GAËL DIAS
ELSA ALVES
Computer Science Department
Beira Interior University
Covilhã, Portugal
ddg@di.ubi.pt, elsalves@zmail.pt

Language-independent informative topic segmentation

Abstract

In this paper, we present an innovative topic segmentation system based on a new informative similarity measure that takes into account word co-occurrence in order to avoid the accessibility to existing linguistic resources such as electronic dictionaries or lexico-semantic databases such as thesauri or ontology. Topic Segmentation is the task of breaking documents into topically coherent multi-paragraph subparts. Topic Segmentation has extensively been used in Information Retrieval and Text Summarization. In particular, our architecture proposes a language-independent Topic Segmentation system that solves three main problems evidenced by previous research: systems based uniquely on lexical repetition that show reliability problems, systems based on lexical cohesion using existing linguistic resources that are usually available only for dominating languages and as a consequence do not apply to less favored languages and finally systems that need previously existing harvesting training data.

1. Introduction

This paper introduces a new technique for improving access to information dividing lengthy documents into topically coherent sections. This research area is commonly called Topic Segmentation and can be defined as the task of breaking documents into topically coherent multi-paragraph subparts. In this paper, we present an innovative topic segmentation system based on a new informative similarity measure that takes into account word co-occurrence in order to avoid the accessibility to existing linguistic resources such as electronic dictionaries or lexico-semantic databases such as thesauri or ontology. In particular, our architecture solves three main problems evidenced by previous research: systems based uniquely on lexical repetition that show reliability problems (Hearst, 1994; Reynar, 1994; Richmond *et al.*, 1997; Yaari, 1997; Sardinha, 2002), systems based on lexical cohesion using existing linguistic resources that are usually available only for dominating languages like English, French or German, and as a consequence do not apply to less favored languages (Morris and Hirst, 1991; Kozima, 1993) and systems that need previously existing harvesting training data (Beeferman *et al.*, 1997).

In order to overcome these drawbacks, we propose a Topic Segmentation system based on a new informative similarity measure that takes into account word co-occurrences automatically acquired from corpora. Our system can be defined as a three step process:

(1) It evaluates the weight of each word in terms of the segmentation task. For that purpose, it uses a combination of three main heuristics: the well-known *tf.idf* measure proposed by (Sparck-Jones, 1972; Salton, 1975), the adaptation of the *tf.idf* measure for sentences, the *tf.isf*, and a new density measure that calculates the density of each word in the text i.e. if the occurrences of the same word are close to each other in the text or not.

(2) For each sentence in the text, it then calculates its similarity with the previous block of *k* sentences and the next block of *k* sentences based on the informative similarity measure that includes the Equivalence Index Association Measure (Muller *et al.*, 1997).

(3) The topic boundaries are then calculated based on the same algorithm proposed by (Hearst, 1994).

3. Weighting Score

Our algorithm is based on the vector space model which determines the similarity of neighboring groups of sentences and places subtopic boundaries between dissimilar blocks. In our specific case, each sentence in the corpus is evaluated in terms of similarity with the previous block of *k* sentences and the next block of *k* sentences (as illustrated in figure 1).

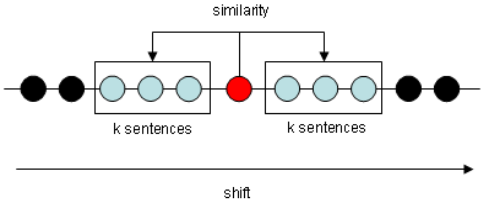

Figure 1: Vector space architecture

We propose a new weighting scheme based on three heuristics: the *tf.idf* measure, the adaptation of the *tf.idf* measure for sentences, the *tf.isf*, and a new density measure that calculates the density of each word in the text.

3.1. The tf.idf Score

The basic idea of the *tf.idf* score (Salton, 1975) is to evaluate the importance of a word within a document based on its frequency and its distribution across a collection of documents. The *tf.idf* score is defined in equation 1.

$$tf.idf(w,d) = \frac{tf(w;d)}{|d|} \times \log_2 \frac{N}{df(w)}$$

Equation 1: *tf.idf* score

However, not all relevant words in a document are useful for Topic Segmentation. For instance, relevant words appearing in all sentences will be of no help for segmenting the text into topics.

3.2. The tf.isf Score

The basic idea of the *tf.isf* score is to evaluate each word in terms of its distribution over the document. Indeed, it is obvious that words occurring in many sentences within a document may not be useful for Topic Segmentation purposes. So, we will define the *tf.isf* to evaluate the importance of a word within a document based on its frequency within a given sentence and its distribution across all the sentences within the document. For that purpose, we will use the *tf.isf* score as a second measure of word relevance (see equation 2).

$$tf.isf(w,s) = \frac{stf(w;s)}{|s|} \times \log_2 \frac{Ns}{sf(w)}$$

Equation 2: *tf.isf* score

However, we can push even further our idea of word distribution. For that purpose, we propose a new density measure that calculates the density of each word in a document.

3.3. The Word Density Score

The basic idea of the word density measure is to evaluate the dispersion of a word within a document. So, very disperse words will not be as relevant as dense words. In order to evaluate the word density, we propose a new measure based on the distance (in terms of words) of all consecutive occurrences of the word in the document. We call this measure *dens* and it is defined in equation 3.

$$dens(w,d) = \sum_{k=1}^{|w|-1} \frac{1}{\ln(dist(occur(k), occur(k+1)) + e)}$$

Equation 3: *dens* score

For any given word *w*, its density *dens(w,d)* in document d, is calculated from all the distances between all its occurrences, |*w*|. So, *occur(k)* and *occur(k+1)* respectively represent the positions in the text of two consecutive occurrences of the word *w* and *dist(occur(k), occur(k+1))* calculates the distance that separates them in terms of words within the document.

3.4. The Weighting Score

The weighting score of any word in a document can be directly derived from the previous three heuristics. A straightforward definition of the weighting score is given in equation 4 where each score is normalized.

$$weight(w,d) = \|tf.idf(w,d)\| + \|tf.isf(w,s)\| + \|dens(w,d)\|$$

Equation 4: *weight* score

4. Similarity Measure

Our methodology is based on the same idea as (Ponte and Croft, 1997) but differs from it as the word co-occurrence information is directly embedded in the calculation of the similarity between blocks of sentences thus proposing a well-founded mathematical model that deals with the word co-occurrence factor. For that purpose, we propose a new informative similarity measure that includes in its definition the Equivalence Index Association Measure proposed by (Muller *et al.*, 1997) defined in Equation 6.

$$EI(w_1, w_2) = p(w_1 \mid w_2) \times p(w_2 \mid w_1) = \frac{f(w_1, w_2)^2}{f(w_1) \times f(w_2)}$$

Equation 6: *Equivalence Index Association Measure*

The Equivalence Index between words w_1 and w_2 is calculated within a context window of any size in order to determine $f(w_1, w_2)$ and from a collection of documents so that we can evaluate the degree of cohesiveness between two words outside the context of the document. So, the basic idea of our informative similarity measure is to integrate into the cosine measure the word co-occurrence factor inferred from a collection of documents with the Equivalence Index association measure as defined in equation 7

$$S_{ij} = infosimba(X_i, X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{jl} \times EI(W_{ik}, W_{jl})}{\sqrt{\sum_{k=1}^{p}\sum_{l=k+1}^{p}(X_{ik} \times EI(W_{ik}, W_{il}))^2} \times \sqrt{\sum_{k=1}^{p}\sum_{l=k+1}^{p}(X_{jk} \times EI(W_{jk}, W_{jl}))^2}}$$

Equation 7: *Informative Similarity Measure*

where $EI(W_{ik}, W_{jl})$ is the Equivalence Index value between W_{ik}, the word that indexes the vector of the document *i* at position *k*, and W_{jl}, the word that indexes the vector of the document j at position *l*. The next step of the application aims at placing subtopic boundaries between dissimilar blocks. For that purpose, we propose a detection methodology based on the standard deviation algorithm proposed by (Hearst, 1994).

5. Topic Boundary Detection

Taking as reference the idea of (Ponte and Croft, 1997) who take into account the preceding and the following contexts of a segment, we calculate the informative similarity of each sentence in the corpus with its surrounding pieces of texts i.e. its previous block of *k* sentences and its next block of *k* sentences as illustrated in figure 1. The basic idea is to know whether the focus sentence is more similar to the preceding block of sentences or to the following block of sentences. In order to evaluate this preference in an elegant way, we propose a score for each sentence in the text in the same manner (Beeferman *et al.*, 1997) compare short and long-range models. Our preference score (*ps*) is defined in equation 8.

$$ps(X_i) = \log \frac{infosimba(X_i, X_{i-1})}{infosimba(X_i, X_{i+1})}$$

Equation 8: *preference score*

In order to illustrate graphically the variations of the *ps* score, we show in figure 2 an experiment made with five texts taken from the web with five different topics using block sentences of size 3 (*k=3*) and a window of the text size for the calculation of the Equivalence Index.

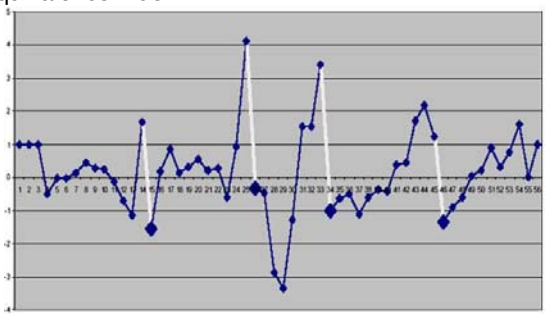

Figure 2: *preference score variation*

In order to better understand the variation of the *ps* score, each time its value goes from positive to negative between two consecutive sentences, there exits a topic shift. We will call this phenomenon a downhill. However, not all downhills identify the presence of a new topic in the text. Indeed, only deep ones must be taken into account. They are represented in white in Figure 2 and represent the correct changes in topic. In order to automatically identify these downhills, and as a consequence the topic shifts, we adapt the algorithm proposed by (Hearst, 1994) to our specific case. So, we propose a threshold that is a function of the average and standard deviation of the downhills depths. For lack of space, we do not present this function.

6. Conclusion, Discussion and Future Work

In this paper, we propose a language-independent unsupervised Topic Segmentation system based on word-co-occurrence that avoids the accessibility to existing linguistic resources such as electronic dictionaries or lexico-semantic databases such as thesauri or ontology. In particular, our architecture proposes a system that solves three main problems evidenced by previous research: systems based uniquely on lexical repetition that show reliability problems, systems based on lexical cohesion using existing linguistic resources that are usually available only for dominating languages and as a consequence do not apply to less favored languages and finally systems that need previously existing harvesting training data. To our point of view, our main contribute to the field is the definition of a new similarity measure, the informative similarity measure, *infosimba*, that proposes a

well-founded mathematical model that deals with the word co-occurrence factor and avoids an extra step in the boundary detection compared to the solution introduced by (Ponte and Croft, 1997).

References

1. Beeferman, D., Berger, A., and Lafferty, J. 1997. *Text segmentation using exponential models*. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing, 35--46. http://citeseer.ist.psu.edu/beeferman97text.html

2. Hearst, M. 1994. *Multi-Paragraph Segmentation of Expository Text*, In Proceedings of the 32nd Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, June, 9--16. http://citeseer.ist.psu.edu/151333.html

3. Kozima, H. 1993. Text *Segmentation Based on Similarity between Words*. In Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics (Student Session), Colombus, Ohio, USA, 286--288. http://citeseer.ist.psu.edu/kozima93text.html

4. Morris, J. and Hirst, G. 1991. *Lexical cohesion computed by thesaural relations as an indicator of the structure of text*, Computational Linguistics 17(1): 21--43. http://acl.ldc.upenn.edu/J/J91/J91-1002.pdf

5. Muller, C., Polanco, X., Royauté, J. and Toussaint, Y. 1997. *Acquisition et structuration des connaissances en corpus: éléments méthodologiques*. Technical Report RR-3198, Inria, Institut National de Recherche en Informatique et en Automatique. http://www.inria.fr/rrrt/rr-3198.html

6. Ponte J.M. and Croft W.B. 1997. *Text Segmentation by Topic*. In Proceedings of the First European Conference on Research and Advanced Technology for Digitial Libraries.120--129. http://citeseer.ist.psu.edu/ponte97text.html

7. Reynar, J.C. 1994. *An Automatic Method of Finding Topic Boundaries*. In Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics (Student Session), Las Cruces, New Mexico, USA. http://citeseer.ist.psu.edu/reynar94automatic.html

8. Richmond, K., Smith, A., and Amitay, E. 1997. *Detecting subject boundaries within text: A language independent statistical approach*. In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP--97), Providence, Rhode Island, August 1-2. 4--54. http://citeseer.ist.psu.edu/154167.html

9. Salton, G., Yang, C.S., and Yu, C.T. 1975. *A theory of term importance in automatic text analysis*. Amer. Soc. Inf. Sc~ 26, 1, 33--44.

10. Sardinha, T.B. 2002. *Segmenting corpora of texts*. *DELTA*, 2002, 18(2), 273--286. ISSN 0102-4450. http://www.scielo.br/pdf/delta/v18n2/v18n2a04.pdf

11. Sparck-Jones, K. (1972). *A statistical interpretation of term specificity and its application in retrieval*. Journal of Documentation, 28(1), 11--21.

12. Yaari, Y. 1997. *Segmentation of expository text by hierarchical agglomerative clustering*. In Proceedings of the Conference on Recent Advances in Natural Language Processing, 59--65. http://arxiv.org/abs/cmp-lg/9709015