors and dividing by the number of vectors in the set, n. As n increases, the document vector approaches the mean vector xmean, which is the average of all term vectors.

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \vec{\mathbf{x}}_i = \vec{\mathbf{x}}_{mean} \tag{1}$$

This means that if we compare the similarity (cosine) between two such random documents, as each document grows longer, the similarity should approach 1, since

$$\frac{\vec{\mathbf{x}}_{mean}\cdot\vec{\mathbf{x}}_{mean}}{\|\vec{\mathbf{x}}_{mean}\|^2} = 1 \tag{2}$$

This is a problem, since it means that the similarity between documents is bound to increase with their length, and regardless of their relatedness. However, if we subtract the mean vector from each of our term vectors, we can remove the bias from the system:

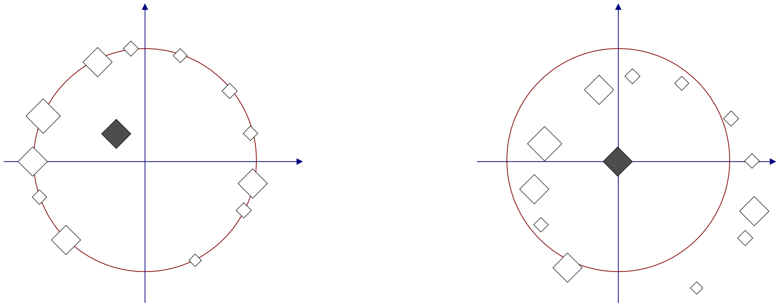
$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_{mean}) = \vec{\mathbf{0}} \tag{3}$$

and $\vec{\mathbf{0}} \cdot \vec{\mathbf{0}} = 0$.

This line of argumentation is illustrated in Figure 1. In the left half of the diagram, the vocabulary of the Random Indexing model is represented, where each 1800-dimensional term vector is represented in two dimensions, and the size of the diamond represents the frequency of the term. All term vectors are normalized, and therefore lie on the unit circle. The mean of all term vectors in the vocabulary is shown as a filled diamond. Since this mean vector does not lie on the origin, a random document vector will not tend to approach the zero vector if its terms are unrelated.

In the right half of Figure 1, the same set of term vectors is represented, so that the mean vector now lies on the origin. Note that the vectors no longer lie on the unit circle.

Figure 1: Term vectors on the unit circle. Translating the vectors brings the mean vector to the origin.



In addition to the fact that this translation of the term vectors remedies the problem we encountered initially, that document similarity tends to increase with document length, it has an additional beneficial side-effect. Subtracting the mean vector has the effect of reducing the magnitude of those term vectors which are close in direction to the mean vector, and increasing the magnitude of term vectors which are most nearly opposite in direction from the mean vector. This means that, when we create a document vector as a sum of term vectors, those terms whose distribution is most distinctive will be given the most weight, while terms which are less picky about what other terms they co-occur with will be given relatively little weight. This achieves the effect of the inverse document frequency weighting done in LSA and the standard IR vector-space model, but without any special machinery. In fact, this property of the model even obviates the need for a stoplist. Since function words' distribution induces term vectors which are very similar to the mean vector, they are given almost no weight when they occur within a document.

This improved document similarity metric produces results competitive with the results achieved using the LSA model in the previous section. The model achieves a maximum accuracy of 70.1%, using a similarity cutoff of about 0.14. This result suggests that Random Indexing can, in fact, be used to produce useful document similarity scores.

It is somewhat surprising that we can achieve such good results in judging document similarity using Random Indexing, given that the training data for the algorithm, newswire text, is not as semantically focused as the encyclopedia text on which LSA is trained. On the other hand, the scalability of Random Indexing allows us to use a much larger corpus, which certainly increases the quality of our results. At the very least, the use of Random Indexing for assessing document similarity should be a welcome alternative to LSA because of the lack of legal restrictions on its use.

3.3 Using multiple sources of evidence

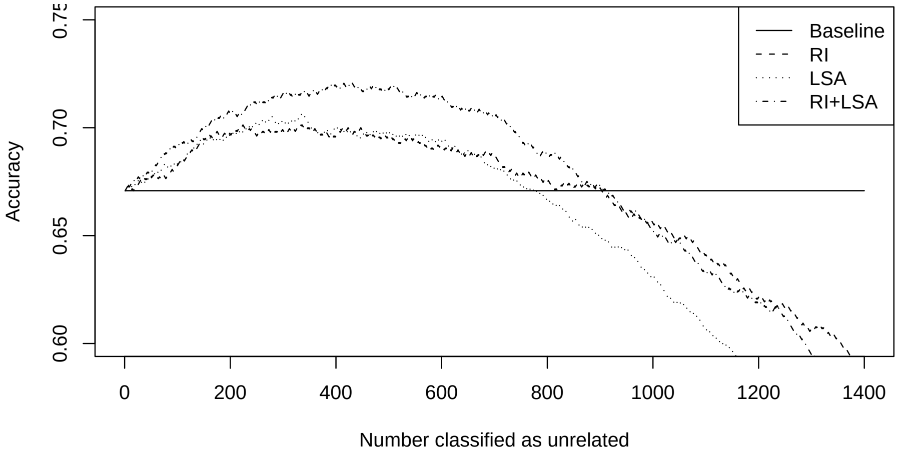
Since the LSA and Random Indexing models of sentence similarity we have constructed are based on different corpora and are mathematically distinct vector-space models, it is a reasonable conjecture that they should make different predictions about the relatedness of certain sentence pairs. As this observation suggests, we can improve upon the results of either model by simply taking the average of the scores each provides. Using this new measure of sentence similarity, the sum of the LSA and Random Indexing scores, we get a maximum accuracy of 72.1%, with a cutoff score of 0.15.

Figure 2 shows the overall classification accuracy of the three models introduced here as a function of the total number of sentences classified as unrelated to the prompt. This graph clearly shows the combined model, which uses the average of the RI and LSA similarity scores, to be superior to the other two. It also supports the claim that there is no substantial difference between the LSA and RI models. While it is true that the LSA model has a higher maximum accuracy value, the two curves track one another very closely throughout the region in which model accuracy exceeds the baseline.

The ROC curves for these models, shown in Figure 3, tell a similar story.[1] In the region of the plot with the lowest false positive rate, which is

[1] This figure takes the basic task to be retrieval of sentences which are not related to

Figure 2: Accuracy in classifying relatedness to prompt using LSA, RI, and both combined



the most critical region for this application, the combined model outperforms the other two. Somewhat surprisingly, however, the Random Indexing model is superior in the region of the plot with the highest rate of true positives. This suggests that Random Indexing might be even more competitive with LSA for applications in which recall is prized over precision.

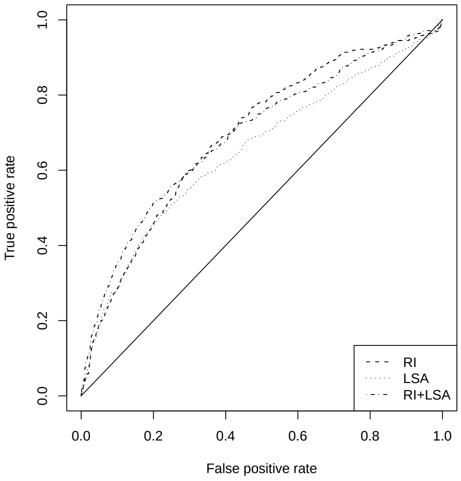
Overall, the performance of all three models on this task is modest, and it is likely that other sources of information will have to be added in order to produce an operationally useful model of this essay coherence dimension. Nevertheless, we believe this problem comprises a useful testbed for the comparison of semantic similarity metrics.

4 Conclusion

We have described the application of vector-based semantic similarity measures to the task of assessing a specific kind of textual coherence in student essays. We further demonstrate that Random Indexing can be applied to this task as an alternative to LSA, although it had not previously been used as a measure of document similarity (but only term similarity). Once the technical hurdles are overcome, Random Indexing may be a superior

the thesis, so that "false positives" are related sentences falsely taken to be unrelated.

Figure 3: ROC curves for RI, LSA, and RI+LSA models



choice for this application and others, because it is not as computationally demanding, and it is not encumbered by the same patent protections. In addition, combining the two methods of semantic similarity calculation can yield results superior to the use of either method individually.

References

Berry, M., Do, T., Krishna, G. O. V., and Varadhan, S. (1993). SVDPACKC (version 1.0) user's guide. University of Tennessee ms.

Burstein, J., Marcu, D., and Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Transactions on Intelligent Systems: Special Issue on Advances in Natural Language Processing*, 181:32–39.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25:285–307.

Hirst, G. and St-Onge, D. (1997). Lexical chains as representation of context for the detection and correction of malapropisms. In Fellbaum, C., editor, *WordNet: An electronic lexical database and some of its applications*. The MIT Press, Cambridge, MA.

Kanerva, P., Kristoferson, J., and Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In Gleitman, L. R. and Josh, A. K., editors, *Proc. 22nd Annual Conference of the Cognitive Science Society*.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Landauer, T. K., Laham, D., and Foltz, P. (1998). Learning human-like knowledge by singular value decomposition: A progress report. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems 10*, pages 45–51. The MIT Press, Cambridge, MA.

Leacock, C., Chodorow, M., and Miller, G. (1998). Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann, San Francisco, CA.

Mann, W. and Thompson, S. (1986). Relational processes in discourse. *Discourse Processes*, 9:57–90.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*, pages 448–453.

Sahlgren, M. (2001). Vector based semantic analysis: Representing word meanings based on random labels. In *Proceedings of the ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*. Helsinki, Finland.

Wiemer-Hastings, P. and Graesser, A. (2000). Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments*, 8(2):149–169.