Topic Segmentation: How Much Can We Do By Counting Words And Sequences of Words

Gaël Dias, Elsa Alves and Célia Nunes
Centre of Human Language Technology and Bioinformatics
University of Beira Interior
ddg@di.ubi.pt, elsalves@zmail.pt, celia@mat.ubi.pt

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. For that purpose, we only use statistics on words and sequences of words based on a set of texts. This solution provides a flexible solution that may narrow the gap between dominating languages and less favored languages thus allowing equivalent access to information.

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 order to provide solutions to access useful information from the ever-growing number of documents on the web, such technologies are crucial as people who search for information are now submerged with unmanageable quantities of text data and most of the time can not find what they are looking for as they can only deal with conveniently-sized packages of information. For that purpose, topic segmentation has extensively been used in information

retrieval and text summarization. In the context of information retrieval, it is clear that some user should prefer a document in which the occurrences of a word or a sequence of words (phrase) are concentrated into one or two paragraphs since such a concentration is more likely to contain a definition of the queried concept and as a consequence the system is more likely to retrieve useful information. This particular research domain is usually called passage retrieval and proposes techniques to extract fragments of texts relevant to a query [Salton et al. 1993] [Kaszkiel and Zobel 1997] [Cormack et al. 1999]. In the context of text summarization, topic segmentation is usually used as the basic text structure in order to apply sentence extraction and sentence compression techniques [Boguraev and Neff 2000] [Angheluta et al. 2002] [Farzindar and Lapalme 2004].

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/phrase co-occurrences automatically acquired from corpora. Our system can be defined as a three step process:

1. It evaluates the weight of each word/phrase 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 et al. 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/phrase in the text i.e. if the occurrences of the same word/phrase 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 [Müller et al. 1997].

3. The topic boundaries are then calculated based on a variation of the algorithm proposed by [Hearst 1994].

Our three-step architecture follows the same ideas as [Phillips 1985] and [Ponte and Croft 1997] using a different methodology. In fact, our common

approach is that topic segmentation should be done "on the fly" on any input text thus avoiding the problems of domain/genre/language-dependent systems that need to be tuned each time one of these parameters changes (domain, genre or language). In fact, we all aim at proposing a language-independent unsupervised architecture.

Although many experiments and systems have been studied in the field of topic segmentation, very few research works have tried to introduce some degree of semantics, intrinsic to language. In this field, the most convincing piece of work has been carried out by [Ferret 2002] who introduces the identification of collocations/phrases into the process of discovering boundaries in texts. By definition phrases are sequences of words which sense is non-compositional i.e. the sense of the overall sequence can not be deduced from the senses of the individual words. As a consequence, phrases embody meaningful sequences of words that are less ambiguous than single words and allow approximating more accurately the contents of texts. In particular, most of the neologisms in technical and scientific domains are realized by phrases. For example, *World Wide Web*, *IP address* and *TCP/IP network* are terminologically relevant phrases that are particularly new in the domain of Computer Science. As a consequence, there has been a growing interest in developing techniques for automatic phrase extraction. In order to extract phrases from text corpora, three main strategies have been proposed in the literature. First, purely linguistic systems [David and Plante 1990] [Bourigault 1996] propose to extract relevant phrases by using techniques that analyze specific syntactical structures in the texts. However, this methodology suffers from its monolingual basis, as the systems require highly specialized linguistic techniques to identify clues that isolate possible candidate phrases. Second, hybrid methodologies [Enguehard 1993] [Justeson 1993] [Daille 1995] [Paio et al. 2003] define co-occurrences of interest in terms of syntactical patterns and statistical regularities. However, by reducing the search space to groups of words that correspond to *a priori* defined syntactical patterns (e.g. Noun+Adjective, Noun+Preposition+Noun), such systems do not deal with a great proportion of phrases. Finally, purely statistical systems [Church and Hanks 1990] [Dunning 1993] [Smadja 1993] [Shimohata 1997] [Tomokiyo and Hurst 2003] extract discriminating phrases from text corpora by means of association measure regularities. As they use plain text corpora and only require the information appearing in texts, such systems are highly flexible and extract relevant units independently from the domain and the language of the input text. However, they emphasize two major drawbacks. On one hand, by relying on *ad hoc* establishment of global thresholds they are prone to error. On the other hand, as they only allow the acquisition of binary associations, these systems must apply bootstrapping techniques to acquire phrases with more than two words. Unfortunately, such techniques have shown their limitations as their retrieval results mainly depend on the identification of suitable two-word phrases for the beginning of the iterative process. In order to overcome the problems previously highlighted by the statistical systems, we conjugate an association measure called the Mutual

Expectation [Dias 2002] with an acquisition process called the GenLocalMaxs [Dias 2002]. On one hand, the Mutual Expectation, based on the concept of Normalized Expectation, evaluates the degree of cohesiveness that links together all the words contained in a sequence of any length. On the other hand, the GenLocalMaxs retrieves the candidate phrases from the set of all the valued phrases by evidencing local maxima of association measure values. This architecture is called SENTA (Software for the Extraction of N-ary Textual Associations) and can be divided into three steps:

1. It segments the input text into positional n-grams i.e. ordered vectors of words.

2. It evaluates the degree of cohesiveness within any positional n-gram by applying the Mutual Expectation association measure.

3. It elects candidate phrases based on the GenLocalMaxs that search for association measure local maxima.

The combination of this association measure with this acquisition process proposes an integrated solution to the problems of bootstrapping techniques and global thresholds defined by experimentation.

Our main purpose in this paper is to do as much as possible in terms of topic segmentation without the introduction of extra data within texts. Thus, we only use statistics on words and sequences of words. It is important to notice that such systems will provide a solution that may narrow the gap between dominating languages and less favored languages in terms of access to information in a globalized world. By so, the economical growth of less favored countries will also be boosted. The paper is divided into four sections. First, we will show the main differences between our work on topic segmentation and the existing ones, in particular the systems proposed by [Phillips 1985] [Ponte and Croft 1997]. We will also compare our phrase extraction process with existing work. Second we will introduce the basic notions of our architecture for phrase extraction. Third, we will show the weighting process of each word/phrase of the input text corpus. Fourth, we will introduce our main innovation i.e. the informative similarity measure. Fifth, we will define how subparts can be elected from the values of the informative similarity measure. Finally, we will evaluate our language-independent system on real data extracted from the web and show that good results can be achieved by just counting words and sequences of words in the study of languages.

2 Related Work

In order to better understand our methodology based on counting words and sequences of words to treat natural language, we must first point at advantages and drawbacks of concurrent approaches. We will first give an overview of the

topic segmentation state-of-the-art and then define the main three approaches applied to phrase extraction.

2.1 Topic Segmentation

[Hearst 1994], [Reynar 1994] and [Sardinha 2002] have proposed different architectures based on lexical item repetition: respectively, TextTiling, Dotplotting and the Link Set Median Procedure. However, it has been proved that systems based on lexical repetition are not reliable when applied to non-technical texts without small controlled vocabularies. For instance, articles in newspapers tend to avoid word repetition. In fact, good writing should avoid word repetition. As a consequence, these techniques can only be applied to technical texts where synonyms rarely exist for a given concept so that word repetition is almost compulsory. In order to avoid these limitations, [Kozima 1993] has proposed an architecture based on a Semantic Network built from the English Dictionary (LDOCE) from which lexical cohesion can be fine-grained induced. First, [Morris and Hirst 1991] had proposed a discourse segmentation algorithm based on lexical cohesion relations called lexical chains using Roget's thesaurus. However, these linguistic resources are not available for the majority of languages so that their application is drastically limited and as a consequence do not apply to less favored languages that may transform themselves into endangered languages. In order to avoid the use of huge linguistic resources, [Beeferman et al. 1997] have proposed a technique for identifying document boundaries using statistical techniques. They built statistical models within a framework which incorporated a number of cues about the story boundaries such as the appearance of particular words before a boundary and the appearance of cue words in the beginning of the previous sentence of a boundary. Unfortunately, this work is limited by the need of previously existing harvesting training data as it proposes a supervised solution to the problem of topic segmentation. Once more, it lacks in flexibility as new training is necessary when the genre/domain/language change. It is clear that unsupervised language-independent techniques that automatically induce some degree of semantics propose a promising solution to solve all the exposed problems. [Phillips 1985] and [Ponte and Croft 1997] have proposed such techniques. [Phillips 1985] proposes to identify a lexical network based on word collocation frequency statistics and cluster analysis. However, he does not propose a classical topic segmentation technique but rather a topic detection system as he does not output boundaries in the text. [Ponte and Croft 1997] proposed a topic segmentation technique based on the Local Content Analysis [Xu and Croft 1996] allowing to substituting each sentence with words and phrases related to it. A pairwise similarity measure is then calculated between all transformed sentences and then introduced into a final score (depending on the length and position of the segment) in order to find at each point in the corpus the best block that maximizes the score function. The important point to focus on is the use of the Local Content Analysis that introduces some degree of semantics to the system without requiring harvesting linguistic resources and thus avoiding the problem of word

repetition. In order to introduce endogenously acquired semantic knowledge, [Ferret 2002] has also proposed to automatically extract collocations/phrases from texts in order to compute semantic similarity measures. Although our approach tends to stand to the basic ideas of these unsupervised methodologies, we differ from them as we clearly pose the problem of word weighting for the specific task of topic segmentation. Indeed, most of the presented systems only rely on frequency and/or the $tf.idf$ measure proposed by [Sparck-Jones 1972] [Salton et al. 1975] of their lexical items. However, we deeply think that better weighting measures can be proposed. For that purpose, we introduce a new weighting score based on three heuristics: the well-known $tf.idf$ 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/phrase in the text. Moreover, in order to introduce a certain degree of semantic knowledge in our system, we propose a new informative similarity measure that includes in its definition the Equivalence Index association measure proposed by [Müller et al. 1997] so that word co-occurrence information is directly embedded in the calculation of the similarity between blocks of sentences. Thus, unlike [Ponte and Croft 1997], we propose a well-founded mathematical model that deals with the word co-occurrence factor. Finally, like classical methodologies, our system then calculates the similarity of each sentence in the corpus with the previous block of k sentences and the next block of k sentences and then elects the best text boundaries based on a variation of the standard deviation algorithm proposed by [Hearst 1994].

2.2 Phrase Extraction

The acquisition of phrases has long been a significant problem in Natural Language Processing, being relegated to the borders of lexicographic treatment. Most of the work in knowledge acquisition has aimed at extracting explicit information from texts (i.e. knowledge about the world) and has generally neglected the extraction of implicit information (i.e. knowledge about the language). For the past ten years, there has been a renewal in phraseology mostly stimulated by full access to large-scale text corpora in machine-readable format. Compound nouns (*Prime minister*), compound names (*Republic of Yugoslavia*), compound determinants (*a number of*), verbal locutions (*to give rise*), adverbial locutions (*as soon as possible*), prepositional locutions (*such as*) and conjunctive locutions (*on the other hand*) share the properties of phrases. In order to test the assumptions made about word flexibility constraints inherent to phrases, a great deal of statistical measures have been proposed in the literature. However, most of them only evaluate the degree of cohesiveness that exists within groups of two words (2-grams) and do not deal with the general case of groups of n words (i.e. n-grams with $n \geq 2$). As a consequence, these mathematical models only allow the acquisition of binary associations and bootstrapping techniques[1] have to be applied to acquire asso-

[1] As a first step, relevant 2-grams are retrieved from the input corpus. Then, n-ary associations may be identified by either (1) gathering overlapping 2-grams or (2) by marking the

ciations with more than two words [Salem 1987] [Smadja 1993] [Smadja 1996] [Shimohata 1997] [Daille 1995]. Unfortunately, such techniques have shown their limitations as their retrieval results mainly depend on the identification of suitable 2-grams for the initiation of the iterative process. In order to overcome the lack of generalization for the case of n individual words, we propose an association measure based on the Normalized Expectation [Dias 2002], the Mutual Expectation [Dias 2002]. As a consequence, its combination with the GenLocalMaxs algortithm provides a solution for the acquisition of n-ary word associations that avoids the definition of global thresholds and does not require bootstrapping techniques.

As we said earlier, our main objective is to exclusively use statistics on words and sequences of words to treat the problem of topic segmentation. This issue is reached by applying, in a first step, our phrase extraction system, called SENTA, to introduce some degree of semantics into texts. In a second step, we apply a new informative similarity measure over vectors of weighted words/phrases which includes lexical cohesion factors in its definition and as a consequence does not need any linguistic extra information but plain text. In the next section, we first introduce our phrase extraction system.

3 Extraction of Phrases Using SENTA

In this section, we present the pre-processing of the input text from which phrases are extracted and then identified. For that purpose, we use the SENTA software [Dias 2002] that is based on three main steps: (1) segmentation of the text into positional n-grams, (2) evaluation of the degree of cohesiveness with the Mutual Expectation association measure and (3) election of the candidate phrases using the GenLocalMaxs algorithm.

3.1 Positional N-grams

A great deal of applied works in lexicography evidence that most lexical relations associate words separated by at most five other words [Sinclair 1974]. And a phrase is a specific lexical relation and so can be defined in terms of structure as a specific n-gram calculated in an immediate span of five words to the left hand side and five words to the right hand side of a focus word. By definition, an n-gram is a vector of n words where each word is indexed by the signed distance that separates it from its associated focus word (i.e. the first word of the vector). Consequently, an n-gram can be contiguous or non-contiguous whether the words involved in the n-gram represent or not a continuous sequence of words in the text. By convention, the focus word is always the first element of the vector and its signed distance is equivalent to zero. We represent an n-gram by the vector $[p_{11}w_1p_{12}w_2...p_{1i}w_i...p_{1n}w_n]$ where p_{1i} (for $i = 2$ to n) denotes the

extracted 2-grams as single words in the text and re-running the system to search for new 2-grams (the process ends when no more 2-grams are identified).

signed distance that separates the word, w_i, from the focus word, w_1 and p_{ii} (for all i), is always equal to zero. As computation is concerned, each word is successively a focus word and all its associated contiguous and non-contiguous n-grams are calculated avoiding duplicates. Finally, each n-gram is associated to its frequency in order to apply the association measure that will evaluate its degree of cohesiveness, the Mutual Expectation.

3.2 Mutual Expectation

By definition, phrases are groups of words that occur together more often than expected by chance. From this assumption, we define an association measure, the Mutual Expectation (*ME*), based on the concept of Normalized Expectation (*NE*).

3.2.1 Normalized Expectation

We define the *NE* of an n-gram as the average expectation of occurring one word in a given position knowing the occurrence of the other $n-1$ words also constrained by their positions. The underlying concept of the *NE* is based on the conditional probability defined in Equation 1.

$$p(X = x|Y = y) \quad = \quad \frac{p(X = x, Y = y)}{p(Y = y)} \tag{1}$$

However, in order to capture in one measure the n conditional probabilities associated to the n events obtained by extracting one word at a time from the n-gram, we introduce the concept of the fair point of expectation (*FPE*). We know that only the n denominators of the n conditional probabilities vary while the n numerators remain unchanged from one probability to another. So, in order to perform the normalization process, we evaluate the gravity center of the denominators thus defining an average event, the *FPE*. Basically, the *FPE* is the arithmetic mean of the n joint probabilities of the sub-$(n-1)$-grams contained in an n-gram and is defined for each n-gram as in Equation 2 where W denotes the n-gram $[p_{11}w_1p_{12}w_2...p_{1i}w_i...p_{1n}w_n]$.

$$FPE(W) = \frac{1}{n}\left(\sum_{i1=1}^{2}\sum_{i2=i1+1}^{3}...\sum_{\substack{i(n-1)=\\ i(n-2)+1}}^{n} p\left(\left[\begin{array}{c} p_{i1i1}w_{i1}p_{i1i2}w_{i2}... \\ p_{i1i(n-1)}w_{i(n-1)} \end{array}\right]\right)\right) \tag{2}$$

Hence, the normalization of the conditional probability, is realized by the introduction of the *FPE* into the general definition of the conditional probability. The resulting measure is called the *NE* and it is proposed as a "fair"

conditional probability as defined in Equation 3 where W denotes the n-gram $[p_{11}w_1p_{12}w_2...p_{1i}w_i...p_{1n}w_n]$.

$$NE(W) = \frac{p(W)}{FPE(W)} \tag{3}$$

3.2.2 Mutual Expectation

Béatrice Daille in [Daille 1995] shows that one effective criterion for phrase identification is frequency. From this assumption, we deduce that between two n-grams with the same *NE*, the most frequent n-gram is more likely to be a phrase. So, the Mutual Expectation of an n-gram is defined in Equation 4, where W denotes the n-gram $[p_{11}w_1p_{12}w_2...p_{1i}w_i...p_{1n}w_n]$, based on its *NE* and its probability of occurrence.

$$ME(W) = p(W) \times NE(W) \tag{4}$$

3.2.3 GenLocalMaxs

Most of the approaches proposed in the literature base their selection process on global association measure thresholds [Church and Hanks 1990] [Daille 1995] [Smadja 1993] [Shimohata 1997]. This is defined by the underlying concept that there exits a limit value of the association measure that allows to decide whether an n-gram is a phrase or not. However, these thresholds are prone to error as they depend on experimentation. Moreover, they highlight evident flexibility constraints as they have to be re-tuned when the type, the size, the domain and the language of the document change. The GenLocalMaxs algorithm [Dias 2002] proposes a more robust, flexible and fine-tuned approach for the election of phrases as it focuses on the identification of local maxima of the association measure values. Let *assoc* be an association measure, W an n-gram, Ω_{n-1} the set of all the *(n-1)*-grams contained in W, Ω_{n+1} the set of all the *(n+1)*-grams containing W and *sizeof(.)* a function that returns the number of words of an n-gram, the GenLocalMaxs is defined as follows:

$\forall x \in \Omega_{n-1}, \forall y \in \Omega_{n+1} \qquad W$ is a phrase if
$(sizeof(W) = 2 \wedge assoc(W) > assoc(y)) \vee$
$(sizeof(W) \neq 2 \wedge assoc(W) \geq assoc(x) \wedge assoc(W) > assoc(y))$

Table 1: GenLocalMaxs Algorithm

The GenLocalMaxs algorithm proposes a theoretically sound acquisition process that does not depend on experimentation and avoids the definition of global thresholds. As a consequence, it overcomes the problems of portability of the existing approaches. Indeed, no tuning is needed in order to run the system and any association measure can be tested.

The first step of our overall architecture is to extract and identify from texts its relevant phrases. This is done by the application of our SENTA software. Once, all candidate phrases have been identified and extracted, they are marked as single words in the text corpus. For instance, the compound noun *Prime Minister* would be marked as the following string *Prime_Minister* so that it can be treated as a single word. By doing so, we identify some degree of semantics carried by texts and narrow the problems of word sense disambiguation. Our second step aims at discovering the topic boundaries within texts based on the normalized corpus, i.e. with the marked phrases.

4 Unsupervised Topic Segmentation System

In this section, we present our topic segmentation system based on an informative similarity measure that takes into account word[2] co-occurrence in order to avoid the accessibility to existing linguistic resources. Although our approach tends to stand to the basic ideas of known unsupervised methodologies such as [Phillips 1985] [Ponte and Croft 1997], we differ from them as we clearly pose the problem of word weighting for the specific task of topic segmentation. Indeed, most of the presented systems only rely on frequency and/or the $tf.idf$ measure proposed by [Sparck-Jones 1972] [Salton et al. 1975] of their lexical items. However, we deeply think that better weighting measures can be proposed. For that purpose, we introduce, in the next section, a new weighting score based on three heuristics: the well-known $tf.idf$ 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/phrase in the text.

4.1 Weighting Score

Our algorithm is based on the vector space model [Salton et al. 1975] which treats documents as vectors of words. The simplest form of the vector space model treats a document (in our case, a sentence or a group of sentences) as a vector which values correspond to the number of occurrences of the words appearing in the document as in [Hearst 1994]. Although [Hearst 1994] showed successful results with this weighting scheme, we strongly believe that the importance of a word in a document does not only depend on its frequency. Indeed, frequency can only be reliable for technical texts where ambiguity is drastically limited and word repetition largely used. But unfortunately, these documents are an exception in the global environment of the internet for example. According to us, two main factors must be taken into account to define the relevance of a word for the specific task of topic segmentation: its semantic importance, based on its frequency but also on its inverse document frequency [Sparck-Jones 1972] [Salton et al. 1975] and its localization in the text. For that purpose, we propose a new weighting scheme based on three heuristics: the well-known $tf.idf$

[2]From now on, we will talk about words and phrases as words generically.

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.

4.1.1 The $tf.idf$ Score

The basic idea of the $tf.idf$ score [Salton et al. 1975] is to evaluate the importance of a word within a document based on its frequency (i.e. frequent words within a document may reflect its content more strongly than words that occur less frequently) and its distribution across a collection of documents (i.e. words that are limited to few documents are useful for discriminating those documents from the rest of the collection). For our specific task, it is important to separate relevant words from meaningless words (usually called stop-words) as the former ones will help us to define topic sentences i.e. sentences that are meaningful for the document. For that purpose, we will use the $tf.idf$ score as a first measure of word relevance. The $tf.idf$ score is defined in Equation 5 where w is a word and d a document.

$$tf.idf(w,d) = \frac{tf(w,d)}{|d|} \times \log_2 \frac{N}{df(w)} \tag{5}$$

For each w in document d, we compute its relative term frequency, i.e. the number of occurrences of w in d, $tf(w,d)$, divided by the number of words in d, $|d|$. We then compute the inverse document frequency of w [Sparck-Jones 1972] by taking the log_2 of the ratio of N, the number of documents in our experiment, to the document frequency of w, i.e. the number of documents in which the word w occurs ($df(w)$). As a result, a word occurring in all documents of the collection will have an inverse document frequency 0 giving it no chance to be a relevant word. On the opposite, a word which occurs very often in one document but in very few other documents of the collection will have a high inverse document frequency as well as a high term frequency and thus a high $tf.idf$ score. Consequently, it will be a strong candidate for being a relevant word within the document. 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. For that purpose, we extend the idea of the $tf.idf$ to sentences.

4.1.2 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. As our objective is to find different topically coherent multi-paragraph subparts, a word appearing evenly in the overall text will not contribute to capture the essence of a specific subpart. 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. The $tf.isf$ score is defined in Equation 6 where w is a word and s a sentence.

$$tf.isf(w,s) = \frac{stf(w,s)}{|s|} \times \log_2 \frac{Ns}{sf(w)} \tag{6}$$

For each w in s, we compute its relative sentence term frequency, that is the number of occurrences of w in s, $stf(w,s)$, divided by the number of words in s, $|s|$. We then compute the inverse sentence frequency of w by taking the log_2 of the ratio of Ns, the number of sentences within the document, to the sentence frequency of w, i.e. the number of sentences in which the word w occurs ($sf(w)$). As a result, a word occurring in all sentences of the document will have an inverse sentence frequency 0 giving it no chance to be a relevant word for topic segmentation. On the opposite, a word which occurs very often in one sentence but in very few other sentences will have a high inverse sentence frequency as well as a high sentence term frequency and thus a high $tf.isf$ score. Consequently, it will be a strong candidate for being a relevant word within the document for the specific task of topic segmentation. However, we can push even further our idea of word distribution. Indeed, a word w occurring 3 times in 3 different sentences may not have the same importance in all cases. Let's exemplify. If the 3 sentences are consecutive, the word w will have a strong influence on what is said in this specific region of the text. On the opposite, it will not be the case if the word w occurs in the first sentence, in the middle sentence and then in the last sentence. It is clear that we must take into account this phenomenon. For that purpose, we propose a new density measure that calculates the density of each word in a document.

4.1.3 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 is defined in Equation 7.

$$dens(w,d) = \sum_{k=1}^{|w|-1} \frac{1}{ln\left(dist\left(occur(k), occur(k+1)\right) + e\right)} \tag{7}$$

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. Thus, by summing their inverse distances, we get a density function that gives higher scores to highly dense words. As a result, a word, the occurrences of which

appear close to one another, will show small distances and as a result a high density. On the opposite, a word, the occurrences of which appear far from each other, will show high distances and as a result a small word density. In particular, if a word occurs only once in the document it must be seen as very dense and receives the value 1.

4.1.4 The Weighting Score

The weighting score of any word in a document can be directly derived from the previous three heuristics. As a matter of fact, by combining these three scores, we deal with the two main factors that must be taken into account to define the relevance of a word for the specific task of topic segmentation: its semantic importance and its localization in the document. A straightforward definition of the weighting score is given in Equation 8 where each score is normalized[3] so that they can be combined.

$$weight(w,d) = \|tf.idf(w,d)\| \times \|tf.isf(w,s)\| \times \|dens(w,d)\| \quad (8)$$

Thus, a relevant word for topic segmentation should evidence a high $tf.idf$ score, a high $tf.isf$ score and a high density score. The next step of the application of the vector space model aims at determining the similarity of neighboring groups of sentences. For that purpose, it is important to define an appropriate similarity measure. That is the objective of our next section.

4.2 Evaluation of Similarity Between Sentences

There are a number of ways to compute the similarity between two documents, in our case, between a sentence and a group of sentences. Theoretically, a similarity measure can be defined as follows. Suppose that $X_i = (X_{i1}, X_{i2}, X_{i3}, , X_{ip})$ is a row vector of observations on p variables associated with a label i. The similarity between two units i and j is defined as $S_{ij} = f(X_i, X_j)$ where f is some function of the observed values. In the context of our work, the application of a similarity measure is straightforward. Indeed, X_i may be regarded as the focus sentence and X_j as a specific block of k sentences, each one being represented as p-dimension vectors, where p is the number of different words within the document and where X_{ib} may represent the weighting score of the b^{th} word in the document also appearing in the focus sentence X_i. Our goal here is to find the appropriate f function that will accurately evaluate the similarity between the focus sentence and the blocks of k sentences. But, before introducing our new informative similarity measure, we will first point at the major drawback of current similarity measures when dealing with text data. We will take the cosine measure as an example although this drawback can be pointed at all known similarity measures.

[3]The normalization is calculated based on the minimum and maximum values of each measure.

4.2.1 The Drawback of Similarity Measures

The cosine similarity determines the angle between the vectors associated to two documents (in our case, the focus sentence and a group of k sentences). As a consequence, the vectors that represent similar documents have a smaller angle between them than those that represent dissimilar documents. The cosine measure is defined in Equation 9.

$$cosine(X_i, X_j) = \frac{\sum_{k=1}^{p} X_{ik}.X_{jk}}{\sqrt{\sum_{k=1}^{p} X_{ik}^2}.\sqrt{\sum_{k=1}^{p} X_{jk}^2}} \quad (9)$$

When applying the cosine similarity between two documents, only the identical indexes of the row vectors X_i and X_j will be taken into account i.e. if both documents do not have words in common, they will not be similar at all and will receive a cosine value of 0. However, this is not tolerable. Indeed, it is clear that both sentences (1) and (2) are similar although they do not share any word in common:

```
(1) Ronaldo defeated the goalkeeper once more.

(2) Real Madrid striker scored again.
```

In fact, good writing rules claim to avoid word repetition. As a consequence, it is clearly improbable that two consecutive blocks of text may share a large number of words in common, unless we deal with very specific domain documents where synonyms rarely exist and vocabulary ambiguity is reduced to its minimum. In order to avoid this problem, previous works have tried to modify the input text so that lexical item repetition could be evidenced. One of the drawbacks of using word repetition to track topic shifts is that words often occur in different inflected forms within texts. For instance, there exist 64 possible inflected forms for any given verb in Portuguese. A solution to avoid this problem is to rely on lemma repetition instead of word repetition. By using a morphological analyzer, it is possible to reduce all inflected forms of the same concept in the text to its lemma and as a consequence, allow taking into account lemma repetition to evaluate similarity between sentences [Phillips 1985] [Hearst 1994] [Reynar 1994] [Yaari 1997]. Unfortunately, such systems are not available for all languages which limits their application. Instead of lemmatizing text and identifying lemma repetition, some works have proposed to rely on the repetition of n-grams of characters [Dai et al. 2003]. For example, without morphology normalization, the words *oceanic* and *oceanographic* have the character sequence *ocean* in common. However, there are also drawbacks to using character n-grams. Some common words are spelled using character sequences that frequently occur in longer words. For instance, the open class word *dent* is a substring of the

unrelated words identifier, *indentation* and *dentist*. Moreover, unrelated words may share features of inflectional and derivational morphology: the verb forms *takes* and *fries* share the same ending *es* but do not have a common root. The most interesting idea to avoid word repetition problems is certainly to identify lexical cohesion relationships between words. Indeed, systems should take into account semantic information that could, for instance, relate Ronaldo to Real Madrid striker. For that purpose, many authors have proposed to computationally identify these relationships (in particular, the synonym relation) using large linguistic resources such as the well-known lexico-semantic database Wordnet [Angheluta et al. 2002] [Moens and De Busser 2003], the Roget's thesaurus [Morris and Hirst 1991] or the English dictionary LDOCE [Kozima 1993]. However, these huge resources are only available for dominating languages and as a consequence do not apply to less favored languages.

4.2.2 The Informative Similarity Measure

A much more interesting research direction is proposed by [Ponte and Croft 1997] that propose a topic segmentation technique based on the Local Content Analysis [Xu and Croft 1996], allowing substituting each sentence with words and phrases related to it. The important point to focus on is the use of the Local Content Analysis that introduces some degree of semantics to the system without requiring harvesting linguistic resources and thus avoiding the problem of word repetition. Our methodology is based on this same idea but differs from it as the word co-occurrence information is directly embedded in the calculation of the similarity between blocks of sentences thus avoiding an extra-step in the topic boundaries discovery. Another direct contribution is that, unlike [Ponte and Croft 1997], we propose a similarity measure 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 [Müller et al. 1997]. Association measures such as the Point-wise Mutual Information [Church and Hanks 1990], the Log-Likelihood ratio [Dunning 1993], the Dice coefficient [Dice 1945], the Kullback-Leibler [Cover and Thomas 1991] or the Jensen-Shannon divergence [Rao 1982] have shown successful results for the discovery of relationships between words. In fact, an association measure can be defined as a measure to evaluate the degree of cohesiveness between words. So, the higher the association measure between words is, the more related the words should be. For the specific task of topic segmentation, we have first implemented the Equivalence Index association measure [Müller et al. 1997] that has shown successful results in our different research works [Silva et al. 1999] [Cleuziou et al. 2003], although any association measure could be used. It is defined in Equation 10.

$$EI(w_1, w_2) = p(w_1|w_2) \times p(w_2|w_1) = \frac{p(w_1, w_2)^2}{p(w_1) \times p(w_2)} \tag{10}$$

The Equivalence Index between words w_1 and w_2 is calculated within

a word-context window of any size in order to determine the probability of co-occurrence between w_1 and w_2 i.e., $(p(w_1, w_2))$ and from a collection of documents so that we can evaluate the degree of cohesiveness between two words outside the context of the document. This collection can be thought as the overall web, from which we are able to infer with maximum reliability the "true" co-occurrence between two words as it is done in [Cleuziou et al. 2003]. So, the basic idea of our informative similarity measure called $infosimba$ is to integrate into the cosine measure the word co-occurrence factor inferred from a collection of documents with the Equivalence Index association measure[4]. This can be done straightforwardly as defined in Equation 11 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.

$$infosimba(X_i, X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{jl}.EI(w_{ik}, w_{jl})}{\sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik}.X_{il}.EI(w_{ik}, w_{il})}.\sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{jk}.X_{jl}.EI(w_{jk}, w_{jl})}} \quad (11)$$

In fact, the informative similarity measure can simply be explained as follows. Let's take the focus sentence X_i and a block of sentences X_j. For each word in the focus sentence, then for each word in the block of sentences, we calculate the product of their weights and then multiply it by the degree of cohesiveness existing between those two words calculated by the Equivalence Index association measure. As a result, the more relevant the words will be and the more cohesive they will be, the more they will contribute for the cohesion within the text and will not contribute to a topic shift. However, any sound similarity measure should guarantee that its values are between 0 and 1. Unfortunately, this is not the case with the $infosimba$. Indeed, counter-examples are easy to find. This situation is due to the fact that the value of the Equivalence Index association measure can not be generalized and may produce unexpected results. In order to propose a mathematically sound similarity measure, we propose two different equations that guarantee that their values stick between 0 and 1. They are presented in Equations 12 and 13. The first similarity measure $S1(X_i, X_j)$ sticks as much as possible to the $infosimba$, being just summed up the numerator of the $infosimba$ to its denominator. Based on the assumption that all our weights are positive numbers, we guarantee that $S1(X_i, X_j)$ will return values between 0 and 1. It is defined in Equation 12 and for notation

[4]Theoretically, any association measure could be applied to the $infosimba$. Here, the Equivalence Index association measure is used as a specification of the $infosimba$ similarity measure.

purposes, we note $A = \sum_{k=1}^{p} \sum_{l=1}^{p} X_{ik}.X_{jl}.EI(w_{ik}, w_{jl})$

$S1(X_i, X_j) =$

$$\frac{A}{\sqrt{\sum_{k=1}^{p} \sum_{l=1}^{p} X_{ik}.X_{il}.EI(w_{ik}, w_{il})}.\sqrt{\sum_{k=1}^{p} \sum_{l=1}^{p} X_{jk}.X_{jl}.EI(w_{jk}, w_{jl}) + A}} \quad (12)$$

The other similarity measure $S2(X_i, X_j)$ uses sums instead of products to normalize its value depending on the length of each sentence. Indeed, our experience in the field tells us that a more drastic normalization can lead to better results in the field of natural language processing. $S2(X_i, X_j)$ is defined in Equation 13.

$S2(X_i, X_j) =$

$$\frac{A}{A + \sum_{k=1}^{p} \sum_{l=1}^{p} (X_{ik}.X_{il}.EI(w_{ik}, w_{il}) + X_{jk}.X_{jl}.EI(w_{jk}, w_{jl}))} \quad (13)$$

Unfortunately, due to lack of time, we will not present the experimental results of these two new similarity measures but instead just the results with the $imfosimba$. It is clear that we expect better results by the application of $S1(X_i, X_j)$ or/and $S2(X_i, X_j)$. The next step of our system 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]. This issue is discussed in the next section.

4.2.3 Topic Boundary Detection

Different methodologies have been proposed to place subtopic boundaries between dissimilar blocks depending on the models used to determine similarity between sentences. [Kozima 1993] and [Hearst 1994] propose a methodology that compares, for a given window size, each pair of adjacent blocks of text according to how similar they are lexically. This method assumes that the more similar two blocks of text are, the more likely it is that the current subtopic continues, and, conversely, if two blocks of text are dissimilar, this implies a change in the subtopic flow. However, in order to accurately determine topic boundaries, their algorithms slightly differ. While [Kozima 1993] defines a change in topic as a valley in the graphical representation of the similarity scores, [Hearst 1994] proposes a more fine-grained algorithm. Boundaries are scored according to the relative depths of the valleys in the plot which results from the similarity values against the sentence gap numbers. Thus, breaks in similarity adjacent to high strong peaks (indicating dense cohesion relations) are considered stronger boundaries than those near lesser peaks. In fact, the actual values of the similarity measures are not taken into account,

but the relative differences are. Thus, the valley depth must exceed a certain threshold to be considered a topic shift. By experimentation, the threshold is a function of the average and standard deviation of the valley depths for each text. Another, interesting methodology is proposed by [Stokes et al. 2002] that take into account sentences instead of blocks of text to determine topic shifts. In particular, [Stokes et al. 2002] propose to segment texts based on the analysis of lexical chains. Thus, they define a boundary strength $w(n, n+1)$ between two consecutive sentences in the text as the product of number of lexical chains whose span ends at sentence n and the number of chains that begin their span at sentence $n+1$. When all boundary strengths between adjacent sentences have been calculated, they then get the mean of all non-zero cohesive strength scores. Similarly to [Hearst 1994], this mean value then acts as the minimum allowable boundary strength that must be exceeded if the end of textual unit n is to be classified as the boundary point between two news stories. [Ponte and Croft 1997] also propose a slightly different methodology from [Kozima 1993] and [Hearst 1994]. They use three different features: the internal similarity which is simply the sum of pairwise similarities within the segment and the left and right external similarity. The left (resp. right) external similarity is the sum of the pairwise similarities of each sentence in the segment of a fixed number of preceding (resp. following) sentences. For each of their size, the segments are then ranked by the internal similarity minus the two external similarities. A Gaussian length model is then combined with a dynamic programming process to find the best topic boundaries. Finally, [Beeferman et al. 1997] propose to evidence shifts in topic by comparing a long-range language model to a short-range language model (the trigram language model). The basic idea is that one might be more inclined towards a boundary when the long-range model suddenly shows a dip in performance compared to the short-range model. Conversely, when the long-range model is consistently assigning higher probabilities to the observed words, a boundary is less likely.

It is difficult to judge any methodology as they differ depending on the research approach. For that purpose, we propose a new methodology based on ideas expressed by different research works. 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. 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 14 where sim is any similarity measure.

$$ps(S_i) = log_2 \frac{sim(S_i, X_{i-1})}{sim(S_i, X_{i+1})} \quad (14)$$

So, if $ps(S_i)$ is positive, it means that the focus sentence S_i is more similar to the previous block of sentences, X_{i-1}. Conversely, if $ps(S_i)$ is negative, it means that the focus sentence S_i is more similar to the following block of sentences, X_{i+1}. In particular, when $ps(S_i)$ is near 0, it means that the focus sentence S_i is similar to both blocks and so we may be in the continuity of a topic. 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. In fact, it means that the previous sentence is more similar to the preceding block of sentences and the following sentence is more similar to the following block of sentences thus representing a shift in topic in the text. However, not all downhills identify the presence of a new topic in the text. Indeed, only deep ones must be taken into account. 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 the standard deviation of the downhills depths. A downhill is simply defined in Equation 15 whenever the value of the *ps* score goes from positive to negative between two consecutive sentences S_i and S_{i+1}.

$$downhill(S_i, S_{i+1}) = ps(S_i) - p(S_{i+1}) \quad (15)$$

Once all downhills in the text have been calculated, the average and standard deviation are evaluated. The topic boundaries are then elected if they satisfy the constraint expressed in Equation 17 where c is a constant to be tuned, $\bar{x}$ the average of all downhills in the text and σ the standard deviation of all downhills.

$$downhill(S_i, S_{i+1}) \geq \bar{x} + c.\sigma \quad (16)$$

Now that we finished the illustration of the architecture of our system, we will show its results on a set of web news documents in the next section. We will see that we can obtain promising results for the discovery of topic boundaries by just counting words and sequences of words.

5 Experiments and Results

Topic segmentation systems [Ferret 2002] [Brants et al. 2002] [Xiang and Hongyuan 2003] have usually been evaluated on [Choi 2000]'s data set that represents the gold standard for evaluation. However, many authors have discussed the validity of this test corpus [Ferret 2002] [Brants et al. 2002] [Xiang and Hongyuan 2003] and [Moens and De Busser 2003] proposed their

own test corpus. Indeed, [Choi 2000]'s data set, also called c99, evidences two major drawbacks: (1) it deals with segments of different domains and (2) lexical repetition is high within each segment. We propose an illustration of the c99 corpus as follows (Directory 3-5, Text 7).

> **Segment 1:**
> The next question is whether board members favor their own social classes in their roles as educational policy-makers. On the whole, it appears that they do not favor their own social classes in an explicit way. Seldom is there an issue in which class lines can be clearly drawn. A hypothetical issue of this sort might deal with the establishment of a free public junior college in a community where there already was a good private college which served the middle-class youth adequately but was too expensive for working-class youth. In situations of this sort the board generally favors the expansion of free education.
>
> **Segment 2:**
> Vincent G. Ierulli has been appointed temporary assistant district attorney, it was announced Monday by Charles E. Raymond, District Attorney. Ierulli will replace Desmond D. Connall who has been called to active military service but is expected back on the job by March 31. Ierulli, 29, has been practicing in Portland since November, 1959.

However, it is clear that the c99 corpus does not apply for an evaluation oriented towards text summarization. Indeed, in this case, the texts must cover a single domain and intra-segment lexical repetitions are not used as much as in the c99 corpus. However, it is likely that there exist inter-segment lexical repetitions which unease the process of boundary detection. This situation is illustrated as follows where the inter-segments lexical repetitions are shown in bold and the intra-segments lexical repetitions are underlined.

> **Segment 1:**
> O avançado brasileiro, novo reforço do **Sporting**, revelou hoje que vai viajar rapidamente para Lisboa, com o objectivo de assinar pelos "**leões**", cumprir os habituais exames médicos e começar a trabalhar às ordens do técnico José Peseiro. "O meu empresário está aí em Lisboa e disse-me que estava tudo acertado. Neste momento eu já me considero como jogador do **Sporting**", realçou Mota, em declarações à Renascença. O ponta-de-lança "canarinho", que está de férias no Brasil, revela que vai precisar de algum tempo para alcançar o mesmo nível físico dos restantes companheiros: "Vou procurar ficar bem fisicamente o mais rapidamente possível para entrar em campo e ajudar o **Sporting** a conquistar mais vitórias." Para concluir, Mota, que vai viajar amanhã rumo a Portugal, admitiu que tem falado com os seus empresários para saber mais informações da cidade e

dos jogadores do **Sporting**: "Tenho falado com os empresários para saber mais do **clube** e dos **jogadores**.".

Segment 2:
O Nacional venceu esta noite na Choupana o **Sporting** por 3-2, na partida que marcou a saıda de Casemiro Mior do **clube** insular. Com este resultado, os "**leões**" desperdiçaram o deslize de FC Porto e também a oportunidade de ascender ao primeiro lugar isolado do pódio. Os primeiros minutos de jogo davam sinais de que o **Sporting** estava a entrar bem no jogo e de pretendia "aceitar" a oportunidade da véspera proporcionada pelo FC Porto, - que foi empatar a Coimbra ante o último classificado (0-0) e voltar assim a reassumir a liderança da SuperLiga. Mas cedo essa imagem foi desfeita, a falta de ideias dos **joagdores** leoninos e a sua consequente ineficácia permitiram à equipa da casa, que pouco fazia para se abeirar da baliza adversària, aproveitar dois erros defensivos e chegar ao golo. Uma falha de Polga à passagem pelo minuto 18 permite a Adriano abrir a contagem na Choupana. Dois minutos volvidos Emerson, livre de marcaço, recebe o esférico e dilata a vantagem, fazendo o 2-0.

By tackling this particular situation, we propose a new challenge compared to other works that have been proposed so far and use test corpora based on multi-domain and multi-genre segments as in [Ferret 2002] [Brants et al. 2002] and [Moens and De Busser 2003]. In fact, the most similar experiment, to our knowledge, is the one proposed by [Xiang and Hongyuan 2003] who use the Mars novel. However, their segments are 2650 words-long while we deal with segments around 100 words each. In fact, we aim at proposing a fine-grained system capable of finding topic boundaries with high precision in a single domain and in short texts. To our knowledge, such a challenge has never been attempted so far. In order to evaluate our system, we propose two distinct experiments. First, we propose an evaluation on a set of web documents about a unique domain using words as the basic textual information. In a second experiment, we show that semantic knowledge automatically acquired from the text, embodied by phrases, can improve previous results. For that purpose, we use the SENTA Software proposed by [Dias 2002] that can be run "on the fly" due to its efficient implementation [Gil and Dias 2003] and flexibility as it does not need any previous knowledge. In order to run our experiments, we built our own corpus by taking from two Portuguese football websites[5] a set of 100 articles of approximatively 100 words each. Then, we built 10 test corpora by choosing randomly 10 articles from our database of 100 articles[6] leading to 10 texts of around 1000 words-long[7]. A classical way of evaluating retrieval systems is to

[5]http://www.abola.pt and http://www.ojogo.pt.

[6]We used the same methodology as [Choi 2000] to build the test corpora although in a smaller scale.

[7]The chosen parameters of our experiments were the following: block size = 2 sentences and EI window context = 10 words.

use Precision, Recall and F-measure that are respectively presented in Equation 17, 18 and 19.

$$Precision = \frac{\text{Number of correct retrieved topic boundaries}}{\text{Number of retrieved topic boundaries}} \tag{17}$$

$$Recall = \frac{\text{Number of correct retrieved topic boundaries}}{\text{Number of total correct topic boundaries}} \tag{18}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{19}$$

We show the results obtained by our system on our test corpus in Table 2.

Without Phrases and c=-1.5											
Texts	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Avg.
Preci.	0,64	0,67	0,80	0,73	0,60	0,73	0,80	0,64	0,60	0,70	0,69
Recall	0,78	0,67	0,89	0,89	0,67	0,89	0,89	0,78	0,67	0,78	0,79
F-mea.	0,70	0,67	0,84	0,80	0,63	0,80	0,84	0,70	0,63	0,74	0,73
With Phrases and c=-2.0											
Texts	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	Avg.
Preci.	0,58	0,73	1,00	0,64	0,64	0,62	0,82	0,64	0,45	0,80	0,69
Recall	0,78	0,89	1,00	0,78	0,78	0,89	1,00	0,78	0,56	0,89	0,84
F-mea.	0,66	0,80	1,00	0,70	0,70	0,80	0,90	0,70	0,50	0,84	0,75

Table 2: Quantitative Results

The results are good considering the challenging task we are facing. Indeed, by using words as basic textual units, the average F-measure reaches 73% being Recall 79% and Precision 69%. After different tuning, the best results were obtained for the value $c = -1.5$. By using phrases, extracted automatically, the results show slight improvements with an average F-measure value of 75% being Recall improved by 5% (84%) and Precision remaining unchanged (69%). In this second experiment, the best results were obtained with $c = -2$. The introduction of phrases allows a bigger number of correct decisions compared to single word processing in some cases (T3 and T7 specifically). However, in other ones, word units work better than with the introduction of phrases like in T9. In fact, when texts gather many small sentences, the *ps* function shows bad behavior. In particular, T9 shows this particularity which is enhanced by the integration of phrases leading to even worse results. In fact, by analyzing T9, we discovered that there were two sentences with 2 words and one sentence with only one word. In any case, these global results hide most of the behavior of our system and a more detailed evaluation is needed. As [Reynar 1994] evidences, Precision and Recall measures are overly strict. By taking into account only Precision and Recall, a hypothesized boundary close to a real segment boundary is

equally detrimental to performance as one far from a boundary. This definitely should not be the case. In order to solve this problem, [Beeferman et al. 1997] proposed a metric that weights exact matches more than near misses and yields a single score. However, [Beeferman et al. 1997] observed that computing this metric requires some knowledge of the collection as parameters have to be tuned and as a consequence, performance comparison on different collections may be difficult. So, up-to-now, there is no standard evaluation measure that the community agrees on. As a consequence, we present, in Table 3, the qualitative results of our system where (1) A stands for the number of exact matches, (2) $\pm n$ stands for the number of boundaries that missed the true boundary for n sentences, (4) > 2 stands for the number of boundaries that missed the true boundary for more than two sentences and (5) F stands for the boundaries that were proposed by the system that do not have any match in the test segmented text i.e. false boundaries.

Without Phrases and c=-1.5										
Texts	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
A	7	6	8	8	6	8	8	7	6	7
± 1	2	2	1	0	2	1	1	2	2	1
± 2	0	0	0	1	0	0	0	0	0	0
> 2	0	0	0	0	0	0	0	0	0	0
F	2	1	1	2	2	2	1	2	2	2
With Phrases and c=-2.0										
Texts	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
A	7	8	9	7	7	8	9	7	5	8
± 1	1	1	0	1	1	1	0	1	2	1
± 2	0	0	0	1	0	0	0	1	0	0
> 2	0	0	0	0	0	0	0	0	0	0
F	4	2	0	2	3	4	2	2	4	1

Table 3: Qualitative Results

The results presented in this section are promising as we deal with a very difficult challenge which is working without any linguistic knowledge, on the basis of small mono-domain texts with many inter-segments lexical repetitions. As we said earlier, to our knowledge, such a challenge has never been attempted so far. Although the quantitative and qualitative results show good figures, some work still need to be done, in particular, with respect to the sizes of the sentences in texts that cause some trouble in the topic boundary extraction. We expect that the new proposed similarity measures may solve this problem.

6 Conclusions and Future Work

In this paper, we have proposed a language-independent unsupervised topic segmentation system based on word-co-occurrences and the identification of phrases that avoids the accessibility to existing linguistic resources such as electronic dictionaries or lexico-semantic databases. In particular, our architecture pro-

poses 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 and emerging languages and finally systems that need previously existing harvesting training data. To our point of view, our main contributions to the field is the definition of a new weighting scheme and a new similarity measure, the informative similarity measure, $infosimba$, 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]. Our evaluation has shown promising results both with word units and phrases. Indeed, by using words as basic textual units, the average F-measure reaches 73% being Recall 79% and Precision 69%. Comparatively, by using phrase identification, the results show slight improvements with an average F-measure value of 75% being Recall improved by 5% (84%) and Precision remaining unchanged (69%). However, the existence of three main parameters (the block size, the window size to calculate the association measure and the topic discovery threshold) may introduce some drawbacks in our solution, although it also provides interesting properties. We will start with the properties. Thanks to the existence of these parameters, fine-tuning of topic segmentation can be done. Indeed, depending on the type of the topic segmentation that is required (topic segmentation inside one main topic text or topic segmentation inside a webpage that contains drastically different news as in electronic newspapers), the adjustment of the parameters may allow a coherent segmentation. However, the existence of parameters is a drawback for totally flexible systems. Indeed, these parameters need to be tuned depending on the wanted application and are usually evaluated by experimentation which introduces partial judgment. It is clear that theoretical work should be carried out in order to avoid the tuning of these parameters; maybe following [Ponte and Croft 1997] and [Beeferman et al. 1997] that propose research directions to avoid the tuning by experimentation. As immediate future work, we intend to test our system in different conditions of topic segmentation in order to find some clues that could help us in the definition of new theories to avoid parameter tuning. We will also experiment different association measures within the informative similarity measure in order to test whether drastically different results may be evidenced. And of course, we will experiment the two proposed similarity measures $S1$ and $S2$ based on a correct normalization. Finally, we strongly think that more work must be done on the automatic boundary detection algorithm. In particular, we are convinced that better algorithms may be proposed based on the transformation of the representation of the ps function into a graph or network. For that purpose, we would like to investigate possible solutions based on statistical mechanics of complex networks [Albert and Barabási 2002]. If the reader is interested in our solution, the system and its evolutions will be available for download as a General Public License at the following address: http://asas.di.ubi.pt.

References

[Salem 1987] Salem, A., "La Pratique des segments répétés", Klincksieck, Paris, (1987).

[Smadja 1993] Smadja, F., "Retrieving Collocations From Text: XTRACT", Computational Linguistics, 19(1), 143-117, (1993).

[Salton et al. 1993] Salton, G., Allan, J. and Buckley, C., "Approaches to passage retrieval in full text information systems", ACM-SIGIR, 4-58, (1993).

[Kaszkiel and Zobel 1997] Kaszkiel, M. and Zobel, J., "Passage retrieval revisited", ACM-SIGIR, 178-185, (1997).

[Cormack et al. 1999] Cormack, G.V., Clarke, C.L.A., Kisman, D.I.E. and Palmer, C.R., "Fast Automatic Passage Ranking. MultiText Experiments for TREC-8", TREC-8, 735-742, (1999).

[Boguraev and Neff 2000] Boguraev, B. and Neff, M., "Discourse Segmentation in Aid of Document Summarization", Hawaii International Conference on System Sciences, Minitrack on Digital Documents Understanding, Maui, Hawaii, IEEE, (2000).

[Angheluta et al. 2002] Angheluta, R., De Busser, R., Moens, M-F., "The Use of Topic Segmentation for Automatic Summarization", Workshop on Text Summarization in Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization, Philadelphia, Pennsylvania, USA, (2002).

[Farzindar and Lapalme 2004] Farzindar, A. and Lapalme, G., "Legal Text Summarization by Exploration of the Thematic Structures and Argumentative Roles", Workshop on Text Summarization Branches Out held in conjunction with ACL 2004, Barcelona, Spain, 27-38, (2004).

[Hearst 1994] Hearst, M., "Multi-Paragraph Segmentation of Expository Text", ACL 1994, Las Cruces, New Mexico, June, 9-16, (1994).

[Reynar 1994] Reynar, J.C., "An Automatic Method of Finding Topic Boundaries", ACL 1994 (Student Session), Las Cruces, New Mexico, June, (1994).

[Richmond et al. 1997] Richmond, K., Smith, A., and Amitay, E., "Detecting Subject Boundaries within Text: A Language Independent Statistical Approach", Second Conference on Empirical Methods in Natural Language Processing, Providence, Rhode Island, USA, August 1-2, 4-54, (1997).

[Yaari 1997] Yaari, Y., "Segmentation of Expository Text by Hierarchical Agglomerative Clustering", Conference on Recent Advances in Natural Language Processing, 59-65, (1997).

[Sardinha 2002] Sardinha, T., "Segmenting Corpora of Texts", DELTA, 18(2), 273-286, ISSN 0102-4450, (2002).

[Morris and Hirst 1991] Morris, J. and Hirst, G., "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text", Computational Linguistics, 17(1), 21-43, (1991).

[Kozima 1993] Kozima, H., "Text Segmentation Based on Similarity between Words", ACL 1993 (Student Session), Colombus, Ohio, USA, 286-288, (1994).

[Beeferman et al. 1997] Beeferman, D., Berger, A., and Lafferty, J., "Text Segmentation Using Exponential Models", Second Conference on Empirical Methods in Natural Language Processing, Providence, Rhode Island, USA, August 1-2, 35-46, (1997).

[Salton et al. 1975] Salton, G., Yang, C.S., and Yu, C.T., "A Theory of Term Importance in Automatic Text Analysis", American Society of Information Science, 26(1), 33-44, (1975).

[Sparck-Jones 1972] Salton, G., Yang, C.S., and Yu, C.T., "A Statistical Interpretation of Term Specificity and its Application in Retrieval", Journal of Documentation, 28(1), 11-21, (1972).

[Müller et al. 1997] Mller, C., Polanco, X., Royaut, J. and Toussaint, Y., "Acquisition et Structuration des Connaissances en Corpus: lments Mthodologiques", Technical Report RR-3198, Inria, Institut National de Recherche en Informatique et en Automatique, France, (1997).

[Phillips 1985] Phillips, M., "Aspects of Text Structure: An Investigation of the Lexical Organisation of Text", North Holland Linguistic Series, North Holland, Amsterdam, (1985).

[Ponte and Croft 1997] Ponte J.M. and Croft W.B., "Text Segmentation by Topic", First European Conference on Research and Advanced Technology for Digitial Libraries, Paris, 120-129, (1997).

[Ferret 2002] Ferret, O., "Using Collocations for Topic Segmentation and Link Detection", International Conference on Computational Linguistics (COLING 2002). Taipei, Taiwan, 26-30 August, (2002).

[David and Plante 1990] David, S. and Plante, P., "Termino Version 1.0", Research Report of Centre d'Analyse de Textes par Ordinateur, Universit du Qubec, Montral, (1990).

[Bourigault 1996] Bourigault, D., "Lexter, a Natural Language Processing Tool for Terminology Extraction", 7th EURALEX International Congress, (1996).

[Enguehard 1993] Enguehard, C., "Acquisition de Terminologie Partir de Gros Corpus", Informatique and Langue Naturelle, 373-384, (1993).

[Justeson 1993] Justeson, J., "Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text", IBM Research Report, RC 18906 (82591) 5/18/93, (1993).

[Daille 1995] Daille, B., "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology", The Balancing Act Combining Symbolic and Statistical Approaches to Language, MIT Press, (1995).

[Paio et al. 2003] Piao, S., Rayson, P., Archer, D., Wilson, A. and McEnery, T., "Extracting Multiword Expressions with a Semantic Tagger", Workshop on Multiword Expressions of ACL, July, Sapporo, Japan, 49-57, (2003).

[Church and Hanks 1990] Church, K.W. and Hanks P., "Word Association Norms Mutual Information and Lexicography", Computational Linguistics, 16(1), 23-29, (1990).

[Dunning 1993] Dunning T., "Accurate Methods for the Statistics of Surprise and Coincidence", Computational Linguistics, 19(1), (1993).

[Shimohata 1997] Shimohata, S., "Retrieving Collocations by Co-occurrences and Word Order Constraints", ACL-EACL, 476-481, (1997).

[Tomokiyo and Hurst 2003] Tomokiyo, T. and Hurst, M., "A Language Model Approach to Keyphrase Extraction", Workshop on Multiword Expressions of ACL July, Sapporo, Japan, (2003).

[Dias 2002] Dias, G., "Extraction Automatique d'Associations Lexicales à Partir de Corpora", Phd Thesis, University of Orléans, France and New University of Lisbon, Portugal, (2002).

[Xu and Croft 1996] Xu, J. and Croft, W.B., "Query Expansion Using Local and Global Document Analysis", Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 4-11, (1996).

[Smadja 1996] Smadja, F., "Translating Collocations for Bilingual Lexicons: A Statistical Approach", Computational Linguistics, 22(1), (1996).

[Sinclair 1974] Sinclair, J., "English Lexical Collocations: A study in Computational Linguistics", Uni Press, Singapore, (1974).

[Dai et al. 2003] Dai, P., Iurgel U. and Rigoll G., "A Novel Feature Combination Approach for Spoken Document Classification with Support Vector Machines", UWorkshop on Multimedia Information Retrieval in conjunction with the 26th annual ACM SIGIR Conference on Information Retrieval, Toronto, Canada, (2003).

[Moens and De Busser 2003] Moens, M-F. and De Busser, R., "Generic Topic Segmentation of Document Texts", ACM SIGIR conference on Documentation. San Francisco, USA, 117-124, (2003).

[Dice 1945] Dice, L.R., "Measures of the Amount of Ecologic Associations Between Species", Journal of Ecology, 26, (1945).

[Cover and Thomas 1991] Cover, T.M. and Thomas, J.A., "Elements of Information Theory", John Wiley and Sons, New York, (1991).

[Rao 1982] Rao, C.R., "Diversity: Its Measurement, Decomposition, Apportionment and Analysis", Indian Journal of Statistics, 44(A), 1-22, (1982).

[Silva et al. 1999] Silva, J., Dias, G., Guilloré, S. and Lopes, J.G.P., "Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units", 9th Portuguese Conference in Artificial Intelligence. Pedro Barahona and Júlio Alferes (eds). Lecture Notes in Artificial Intelligence number 1695, Springer-Verlag, Universidade de Évora, Évora, Portugal, September, 113-132, (1999).

[Cleuziou et al. 2003] Cleuziou G., Clavier V. and Martin L., "Une Méthode de Regroupement de Mots Fonde sur la Recherche de Cliques dans un Graphe de Cooccurrences", Rencontres Terminologie et Intelligence Artificielle, LIIA - ENSAIS, Strasbourg, France, 179-182, (2003).

[Stokes et al. 2002] Stokes, N., Carthy, J. and Smeaton, A.F., "Segmenting Broadcast News Streams Using Lexical Chains", Starting Artificial Intelligence Researchers Symposium, (1), 145-154, (2002).

[Xiang and Hongyuan 2003] Xiang, J. and Hongyuan, Z., "Domain-independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming", ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, 322-329, (2003).

[Brants et al. 2002] Brants, T., Chen, F. and Tsochantaridis, I., "Topic-Based Document Segmentation with Probabilistic Latent Semantic Analysis", CIKM International Conference on Information and Knowledge Management, McLean, Virginia, USA, 211-218, (2002).

[Choi 2000] Choi, F.Y.Y., "Advances in Domain Independent Linear Text Segmentation", NAACL, Seattle, USA, (2000).

[Gil and Dias 2003] Gil, A. and Dias, G., "Using Masks, Suffix Array-based Data Structures and Multidimensional Arrays to Compute Positional Ngram Statistics from Corpora", Workshop on Multiword Expressions of ACL, Sapporo, Japan, July, 25-33, (2003).

[Albert and Barabási 2002] Albert, R. and Barabsi, A-L., "Statistical Mechanics of Complex Networks", Reviews of Modern Physics, 74, The American Physical Society, January, (2002).