Ordering Sentences According to Topicality

Ilana Bromberg
The Ohio State University
bromberg@ling.ohio-state.edu

Abstract

This paper addresses the problem of finding or producing the best ordering of the sentences in a text. I focus on using semantic properties of the words, as well as the high-level structure of each of the texts, to produce or choose the best ordering. In choosing an original sentence ordering from a group of sentential permutations of the same text, information about the topic of the text is helpful when added to information about coherence. A term distribution method of determining the semantic properties of words is successfully used to locate the most topical sentence in each text.

1 Introduction

As noted in Barzilay et al. (2002), the quality of a single- or multi-document summarization is greatly affected by the order in which the information is presented. Barzilay et al. (2002) demonstrated that the chronological order of event-related information and topical relatedess are important factors in generating a good ordering. Topical relatedness is generally used in multi-document summarization to group together sentences that relate to the same theme, in part so that the same information from different source documents is not repeated within the summary (cf. McKeown et al. (1999)). The goal of this paper is to incorporate topicality into the surface ordering of the generated text.

In Barzilay and Lapata (2005), the researchers train several different models to select the best ordering from among sentential permutations of a given text. Adjacent pairs of sentences are measured for local cohesion using features of syntax, salience, coreference, and/or semantics. The overall coherence of the text is based on these sentence-pair measurements, with greater coherence values preferred. At best, their models succeed in choosing an original ordering over a permutation 90.4% of the time. Their baseline model for measuring coherence, which takes into account only semantic features, succeeds 72.1% of the time.

The texts in question are newspaper articles, a domain with a rather specific overall structure. In most newspaper articles, it is reasonable to say that the information at the beginning of the text is very relevant to the topic of the article, and as one reads further, the details become less relevant to the topic. This rhetorical structure is described in tutorials on writing newspaper articles, such as in Nelson (2005): "The story should start with the 'lead paragraph' which is the summary of the story... The lead paragraph should include the who, what, when, where, and why of the story."

In this paper, I propose a method of using this topicalized structure of newspaper text to try to automatically construct or choose the correct ordering of the texts. A numerical semantic representation of each sentence in the article is created via a term distribution analysis. From these numeric representations, an approximation of the topic of the article is calculated. This topic is taken into account in judging the sentence order of a text. I will try to improve upon the baseline results in Barzilay and Lapata (2005) by combining their semantic coherence metric with this topicality metric.

The next section discusses the nature of the data and the type of semantic analysis performed. Sec-

tion 3 reports on two sets of experiments designed to incorporate topicality in the process of choosing or creating the best sentence ordering. The paper ends with a discussion of the results.

2 The Data

The data are 74 texts in the AP Natural Disaster corpus[1], one of the corpora used in Barzilay and Lapata (2005). These are newspaper articles concerning such events as earthquakes and hurricanes. The number of sentences in each text ranges from 4 to 24, with an average of 10.2 sentences.

2.1 Semantic Representations

The semantic representations of the sentences are produced using Latent Semantic Analysis (Landauer et al. (1998)). This method exploits the hypothesis that words with similar meanings tend to be found in similar contexts. By recording the distribution of the words in large corpus, relationships of the words in question can be quantified. A numeric vector of a given dimensionality is calculated for each word based on its distribution, and these vectors are compared to determine the words' semantic similarity or disparity.

The first step of employing this method is to extract a word list from the 74 texts. Stopwords, numbers, and other uninformative words (such as file names present within the texts) are removed from the word list, and punctuation is removed. The resulting list has 3626 words. I then measure the distribution of each of these words among the documents of a larger corpus. For this, I chose the Reuters Corpus, which is a very large corpus of newspaper text.[2] I create a term-by-document matrix using 3626 terms from the Natural Disasters corpus and 806,799 documents from the Reuters corpus.

This large matrix undergoes singular value decomposition (SVD), producing three smaller matrices, that together are equally as expressive as the original large matrix. The dimensionality of these smaller matrices is chosen by the researcher. For this exercise, I chose three dimensions to work with: 50, 100, and 200. The study in Barzilay and Lapata (2005) uses 100 dimensions. In addition to 100, I chose a smaller and a larger dimensionality to see if the results differed. Of the three matrices that result from each implementation of SVD, I employ the one that describes each of the 3626 terms in the reduced dimensionality.

Each word is thus represented by a numeric vector of the same size. To create sentence vectors, the word vectors of each word in a sentence are summed to create a single numeric vector for that sentence. The length of the sentence affects the value of that sentence's vector. Furthermore, word repetition is taken into account.

A second way of deriving the sentence vectors is to calculate the average of the word vectors that comprise it. In this case, the differing lengths of the sentences are normalized.

Once sentence vectors have been calculated, they are compared using cosine distance. This is a measure of similarity, such that a distance of 1 signifies maximally similar vectors, and a distance of 0 signifies maximally dissimilar vectors.

Note that the cosine distance between two sentence vectors does not change based on whether the sum or average calculation is used. However, the calculation of the centroid, as addressed in the following section, is affected by this choice. In Experiment Set 1, sentence vectors are always calculated by summing the word vectors. In Experiment Set 2, I perform experiments with sentence vectors derived from both summed word vectors and averaged word vectors, and report the results separately.

2.2 Calculating the Centroid

A vector representing the average of all of the sentence vectors is called a centroid. I hypothesize that this centroid represents the topic of the text. Using cosine distance, I calculate the distance of each sentence vector from this centroid, thus giving a measurement of that sentence's topicality. This measure of topicality is employed in the following experiments to determine the quality of different sentence orderings.

[1] This corpus and the sentential permutations are available at http://people.csail.mit.edu/regina/coherence/

[2] The Reuters Corpus is available at: http://trec.nist.gov/data/reuters/reuters.html

Table 1: Original Text Orderings Chosen over Multiple Permutations

Dimensionality	50	100	200
Experiment 1A: Coherence Only	Percent Correct *(Number of 1833)*		
With Title and Header	**64.8** *(1187)*	59.3 *(1087)*	58.3 *(1069)*
Without Title and Header	60.2 *(1104)*	53.0 *(972)*	51.8 *(950)*
Experiment 1B: Coherence & Topicality			
With Title and Header	51.9 *(952)*	51.6 *(946)*	51.4 *(942)*
Without Title and Header	65.0 *(1191)*	62.5 *(1146)*	**66.1** *(1212)*

3 Experiments

3.1 Experiment Set 1: Choosing the Original Ordering

The first set of experiments replicate and expand on the work done in Barzilay and Lapata (2005). The task is to assign a score to each original text and up to 20 of its sentential permutations. After assigning scores, the algorithm counts how often the original texts have higher scores than their permutations. I build two models to assign scores; in Experiment 1A, only local cohesion is calculated, and in Experiment 1B, the notion of topicality as reflected by the semantic centroid is also taken into account.

3.1.1 Experiment 1A

To replicate the original experiment, I begin with the LSA model on its own, without introducing any notion of topicality. Sentence vectors are created as explained above, using only word vector sums. For each text and its permutations, I compute the cosine similarity between adjacent sentences in the text. The average of these cosine similarities is the overall coherence measure of the text. When comparing two texts, the one with the larger average score is considered the more coherent text. Table 1 shows the extent to which the original orderings are considered more coherent than their permutations.

In briefly looking at the texts, I found that the first sentence listed contains the title and header information, not the lead sentence of the article. Therefore, I performed this experiment on texts with and without this title and header included. The best result for Experiment 1A is produced when the word and sentence vectors are built with 50 dimensions, and the title line remains in the text. In this case, the algorithm chooses the original ordering over its permutations 64.8% of the time, which represents the best baseline result.

3.1.2 Experiment 1B

The second experiment incorporates topicality in the following way: The centroid of the text is calculated, and the cosine distance between the first sentence in the text and the centroid is measured. This cosine similarity score is added to the coherence score as described in Experiment 1A. If the first sentence is close to the centroid, the score will increase by more than if the first sentence is not central. Again, the preferred ordering is the one with the higher overall score. Table 1 shows the results of this model for each SVD dimension, and for texts with and without the title line removed. The best result occurs when the word and sentence vectors are built with 200 dimensions, and the title line is removed from the text. In this case, the algorithm chooses the original ordering over its permutations 66.1% of the time.

The results of these experiments can be fitted to a binomial curve, such that the mean result would be 916.5 correct with a standard deviation of 21.4. The best results of Experiment 1A and 1B, 1187 and 1212 correct respectively, fall well outside three standard deviations, and can be considered significantly better than chance.

3.2 Experiment Set 2

In the next set of experiments, I further expand on the ordering task by creating an ideal ordering of each set of sentences, based only on their semantic representations, and taking into account topicality. The quality of the ordering I create is determined by comparing it to the original ordering.

Figure 1: Distribution of Sentences Closest to the Centroid, 100 Dimensions, *Word Vectors Summed*

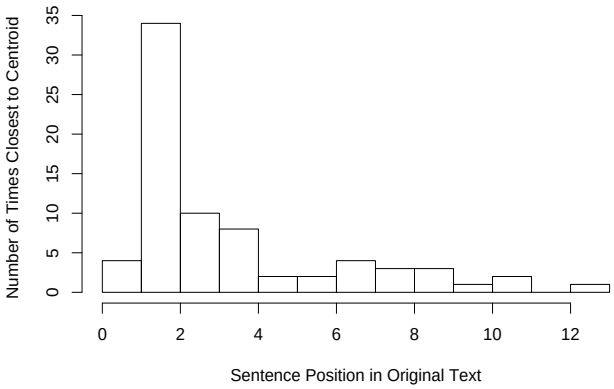
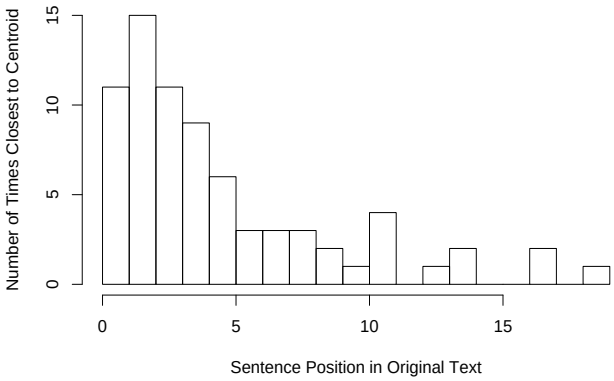


Figure 2: Distribution of Sentences Closest to the Centroid, 100 Dimensions, *Word Vectors Averaged*



3.2.1 Determing the Lead Sentence

The first step of the ordering task is always to choose the lead sentence. This is done by determining which of the sentence vectors is closest to the centroid vector, and thus most representative of the topic. Figure 1 shows the number of times each sentence position (i.e., the position of the sentence in the original text) was chosen as the most topical sentence. It is clear that the second sentence of the original texts is chosen as most topical more often than any other. Recall that the second sentence of the original text is actually the lead sentence of the article, since the title and header appear before it. The distribution shown is for calculations made with vectors of 100 dimensions, and sentence vectors as word vector sums. Very similar distributions occur when using vectors of 50 and 200 dimensions.

A slightly different distribution occurs when using word vector averages to compute the sentence vectors. As seen in Figure 2, sentence one (the header and title) is often chosen as the most topical, but still not as often as sentence two (the actual lead sentence). Table 2 shows, for all of the analyses (dimensionalities and sentence vector representations), how often the second sentence was chosen as most topical.

Figures 1 and 2 and Table 2 show the effects of the differing sentence representations. The length of sentences plays a role in determining the placement of the centroid and the distance of each sentence to the centroid. A 'long' sentence vector calculated as the sum of many word vectors has greater magnitude and more 'pull' on the centroid than a 'short' sentence vector. In the texts, the lead sentence is usually among the longest of the article. Consequently, when word vectors are summed, the actual lead sentence is found closest to the centroid more often than when the word vectors are averaged. However, there is more at work here than length, as depicted in Figure 2. Figure 2 shows that sentence two is still most often the closest to the centroid, in an environment where sentence length is not a factor in determining the closeness of each sentence to the centroid. Furthermore, the sentences at the start of the article, including the title and header, are found closest to the centroid more often than later sentences. This supports the assumptions that the beginning of the article is more topical than the end, and the centroid vector is a good approximation of the topic.

The dimensionality of the word and sentence vectors also plays a role. The vectors of larger dimensionality produce a larger correct percentage for the choice of first sentence.

To judge the effectiveness of this method against a baseline, I compare the best result to the results

Table 2: How often is Sentence 2 chosen as most topical? Number correct out of 74 trials.

	Dimensionality of Vectors		
	50	100	200
Word vector sums	29	34	40
Word vector avgs	14	15	16

achieved by simply assuming that the longest sentence in the article is the lead sentence. Table 3 shows this comparison. Choosing the longest sentence produces the correct result 38 times, compared to the semantic analysis method's 40 correct choices. The two methods both choose correctly for 25 of the texts. When the two methods disagree, the semantic analysis method chooses correctly more often than the baseline method. I learn from Table 3 that a good method for choosing the lead sentence might be one that takes into account both the topicality as measured here, and also the length of the sentence relative to others in the text.

Table 3: How does the semantic analysis method (best result) differ from picking the longest sentence?

		Closest to Centroid	
		Lead	Non-Lead
Longest sentence	Lead	25	13
	Non-Lead	15	21

In summary, ranking sentences by their closeness to a semantic centroid is a generally successful way of finding the lead sentence in a text, or the most topical information in that text.

3.2.2 Ordering the Remaining Sentences

The method of ordering the remaining sentences differs for each of three experiments.

In Experiment 2A, the sentences of each text are placed in order of closest to farthest from the centroid.

In Experiment 2B, the centroid is recalculated for each sentence, following the removal of the sentence judged closest. That is, the centroid is calculated over the whole set, the closest sentence is placed first in the order, and is removed from the set. The centroid is recalculated, and the closest sentence to that new centroid is chosen as second. This process is repeated until all sentences are accounted for.

In Experiment 2C, the cosine similarity between each of the sentences is calculated. After the first sentence is chosen based on its distance to the centroid, each subsequent sentence is chosen based on its similarity to the one previously chosen. This is meant to mimic the cohesion metric that was the baseline in the work of Barzilay and Lapata (2005).

The orderings that are produced by each of these experiments are judged, using Kendall's tau, by comparing them to the original orderings. Kendall's tau is a correlation statistic used to judge rank data. The value of Kendall's tau ranges between -1 and 1, with -1 or 1 as perfectly negative or positive correlations, respectively, and 0 representing no correlation (Kendall (1948)). The correlations shown in Table 4 are average correlations over the orderings of all texts. It is clear that none of the methods of ordering the texts is successful. The experimental orderings are not correlated with the original orderings, regardless of vector dimensionality, or whether sentence vectors are created through word vector sums or averages.

4 Discussion

Using Latent Semantic Analysis to find the most topical sentence works reasonably well when sentence vectors are calculated by taking the sum of the word vectors included in that sentence. In order for this to work optimally, the title and header should not be included in the calculation. A higher dimensionality for SVD leads to better results. Using 200 dimensions to describe each sentence, the lead in sentence is chosen 54.1% of the time with this method. Because the texts all have 4 or more sentences, this figure is above chance. Furthermore, it is a more successful method than a baseline of choosing the longest sentence as most topical.

The results reported for Experiment 1A should mirror those of Barzilay and Lapata (2005), however, the scores reported here are somewhat lower than the previously reported results. This may be due to differences in the method, including the manually edited word list used here, as well as the use of different newspaper corpora for building the term-

Table 4: Average correlation of experimental orderings to original orderings

	Experiment 2A			Experiment 2B			Experiment 2C		
Dimensions:	50	100	200	50	100	200	50	100	200
Sums	-.139	-.126	-.105	-.215	-.197	-.182	-.268	-.277	-.236
Averages	-.182	-.176	-.137	-.206	-.215	-.208	-.155	-.171	-.096

by-document matrix. It seems that the title sentence and lead sentence are highly cohesive in these texts, as can be seen in the difference between the results of the two parts of Experiment 1A, in Table 1.

Leaving out the title sentence allows a greater ability to estimate the most topical sentence in the text. Doing so in Experiment 1B leads to better results. The originally ordered text is chosen over its permutation 66.1% of the time when both coherence and topicality are taken into account.

5 Conclusion and Future Work

The results of Experiment 1B show that the baseline coherence measure in Barzilay and Lapata (2005) can be improved by taking into account topicality. Future work may show that adding topicality to the syntactic measures also incorporated in Barzilay and Lapata (2005) improves the ordering preference models.

While the methods described in Experiment Set 2 did not often produce the original orderings, it may be that the order produced was often comprehensible. As shown in Barzilay et al. (2002), more than one order of information may be acceptable for a given text. A study involving human ratings of the produced orderings is another possible way of testing the results of the various experiments.

References

Regina Barzilay and Mirella Lapata. 2005. Modeling Local Coherence: An Entity-Based Approach. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 141–148, Ann Arbor, MI.

Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Morris H DeGroot. 1986. *Probability and Statistics*. Addison Wesley Publishing Company, 2nd Edition.

M. Kendall. 1948. *Rank Correlation Methods*. Charles Griffin & Company Limited.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin. 1999. Towards Multidocument Summarization by Reformulation: Progress and Prospects. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*.

P. Nelson. 2005. How to Write a Newspaper Article.