Application of Latent Semantic Analysis to Article Coherence Evaluation

Henrique Stumm Rocha

March 2021

1 Introduction

Latent Semantic Analysis is a method to extract information from text that was first proposed by Scott Deerwester in 1988 [**?**], and has since found many applications in natural language processing (NLP). These applications include, but are not limited to:

1. Topic Modeling: extracting clusters of words from a text that correspond to a given topic, classifying texts according to topic
2. Coherence Evaluation: comparing coherence between different bodies of text
3. Information Retrieval: Searching for specific information in a passage

The second application cited above has the potential to be useful in a number of situations. In particular, evaluating coherence in news articles can be beneficial towards producing better articles in journalistic settings. In this context, Latent Semantic Analysis can be a helpful tool.

The main goal of this paper is to investigate the correlation between what humans perceive as coherent and the scoring of the LSA algorithm, by comparing the scores of a collection of news articles and nonsensical text. I collect news articles from reddit links and use a pairwise method to create the coherence score for a large block of text. The scores were documented and the same method was applied to a collection of mostly grammatically correct incomprehensible text.

In the subsequent section, we review the mechanism behind latent semantic analysis

2 Latent Semantic Analysis

First, the algorithm extracts the term-document matrix of a text. A term-document matrix is a table in which the frequencies of each word of a passage of text are correlated with specific passages. In Table 1 The columns correspond to the individual passages,

and the rows correspond to the words. Stopwords (prepositions, articles, and other words that do not convey meaning by themselves) are commonly removed to improve model comprehension. To illustrate the deriving of a term-document matrix, consider the following:

	d1	d2	d3	d4
t1	1	0	2	0
t2	3	0	0	2
t3	9	1	3	5

Table 1: Example table

The matrix in Table 1 represents that there are 9 occurrences of term 3, 3 occurrences of term 2 and 1 occurrence of term 1 in document 1, for example. Subsequently, Singular Value Decomposition is applied to separate the term-document matrix X into 3 different factors:

$$X = U \cdot \Sigma \cdot V^T$$

Following suit, the dimensionality of the matrix decomposition is reduced by discarding the rows of U and the columns of V^T corresponding to the smallest singular values in Σ. The reason for this is that, as elaborated in **?** [**?**], a classification of text by word frequency usually has a higher dimensionality than the semantic space from which the writer derives the meaning. That is consistent with the fact that multiple words contribute to the general meaning a certain passage is trying to convey. Thus, by dropping the associations that least contribute to the formation of the X matrix, one can approximate the relationships inherent to semantic space.

The vectors related to individual documents can be compared by cosine similarity to produce a metric that is well correlated with human subjective measures of coherence. As a latent model, it can do so without any training We apply such method to news articles in the following section

3 Methodology

We chose reddit as a platform from which to acquire news links because it has a relatively easy to use API, from which individual post content can be extracted automatically. Subreddit r/news contains posts consisting of links to news sources in the web, so we created a scraper to follow them and parse the text contained in the linked articles. Articles were collected from 4 news sources very prevalent among links on the subreddit r/news: Reuters, New York Times, CNN and BBC News. Their paragraphs were separated into individual strings and the entire collection of strings pertaining to a single article was converted to a term-document matrix. Vectors that corresponded to adjacent paragraphs in the matrix V^T had their pairwise coherence evaluated with

cosine similarity. The coherence scores of these pairs was averaged out over an entire news source, producing the following results. k corresponds to the final number of dimensions used in the calculation:

4 Results and Discussion

	Control Group [1]	CNN (30 articles)	Reuters (28 articles)
k = 20	0.66	0.730	0.68
k = 10	0.66	0.738	0.672
k = 5	0.673	0.764	0.70

	BBC News (18 articles)	New York Times (13 articles)
k = 20	0.70	0.71
k = 10	0.70	0.70
k = 5	0.74	0.760

Table 2: Average cosine similarity between adjacent paragraphs decomposed with LSA. Values of k = 5, 10 and 20 were tested for all the news sources cited above, extracted from r/news

As can be seen, the 5-dimensional case was the one that approximated semantic space the most, from the wider differences between the well-structured articles and the control group.

5 Conclusion

Latent Semantic Analysis appears to be capable of finding the difference between the coherence of the news articles and the control group, particularly in the 5-dimensional case. This indicate that using latent semantic analysis coupled with cosine similarity for pairs of similarly sized passages in texts can yield a good approximation for human measures of language coherence for journalism. Further work will analyze articles on a case-by-case manner and compare that with human ratings to understand what exactly contributes to readability and logic. Deep Learning may prove to be much more efficient for such a task, given that it has achieved success in many other areas of NLP in recent years.

Additional work will also attempt to associate the latent semantic algorithm with Deep Learning. I hypothesize that by incorporating the latent semantic coherence between the an external database and the output of a more robust artificial intelligence training algorithm into the loss function of such algorithm, training times could be sped up. A weighted latent semantic analysis transform could also be used as an additional

parameter for language models, provided that the operation is differential.

As of 2021, this approach to natural language processing has largely been overshadowed by large deep learning models, particularly transformers. Additional developments will attempt to utilize these algorithms for this reason. Furthermore, Latent Semantic Analysis does not leverage the full semantic qualities of natural language to create its semantic space abstraction. It only uses the term-passage relative frequencies. More powerful models would be able to infer relationships from additional data, such as word order. In fact, transformer deep learning models can use positional data for inference, unlike the approach utilized in this article.