A Non-Linear Topic Detection Method for Text Summarization Using Wordnet

Carlos N. Silla Jr.[1*], **Celso A. A. Kaestner**[1] , **Alex A. Freitas**[2]

[1]Pontifícia Universidade Católica do Paraná
Rua Imaculada Conceição 1155 – 80215-901 Curitiba - PR

[2]University of Kent
Canterbury CT2 7NZ, UK

{silla,kaestner}@ppgia.pucpr.br, A.A.Freitas@ukc.ac.uk

Abstract. *This paper deals with the problem of automatic topic detection in text documents. The proposed method follows a non-linear approach. The method uses a simple clustering algorithm to group the semantically-related sentences. The distance between two sentences is calculated based on the distance between all nouns that appear in the sentences. The distance between two nouns is calculated using the Wordnet thesaurus. An automatic text summarization system using a topic strength method was used to compare the results achieved by the Text Tiling Algorithm and the proposed method. The obtained initial results shows that the proposed method is a promising approach.*

Resumo. *Este trabalho trata do problema de detecção automática de tópicos em documentos. O método proposto utiliza uma abordagem nova, não-linear. Um algoritmo simples de agrupamento é utilizado para agrupar as sentenças relacionadas semanticamente. A distância entre duas sentenças é calculada com base na distância entre todos os substantivos que aparecem nas sentenças. A distância entre os substantivos é calculada utilizando o thesaurus Wordnet. Para avaliar a performance desta proposta foi implementado um sumarizador automático de textos que utiliza um método baseado na força de cada tópico e no algoritmo Text Tiling. Os resultados iniciais obtidos com o método proposto são promissores.*

1. Introduction

Automatic text summarization is one important task of the Text Mining field: given a text, one wishes to obtain a summary that can satisfy the specific needs of the user [Luhn, 1958]. The main objective is to reduce the reading time of the original text but maintaining the main ideas of the text. The produced summary should allow the reader to answer questions about the subjects in the given text or work as a reference pointer to parts of the original text.

---

[*]Bolsista PIBIC - CNPq

This paper describes a new topic detection method that follows a non-linear approach. One of the summary systems on the literature uses the Text Tiling Algorithm [Hearst, 1997], which follows a linear approach to detect topics in a given text: the topics are detected in the same order in which they appear in the original text. However, when dealing with multi-document summarization [Stein et al., 2000] new methods for relating topics are needed, because we cannot follow a single linear order.

This work presents an alternative to the Text Tiling Algorithm. A non-linear method for topic detection is proposed. It uses a simple clustering algorithm to group semantically-related sentences using the knowledge attained from Wordnet. To evaluate the practical results of this method, a topic strength summarizer for single documents was implemented: it will be referred from now on as Non-Linear Topical TF-ISF. The results achieved by this approach have been compared with the ones shown in [Larocca Neto, 2002]. This work uses the Text Tiling Algorithm to detect the topics in a given text and a summarizer based on topic strength. It will be referred to here as Topical TF-ISF. For evaluation we employ a collection of 30 documents extracted from the Ziff-Davis TIPSTER base [Mani et al., 1998].

This article is organized as follows: section 2 presents a brief explanation of the Topical TF-ISF method; in section 3 the proposed method is explained; section 4 presents the tests and the computational results; and finally in section 5 we present the conclusions and future research.

2. Linear Topic Detection Method

The original Text Tiling algorithm was presented by [Hearst, 1993]. It is used for partitioning full-length documents into coherent multi-paragraph units. The layout of text tiles is meant to reflect the pattern of subtopics contained in an expository text. The approach uses lexical analysis based on TF-IDF (Term Frequency - Inverse Document Frequency) a commonly used metric in Information Retrieval [Salton et al., 1996].

The algorithm is a two step process; first, all pairs of adjacent blocks of text (usually 3-5 sentences long) are compared and assigned a similarity value; then the resulting sequence of similarity values, after being graphed and smoothed, is examined for peaks and valleys. High similarity values, implies that the adjacent block cohere well, tend to form peaks and low similarity values, indicate a potential boundary between tiles, creating a valley.

An extractive text summarization algorithm based on topic strength was presented by [Larocca Neto et al., 2000a]. The basic ideas of the proposed algorithm are as follows. Initially the document is partitioned into topics using the Text Tiling algorithm. Then for each topic the algorithm computes a metric of its relative importance in the document. This measure is computed by using the notion of TF-ISF (Term-Frequency - Inverse Sentence Frequency) [Larocca Neto et al., 2000b] which is an adaptation of the TF-IDF measure. After that the algorithm with determine how many sentences must be selected from each topic using a topic strength formula. The sentences selected from each topic are the ones closer to the centroid of the corresponding topic.

3. The Non-Linear Topic Detection Method

3.1. Pre-Processing

The pre-processing consists of two steps: first the document is tagged using Brill's Part of Speech Tagger [Brill, 1992]. After that the nouns of each sentence are extracted from the document, creating a new representation of the document that contains only nouns. If for some reason there are any sentences that doesn't have any nouns, they will be discarded during this phase. For example, consider the sentence: "A WELL-STOCKED MACHINE". After the tagging it will look like: "A/DT WELL-STOCKED/VBD MACHINE/NNP". Then it will be represented only by the nouns of the sentence, resulting in: "MACHINE". The motivation for representing a sentence only by its nouns is that nouns typically have a richer semantics than other parts of speech.

3.2. Creating the Distance Matrix

Now that the document is represented only by nouns, the sentences will be grouped by their semantic similarity, based on a distance matrix M where each cell M_{xy} contains the distance between sentence x and sentence y. (This kind of distance matrix is computed in several clustering algorithms [Manning and Schutze, 2001]).

The semantic distance between two words using Wordnet [Miller et al., 1990] can be calculated in several ways [Budanitsky, 2001]. However in this work, since the document is represented only by nouns, the distance between two nouns is obtained by the hypernym relation. One of the problems using this approach is that the hypernym relation in Wordnet is not well distributed: for example in the botanical domain the taxonomy is more fine-grained than in other domains. For that reason the normalized distance shown in (1) was used.

$$\text{Normalized Distance} = \frac{\text{Dist.}(W_i, \text{DCA})}{\text{Dist.}(W_i, \text{Root})} + \frac{\text{Dist.}(W_j, \text{DCA})}{\text{Dist.}(W_j, \text{Root})} \quad (1)$$

Where:

- W_i and W_j are the i-th noun and the j-th noun of the first and second sentences whose distance is being computed, respectively.
- DCA is the deepest common ancestral between W_i and W_j.
- Root is the common unique beginner between the two nouns.

For example: Let W_i be *cat* and W_j be *dog*. Their deepest common ancestral (DCA) is *Carnivore* and their common unique beginner is *Entity, Something*. This formula can only be used if the two nouns have the same Unique Beginner; to solve this problem we established that in the other cases the distance will be set to the maximum distance plus 0.1. The procedure used to calculate the distance between two sentences is presented in Figure 1.

The procedure calculates the distance between sentence x (S_x) and sentence y (S_y). However the relationship between the two sentences will not be always symmetric: for example, if sentence x is represented by (cat, dog) and sentence y is represented by (car, cow). In this example the distance between sentence x and y will be different

For each $W_i \in S_x$ do
    For each $W_j \in S_y$ do
        Normalized Distance(W_i, W_j)
    End For
    /* Dist (W_i, S_x, S_y) denotes the distance between sentences
       S_x and S_y with respect to the word W_i */
    Dist (W_i, S_x, S_y) = Min(Dist(W_i, W_j))
End For
Dist(S_x, S_y) = $\frac{\Sigma_{i=1}^{n} W_i}{n}$
Where:

- The normalized distance is given by (1).
- n is the number of words in sentence S_x.
- Min(Dist(W_i, W_j)) is the smallest value between the word W_i and all words of sentence y.

Figure 1: Procedure used to calculate the distance between two sentences.

from the distance between sentence y and sentence x. To overcome this problem the two sentences are permuted and the procedure is used again.

The procedure will produce two distance values: the final value stored in the distance matrix will be the arithmetic mean between Dist.(S_x, S_y) and Dist.(S_y, S_x). This procedure will be repeated until the matrix distance is completely known.

3.3. Clustering the Sentences by Semantic Similarity

Using the distance matrix, a simple and fast clustering algorithm will be used to group sentences by semantic similarity. (We did not use a classical clustering algorithm, such as k-means, because they usually assume that the coordinates of each cluster centroid [Duda et al., 2001] can be computed as the average of the coordinates of all the examples belonging to the cluster, which is not the case in our application involving sentences words. The simple clustering algorithm described here is customized for this example representation.

To start the algorithm the number of clusters [Manning and Schutze, 2001] will be the equivalent to 10% or 20% of the total number of sentences in the given document, this value will depend on the compression rate desired for the summary. Let K be the number of clusters. Then the K closest pairs of sentences will be selected from the distance matrix to represent the K initial clusters. Each initial cluster will then consist of the union of the sets of words representing each of the two sentences allocated to the cluster.

After the initial clusters are set, the procedure presented in Figure 2 will be applied to cluster the sentences. The update cluster function will concatenate the sentences representing the cluster and the newly added sentence, i.e., the set of words representing the sentence added to the cluster will be added to the set of words representing the cluster. In this procedure we don't use the sentence appearance order in the text; for that reason we call our approach as "Non-Linear" in contrast with the linear approach followed in [Larocca Neto et al., 2000a].

```
Repeat
 Calculate the distance between all sentences and clusters.
 Select the pair (sentence,cluster) with the smallest
   distance value.
 Add the selected sentence to the cluster.
 Update the cluster.
Until all sentences have been clustered.
```

Figure 2: Procedure used to cluster the sentences.

Table 1: Results for Manually Made Summaries with 10% Compression

Method	Precision / Recall
Random Sentences	0.097222 ± 0.017737
Topical TF-ISF	0.195278 ± 0.029753
Non-Linear Topical TF-ISF	0.192023 ± 0.130647

4. Computational Results

To evaluate the performance of the Non-Linear Topical TF-ISF against the Topical TF-ISF we implement several tests. We used a data set composed of 30 documents from the ZIF-Davis TIPSTER base [Mani et al., 1998], with a set of "ideal" summaries created by a linguist expert [Larocca Neto et al., 2002]. The generated summaries have a compression rate of 10% and 20%. We employ the classical precision / recall metrics from Information Retrieval [Baeza-Yates and Ribeiro-Neto, 1999] as evaluation metric. In our case of text summarization, since the size of the ideal summaries and the generated ones are the same, precision is equal to recall.

Table 1 shows the computational results of the proposed method against the "ideal" summaries with compression rate of 10%. It also compares this method with the results achieved by the Topical TF-ISF [Larocca Neto, 2002] and the random sentences method, which is used as a base line. This results shows that the precision / recall for summaries with compression rate of 10% generated by the Non-Linear Topical TF-ISF are close to the ones obtained by the Topical TF-ISF method and are significantly better than the base-line. Figure 3 shows an example of one of the produced summaries using a compression rate of 10%.

Table 2 shows the computational results of the proposed method against the "ideal" summaries with compression rate of 20%. The results obtained are once again close to the ones achieved by the Topical TF-ISF, and are much better than the random sentences approach. The absolute values seems to be low; however these results are in conformance with the experiments realized by [Mitra et al., 1997] where even human judges have a low agreement on which sentences must belong to the summary.

Although the Non-Linear Topical TF-ISF achieved slightly worse results than the the Topical TF-ISF, the advantage of using a non-linear approach is that it can be used in many other document applications, like multi-document summarization and clipping. This makes our proposal an interesting approach.

```
The first area is basic development tools: a language for object
programming (for example, c++ and object pascal), robust class
libraries of foundation classes, environment interfaces,
relatively common domain-specific problem solving (compound
document processing), and application frameworks.[10]

As a normative condition, a database abstraction at the core of
the environment should be able to support projects ranging from
very small ones to corporate-wide libraries.[18]

One important concept is extending the hypertext paradigm to
encode semantic information in the database, analogous to the
way attribute grammars encode semantic content in a language
specification.[25]

Each object in the database is an instance of some class, whose
code is available to process requests made on it.[27]

The arm covers c++ 2.1, along with the two major experimental
areas--templates and exception handling.[49]
```

Figure 3: An example of one of the Produced Summaries

Table 2: Results for Manually Made Summaries with 20% Compression

Method	Precision / Recall
Random Sentences	0.194918 ± 0.018539
Topical TF-ISF	0.336820 ± 0.020626
Non-Linear Topical TF-ISF	0.297404 ± 0.122434

5. Conclusions and Future Research

This work presents a new non-linear topic detection method that can be used in many text mining applications. The proposed method has been evaluated in the field of single document text summarization. We use a topic strength method for selecting and identifying the most important topics and determining how many sentences to select from each topic.

Although the results achieved by the Non-Linear Topical TF-ISF are slightly worse than the ones achieved by the Topical TF-ISF, in our experiment of single document summarization, the advantage of using the proposed method is that it can be used in other applications like clipping, multi-document summarization and others.

The results obtained in this work also indicate that a better method for selecting sentences from topics is also needed. There are many issues to deal when performing multi-document summarization but this approach seems to be a step in the right direction. The proposed method could also be used for other languages if there is a Wordnet version available for that language.

In future research we intend to use the method as part of an information retrieval system to automatically retrieve web documents and perform multi-document summarization and clipping.

References

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley.

Brill, E. (1992). A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento, IT.

Budanitsky, A. (2001). Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley-Interscience.

Hearst, M. A. (1993). Texttiling: A quantitative approach to discourse segmentantion. Technical Report 93/24, University of California, Berkeley.

Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Larocca Neto, J. (2002). Contribution to the study of automatic text summarization techniques (in portuguese). Master's thesis, Pontifical Catholic University of Paraná.

Larocca Neto, J., Freitas, A. A., and Kaestner, C. A. A. (2002). Automatic text summarization using a machine learning approach. In *XVI Brazilian Symposium on Artificial Intelligence*, pages 205–215, Porto de Galinhas, PE, Brazil.

Larocca Neto, J., Santos, A. D., Kaestner, C. A. A., and Freitas, A. (2000a). Generating text summaries through the relative importance of topics. In *Proc. Int. Joint Conf.: IBERAMIA-2000 (7th Ibero-American Conf. on Artif. Intel.) & SBIA-2000 (15th Brazilian Symp. on Artif. Intel.)*, pages 301–309, Sao Paulo, SP, Brazil.

Larocca Neto, J., Santos, A. D., Kaestner, C. A. A., and Freitas, A. A. (2000b). Document clustering and text summarization. In *Proc. 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining (PADD-2000)*, pages 41–55, London: The Practical Application Company.

Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(92):159–165.

Mani, I., House, D., Klein, G., Hirschman, L., Obrsl, L., Firmin, T., Chrzanowski, M., and Sundheim, B. (1998). The tipster summac text summarization evaluation. MITRE Technical Report MTR 98W0000138, The MITRE Corporation.

Manning, C. D. and Schutze, H. (2001). *Foundations of Statistical Natural Language Processing*. The MIT Press.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Five papers on wordnet. Technical Report Cognitive Science Laboratory Report 43, Princenton University.

Mitra, M., Singhal, A., and Buckley, C. (1997). Automatic text sumarization by paragraph extraction. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pages 31–36, Madrid, Spain.

Salton, G., Allan, J., and Singhal, A. (1996). Automatic text decomposition and structuring. *Information Processing and Management*, 32(2):127–138.

Stein, G. C., Bagga, A., and Wise, G. B. (2000). Multi-document summarization: Methodologies and evaluations. In *Proceedings of the 7th Conference on Automatic Natural Language Processing (TALN'00)*, pages 337–346, Lausanne, Switzerland.