Using Thematic Information in Statistical Headline Generation

Stephen Wan
Center for Language Technology
Macquarie University
Sydney, Australia
swan@ics.mq.edu.au

Mark Dras
Center for Language Technology
Macquarie University
Sydney, Australia
madras@ics.mq.edu.au

Cécile Paris
CSIRO Mathematical and Information Sciences
Locked Bag 17
North Ryde 1670
Sydney, Australia
Cecile.Paris@csiro.au

Robert Dale
Center for Language Technology
Macquarie University
Sydney, Australia
rdale@ics.mq.edu.au

Abstract

We explore the problem of single sentence summarisation. In the news domain, such a summary might resemble a headline. The headline generation system we present uses Singular Value Decomposition (SVD) to guide the generation of a headline towards the theme that best represents the document to be summarised. In doing so, the intuition is that the generated summary will more accurately reflect the content of the source document. This paper presents SVD as an alternative method to determine if a word is a suitable candidate for inclusion in the headline. The results of a recall based evaluation comparing three different strategies to word selection, indicate that thematic information does help improve recall.

1 Introduction

Ours is an age where many documents are archived electronically and are available whenever needed. In the midst of this plethora of information, the successful completion of a research task is affected by the ease with which users can quickly identify the relevant electronic documents that satisfy their information needs. To do so, a researcher often relies on generated summaries that reflect the contents of the original document.

We explore the problem of single sentence summarisation, the primary focus of this paper. Instead of identifying and extracting the most important sentence, we generate a new sentence from scratch. The resulting sentence summary may not occur verbatim in the source document but may instead be a paraphrase combining key words and phrases from the text.

As a precursor to single sentence summarisation, we first explore the particular case of headline generation in the news domain, specifically English news. Although headlines are often constructed to be sensationalist, we regard headline generation as an approximation to single sentence summarisation, given that a corpus of single sentence summaries does not exist.

Our system re-uses words from the news article to generate a single sentence summary that resembles a headline. This is done by selecting and then appending words from the source article. This approach has been explored by a number of researchers (eg. see Witbrock and Mittal, 1999; Jin and Hauptmann, 2002) and we will describe their work further in the next section. In existing approaches, a word is selected on the basis of two criteria: how well it acts as a summary word, and how grammatical it will be given the preceding summary words that have already been chosen.

The purpose of this paper is to present work which investigates the use of Singular Value Decomposition (SVD) as a means of determining if a word is a good candidate for inclusion in the headline.

To introduce the notion of using SVD for single sentence summarisation in this paper, we examine the simplest summarisation scenario. Thus, presently we are only concerned with

single document summarisation. In addition, we limit the focus of our discussion to the generation of generic summaries.

In the remainder of this paper, we describe our motivation for using SVD by describing difficulties in generating headlines in Section 2. In Section 3, as motivation for our approach, we illustrate how words can be used out of context, resulting in factually incorrect statements. Section 4 provides an overview of related work. In Section 5, we give a detailed description of how we generate the sentence summary statistically and how we use SVD to guide the generation process. In Section 6, we present our experimental design in which we evaluated our approach, along with the results and corresponding discussion. Finally, in Section 7, we present our conclusions and future work.

2 The Veracity of Generated Summaries

Berger and Mittal (2000) describe limitations to the generation of headlines by recycling words from the article. One such limitation is that the proposition expressed by the generated summary is not guaranteed to reflect the information in the source text. As an example, they present two sentences of differing meaning which uses the same words. We present their example in Example 1, which illustrates the case in which the subject and object are swapped.

The dog bit the postman
The postman bit the dog.

Example 1. An example of different propositions presented in two sentences which use the same words.

However, we believe that the veracity of the generated sentence, with respect to the original document, is affected by a more basic problem than variation in word order. Because words from any part of a source document can be combined probabilistically, there is a possibility that words can be used together out of context. We refer to this as *Out-of-Context* error. Figure 1 presents an example of a generated headline in which the adverb wrongly reports stock price movement. It also presents the actual context in which that adverb was used.

Generated headline
"singapore stocks shares rebound""

Actual headline:
"Singapore shares fall, seen higher after holidays."

Original context of use of 'rebound':
"Singapore shares closed down below the 2,200 level on Tuesday but were expected to *rebound* immediately after Chinese Lunar New Year and Muslim Eid Al-Fitr holidays, dealers said."

Figure 1. An error in the generated headline due to a word being re-used out of context.

Out-of-Context errors arise due to limitations in the two criteria for selecting words mentioned in Section 1. While, for selection purposes, a word is scored according to its goodness as candidate summary word, word order is determined by a notion of grammaticality, modelled probabilistically using ngrams of lexemes. However, the semantic relationship implied by probabilistically placing two words next to each other, for example an adjective and a noun, might be suspect. As the name "Out-of-Context" suggests, this is especially true if the words were originally used in non-contiguous and unrelated contexts. This limitation in the word selection criteria can be characterized as being due to a lack of long distance relationship information.

3 Our Approach to "Encouraging Truth"

In response to this limitation, we explore the use of a matrix operation, *Singular Value Decomposition* (SVD) to guide the selection of words. Although our approach still does not guarantee factual correctness with respect to the source document, it has the potential to alleviate the Out-of-Context problem by improving the selection criteria of words for inclusion in the generated sentence, by considering the original contexts in which words were used. With this improved criteria, we hope to "encourage truth" by incorporating long distance relationships between words. Conceptually, SVD provides an analysis of the data which describes the relationship between the distribution of words and sentences. This analysis includes a

grouping of sentences based on similar word distributions, which correspond to what we will refer to here as the main *themes* of the document.[1] By incorporating this information into the word selection criteria, the generated sentence will "gravitate" towards a single theme. That is, it will tend to use words from that theme, reducing the chance that words are placed together out of context.

By reflecting the content of the main theme, the summary may be *informative* (Borko, 1975). That is, the primary piece of information within the source document might be included within the summary. However, it would remiss of us to claim that this quality of the summary is guaranteed. In general, the generated summaries are at least useful to gauge what the source text is about, a characteristic described by Borko as being *indicative*.

Figure 2 presents the generated summary using SVD for the same test article presented in Figure 1. In this case, the summary is informative as not only are we told that the article is about a stock market, but the movement in price in this example is correctly determined.

Generated headline using SVD:
"singapore shares fall"

Figure 2. The headline generated using an SVD-based word selection criterion. The movement in share price is correct.

4 Related Work

As the focus of this paper is on statistical single-sentence summarisation we will not focus on preceding work which generates summaries greater in length than a sentence. We direct the reader to Paice (1990) for an overview of summarisation based on sentence extraction. Examples of recent systems include Kupiec et al. (1995) and Brandow et al. (1995). For examples of work in producing abstract-like summaries, see Radev and McKeown (1998), which combines work in information extraction and natural language processing. Hybrid methods for abstract-like summarisation which combine statistical and symbolic approaches have also been explored; see, for example, McKeown et al. (1999), Jing and McKeown (1999), and Hovy and Lin (1997).

Statistical single sentence summarisation has been explored by a number of researchers (see for example, Witbrock and Mittal, 1999; Zajic et al., 2002). We build on the approach employed by Witbrock and Mittal (1999) which we will describe in more detail in Section 3.

Interestingly, in the work of Witbrock and Mittal (1999), the selection of words for inclusion in the headline is decided solely on the basis of corpus statistics and does not use statistical information about the distribution of words in the document itself. Our work differs in that we utilise an SVD analysis to provide information about the document to be summarized, specifically its main theme.

Discourse segmentation for sentence extraction summarisation has been studied in work such as Boguraev and Neff (2000) and Gong and Liu (2001). The motivation behind discovering segments in a text is that a sentence extraction summary should choose the most representative sentence for each segment, resulting in a comprehensive summary. In the view of Gong and Liu (2001), segments form the main themes of a document. They present a theme interpretation of the SVD analysis, as it is used for discourse segmentation, upon which our use of the technique is based. However, Gong and Liu use SVD for creating sentence extraction summaries, not for generating a single sentence summary by re-using words.

In subsequent work to Witbrock and Mittal (1999), Banko *et al*. (2000) describe the use of information about the position of words within four quarters of the source document. The headline candidacy score of a word is weighted by its position in one of quarters. We interpret this use of position information as a means of guiding the generation of a headline towards the central theme of the document, which for news articles typically occurs in the first quarter. SVD potentially offers a more general mechanism for handling the discovery of the

[1] *Theme* is a term that is used in many ways by many researchers, and generally without any kind of formal definition. Our use of the term here is akin to the notion that underlies work on text segmentation, where sentences naturally cluster in terms of their 'aboutness'.

central themes and their positions within the document.

Jin et al. (2002) have also examined a statistical model for headlines in the context of an information retrieval application. Jin and Hauptmann (2001) provide a comparison of a variety of learning approaches used by researchers for modelling the content of headlines including the Iterative Expectation-Maximisation approach, the K-Nearest neighbours approach, a term vector approach and the approach of Witbrock and Mittal (1999). In this comparison, the approach of Witbrock and Mittal (1999) fares favourably, ranking second after the term vector approach to title word retrieval (see Jin and Hauptmann, 2001, for details). However, while it performs well, the term vector approach Jin et al. (2002) advocate doesn't explicitly try to model the way a headline will usually discuss the main theme and may thus be subject to the Out-of-Context problem.

Finally, for completeness, we mention the work of Knight and Marcu (2000), who examine single sentence compression. Like Witbrock and Mittal (1999), they couch summarisation as a noisy channel problem. Under this framework, the summary is a noise-less source of information and the full text is the noisy result. However, in contrast to our approach, Knight and Marcu (2000) handle parse trees instead of the raw text. Their system learns how to simplify parse trees of sentences extracted from the document to be summarized, to uncover the original noise-less forms.

5 Generating a Single Sentence Summary

In this section, we describe our approach to single sentence summarisation. As mentioned earlier, our approach is based on that of Witbrock and Mittal (1999). It differs in the way we score words for inclusion in the headline. Section 5.1 presents our re-implementation of Witbrock and Mittal's (1999) framework and introduces the Content Selection strategy they employ. Section 5.2 describes our extension using SVD resulting in two alternative Content Selection strategies.

5.1 Searching for a Probable Headline

We re-implemented the work described in Witbrock and Mittal (1999) to provide a single sentence summarisation mechanism. For full details of their approach, we direct the reader to their paper (Witbrock and Mittal, 1999). A brief overview of our implementation of their algorithm is presented here.

Conceptually, the task is twofold. First, the system must select *n* words from a news article that best reflect its content. Second, the best (grammatical) word ordering of these *n* words must be determined. Witbrock and Mittal (1999) label these two tasks as *Content Selection* and *Realisation*. Each of these criteria are scored probabilistically, whereby the probability is estimated by prior collection of corpus statistics.

To estimate Content Selection probability for each word, we use the Maximum Likelihood Estimate (MLE). In an offline training stage, the system counts the number of times a word is used in a headline, with the condition that it occurs in the corresponding news article. To form the probability, this frequency data is normalised by the number of times the word is used in articles across the whole corpus. This particular strategy of content selection, we refer to this as the Conditional probability.

The Realisation criterion is determined simply by the use of bigram statistics, which are again collected over a training corpus during the training stage. The MLE of the probability of word sequences is calculated using these bigram statistics. Bigrams model the grammaticality of a word given the preceding word that has already been chosen.

It should be noted that both the Content Selection and Realisation criteria influence whether a word is selected for inclusion in the headline. For example, a preposition might poorly reflect the content of a news article and score a low Content Selection probability. However, given the context of the preceding word, it may be the only likely choice.

In both the training stage and the headline generation stage, the system employs the same preprocessing. The preprocessing, which

mirrors that used by Witbrock and Mittal (1999), replaces XML markup tags and punctuation (except apostrophes) with whitespace. In addition, the remaining text is transformed into lower case to make string matching case insensitive. The system performs tokenisation by using whitespace as a word delimiter.

In Witbrock and Mittal's approach (1999), the headline generation problem reduces to finding the most probable path through a bag of words provided by the source document, essentially a search problem. They use the beam search variety of the Viterbi algorithm (Forney, 1973) to efficiently search for the headline. In our implementation, we provided the path length as a parameter to this search mechanism. In addition, we used a beam size of 20.

To use the Viterbi algorithm to search for a path, the probability of adding a new word to an existing path is computed by combining the Content selection probability, the Realisation probability and the probability of the existing path, which is recursively defined. Combining each component probability is done by finding the logs of the probabilities and adding them together. The Viterbi algorithm sorts the paths according to the path probabilities, directing the search towards the more probable word sequences first. The use of repeated words in the path is not permitted.

5.2 Using Singular Value Decomposition for Content Selection

As an alternative to the Conditional probability, we examine the use of SVD in determining the Content Selection probability. Before we outline the procedure for basing this probability on SVD, we will first outline our interpretation of the SVD analysis, based on that of Gong and Liu (2001). Our description is not intended to be a comprehensive explanation of SVD, and we direct the reader to Manning and Schütze (2000) for a description of how SVD is used in information retrieval.

Conceptually, when used to analyse documents, SVD can discover relationships between word co-occurrences in a collection of text. For example, in the context of information retrieval, this provides one way to retrieve additional documents that contain synonyms of query terms, where synonymy is defined by similarity of word co-occurrences. By discovering patterns in word co-occurrences, SVD also provides information that can be used to cluster documents based on similarity of themes.

In the context of single document summarisation, we require SVD to cluster sentences based on similarities of themes. The SVD analysis provides a number of related pieces of information relating to how words and sentences relate to these themes. One such piece of information is a matrix of scores, indicating how representative the sentence is of each theme. Thus, for a sentence extraction summary, Gong and Liu (2001) would pick the top n themes, and for each of these themes, use this matrix to choose the sentence that best represents it.

For single sentence summarisation, we assume that the theme of the generated headline will match the most important theme of the article. The SVD analysis orders its presentation of themes starting with the one that accounts for the greatest variation between sentences. The SVD analysis provides another matrix which scores how well each word relates to each theme. Given a theme, scores for each word, contained in a column vector of the matrix, can then normalised to form a probability. The remainder of this section provides a more technical description of how this is done.

To begin with, we segment a text into sentences. Our sentence segmentation preprocessing is quite simple and based on the heuristics found in Manning and Schütze (2000). After removing stopwords, we then form a terms by sentences matrix, A. Each column of A represents a sentence. Each row represents the usage of a word in various sentences. Thus the frequency of word t in sentence s is stored in the cell A_{ts}. This gives us an $t * s$ matrix, where $t \geq s$. That is, we expect the lexicon size of a particular news article to exceed the number of sentences. For such a matrix, the SVD of A is a process that provides the right hand side of the following equation:

$$A = U.\Sigma.\ V^{transpose}$$

where U is a $t * r$ matrix, Σ is an $r * r$ matrix, and V is an $s * r$ matrix. The dimension size r is the rank of A, and is less than or equal to the number of columns of A, in this case, s. The matrix Σ is a diagonal matrix with interesting properties, the most important of which is that the diagonal is sorted by size. The diagonal values indicate the variation across sentences for a particular theme, where each theme is represented by a separate diagonal element. The matrix V indicates how representative a sentence is of a score. Similarly the matrix U indicates how related to the themes each word is. A diagram of this is presented in Figure 3.

Before describing how we use each of these matrices, it is useful to outline what SVD is doing geometrically. Each sentence, a column in the matrix A, can be thought of as an object in t dimensional space. SVD uncovers the relations between dimensions. For example, in the case of text analysis, it would discover relationships between words such as synonyms. In a trivial extreme of this case where two sentences differ only by a synonym, SVD would ideally discover that the two synonyms have very similar word co-occurrences. In the analysis matrices of U, Σ and V, the redundant dimensions corresponding to these highly similar words might be removed, resulting in a reduced number of dimensions, r, required to represent the sentences.

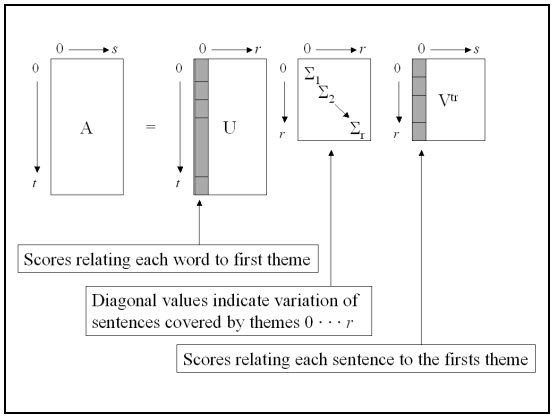

Figure 3. A diagram of our interpretation of the SVD matrices as it relates to single sentence summarisation.

Of the resulting matrices, V is an indication of how each sentence relates to each theme, indicated by a score. Thus, following Gong and Liu (2001), a plausible candidate for the most important sentence is found by taking the first column vector of V (which has s elements), and finding the element with the highest value. This sentence will be the one which is most representative of the theme. The index of that element is the index of the sentence to extract.

However, our aim is not to extract a sentence but to utilise the theme information. The U matrix of the analysis provides information about how well words correspond to a particular theme. We examine the first column of the U matrix, sum the elements and then normalize each element by the sum to form a probability. This probability, which we refer to as the SVD probability, is then used as the Content Selection probability in the Viterbi search algorithm.

As an alternative to using the SVD probability and the Conditional Probability in isolation, a Combined Probability is calculated using the harmonic mean of the two. The harmonic mean was used in case the two component probabilities differed consistently in their respective orders of magnitude. Intuitively, when calculating a combined probability, this evens the importance of each component probability.

To summarize, we end up with three alternative strategies in estimating the Content Selection Probability: the Conditional Probability, the SVD Probability and the Combined Probability.

6 Experiments

6.1 Data

In our experiments, we attempted to match the experimental conditions of Witbrock and Mittal (1999). We used news articles from the first six months of the Reuters 1997 corpus (Jan 1997 to June 1997). Specifically, we only examined news articles from the general Reuters category (*GCAT*) which covers primarily politics, sport and economics. This category was chosen not because of any particular domain coverage but because other categories exhibited frequent use of tabular presentation. The GCAT category contains in excess of 65,000 articles. Following Witbrock and Mittal (1999), we randomly selected 25,000 articles for training and a further

1000 articles for testing, ensuring that there was no overlap between the two data sets. During the training stage, we collected bigrams from the headline data, and the frequency of words occurring in headlines.

6.2 Experiment Design

We conducted an evaluation experiment to compare the performance of the three Content Selection strategies that we identified in Section 5: the Conditional probability, the SVD probability, and the Combined probability. We measure performance in terms of recall, i.e. how many of the words in the actual headline match words in the generated headline.[2] The recall metric is normalised to form a percentage by dividing the word overlap by the number of words in the actual headline.

For each test article, we generated headlines using each of the three strategies. For each strategy, we generated headlines of varying lengths, ranging from length 1 to 13, where the latter is the length of the longest headline found in the test set. We then compared the different strategies for generated headlines of equal length.

To determine if differences in recall scores were significant, we used the Wilcoxon Matched Pairs Signed Ranks (WMPSR) test (Seigel and Castellan, 1988). In our case, for a particular pair of Content Selection strategies, the alternate hypothesis was that the choice of Content Selection strategy affects recall performance. The null hypothesis held that there was no difference between the two content selection strategies. Our use of the non-parametric test was motivated by the observation that recall scores were not normally distributed. In fact, our results showed a positive skew for recall scores. To begin with, we compared the recall scores of the SVD strategy and the Conditional strategy in one evaluation. The strategy that was found to perform better was then compared with the Combined strategy.

In addition to the recall tests, we conducted an analysis to determine the extent to which the SVD strategy and the Conditional probability strategy were in agreement about which words to select for inclusion in the generated headline. For this analysis, we ignored the bigram probability of the Realisation component and just measured the agreement between the top *n* ranking words selected by each content selection strategy. Over the test set, we counted how many words were selected by both strategies, just one strategy, and no strategies. By normalising scores by the number of test cases, we determine the average agreement across the test set. We ran this experiment for a range of different values of N, ranging from 1 to 13, the length of the longest headline in the test set.

6.3 Results

6.3.1 Recall Comparison

The results for the comparison of recall scores are presented in Table 1 and Table 2. Table 1 shows results of the WMPSR test when comparing the SVD strategy with the Conditional strategy.[3] Since the Conditional strategy was found to perform better, we then compared this with the Combined strategy, as shown in Table 2. From Table 1, it is clear that, for all sentence lengths, there is a significant difference between the SVD strategy and the Conditional strategy, and so we reject the null hypothesis. Similarly, Table 2 shows that there is a significant difference between the Conditional strategy and the Combined strategy, and again we reject the null hypothesis. We conclude that SVD probability alone is outperformed by the Conditional probability; however, using both probabilities together leads to a better performance.

[2] Word overlap, whilst the easiest way to evaluate the summaries quantitatively, is an imprecise measure and must be interpreted with the knowledge that non-recall words in the generated headline might still indicate clearly what the source document is about.

[3] The performance of our Conditional strategy is roughly comparable to the results obtained by Banko, Mittal and Witbrock (2000), in which they report recall scores between 20% to 25%, depending on the length of the generated headline.

Sentence Length	Average Recall : SVD	Average Recall : Cond.	Probability	Reject H_0
1	03.68%	03.98%	p ≤ 0.0	yes
2	07.02%	06.97%	p ≤ 0.5	yes
3	10.05%	11.44%	p ≤ 0.0	yes
4	12.39%	13.90%	p ≤ 0.0	yes
5	14.21%	15.73%	p ≤0.0	yes
6	15.57%	17.84%	p ≤1.1e-05	yes
7	16.59%	19.14%	p ≤ 1.8e-07	yes
8	17.74%	20.30%	p ≤ 1.3e-07	yes
9	18.74%	21.33%	p ≤ 1.3e-06	yes
10	19.73%	22.44%	p ≤ 1.0e-06	yes
11	20.19%	23.50%	p ≤ 2.2e-10	yes
12	20.85%	24.54%	p ≤ 4.4e-13	yes
13	21.13%	25.13%	p ≤ 1.4e-12	yes

Table 1. A comparison of recall scores for the SVD strategy and the Conditional strategy.

Sentence Length	Average Recall : Cond	Average Recall : Combined	Probability	Reject H_0
1	03.98%	04.05%	p ≤ 0.1305	yes
2	06.97%	08.60%	p ≤ 2.8e-13	yes
3	11.44%	12.34%	p ≤ 0.0007	yes
4	13.90%	15.44%	p ≤ 8.5e-09	yes
5	15.73%	17.33%	p ≤ 1.9e-09	yes
6	17.84%	18.72%	p ≤ 0.0003	yes
7	19.14%	20.34%	p ≤ 1.3e-05	yes
8	20.30%	21.48%	p ≤ 2.9e-06	yes
9	21.33%	22.60%	p ≤ 4.0e-06	yes
10	22.44%	23.82%	p ≤ 1.2e-06	yes
11	23.50%	24.56%	p ≤ 0.0003	yes
12	24.54%	25.44%	p ≤ 0.0008	yes
13	25.13%	26.37%	p ≤ 8.6e-06	yes

Table 2. A comparison of recall scores for the Conditional strategy and the Combined strategy.

6.3.2 Agreement between Strategies

The agreement between strategies is presented in Table 3. Interestingly, of the words recalled, the majority have only been selected by one content selection strategy. That is, the set of words recalled by one content selection strategy do not necessarily subsume the set recalled by the other. This supports the results obtained in the recall comparison in which a combined strategy leads to higher recall. Interestingly, the last column in the table shows that the potential combined recall is greater than the recall achieved by the combined strategy; we will return to this point in Section 6.4.

Sentence Length	Selected by neither method	Selected by only 1 method	Selected by both methods	Total Recall
1	91.6%	8.0%	0.3%	8.3%
2	84.7%	14.1%	1.0%	15.1%
3	79.9%	17.5%	2.5%	20.0%
4	76.6%	19.3%	3.9%	23.2%
5	73.8%	21.0%	5.1%	26.1%
6	71.4%	22.1%	6.4%	28.5%
7	69.6%	22.4%	7.8%	30.2%
8	67.9%	22.9%	9.1%	32.0%
9	66.4%	23.2%	12.3%	35.5%
10	65.0%	23.5%	11.3%	34.8%
11	63.9%	23.6%	12.3%	35.9%
12	63.0%	23.6%	13.2%	36.8%
13	62.1%	23.5%	14.3%	37.8%

Table 3. Agreement of words chosen between the SVD strategy and the Conditional probability strategy to content selection

6.4 Discussion

The SVD strategy ultimately did not perform as well ass we might have hoped. There are a number of possible reasons for this.

1. Whilst using the Combined probability did lead to a significantly improved result, this increase in recall was only small. Indeed, the analysis of the agreement between the Conditional strategy and the SVD strategy indicates that the current method of combining the two probabilities is not optimal and that there is still considerable margin for improvement.

2. Even though the recall of the SVD strategy was poorer by a only a few percent, the lack of improvement in recall is perplexing, given that we expected the thematic information to ensure words were used in correct contexts. There are several possible explanations, each warranting further investigation. It may be the case that the themes identified by the SVD analysis were quite narrow, each encompassing only small number of sentences. If this is the case, certain words occurring in sentences outside the theme would be given a lower probability even if they were good headline word candidates. Further investigation is necessary to determine if this is a short-coming of our SVD strategy or an artefact of the domain. For example, it might be the case that the sentences of news articles are already thematically quite dissimilar.

3. One might also question our experimental design. Perhaps the kind of improvement brought about when using the SVD probability cannot be measured by simply counting recall. Instead, it may be the case that an evaluation involving a panel of judges is required to determine if the generated text is qualitatively better in terms of how faithful the summary is to the information in the source document. For example, a summary that is more accurate may not necessarily result in better recall. Finally, it is conceivable that the SVD strategy might be more sensitive to preprocessing stages such as sentence delimitation and stopword lists, which are not necessary when using the Conditional strategy.

Despite these outstanding questions, there are pragmatic benefits when using SVD. The conditional strategy requires a paired training set of summaries and source documents. In our case, this was easily obtained by using headlines in lieu of single sentence summaries. However, in cases where a paired corpus is not available for training, the SVD strategy might be more appropriate, given that the performance does not differ considerably. In such a situation, a collection of documents is only necessary for collecting bigram statistics.

7 Conclusion

Combining both the SVD probability and Conditional probability marginally improves recall, lending support to the intuition that thematic information may help generate better single sentence summaries. However, there are still many unanswered questions. In future work, we intend to investigate these techniques in a domain other than news text so that we can draw conclusions as to how well these strategies generalise to other genres. We also intend to conduct user evaluations to gauge the quality of the generated summaries for both the Conditional and the SVD strategies. Indeed, a user-based evaluation would be extremely helpful in determining if the thematic information provided by the SVD strategy does help improve the veracity of the generated summaries.

References

Banko M., Mittal V., and Witbrock M. (2000) Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

Boguraev B., and Neff M. (2000) Discourse segmentation in aid of document summarization. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS- 33)*, *Minitrack on Digital Documents Understanding*. Maui, Hawaii: IEEE.

Borko, H., and Bernier, C. (1975) *Abstracting Concepts and Methods*. New York: Academic Press.

Brandow, R., Mitze, K., and Rau, L. (1995) Automatic condensation of electronic publications by sentence selection. In *Information Processing and Management*, 31(5), pages 675-685.

Forney G. D. (1973) The Viterbi Algorithm. In the *Proceedings of the IEEE,* pages 268-278.

Gong Y., and Liu, X. (2001) Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In the *Proceedings SIGIR 2001*: pages 19-25.

Hovy, E. and Lin, C. (1997) Automated text summarization in SUMMARIST. In the *Proceedings of ACL-EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 18-24.

Jin, R., and Hauptmann, A. (2001) Learning to Select Good Title Words: An New Approach based on Reversed Information Retrieval. In the *Proceedings of the Eighteen International Conference on Machine Learning (ICML 2001),* Williams College,MA, June 28-July 1.

Jin, R., Zhai, C., and Hauptmann, A. (2002) Title language model for information retrieval. In the *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002),* Tampere, Finland, August 11-15.

Jing, H., and McKeown, K. (1999) The decomposition of human-written summary sentences. In the *Proceedings of the 22nd Conference on Research and Development in Information Retrieval (SIGIR--99)*.

Knight, K. and Marcu, D. (2000) Statistics-based summarization---Step one: Sentence compression. In *Proceedings of AAAI-2000*.

Kupiec, J., Pedersen, J., and Chen, F. (1995) A Trainable Document Summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in*

Information Retrieval. Fox, E., Ingwersen, P., and Fidel, R. (Editors), pages 68—73.

Manning C. and Schütze H. (2000) *Foundations of Statistical Natural Language Processing*. MIT Press: Cambridge MA.

Marcu, D. (2000) *The Theory and Practice of Discourse Parsing and Summarization.* Cambridge: The MIT Press.

McKeown, K., Klavans, J., Hatzivassiloglou, V., Barzilay, R., and Eskin, E. (1999) Towards multidocument summarization by reformulation: Progress and prospects. In the *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI--99)*.

Paice, C. (1990) Constructing Literature Abstracts by Computers: Techniques and Prospects. In *Information Processing and Management*, Vol. 26, No. 1, pages 171–186.

Radev, D. and McKeown, K. (1998) Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469-500, September.

Siegel, Sidney and Castellan, Jr. N. John. (1988) *Nonparametric Statistics For The Behavioral Sciences*. McGraw-Hill, Inc., second edition.

Witbrock, M., and Mittal, V. (1999) Ultrasummarization: A statistical approach to generating highly condensed non-extractive summaries. In the *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR '99).*

Zajic D., Door B., and Schwartz R. (2002) Automatic Headline Generation for Newspaper Stories. In the *Proceedings of the Document Understanding Conference (DUC 2002)*.