

# Text Summarization Using Neural Networks

KHOSROW KAIKHAH, Ph.D.  
Department of Computer Science  
Texas State University  
San Marcos, Texas 78666  
[kk02@TxState.edu](mailto:kk02@TxState.edu)

**Abstract:** — A new technique for summarizing news articles using a neural network is presented. A neural network is trained to learn the relevant characteristics of sentences that should be included in the summary of the article. The neural network is then modified to generalize and combine the relevant characteristics apparent in summary sentences. Finally, the modified neural network is used as a filter to summarize news articles.

**Keywords:** adaptive clustering, feature fusion, neural networks, pruning, text summarization

## 1 Introduction

With the explosion of the World Wide Web and the abundance of text material available on the Internet, text summarization has become an important and timely tool for assisting and interpreting text information. The Internet provides more information than is usually needed. Therefore, a twofold problem is encountered: searching for relevant documents through an overwhelming number of documents available, and absorbing a large quantity of relevant information. Summarization is a useful tool for selecting relevant texts, and for extracting the key points of each text. Some articles such as academic papers have accompanying abstracts, which make them easier to decipher their key points. However, news articles have no such accompanying summaries, and their titles are often not sufficient to convey their important points. Therefore, a summarization tool for news articles would be extremely useful, since for a given news topic or event, there are a large number of available articles from the various news agencies and newspapers. Because news articles have a highly structured document form, important ideas can be obtained from the text simply by selecting sentences based on their attributes and locations in the article.

We propose a machine learning approach that uses artificial neural networks to produce summaries of arbitrary length news articles. A neural network is trained on a corpus of articles. The neural network is then modified, through feature fusion, to produce a summary of highly ranked sentences of the article.

Through feature fusion, the network discovers the importance (and unimportance) of various features used to determine the summary-worthiness of each sentence.

## 2 Features

Each document is converted into a list of sentences. Each sentence is represented as a vector  $[f_1, f_2, \dots, f_7]$ , composed of 7 features.

**Table 1: Seven Features of a Document**

$f_1$	Paragraph follows title
$f_2$	Paragraph location in document
$f_3$	Sentence location in paragraph
$f_4$	First sentence in paragraph
$f_5$	Sentence length
$f_6$	Number of thematic words in the sentence
$f_7$	Number of title words in the sentence

Features  $f_1$  to  $f_4$  represent the location of the sentence within the document, or within its paragraph. It is expected that in structured documents such as news articles, these features would contribute to selecting summary sentences. Brandow et al. in [1] have shown that summaries consisting of leading sentences outperform most other methods in this domain, and Baxendale in [2] demonstrated that sentences located at the beginning and end of paragraphs are likely to be good summary sentences.

Feature  $f_5$ , sentence length, is useful for filtering out short sentences such as datelines and author names commonly found in news articles. We also anticipate that short sentences are unlikely to be included in summaries [3]. Feature  $f_6$ , the number of thematic words, indicates the number of thematic words in the sentence, relative to the maximum possible. It is obtained as follows: from each document, we remove all prepositions, and reduce the remaining words to their morphological roots [4]. The resultant *content words* in the document are counted for occurrence. The top 10 most frequent content words are considered as thematic words. This feature determines the ratio of thematic words to content words in a sentence. This feature is expected to be important because terms that occur frequently in a document are probably related to its topic [5]. Therefore, we expect a high occurrence of thematic words in salient sentences. Finally, feature  $f_7$  indicates the number of title words in the sentence, relative to the maximum possible. It is obtained by counting the number of matches between the content words in a sentence, and the words in the title. This value is then normalized by the maximum number of matches. This feature is expected to be important because the salience of a sentence may be affected by the number of words in the sentence also appearing in the title. These features may be changed or new features may be added. The selection of features plays an important role in determining the type of sentences that will be selected as part of the summary and, therefore, would influence the performance of the neural network.

### 3 Text Summarization Process

There are three phases in our process: neural network training, feature fusion, and sentence selection. The first step involves training a neural network to recognize the type of sentences that should be included in the summary. The second step, feature fusion, prunes the neural network and collapses the hidden layer unit activations into discrete values with frequencies. This step generalizes the important features that must exist in the summary sentences by fusing the features and finding trends in the summary sentences. The third step, sentence selection, uses the modified neural network to filter the text and to select only the highly ranked sentences. This step controls the selection of the summary sentences in terms of

their importance. These three steps are explained in detail in the next three sections.

### 3.1 Neural Network Training

The first phase of the process involves training the neural networks to learn the types of sentences that should be included in the summary. This is accomplished by training the network with sentences in several test paragraphs where each sentence is identified as to whether it should be included in the summary or not. This is done by a human reader. The neural network learns the patterns inherent in sentences that should be included in the summary and those that should not be included. We use a three-layered feedforward neural network, which has been proven to be a universal function approximator [6]. It can discover the patterns and approximate the inherent function of any data to an accuracy of 100% , as long as there are no contradictions in the data set. Our neural network consists of seven input-layer neurons, six hidden-layer neurons, and one output-layer neuron. We use a conjugate gradient method where the energy function is a combination of error function and a penalty function. The goal of training is to search for the global minima of the energy function. The addition of the penalty function drives the associated weights of unnecessary connections to very small values while strengthening the rest of the connections. Therefore, the unnecessary connections and neurons can be pruned without affecting the performance of the network.

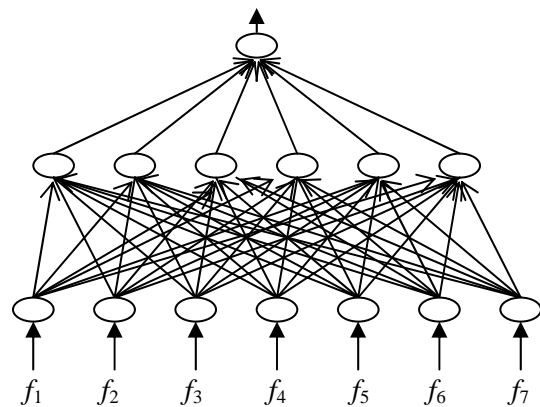


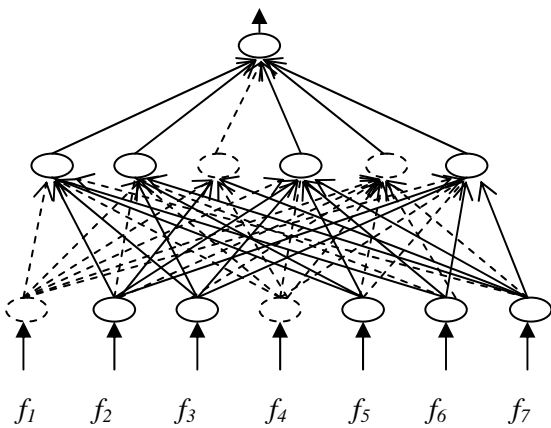
Figure 1: The Neural Network after Training

### 3.2 Feature Fusion

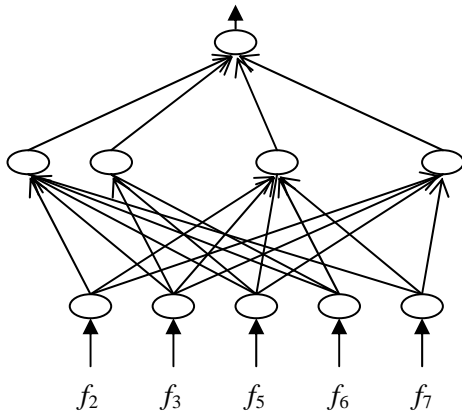
Once the network has learned the features that must exist in summary sentences, we need to discover the trends and relationships among the features that are

inherent in the majority of sentences. This is accomplished by the feature fusion phase, which consists of two steps: 1) eliminating uncommon features; and 2) collapsing the effects of common features.

Since the neural network is trained using a penalty function in phase I, the connections having very small weights after training can be pruned without affecting the performance of the network. As a result, any input or hidden layer neuron having no emanating connections can be safely removed from the network. In addition, any hidden layer neuron having no abutting connections can be removed. This corresponds to eliminating uncommon features from the network.



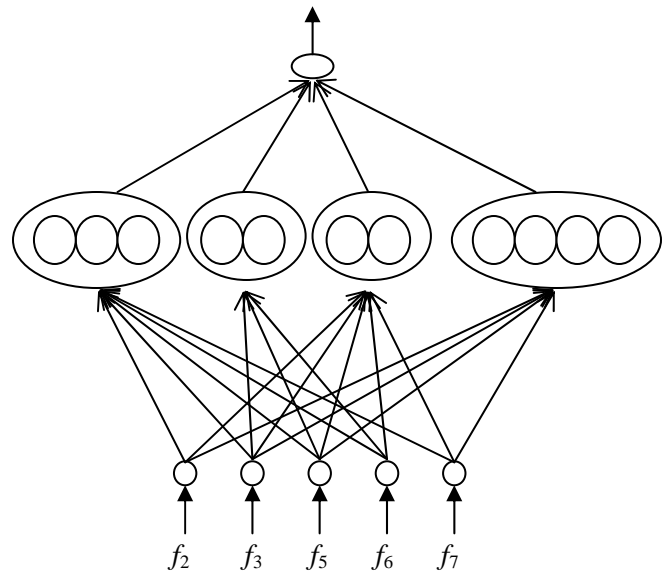
**Figure 2(a): The Neural Network after Pruning**



**Figure 2(b): The Neural Network after Pruning**

The hidden layer activation values for each hidden layer neuron are clustered utilizing an adaptive clustering technique. Each cluster is identified by its centroid and frequency. The activation value of each hidden layer neuron is replaced by the centroid of the

cluster, which the activation value belongs to. This corresponds to collapsing the effects of common features.



**Figure 3: The Neural Network after Feature Fusion**

The combination of these two steps corresponds to generalizing the effects of features, as a whole, and providing control parameters for sentence ranking.

### 3.3 Sentence Selection

Once the network has been trained, pruned, and generalized, it can be used as a tool to filter sentences in any paragraph and determine whether each sentence should be included in the summary or not. This phase is accomplished by providing control parameters for the radius and frequency of hidden layer activation clusters to select highly ranked sentences. The sentence ranking is directly proportional to cluster frequency and inversely proportional to cluster radius. Only sentences that satisfy the required cluster boundary and frequency are selected as high-ranking summary sentences.

## 4 Results and Analysis

We used 50 news articles from the Internet with various topics such as technology, sports, and world news to train the network. Each article consists of 16 to 89 sentences with an average of 29 sentences. The entire set consists of 1,435 sentences. Every sentence is labeled as either a summary sentence or an

unimportant sentence by a human reader. A total of 389 sentences were labeled as summary sentence with an average of 6 sentences per article. After the feature fusion phase,  $f_1$  (Paragraph follows title) and  $f_4$  (First sentence in the paragraph) were removed as well as two hidden layer neurons. The removal of  $f_1$  feature is understandable, since most of the articles did not have sub-titles or section headings. Therefore, the only paragraph following a title would be the first paragraph, and this information is already contained in feature  $f_2$  (paragraph location in document). The removal of  $f_4$  feature indicates that the first sentence in the paragraph is not always selected to be included in the summary. We then used the same 50 news articles as a test set for the modified network. The accuracy of the modified network ranged from 94% to 100% with an average accuracy of 96.4% when compared to the summaries of the human reader. That is, the network was able to select all sentences that were labeled as summary sentence in most of the articles. However, the network missed one to two sentences in six of the articles and selected one sentence that was not labeled as summary in each of five articles. The performance of the text summarization process depends heavily on the style of the human reader and to what the human reader deems to be important to be included in the summary.

A human reader and the modified network summarized 10 different news articles, independently. The performance of the modified network was 96% accurate when compared with the human reader's summaries. For each article, 4 to 9 sentences were selected by the network as summary sentences with an average of 5 sentences. Both the human reader and the modified network summarized 7 of the articles exactly the same. For two of the articles, the modified network missed one sentence in each article; and for one of the articles, the modified network included a sentence that was not selected by the human reader.

## 5 Conclusions

Our text summarization method performed well on the test paragraphs with accuracy of 96%. The selection of features as well as the selection of summary sentences by the human reader from the training paragraphs plays an important role in the performance of the network. The network is trained

according to the style of the human reader and to which sentences the human reader deems to be important in a paragraph. This, in fact, is an advantage our approach provides. Individual readers can train the neural network according to their own style. In addition, the selected features can be modified to reflect the reader's needs and requirement.

## References

- [1] R. Brandow, K. Mitze and L. Rau, "Automatic condensation of electronic publications by sentence selection", *Information Processing and Management*, vol. 31(5), pp. 675-685, 1995.
- [2] P.B. Baxendale, "Machine-Made Index for Technical Literature: An Experiment" *IBM Journal of Research and Development*, vol. 2(4), pp. 354-361, 1958.
- [3] J. Kupiec, J. Pederson and F. Chen, "A Trainable Document Summarizer", *Proceedings of the 18<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, Washington, pp. 68-73, 1995.
- [4] M. Porter, "An algorithm for suffix stripping", *Program*, vol. 14(3), pp. 130-137, 1980.
- [5] H.P. Luhn, "The Automatic Creation of Literature Abstracts", *IBM Journal for Research and Development*, vol. 2(2), pp. 159-165, 1958.
- [6] M.R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems", *Journal of Research of the National Bureau of Standards*, vol. 49, pp. 409-436, 1952.
- [7] I. Mani, *Automatic Summarization*, John Benjamins Publishing Company, pp. 129-165, 2001.
- [8] W.T. Chuang and J. Yang, "Extracting sentence segments for text summarization: a machine learning approach", *Proceedings of the 23<sup>rd</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, pp. 152-159, 2000.
- [9] C-Y. Liu and E. Hovy, "Identifying Topics by Position", *Proceedings of the 5<sup>th</sup> Conference on Applied Natural Language Processing (ANLP-97)*, Seattle, Washington, pp. 283-290, 1997.
- [10] K. Sparck Jones, "A Statistical Interpretation of Team Specificity and its Application in Retrieval", *Journal of Documentation*, vol. 28(1), pp. 11-21, 1972.