Abstract

Latent Semantic Analysis (LSA) has been successfully used in a number of information retrieval, document visualization and summarization applications. LSA semantic spaces are normally created from large corpora that reflect an assumed background knowledge. However the right size and coverage of the background knowledge for each application are still open research questions. Moreover, LSA’s computational cost is directly related to the size of the corpus, making the technique inviable in many cases. This paper introduces a technique for creating semantic spaces using a single document and no background knowledge, which cuts computational cost and is domain independent. Single document semantic spaces’ reliability was evaluated on a collection of student essays. Several semantic spaces generated from large corpora and single documents were used to compare how essays are represented. The distance between consecutive sentences in the essays changes between semantic spaces, but the rank of the distances is preserved. The results show that high correlations (0.7) of ranked distances between sentences can be achieved on the different spaces for the weight schemes evaluated. This has important implications for the applications discussed.

Keywords: Single Document Semantic Space, Latent Semantic Analysis, LSA, background knowledge, corpus size

1 Introduction

A common problem encountered in information retrieval, document analysis and visualization applications is that people use words for their collective meaning and not just for the literal term. Linguistically the difficulties introduced are explained by the synonymy and polysemy problems. The former refers to the many ways of expressing the same concept, where people adapt their vocabulary based on the topic being discussed, or on the particular background knowledge (both of the writer and/or the reader). The latter refers to the many meanings that a word can have, meanings that humans are able to disambiguate using information about the topic being discussed or other contextual information. Synonymy and polysemy are known to affect the accuracy of computer systems that use terms (instead of concepts) as the main way of representing information.

In information retrieval in particular synonymy affects recall and polysemy affects precision.

Latent Semantic Analysis (LSA) is a statistical dimensionality reduction technique proposed by Deerwester et al. (Deerwester et al. 1990) to address these issues by indexing documents based on ‘concepts’ rather than terms. This requires a semantic representation for the corpus of documents and queries. LSA starts with a term-document matrix and uses Singular Value Decomposition (SVD) to create a semantic space where the distances between terms and/or documents reflect a ‘semantic’ proximity. When LSA is used for information retrieval tasks user queries are projected in the semantic space as pseudo documents.

The ‘concepts’ in this semantic space are features of the set of documents used to create the space. For example, the collection of books that kids in primary school are likely to have read can be used to create the semantic space that best fits the way in which primary school kids communicate. This ‘shared’ knowledge representation is useful in information retrieval and learning technologies. For example, LSA has been used to analyze the quality of students’ essays, using papers and books that the student should have read as background knowledge and the essays as queries. In fact, systems such as e-Rater and Intelligent Essay Assessor that use LSA for automatic essay grading (Miller 2003, Burstein et al. 2003) use semantic spaces to mark essays by calculating the distances between previously marked essays and the one to be marked. These automatic grading tools have shown to be very reliable (Shermis & Burstein 2003) and are commercially used in standardized tests.

LSA has also been used to measure deeper quality patterns in essays, such as discourse coherence (Foltz 2007). By measuring the distances between consecutive sentences and/or paragraphs, possible breaks in coherence are identified. These measures have been found to correlate positively with the quality of essays (Graesser et al. 2004).

In the above applications the concepts in a single essay are described as a function of the shared concepts in the semantic space created from the background knowledge. However, the right size and coverage of the background knowledge for each application are still open research questions. Moreover, LSA’s computational cost is directly related to the size of the corpus, making it many times inaccessible for the final user.

Other dimensionality reduction techniques that do not make any assumptions regarding the background knowledge are more appropriate in some applications. For example, Gong and Liu (Gong & Liu 2001) proposed a generic text summarization technique that aims to summarize a document by selecting sentences that are important, and yet different from each other. They used a variation on LSA that creates a semantic space based on the single document to be summa-
rized. This approach was shown to be useful on sets of documents that cover a broad set of topics (i.e. newsgroups). Other applications that would benefit from this approach are automatic concept mapping on essays (Villalon & Calvo 2008), and clustering search results (Osiniski 2006), both covering broad sets of topics. The distances between documents or sentences produced by these two types of LSA spaces will be different, but this paper studies what the approaches have in common and shows how the ranked lists of consecutive sentences mapped by the different approaches correlate.

This paper contributes an analysis of semantic spaces generated with a single document, and how they compare with those generated from small and large corpora. The results provide evidence that the different approaches produce similar ranked list of distances between consecutive sentences, however single document semantic spaces require much less computational power.

Section 2 presents LSA and previous work analyzing the effect of small corpora including single document semantic spaces. Section 3 describes the implementation issues in single document semantic spaces, while Section 4 describes the methodology and results of the evaluation. Our evaluation used real-world essay corpora and compared how different spaces produce different ranked distances between consecutive sentences. Section 5 presents a discussion of its implications.

2 Previous work

This section briefly describes the mathematical background of LSA, and discusses the two areas of previous work that are relevant for this study: Studies analyzing the effect of the corpus size, particularly very small corpora, and the use of LSA in a single document to obtain meaningful patterns.

2.1 Latent Semantic Analysis

LSA defines the semantic space of a term-by-document (or term-by-sentence) matrix \( \mathbf{X} \in \mathbb{R}^{n \times m} \) (that can be called the knowledge base) by decomposing it using Singular Value Decomposition as:

\[
\mathbf{X}_k = \mathbf{U}_k \Sigma_k \mathbf{V}_k^T
\]

where \( \mathbf{U}_k \in \mathbb{R}^{n \times k} \), \( \Sigma_k \in \mathbb{R}^{k \times k} \), \( \mathbf{V}_k \in \mathbb{R}^{m \times k} \) and \( k < \min(m, n) \).

In this representation the \( m \) columns \( \mathbf{X}_k \) represent the weighted term-frequency vectors (of size \( n \)) of each of the documents used to create the semantic space. The column vectors of orthonormal matrices \( \mathbf{U}_k \) and \( \mathbf{V}_k \) are the left and right singular vectors respectively. \( \Sigma_k \) the non-negative diagonal matrix of the \( k \) biggest singular values sorted in descending order. The rows of \( \mathbf{U}_k \) and \( \mathbf{V}_k \) can be interpreted as the coordinates of points that represent terms and documents respectively in the \( k \) dimensional space.

If new documents need to be represented in this semantic space, they can be represented as \( \mathbf{d} \in \mathbb{R}^n \) and projected on the \( k \)-dimensional space as:

\[
\hat{\mathbf{d}} = \mathbf{d}^T \mathbf{U}_k \Sigma_k^{-1} \mathbf{V}_k^T
\]

The result \( \hat{\mathbf{d}} \) is a \( k \)-dimensional vector that can be compared with other documents in the original knowledge base corpora or with other query documents. The criteria to decide the value of \( k \) still remains an open question for LSA and is usually set for individual experiments (Dumais 1991, Landauer et al. 1998, Haley et al. 2005).

2.2 Use of small corpora

Most studies on LSA have used large corpora to create semantic spaces, however the right size for a corpus remains an open question. In a recent report by Giesbers et al. (Giesbers et al. 2006) they argued that it is not clear what a small or large corpus is, the same applies for minimum or maximum size.

In LSA’s seminal paper, Deerwester suggested that a reasonable size for a corpus should be between 1,000-2,000 documents and 5,000-7,000 terms (Deerwester et al. 1990), this is not surprising because LSA’s original purpose was to improve IR accuracy by finding the underlying concept in the words of a user query. The technique’s accuracy is based on redundancy in the corpora, the premise being that terms that appear together are conceptually related.

Gong and Liu (Gong & Liu 2001) serendipitously used a single document to generate a semantic space for automatic summarization but did not compare it to other techniques or analyze the trade-offs involved. The technique has been used to create visualizations (Stephen O’Rourke 2009), labeled clusters (Osiniski 2006), and automatic feedback for students (Villalon et al. 2008) but again the technique itself was not analyzed in detail and questions remain on how it works.

In the summarization work, semantic spaces created from a single document were used to extract the most relevant sentences in a document identifying topics in the document and selecting the most representative sentences for each topic. Semantic spaces were created using each of the document’s paragraphs or sentences, with each singular vector obtained from the SVD representing a different topic (Steinberger et al. 2007).

2.3 Measuring coherence with LSA

One important aspect of the quality of essays is coherence (or cohesion), which reflects how the author links related pieces of information to create the essay’s structure. Foltz explains that for a text to be coherent it requires a high quality “overlap and transitions of the meaning as it flows across the discourse”, and LSA is able to model this phenomenon “by measuring the semantic similarity of one section of text to the next” (Foltz 2007).

The measurement of coherence using LSA was first proposed by Foltz et al. (7), in their study they calculated the essay’s coherence as the average distance between consecutive sentences. They found that the LSA measure correlated better with human comprehension of the text than other automatic measures for coherence such as word overlap and readability indices. Another study by Higgins et al. (?) used an LSA like algorithm to calculate coherence as a quality measure for automatic essay grading. They calculated the semantic distance between sentences, between discourse segments (usually paragraphs) and the prompt (text of the essay question), and between each discourse segment and the essay thesis (one of the discourse segments). They found that all three aspects correlated higher than a random baseline. Another study by Graesser et al. used LSA to calculate several distances between sentences, paragraphs and the whole document, including consecutive sentences and consecutive paragraphs to measure coherence and provide feedback to students (Graesser et al. 2004). These measures were used to reliably identify the writing style of different authors (7). Many other examples of measuring coherence with LSA can be found in (Foltz 2007).
the limits have to be set as a function of the essay content.

Other factor analysis techniques that perform dimensionality reduction, such as Principal Component Analysis, have defined criteria to find $k$. These criteria can be classified in three: Ad-hoc but intuitively plausible, and statistically based, both with and without distributional assumptions (?). The first one includes selecting $k$ based on the singular values in $\Sigma$, like a percentage of the cumulative variance, or all values above 1 (Kaiser’s rule). The second includes all singular values until the data fits an assumed distribution. The third one also includes singular values until a criterion such as cross-validation or bootstrapping is fulfilled.

However, deciding an appropriate value of $k$ for single document semantic spaces presents added complexities. In LSA there are no theoretical frameworks that explain the distribution of singular values in text corpora, least of all in single document spaces. Moreover, short essays usually have a lower number of sentences than different terms, and each one has a different length. Therefore $k$ can only take values between 1 (only one dimension) and the total number of sentences in the essay (no dimensionality reduction).

In order to study how dimensionality reduction in single document spaces affects its passage distances, $k$ will be evaluated from its minimum to its maximum values. Therefore $k$ was defined as a percentage of the maximum number of dimensions, which corresponds to the minimum between the number of passages and the number of words, varying from 5% to 100% in steps of 5% each.

### 4 Evaluation

Essays (N=43) collected as a diagnostic activity for first year university students were used to evaluate different semantic spaces. The essays were written in an activity where students first read three short papers on the topic of English as a global language, and then answered two questions: Is English becoming a global language? Is this a positive development? The three readings were 874, 812 and 888 words long respectively (888 average). Distances between consecutive sentences were calculated for each essay using different semantic spaces created from different background knowledge: Single document, large corpora and prescribed readings for the activity. Using the notation defined earlier, each of these three semantic spaces will produce different $U_k$ and $\Sigma_k^{-1}$.

Each sentence in an essay ($d$) was projected on the particular semantic space being studied, producing a $\hat{d}$ vector. Having the coordinates of each sentence on a semantic space, the distance between each consecutive $\hat{d}$ was then calculated using a cosine function.

The distances change for LSA spaces generated with different background knowledge, but for the applications of concern in this study only the relationships (relative distances) between the sentences are required. The distances were then ranked, and the ranking correlation was calculated using Spearman’s $\rho$. The statistical significance was calculated using a permutation-test (also known as randomization-test and exact test) (Zar 1972). All statistics were calculated using the JSC Java library for statistical computation. The correlation for the collection of essays was calculated as the average of the correlations for each essay.

1. The Snowball analyzer was used, which in turn uses the Porter’s stemmer.
2. For the single document space there’s no need to project the sentences because they actually formed the space.
3. http://www.jsc.nildram.co.uk/
Spearman’s correlation reliability is affected by ties groups (two or more distances with the same value), because these distances are interchangeable in the ranking without affecting the correlation (Zar 1972). The bigger the group of ties, the bigger the effect on the reliability, also more than one group of ties can occur in each document; hence the average size of ties groups for each document was also analyzed using the same conditions as for the distances (when no ties were found in a document, its average ties group size was assigned to 1).

### 4.1 Large corpora

The distances between consecutive sentences of each essay were calculated with LSA spaces produced by six different corpora in the Colorado LSA website (Dennis 2007). The corpora used included: five sets produced by TASA (Touchstone Applied Science Associates, Inc.) 3rd Grade (6,974 documents and 29,315 terms), 6th Grade (17,949 documents and 55,105 terms), 9th Grade (22,211 documents and 63,582 terms), 12th Grade (28,882 documents and 92,409 terms); and ‘Psych’ (13,902 documents and 55,105 terms), 12 Grade (28,882 documents and 92,409 terms); and ‘Psych’ (13,902 documents and 30,119 terms). For each corpora the first 300 dimensions (k) were preserved.

Table 2 shows the average rank correlation of the 43 essays projected on the different TASA (Touchstone Applied Science Associates, Inc.) corpus and ‘Psych’. Since the reading materials that students in 6th year are supposed to know include those of 3rd year, every TASA corpus includes those of lower literacy levels, and this can be seen in how the bigger the difference in content included the smaller the correlation.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>06th Grade</th>
<th>09th Grade</th>
<th>12th Grade</th>
<th>College</th>
<th>Psych</th>
</tr>
</thead>
<tbody>
<tr>
<td>03rd Gr</td>
<td>0.93</td>
<td>0.9</td>
<td>0.88</td>
<td>0.85</td>
<td>0.79</td>
</tr>
<tr>
<td>06th Gr</td>
<td>0.97</td>
<td>0.95</td>
<td>0.92</td>
<td>0.9</td>
<td>0.8</td>
</tr>
<tr>
<td>09th Gr</td>
<td>0.97</td>
<td>0.95</td>
<td>0.9</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>12th Gr</td>
<td>0.98</td>
<td>0.98</td>
<td>0.98</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>College</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.81</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Spearman’s rank correlation for each LSA colorado corpora

The average distance between consecutive sentences dropped faster for both LogEntropy and TFIdf. Global weights (Entropy and Idf) are used to smooth the effect of rare terms and/or too long documents in large corpora, however in single documents semantic spaces most terms have low frequencies (therefore there are no rare terms), and the length of sentences is limited (therefore there are no long documents). Therefore using them produced an information loss that made the sentences more dissimilar.

The final analysis was the average ties group size calculated for all k values and weighting schemes. Figure ?? shows that with k = 5%, an average group size of 9 ties was produced, which corresponds to almost 50% of the sentences. As more dimensions were added the groups of ties diminish, almost disappearing between k = 20% and k = 50%, and starting to grow again reaching an average of 4 when all dimensions were included. This phenomenon was caused by the distances between sentences, because ties are produced by too similar or too dissimilar sentences.

The final analysis was the average Spearman’s correlation ρ of ranked distances for the 43 essays projected both on the single document and the TASA 1st year college semantic spaces, calculated for all k values and term weighting schemes. The correlation divided by the average ties group size was also calculated (ρ in the graph), to illustrate the loss of reliability in the calculation of ρ. Figure 1 shows that the TF weighting scheme is the best, with values above 0.6 when k = 40%, and continued growing as dimensions were added. The adjusted ρ started falling around k = 55% due to the appearance of ties, however the average ties group size remained below 2 (the average ties group size for the large corpora) until k = 80%, when ρ reached 0.7. Statistical significance was also calculated but not shown in the graph. For both LogEntropy and TFIdf it was never below 0.05 (95% confidence), until almost all dimensions were included (k = 90%), results that due to the ties produced were not reliable to make a fair comparison. TF or the other hand rapidly achieved a significance below 0.05 and it was stable between k = 35% and k = 75%.

### 4.3 Prescribed readings

The distances between consecutive sentences were also calculated using a semantic space created from a small corpus containing the prescribed readings (126 sentences and 199 terms). The single document space correlated with the readings space almost identically than with the TASA corpus (0.71, p < 0.001) even though the readings space was three orders of magnitude smaller than the TASA corpus. This could be explained by the specificity of the readings, because the essays were expected to have similar content to them. Finally the correlation between the readings space and the TASA corpus was slightly higher (0.76, p < 0.001) than the single document corpus. This could be explained by associations between terms that were explicit both in the readings and the TASA corpus, but implicit in the students’ essay.
5 Conclusion

This paper questioned the premise that a knowledge base is needed to generate semantic spaces used in summarization, visualization and other applications. Different approaches to constructing semantic spaces were evaluated.

First we showed that the distances between consecutive sentences change significantly between the representations produced by different background knowledge. This makes certain applications unfeasible or more demanding because customized knowledge bases are required.

The results show that the collection used to create the semantic spaces must be taken into account for each application. For example, the differences between using a collection of documents such as those read by a 3rd year primary school student and those read by college student are not substantial ($\rho = 0.85$) while they are significant when using a collection of medical papers (a topic not related to the topic of the essays).

We also evaluated how the inter-sentence distance on a semantic space generated only using the sentences of the essay correlates to those generated with the College database. We found that good rank correlation can be achieved (up to 0.7) using a variable number of dimensions, 75% of the maximum possible.

References


Figure 2: Average ties groups size vs $k$


Figure 3: Average Spearman $\rho$ and $\rho'$ for the 43 essays projected on both the single document and TASA 1st year college semantic spaces vs $k$. 