Unsupervised Topic Segmentation Based on Word Co-occurrence and Multi-Word Units for Text Summarization

Gaël Dias
Centre of Human Language Technology and Bioinformatics
University of Beira Interior
+351 275 319 891
ddg@di.ubi.pt

Elsa Alves
Natural Language Research Group
Department of Computer Science
New University of Lisbon
+351 21 294 85 36
elsalves@zmail.pt

ABSTRACT
Topic Segmentation is the task of breaking documents into topically coherent multi-paragraph subparts. In particular, Topic Segmentation is extensively used in Passage Retrieval and Text Summarization to provide more coherent results by taking into account raw document structure. However, most methodologies are based on lexical repetition that show evident reliability problems or rely on harvesting linguistic resources that are usually available only for dominating languages and do not apply to less favored and emerging languages. Moreover, most systems have been evaluated using Choi's data set [1] which is biased for systems using mostly lexical repetition. As a consequence, these systems are not tested in real-world environments and their application may prove worst results than presented in the literature. In order to tackle all these drawbacks, we present an innovative Topic Segmentation system based on a new informative similarity measure based on word co-occurrences and evaluate it on a set of web documents within which Multiword Units have previously been identified.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods.*

General Terms
Algorithms, Measurement, Experimentation.

Keywords
Unsupervised Topic Segmentation, Evaluation on Single Domain Web Documents, Text Summarization, Passage Retrieval.

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.

ELECTRA Workshop -Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications Salvador, Brazil, August 19, 2005, in association with SIGIR 2005

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 texts.
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 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 [2][3][4]. 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 [5][6][7].
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. In particular, our architecture solves three main problems evidenced by previous research. First, systems based uniquely on lexical repetition show reliability problems [8][9][10][11][12] as common writing rules prevent from using lexical repetition. Second, systems based on lexical cohesion, using existing linguistic resources that are usually only available for dominating languages like English, French or German, do not apply to less favored and emerging languages [13][14]. Third, systems that need previously existing harvesting training data [15] do not adapt easily to new domains as training data is usually difficult to find or build depending on the domain being tackled. Instead, our architecture proposes a language-independent unsupervised solution, similar to [16][17], defending that Topic Segmentation should be done "on the fly" on any 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 order to show the results of our system in real-world conditions, we propose two different evaluations on a set of web documents: (1) one based only on words and (2) one based on the set of documents within which multiword units have previously been identified "on the fly". Unlike other methodology that have been evaluated on Choi's data set [1] which relies on small texts of different domains within which lexical repetition is high, we propose an evaluation on real-world texts where lexical distribution does not overuse repetition. In particular, we show

that the introduction of semantic information into the set of documents such as Multiword Units leads to better results.
This paper is divided into five sections. First, we show the main differences between our work and the existing ones, in particular the systems proposed by [16] and [17]. Second, we show the weighting process of each word of the input text corpus. Third, we introduce our main innovation i.e. the informative similarity measure. Fourth, we define how subparts can be elected from the values of the informative similarity measure. And fifth, we propose an evaluation on a real-world situation using "on the fly" identification of Multiword Units.

2. RELATED WORK

[8], [9] and [12] have proposed different architectures based on lexical item[1] 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, [14] 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, [13] 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 and emerging languages.
In order to avoid the use of huge linguistic resources, [15] have proposed a technique for identifying document boundaries using statistical techniques. So, 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. [16] and [17] have proposed such techniques. [16] 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. [17] propose a Topic Segmentation technique based on the Local Content Analysis [18] 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 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 reducing the problem of word repetition. In order to introduce endogenously acquired semantic knowledge, [19] has also proposed to automatically extract collocations from texts in order to compute semantic similarity measures[2].
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 [20][21] 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 in the text. Moreover, in order to introduce a certain degree of semantics in our system, we propose a new informative similarity measure that includes in its definition the Equivalence Index Association Measure proposed by [22] so that word co-occurrence information is directly embedded in the calculation of the similarity between blocks of sentences. Thus, unlike [17], 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 the standard deviation algorithm proposed by [8].

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.
The simplest form of the vector space model treats a document (in our case, a sentence or a group of sentences) as a vector whose values correspond to the number of occurrences of the words appearing in the document as in [8]. Although [8] 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 (*idf*) [20][21] and its distribution across the text. For that purpose, we propose a new weighting scheme 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 in the text.

[1] A lexical item can be a sequence of characters, a stem, a morphological root, a word or an ngram.

[2] We will show in our final section that this methodology proves to lead to encouraging results.

3.1 The *tf.idf* Score

The basic idea of the *tf.idf* score [21] 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 meaning more strongly than words that occur less frequently) and its distribution across a collection of documents (i.e. terms that are limited to a few documents are useful for discriminating those documents from the rest of the collection). The *tf.idf* score is defined in equation 1 where *w* is a word and *d* a document.

$$tf.idf(w,d) = \frac{tf(w;d)}{|d|} \times \log_2 \frac{N}{df(w)} \quad (1)$$

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* [20] 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)*).

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 to segment the text into topics. For that purpose, we extend the idea of the *tf.idf* to sentences.

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. The *tf.isf* score is defined in equation 2 where *w* is a word and *s* a sentence.

$$tf.isf(w,s) = \frac{stf(w;s)}{|s|} \times \log_2 \frac{Ns}{sf(w)} \quad (2)$$

For each *w* in *s*, we compute its relative sentence term frequency, that is the number of occurrences of *w* in *s*, s*tf(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.

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 of all consecutive occurrences of the word in the document. We call this measure *dens* and is defined in equation 3.

$$dens(w,d) = \sum_{k=1}^{|w|-1} \frac{1}{\ln(dist(occur(k), occur(k+1)) + e)} \quad (3)$$

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.

3.4 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 distribution across the document. A straightforward definition of the weighting score is given in equation 4 where each score is normalized so that they can be combined.

$$weight(w,d) = \left\| tf.idf(w,d) \right\| \times \left\| tf.isf(w,s) \right\| \times \left\| dens(w,d) \right\| \quad (4)$$

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. 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 $Xi = (X_{i1}, X_{i2}, X_{i3}, ..., X_{ip})$ 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. Most applications in Natural Language Processing have used the cosine similarity measure. However, we will show that it evidences problems, like all other similarity measures proposed so far.

4.1 The Drawback of Similarity Measures

The cosine similarity (Equation 5) determines the angle between the vectors associated to two documents (in our case, the focus sentence and a group of *k* sentences). However, 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*.

The most interesting idea to avoid word repetition problems is certainly to identify lexical cohesion relationships between words.

$$S_{ij} = \cos(X_i, X_j) = \frac{\sum_{k=1}^{p} X_{ik} \times X_{jk}}{\sqrt{\sum_{k=1}^{p} X_{ik}^2} \times \sqrt{\sum_{k=1}^{p} X_{jk}^2}} \quad (5)$$

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 Wordnet [6][23], Roget's thesaurus [13] or LDOCE [14]. However, these huge resources are only available for dominating languages and as a consequence do not apply to less favored languages.

4.2 The Informative Similarity Measure

A much more interesting research direction is proposed by [17] that propose a Topic Segmentation technique based on the Local Content Analysis [18], allowing substituting each sentence with words and phrases related to it. 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 [17], we propose 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 (*EI*) proposed by [22] that has shown successful results in our different research works [24] [25]. It is defined in equation 6.

$$EI(w_1, w_2) = p(w_1 \mid w_2) \times p(w_2 \mid w_1) = \frac{f(w_1, w_2)^2}{f(w_1) \times f(w_2)} \quad (6)$$

The Equivalence Index between words w_1 and w_2 is calculated within a word-context window in order to determine the frequency between w_1, and w_2 ($f(w_1, w_2)$) and from a collection of documents so that we can evaluate the degree of cohesiveness between two words outside the context of the document. 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 [24].

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. This can be done straightforwardly as defined in equation 7 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*. 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 *EI*. 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 for a topic shift.

$$S_{ij} = \text{infosimba}(X_i, X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{jl} \times EI(W_{ik}, W_{jl})}{\sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{il} \times EI(W_{ik}, W_{il})} \times \sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{jk} \times X_{jl} \times EI(W_{jk}, W_{jl})}} \quad (7)$$

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 [8].

5. TOPIC BOUNDARY DETECTION

Different methodologies have been proposed to place subtopic boundaries between dissimilar blocks depending on the models used to determine similarity between blocks of sentences [8] [14] [15][17][26]. In fact, 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. Taking as reference the idea of [17] 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 way [15] compare short and long-range models. Our preference score (*ps*) is defined in equation 8.

$$ps(S_i) = \log_2 \frac{\text{infosimba}(S_i, X_{i-1})}{\text{infosimba}(S_i, X_{i+1})} \quad (8)$$

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 X_i is similar to both blocks and so we may be in the continuity of a topic. In order to illustrate the variations of the *ps* score, we show, in Figure 1, an experiment made with five texts taken from the web with five different topics.

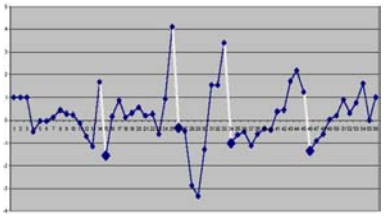

Figure 1: Preference score variation

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. 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 deeper ones must be taken into account. They are represented in white in Figure 1 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 [8] 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 9 whenever the value of the *ps* score goes from positive to negative between two consecutive sentences S_i and S_{i+1}.

$$downhill(S_i, S_{i+1}) = ps(S_i) - ps(S_{i+1}) \quad (9)$$

Once all downhills in the text have been calculated, their mean $\bar{x}$ and standard deviation σ are evaluated. The topic boundaries are then elected if they satisfy the constraint expressed in equation 10 where *c* is a constant to be tuned.

$$downhill(S_i, S_{i+1}) \geq \bar{x} + c\sigma \quad (10)$$

By applying this threshold, we obtain promising results for the discovery of topic boundaries for the specific case of web news segmentation. We illustrate these results in the next section.

6. RESULTS

Topic Segmentation systems [19][27][28] have usually been evaluated on [1]'s data set that represents the gold standard for evaluation. However, many authors have discussed the validity of this test corpus [19][23][27][28] and proposed their own test corpus. Indeed, [1]'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 in Figure 2.

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.

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.

Figure 2: Example of the C99 corpus (Directory 3-5, Text 7)

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 in Figure 3 where the inter-segments lexical repetitions are covered in yellow and the intra-segments lexical repetitions are covered in red. 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 [19][23][28]. In fact, the most similar experiment, to our knowledge, is the one proposed by [27] 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.

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 companheiro: «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.».

O Nacional venceu esta noite na Choupana o Sporting por 3-2, na partida que marcou a saída 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 jogadores leoninos e a sua consequente ineficácia permitiram à equipa da casa, que pouco fazia para se abeirar da baliza adversária, 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 marcação, recebe o esférico e dilata a vantagem, fazendo o 2-0.

Figure 3: Our Test corpus

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 Multiword Units, can improve previous results. For that purpose, we use the SENTA Software proposed by [29] that can be run "on the fly" due to its efficient implementation [30] 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 soccer websites[3] 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[4] leading to 10 texts of around 1000 words-long[5].

A classical way of evaluating retrieval systems is to use Precision, Recall and F-measure. So, we show the results obtained by our system on our test corpus in Table 1.

Table 1. Quantitative Results

	Without multiword units		With multiword units	
	Measures	c=-1.5	Measures	c=-2
T1	Precision	0,64	Precision	0,58
	Recall	0,78	Recall	0,78
	F-measure	0,70	F-measure	0,66
T2	Precision	0,67	Precision	0,73
	Recall	0,67	Recall	0,89
	F-measure	0,67	F-measure	0,80
T3	Precision	0,80	Precision	1,00
	Recall	0,89	Recall	1,00
	F-measure	0,84	F-measure	1,00
T4	Precision	0,73	Precision	0,64
	Recall	0,89	Recall	0,78
	F-measure	0,80	F-measure	0,70
T5	Precision	0,60	Precision	0,64
	Recall	0,67	Recall	0,78
	F-measure	0,63	F-measure	0,70
T6	Precision	0,73	Precision	0,62
	Recall	0,89	Recall	0,89
	F-measure	0,80	F-measure	0,73
T7	Precision	0,80	Precision	0,82
	Recall	0,89	Recall	1,00
	F-measure	0,84	F-measure	0,90
T8	Precision	0,64	Precision	0,64
	Recall	0,78	Recall	0,78
	F-measure	0,70	F-measure	0,70
T9	Precision	0,60	Precision	0,45
	Recall	0,67	Recall	0,56
	F-measure	0,63	F-measure	0,50
T10	Precision	0,70	Precision	0,80
	Recall	0,78	Recall	0,89
	F-measure	0,74	F-measure	0,84
Average	Precision	0,69	Precision	0,69
	Recall	0,79	Recall	0,84
	F-measure	0,73	F-measure	0,75

The results are surprisingly good considering the challenging task we were 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 Multiword Unit 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%). In this second experiment, the best results were obtained with c=-2.The introduction of Multiword Units 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 Multiword Units like in T9. In fact, when texts gather many small sentences, the *ps(.)* function show bad behavior. In particular, T9 shows this particularity which is enhanced by the integration of Multiword Units 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[6].

In any case, these global results hide most of the behavior of our system and a more detailed evaluation is needed.

Table 2. Qualitative Results

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

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

[4] We used the same methodology as [1] to build the test corpora although in a smaller scale.

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

[6] We are already working on a normalization measure that takes into account sentence length.

As [9] 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, [15] proposed a metric that weights exact matches more than near misses and yields a single score. However, [15] 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 2, the qualitative results of our system where (1) A stands for the number of exact matches, (2) ± *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.

We can see from these results, which by taking into account, as correct boundaries, all A and near misses ± 1, that we would obtain between 84% and 89% F-measure as shown in Table 3.

Table 3. Estimated Results

Without multiword units		With multiword units	
Precision	0,83	Precision	0,77
Recall	0,95	Recall	0,93
F-measure	0,89	F-measure	0,84

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.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a language-independent unsupervised Topic Segmentation system based on word-co-occurrences that avoids the accessibility to existing linguistic resources such as electronic dictionaries or lexico-semantic databases. 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 and emerging languages and finally systems that need previously existing harvesting training data. To our point of view, our main contribution 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 [17]. Our evaluation has shown promising results both with word units and Multiword Units. Indeed, by using words as basic textual units, the average F-measure reaches 73% being Recall 79% and Precision 69%. Comparatively, by using Multiword Unit 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 [17] and [15] 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. 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 [33]. The system and its evolutions will be available for download as a GPL license at the following address: http://asas.di.ubi.pt.

8. REFERENCES

[1] Choi, F.Y.Y. 2000. Advances in Domain Independent Linear Text Segmentation. In Proceedings of NAACL'00, Seattle, April 2000. ACL.

[2] Salton, G., Allan, J. and Buckley, C. 1993. Approaches to passage retrieval in full text information systems. In Proceedings of ACM-SIGIR. 4--58.

[3] Kaszkiel, M. and Zobel, J. 1997. Passage retrieval revisited. In Proceedings of ACM-SIGIR, 178--185.

[4] Cormack, G.V., Clarke, C.L.A., Kisman, D.I.E. and Palmer, C.R. 1999. Fast Automatic Passage Ranking. MultiText Experiments for TREC-8. In Proceedings of TREC-8. 735-742.

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

[6] Angheluta, R., De Busser, R., Moens, M-F. 2002. The Use of Topic Segmentation for Automatic Summarization. In Workshop on Text Summarization in Conjunction with the

ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization. July 11-12, Philadelphia, Pennsylvania, USA.

[7] Farzindar, A. and Lapalme, G. 2004. Legal text summarization by exploration of the thematic structures and argumentative roles. In Text Summarization Branches Out Conference held in conjunction with ACL 2004, Barcelona, Spain, 27-38

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Sardinha, T.B. 2002. Segmenting corpora of texts. DELTA, 2002, 18(2), 273--286. ISSN 0102-4450.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

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

[19] Ferret, O. 2002.*Using Collocations for Topic Segmentation and Link Detection*. In Proceedings of COLING 2002, 19th International Conference on Computational Linguistics, August 24 - September 1, 2002, Howard International House and Academia Sinica, Taipei, Taiwan.

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

[21] 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.

[22] 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

[23] Moens, M-F. and De Busser, R. 2003. Generic topic segmentation of document texts. In Proceedings of the 24th annual international ACM SIGIR conference on Documentation. San Francisco, USA. 117--124.

[24] Cleuziou G., Clavier V., Martin L. 2003. Une méthode de regroupement de mots fondée sur la recherche de cliques dans un graphe de cooccurrences. In Proceedings of the 5èmes rencontres Terminologie et Intelligence Artificielle), LIIA - ENSAIS ed., pages 179--182, Strasbourg, France.

[25] Silva, J., Dias, G., Guilloré, S. and Lopes, J.G.P. 1999. Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In the 9th Portuguese Conference in Artificial Intelligence. Pedro Barahona and Júlio Alferes (eds). Lecture Notes in Artificial Intelligence nº1695, Springer-Verlag, Universidade de Évora, Évora, Portugal, September, 113--132.

[26] Stokes, N., Carthy, J. and Smeaton, A.F. 2002. Segmenting Broadcast News Streams Using Lexical Chains. In Proceedings of 1st Starting AI Researchers Symposium (STAIRS 2002), volume 1, pp.145--154.

[27] Xiang, J. and Hongyuan, Z. 2003. Domain-independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming. In proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada. pp.322—329.

[28] Brants, T., Chen, F. and Tsochantaridis, I. 2002. Topic-Based Document Segmentation with Probabilistic Latent Semantic Analysis. In Proceedings of the CIKM 11th International Conference on Information and Knowledge Management, McLean, Virginia, USA. Pp.211-218.

[29] Dias, G., Guilloré, S., Bassano, J.C. and Lopes, J.G.P. 2000. *Extraction Automatique d'unités Lexicales Complexes: Un Enjeu Fondamental pour la Recherche Documentaire*. In Traitement Automatique des Langues, Vol 41:2, Christian Jacquemin (eds). Paris, France. pp. 447-473. ISBN: 2-7462-0225-5.

[30] Gil, A. & Dias, G. (2003). *Using Masks, Suffix Array-based Data Structures and Multidimensional Arrays to Compute Positional Ngram Statistics from Corpora*. In proceedings of the Workshop on Multiword Expressions of the 41st Annual Meeting of the Association of Computational Linguistics, Sapporo, Japan, July 7-12. pp. 25-33. ISBN: 1-932432-20-5.

[31] Albert, R. and Barabási, A-L. 2002. Statistical mechanics of complex networks. In Reviews of Modern Physics. Vol 74. The American Physical Society. January.