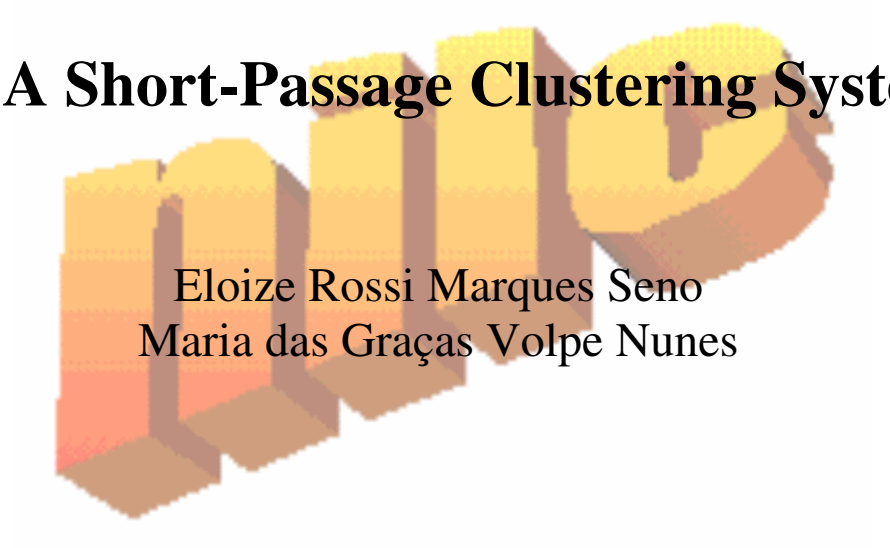
Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

SiSPI: A Short-Passage Clustering System



Eloize Rossi Marques Seno
Maria das Graças Volpe Nunes

NILC-TR-08-01

Janeiro, 2008

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Abstract

We describe SiSPI, a clustering tool based on an unsupervised and incremental approach which aims at arranging short passages from one or multiple documents written in Brazilian Portuguese into clusters. In order to identify similar passages, SiSPI makes use of a statistical model, named TF-ISF (Term Frequency - Inverse Sentence Frequency). By grouping similar passages into the same cluster, SiSPI enables a subsequent alignment/fusion component to transform each cluster into a single sentence by fusing common information. We present a pilot experiment which evaluates the system performance in the news domain. The results obtained suggest that SiSPI has potential.

This work is support by CNPq

Summary

Figures

Tables

1. Introduction

This report introduces SiSPI, an acronym for *Similar Short Passages Identifier*, which focuses on the problem of identifying similar short passages from one or multiple documents related to the same subject (or topic) and then grouping them into clusters.

Short passages clustering has many applications in Natural Language Processing - NLP such as multidocument summarization - grouping paragraphs and sentences to be reformulated to compose a summary (Hatzivassiloglou et al., 2001), monodocument summarization - guiding the sentence selection process (Wang et al., 2003; Hu et al., 2004), digital libraries - clustering sentences to facilitate information access (Tombros et al., 2004), ontology enhancement with concepts or relationships identified by sentence clustering (Schaal et al., 2005) and spoken language understanding systems - grouping sentences for semantic decoding (Ye and Young, 2006).

In this work, we refer to short passages only as sentences. Sentence clustering is performed as a primary step towards aligning and fusing common information among similar sentences (e.g., paraphrases and synonyms). In other words, SiSPI is the first module of a sentence fusion system which aims at producing a single sentence by combining information from several similar sentences. A fusion system is useful, for instance, to treat redundancy in multidocument summarization systems (Barzilay and McKeown, 2005) and to formulate answers in question answering systems (Xie and Liu, 2005).

SiSPI has been developed to treat documents written in Brazilian Portuguese. It is based on an unsupervised and incremental clustering approach that is combined to a statistical model in order to identify and to group semantically similar sentences. More specifically, it makes use of Salton et al.'s vector space model (Salton et al., 1975) in which each cluster is depicted by a weight vector that represents the relevance of the words in that cluster. The weight of each word is computed by using the TF-ISF measure - Term Frequency Inverse *Sentence* Frequency (Larocca Neto et al., 2000), which is an adaptation of the standard TF-IDF measure - Term Frequency Inverse *Document* Frequency from Information Retrieval (Salton, 1989). Similarity between a sentence and a cluster is given by the cosine distance between the term frequency vector of the corresponding sentence and the vector with the highest TF-ISF values of the corresponding cluster, named centroid (see Section 3).

The notion of similarity is a key concept to SiSPI, aiming at identifying sets of highly semantically-related sentences from a collection of documents. The similarity definition used in this work follows the Hatzivassiloglou et al.'s definition (Hatzivassiloglou et al., 1999), which has been proposed specifically to the task of common information fusion. Thus, we regard two sentences as similar if they refer to the same concept, actor, object or action. Moreover, the actor or object must accomplish the same action in both units or be subject of the same description. Figure 1 presents four sentences extracted from the experimental corpus (see Section 4), all referring to the TAM airplane crash. While sentences (a), (b) and (c) focus on the crash description by presenting details about how it happened, sentence (d) emphasizes the biggest air crash in the history of the country (Brazil). Therefore, we consider only sentences (a), (b) and (c) as similar.

(a) Um avião da TAM com capacidade para 170 passageiros derrapou na pista do Aeroporto de Congonhas, na Zona Sul de São Paulo, atravessou uma avenida e bateu em um prédio de carga e descarga da companhia aérea. *(One TAM airplane with a capacity of 170 passengers skidded on Congonhas Airport runway, in the south zone of São Paulo, crossed an avenue and crashed into a warehouse building of the air company.)*

(b) O avião da TAM com 176 pessoas a bordo derrapou na pista do Aeroporto de Congonhas, em São Paulo, atravessou a avenida Washington Luiz e bateu em um prédio da TAM Express. *(TAM airplane with 176 people on-board skidded on Congonhas Airport runway, in São Paulo, crossed Washington Luiz avenue and crashed into the building of TAM Express.)*

(c) O vôo 3054 da TAM com passageiros a bordo derrapou na noite desta terça-feira enquanto pousava no aeroporto de Congonhas (zona sul de São Paulo) e bateu contra um depósito da empresa que fica do lado oposto da avenida Washington Luiz. *(The flight number 3054 operated by TAM with passengers on-board skidded on Tuesday evening while it was landing on Congonhas airport (south zone of São Paulo) and crash into a store of the company that is located on the opposite side of Washington Luiz avenue.)*

(d) Um acidente com um Airbus da TAM que se chocou com dois prédios e um posto de gasolina na terça-feira após não conseguir frear quando pousava no Aeroporto de Congonhas pode ter sido o maior desastre aéreo da história do país se for confirmada a morte de todas as 176 pessoas que estavam a bordo. *(A crash involving a TAM Airbus which collided with 2 buildings and a gas station on Tuesday after not being able to break when it was landing on Congonhas airport can be the biggest air disaster in the history of the country if the death of all 176 people were on-board is confirmed.)*

Figure 1: Examples of similar and non-similar sentences

The remainder of this report is organized as follows. Some related works are described next in Section 2 and the proposed clustering system is described in Section 3. An experimental evaluation of SiSPI is presented in Section 4, and some final remarks are presented in Section 5.

2. Related Work

Various techniques for detecting similar short passages have been proposed in the literature recently. Most of them base on unsupervised approaches (clustering methods) and rely on statistics of words in common (Hu et al., 2004; Schaal et al., 2005; Ye and Young, 2006). In general, they make use of Salton et al.'s vector space model (Salton et al., 1975) and of some statistical similarity measure to identify similar passages. Schaal et al. (2005), for example, use the TF-IDF model, which is widely utilized for document clustering (e.g., Radev et al., 1999; Larocca Neto et al., 2000), in order to cluster sentences and paragraphs for ontology enhancement. In that work, each passage is represented by a weight vector over a feature space of concepts from ontology. The weights are computed by using the TF-IDF product of the term calculated for a feature. Similarity between short passages is given by the cosine distance (normalized inner product) of the corresponding vectors. The clustering algorithm employed in that work is Bi-Secting K-means, which is a variation of K-means algorithm. In Hu et al. (2004), the vector representation is used to cluster paragraphs in order to facilitate the sentence selection process in automatic summarization systems. A statistical measure, similar to TF-IDF, is used to determine the relevance of each term of

a paragraph. Similarity between paragraphs is measured by using the Euclidean distance and the clustering method employed is K-medoids, which is also a variation of K-means algorithm. Since in those works the task of identifying similar short passages is an intermediate process, they do not present results assessing specifically the clustering method performance.

Despite those works focus on the detection of short passages, there is a major difference with respect to their concept of similar passages and ours. That is, the concept of similarity used here is more restrict than the one used in earlier works, as presented in the previous section. Regarding the notion of similarity, our work is more similar to Hatzivassiloglou et al.'s work (Hatzivassiloglou et al., 1999). However, they utilize a supervised approach which is based on deep linguistic knowledge. More specifically, Hatzivassiloglou et al. make use of a rule induction method, called RIPPER, which combines 43 linguistics features in order to classify paragraph pairs as similar or non-similar. Such features include morphological, syntactic and semantic information (e.g., verb-object and subject-verb relations, noun phrases, proper nouns, synonyms). RIPPER has been trained with a corpus of 10.345 manually-classified paragraph pairs. Using three-fold cross-validation, the algorithm included 11 out of the 43 features in the induced set of rules and obtained 45.6% F-measure. In a subsequent experiment, reported by Hatzivassiloglou et al. (2001), a log-linear regression model has been based on a more refined set of those features. In addition, they have used a co-reference resolution component that allows comparing multiple forms of the same name. This model resulted in a performance increase of 51.0% F-measure compared to RIPPER.

Hatzivassiloglou et al. (1999; 2001) also present some experiments using the TF-IDF model to cluster paragraphs of documents written in English, which have been performed by using the same data set used by the RIPPER classifier and by the regression model. The TF-IDF model has obtained 36.7% F-measure on average, while a more sophisticated version of TF-IDF that uses the word stems and a list of irrelevant words has obtained 36.3% F-measure on average. These results may indicate that the TF-IDF model is not appropriate to identify highly semantically-related passages and that a more specific model is required to treat short passages, for instance, the TF-ISF model that has been proposed to treat sentence rather than documents. Based on this hypothesis, our system makes use of the TF-ISF model, as it will be explained in next section.

3. System Description

In this section we present SiSPI architecture and we describe the clustering approach employed by this system.

3.1 Architecture

SiSPI system is composed by two main processing modules named Sentence Splitting and Sentence Clustering (Figure 2). The former is responsible for splitting each document of a collection into sentences by using a textual-segmentation tool called SENTER (Pardo, 2006). The latter is responsible for identifying and clustering similar sentences. During this process, SiSPI makes use of a Brazilian Portuguese stemmer (Caldas Jr. et al., 2001) which identifies the word stems, thus allowing to consider words with the same stem, but with different flections (e.g. *venceram* and *venceu* - won). In addition, it utilizes a stoplist, that is, a list of common words that are irrelevant

for the processing (e.g. articles, pronouns and prepositions). The use of the stemmer and of the stoplist may contribute to enhance the similarity detection process performance, since the stemmer allows to identify words of the same semantic class and the stoplist eliminates the words that generate noise in that process. As a result of the clustering process, the system produces several files of sentence clusters.

It is worth noting that SiSPI is domain independent, for it is based only on lexical information. It is also weakly language-dependent, for it just uses a stemmer and a stoplist and it does not utilizes any deep linguistic knowledge (e.g., syntactic and semantic information).

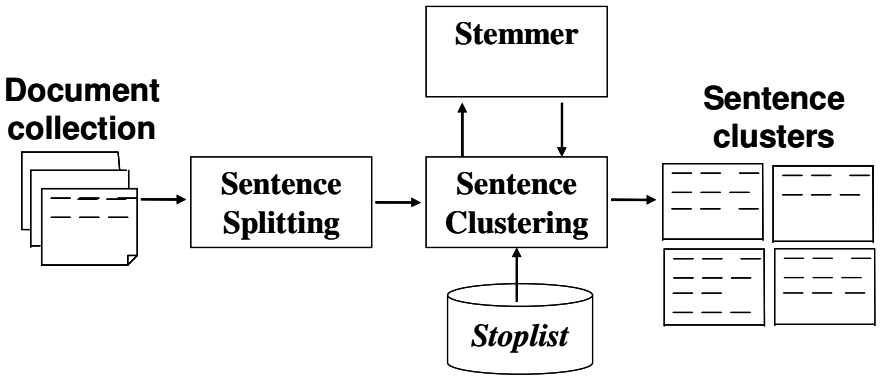


Figure 2: SiSPI architecture

3.2 Clustering Approach

Clustering comprises both a similarity metric and a clustering method. There are various clustering algorithms in the literature and they can be classified as hierarchical (e.g. Single-link (Sneath and Sokal, 1973)) or non-hierarchical (e.g. Single-pass (Van Rijsbergen, 1979), K-means (MacQueen, 1967). While the complexity of the former ones is $O(n^2 log(n))$, in which n is the number of elements to be clustered, the complexity of the latter is generally linear. For instance, the space complexity of Single-pass algorithm is $O(n)$ and the time complexity is $O(n \log n)$. Due to the simplicity and the effectiveness of Single-pass, it has become one of the most popular clustering algorithms, mainly among the Information Retrieval community (e.g. Radev et al., 1999; Klampanos et al., 2006). So, in order to achieve high efficiency of our method, we have also chosen Single-pass.

As the name suggests, Single-pass requires a single sequential pass over the set of sentences to be clustered. It is an incremental clustering algorithm in which the clusters are created incrementally at each iteration. The general schema of Single-pass method to treat sentences is shown in Figure 3.

Intput: A set $D = <d_1,\ldots,d_n>$ with n documents where each $d_i = <s_1,\ldots,s_m>$ with m sentences for n and m >= 1.
Output: A set $C = <c_1,\ldots,c_x>$ with x clusters where each $c_i = <s_1,\ldots,s_y>$ with y sentences for y >= 1.

Step 1: Set the initial set C of clusters to be empty
Step 2: Select a sentence s_i of a document d_i following a given order
If C is empty
then add the first cluster to C by inserting a single element s_i
else compare s_i (treated as a new cluster with only one element) with all the clusters in C
If the similarity between s_i and any cluster in C is above a predetermined threshold
then place s_i within the closest cluster in C
else add the new cluster to C
Step 3: Repeat the step above until all the sentences of all documents are processed.

Figure 3: General schema of Single-pass method to cluster sentences (Van Rijsbergen, 1979)

Initially, the algorithm creates the first cluster by selecting the first sentence of a document collection to be clustered. Then, this first cluster starts the work of clustering all sentences from that collection. At each iteration, the algorithm decides on whether a newly selected sentence should be placed in an already created cluster or be placed in a new one. This decision is made according to a condition specified by the similarity function employed, that is, a previously determined similarity threshold. The similarity threshold is a value in the range of 0 to 1, which is derived experimentally (see Section 4).

In this work, the similarity function is the cosine coefficient (Salton, 1989) applied to the term frequency vector of a sentence and to the vector that represents the most important terms of a cluster, named centroid. The larger the similarity value between the vectors, the more similar the sentence is to that cluster. It is worth noting that in SiSPI each sentence belongs to a single cluster, that is, the cluster most similar to it.

The calculation of a cluster centroid is based on the TF-ISF (Term Frequency Inverse Sentence Frequency) values of the corresponding words of that cluster (Larocca Neto et al., 2000). The TF-ISF value of a word w of a cluster c, denoted *TF-ISF(w,c),* is given by the following formula:

$$(1) \qquad TF\text{-}ISF(w,c) = TF(w,c) * ISF(w)$$

where *TF(w,c)* depictes the number of times the word w occurs in cluster c, *i.e.,* the frequency of w in c. As we have shown in Section 3, we use the word stem and a

stoplist in order to compute the word frequency. The higher *TF(w,c)*, the more representative of the cluster *c* the word *w* is. The inverse sentence frequency of a word *w*, denoted *ISF(w)*, is given by Formula 2, where *S* is the total of sentences in the current cluster and *SF(w)* is the sentence frequency of the cluster in which *w* occurs.

$$(2) \qquad ISF(w) = 1 + \log (|S| / SF(w))$$

According to Formula 2, the *ISF* of a word *w* is high if *w* is present in few sentences of a cluster, meaning that *w* has a great cluster-discriminating power. On the other hand, the *ISF* of a word *w* is low if *w* occurs in a variety of sentences of a cluster, indicating that *w* has a little cluster-discriminating power.

In order to a word be representative of a given cluster it must have both a high TF value and a high ISF value (therefore, a high TF-ISF value). Thus, only the words with highest TF-ISF scores are selected to represent the cluster centroid. The number of words to be selected is a predetermined parameter. This parameter is also derived experimentally, as it will be explained in the next Section.

4. Experimental Evaluation

The existing measures in the literature to assess the quality of a clustering method may be divided into external and internal measures (Steinback et al., 2000). External quality measures evaluates how good the clusters produced by a given algorithm are, by comparing them to reference clusters (typically manually classified clusters). So, this kind of evaluation can be carried on only if the class for each sentence of a document set is determined a priori. On the other hand, internal quality measures do not make use of any external knowledge and assess only the cohesiveness of a clustering solution, i.e., how similar the elements of each cluster are. If the purpose is to measure the goodness of a solution or the effectiveness of the clustering method, external measures are more appropriated.

In this experiment, SiSPI has been assessed by three external quality measures that will be presented in Section 4.2. Next, we describe the corpus used for evaluation.

4.1 The Evaluation Corpus

The evaluation corpus was composed by 20 collections of news articles written in Brazilian Portuguese, with 3.6 documents per collection on average, all on the same subject. This corpus has been manually collected from several web news agencies and totalizes 1.153 sentences from 71 documents.

Aiming at creating a reference clustering corpus, each sentence of a document collection has been manually classified by the first author of this work, according to the similarity definition presented in Section 1. In cases where there was more than one possible class to the same sentence, only one has been chosen, since in SiSPI each sentence is added to a single cluster (see Section 3). Decisions about the best class to be chosen were based on semantic similarity (that is, the cluster which was most semantically similar to that sentence) or randomly, in cases where clusters were considered equally similar to that sentence. Henceforth, we will refer to manual classifications consisting of *classes* and automatic clustering consisting of *clusters*.

4.2 The Evaluation Measures

The first measure is the widely used F-measure (Fung et al., 2003), which was used to assess the accuracy of the produced clustering solution. This metric combines two other metrics called Precision and Recall.

Let N be the total number of sentences to be clustered, K the set of classes, C the set of clusters and n_{ij} the number of sentences of the class $k_i \in K$ that are present in cluster $c_j \in C$. The Precision, Recall and F-measure for k_i and c_j, denoted $P(k_i,c_j)$, $R(k_i,c_j)$ and $F(k_i,c_j)$ respectively, are computed by formulas 3, 4 e 5.

$$(3) \qquad P(k_i,c_j) = \frac{n_{ij}}{|c_j|}$$

$$(4) \qquad R(k_i,c_j) = \frac{n_{ij}}{|k_i|}$$

$$(5) \qquad F(k_i,c_j) = \frac{2 * R(k_i,c_j) * P(k_i,c_j)}{R(k_i,c_j) + P(k_i,c_j)}$$

Precision means the number of sentences of cluster c_j which belong to the class k_i, thus measuring the homogeneity of cluster c_j with respect to class k_i. Similarly, Recall indicates the portion of sentences from class k_i that are present in cluster c_j, thus measuring how complete cluster c_j is with respect to class k_i.

Intuitively, $F(k_i,c_j)$ measures the quality of cluster c_j in describing the class k_i, by calculating the harmonic mean between Recall and Precision of cluster c_j regarding class k_i. The F-measure for each class over the entire data set is based on the cluster that best describes each class k_i, i.e., the one that maximizes $F(k_i,c_j)$ for all j. Thus, the overall F-measure of a clustering solution S, denoted $F(S)$, is calculated by using the weighted sum of such maximum F-measures for all classes, according to Formula 6.

$$(6) \qquad F(S) = \sum_{k_i \in K} \frac{|k_i|}{N} \max_{c_j \in C} \{F(k_i,c_j)\}$$

$F(S)$ values range from 0 to 1, in which a larger value indicates a higher accuracy of a clustering solution.

The second metric employed is Entropy (Steinback et al., 2000). It measures how well each cluster is organized, i.e., how the various classes of sentences are distributed in each cluster. A perfect clustering solution will be the one in which all its clusters contain sentences from a single class only. In this case the Entropy is zero.

The calculation of Entropy is based on the class distributions of each cluster. This is exactly what is done by Precision metric. In fact, Precision represents the probability of a sentence chosen randomly from cluster c_j to belong to class k_i. Hence, the Entropy of a cluster c_j, denoted $E(c_j)$, can be calculated by Formula 7.

$$(7) \qquad E(c_j) = -\sum_{k_i} P(k_i,c_j) \log P(k_i,c_j)$$

The Entropy of a whole clustering solution S, denoted $E(S)$, is given by the sum of the individual cluster entropies weighted by the size of the cluster as Formula 8 shows. $E(S)$ values are ≥ 0. The smaller the $E(S)$, the better the clustering solution is.

$$(8) \qquad E(S) = \sum_{c_j} \frac{|c_j|}{N} E(c_i)$$

The third metric used was Purity (Rosell et al., 2004), which indicates the percentual of a given cluster that the largest class of sentences assigned to it represents (i.e., the majority class). In other words, Purity represents the largest class distribution of a given cluster. Thus, the Purity of cluster c_j, denoted $P(c_j)$, is defined by class k_i that maximizes the Precision of that cluster, as Formula 9 shows.

$$(9) \qquad P(c_j) = max\ k_i \{P(k_i,c_j)\}$$

The overall Purity of a clustering solution, denoted $P(S)$, is obtained as a weighted sum of the individual cluster purities and is given by Formula 10.

$$(10) \qquad P(S) = \sum_{c_j \in C} \frac{|c_i|}{N} P(c_j)$$

$P(S)$ is a value in the range of 0 to 1. The larger the value of purity, the better the clustering solution is.

It is interesting to note that the Entropy and Purity metrics measure the goodness of a clustering solution, while F-measure represents the effectiveness of the clustering method. In next section we present the goodness and effectiveness results for SiSPI.

4.3 Experimental Results

Two parameters are relevant to determine the success of SiSPI: centroid size and similarity threshold. The first one is the number of words that better represent a cluster and it is used to measure the similarity between a cluster and a candidate sentence to be added to that cluster. The second one is the similarity threshold which determines if a sentence should be added to an existing cluster or if a new cluster should be created.

In order to assess the influence of the centroid size in the system performance, a first experiment has been performed with four different configurations of centroids: 5, 10, 15 and 20 words. For this experiment, a similarity threshold of 0.4 (empirically determined) has been used. The average values of each measure presented in the previous section obtained for each configuration for all document collections are depicted in Table 1.

Table 1: Average results obtained for the 20 document collections by four different centroid configurations

Centroid size in words	Entropy	F-measure	Purity
5	0.101	0.860	0.917
10	0.106	0.863	0.912
15	0.101	0.864	0.913
20	0.106	0.863	0.913

In general, there is not significant difference among the results obtained with one or other configuration. Moreover, the best values for each metric differ from the configuration used. For instance, the best values for Entropy have been obtained for 5 and 15 words, while the best value for F-measure has been obtained for 15 words and the best value for Purity has been obtained for 5 words. Therefore, we can say that a 5-word centroid is the best configuration regarding Entropy and Purity metrics, while a 15-word centroid is the best configuration with respect to Entropy and F-measure metrics. As F-measure is a more complete metric than Purity (Purity only measures the homogeneity of a clustering solution and does not address the question of whether all elements of a given class are present in a single cluster) we have preferred to use the configuration with the highest F-measure instead of the highest Purity. So, in the following experiments we have used a 15-word centroid. This value is close to the one employed in similar tasks, for instance, document clustering (Radev et al., 1999), whose experiments show that a 10-word centroid is enough to give a clear idea of what each cluster is about.

Aiming at identifying the similarity threshold that best describes the evaluation corpus, SiSPI has been assessed with several different threshold configurations that range from 0.1 to 1. The average values obtained for each configuration of all 20 document collections are shown in Table 2.

Table 2: Average results obtained for the 20 document collections for different similarity threshold configurations

Similarity threshold	**Entropy**	**F-measure**	**Purity**
0.1	1.759	0.348	0.315
0.2	0.900	0.603	0.564
0.3	0.319	0.805	0.804
0.4	0.101	0.864	0.913
0.5	0.043	0.873	0.988
0.6	0.013	0.843	0.950
0.7	0.004	0.828	0.954
0.8	0.003	0.830	1.000
0.9	0.002	0.798	0.952
1.0	0.002	0.786	0.951

The Entropy values improve in a considerably way as the threshold increases. This also happens with F-measure and Purity values, but until a given point. F-measure values achieve its maximum at a threshold of 0.5 and then decreases smoothly, while Purity values increase until a threshold of 0.5 and then become unstable. These variations can be better seen in the chart of Figure 4.

Specifically regarding Entropy and Purity values, these can be justified by the fact that whereas the threshold increases, the number of clusters also grows (see Figure 5) in a way that they become more homogeneous (i.e., the variety of classes in each cluster tend to decrease). Moreover, as in the evaluation corpus there are many non-similar sentences, the tendency is that these values increase even more, once many clusters contain only one sentence. With respect to F-measure, we believe that in spite

of the cluster tendency to become more homogenous (increasing the precision), as the threshold increases, it becomes harder to identify the similarity among those sentences that are similar in meaning (semantic equivalence) but different lexically, as the case of paraphrases. Hence, the recall values tend to decrease.

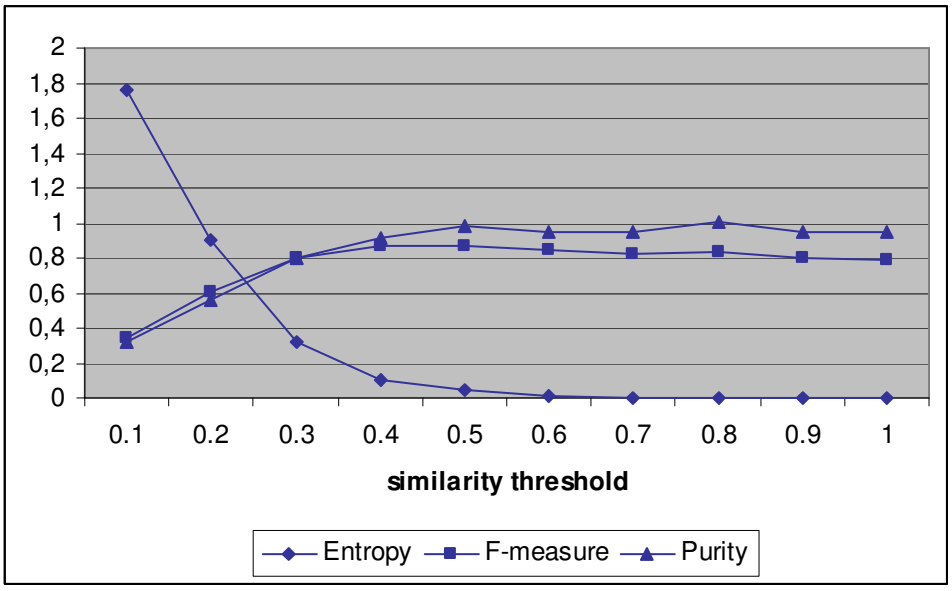


Figure 4: Average results obtained for each measure with different similarity thresholds

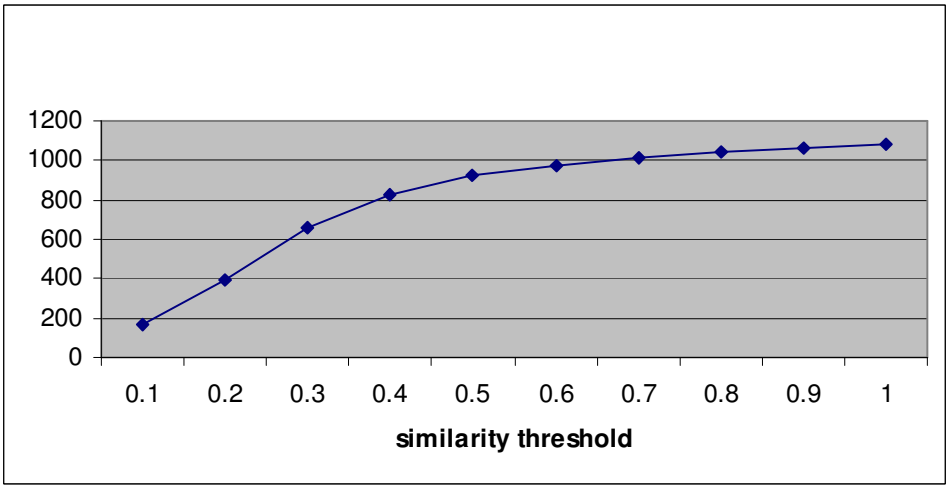


Figure 5: Total number of clusters obtained for each similarity threshold

5. Final Remarks

This report has presented SiSPI, a sentence clustering system which for our best knowledge is the first system proposed to Brazilian Portuguese.

SiSPI is domain independent and, in spite of being developed to treat documents written in Portuguese, its statistical clustering approach allows it to be easily extended to other languages.

We have also presented a preliminary evaluation of the proposed system to the news domain, in which it has obtained a satisfactory performance. Regarding other

existing methods for English language, for instance, which employ more complex techniques and make use of deep linguistic information, SiSPI performance measured in terms of F-measure was higher than those achieved for those works (as presented in Section 2). Despite its good performance, the system can have been penalized in cases which there were more than one possible class to the same sentence and the system choice was different from the human decision, since in SiSPI each sentence belongs to only one cluster. A more careful analysis is necessary to verify the influence of these factors in the system performance.

References

Barzilay, R. and McKeown, K. (2005). Sentence Fusion for Multi-document News Summarization. *Computational Linguistics*, Vol. 31, nº 3, pp. 297-327.

Caldas Junior, J.; Imamura, C.Y.M.; Rezende, S.O. (2001). Avaliação de um Algoritmo de Stemming para a Língua Portuguesa. In the *Proceedings of the 2nd Congress of Logic Applied to Technology*, Vol. 2, pp. 267-274.

Larocca Neto, J.; Santos, A.D.; Kaestner, C.A.A.; Freitas, A.A. (2000). Document Clustering and Text Summarization. *In Proceedings of the 4th International Conference Practical Applications of Knowledge Discovery and Data Mining – PAAD' 2000*, pp. 41-55.

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the 5-th Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley, University of California Press, Vol. 1, pp. 281-297.

Fung, B.C.M.; Wang, K.; Ester, M. (2003). Hierarchical Document Clustering using Frequent Itemsets. In *Proceedings of the SIAM International Conference on Data Mining*.

Hatzivassiloglou, V.; Klavans, J. L.; Eskin, E. (1999). Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning. In *Proceedings of Empirical Methods in Natural Language Processing and Very Large Corpora – EMNLP' 99,* pp. 203-212.

Hatzivassiloglou, V.; Klavans, J.L.; Holcombe, M.L.; Barzilay, R.; Kan, M.; McKeown, K.R. (2001). SimFinder: A Flexible Clustering Tool for Summarization. In *Proceedings of the Workshop on Automatic Summarization at NAACL' 2001,* pp. 41-49.

Hu P.; He, T.; Ji, D. (2004). Chinese Text Summarization Based on Thematic Area Detection. In *Proceedings of the Workshop on Text Summarization Branches Out at ACL' 04*, pp. 112-119.

Klampanos, I.A.; Jose, J.M.; van Rijsbergen, C.J.K. (2006). Single-pass Clustering for Peer-to-Peer Information Retrieval: The Effect of Document Ordering. In *Proceedings of First International Conference on Scalable Information Systems – INFOSCALE' 2006*, Vol. 152, pp. 36-43.

Pardo, T.A.S. (2006). *SENTER: Um Segmentador Sentencial Automático para o Português do Brasil*. Série de Relatórios do NILC. NILC-TR-06-01. São Carlos-SP, Janeiro, 6p.

Radev, D.R.; Hatzivassiloglou, V.; McKeown, K.R. (1999). A Description of the CIDR System as Used for TDT-2. In *Proceedings of the DARPA Broadcast News Workshop*.

Rosell, M.; Kann, V.; Litton, J. (2004). Comparing Comparisons: Document Clustering Evaluation Using Two Manual Classifications. In *Sangal R, Bendre SM, eds.*

Proceedings of the International Conference on Natural Language Processing – ICON'2004. Allied Publishers Private Limited, pp. 207-216.

Salton, G.; Wong, A.; Yang, C.S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM,* Vol 18, nº 11, pp. 613-620.

Salton, G. (1989). *Automatic Text Processing. The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley.

Schaal, M.; Müller, R.M.; Brunzel, M.; Spiliopoulou, M. (2005). RELFIN - Topic Discovery for Ontology Enhancement and Annotation. In *The Semantic Web: Research and Applications.* Lecture Notes in Computer Science, Springer, Berlin, pp. 608-622.

Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy*. Freeman, London, UK.

Steinbach, M.; Karypis, G.; Kumar, V. (2000). A comparison of document clustering techniques. In *Proceedings of the International Conference on Knowledge Discovery & Data Mining - KDD'00*, 2000.

Tombros, A.; Jose, J.M., Ruthven, I. (2004). Clustering Top-Ranking Sentences for Information Access. In *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science, Springer, Berlin, pp. 523-528.

Van Rijsbergen, C.J. (1979). *Information Retrieval.* 2nd edition, Butterworths, Massachusetts.

Wang, J.; Zhou, S.; Hu, Y. (2003). Sentences Clustering Based on Automatic Summarization. In *Proceedings of the Second International Conference on Machine Learning and Cybernetics,* pp. 57-62.

Xie, N. and Liu, W. (2005). An Answer Fusion Model for Web-based Question Answering. In *Proceedings of the First International Conference on Semantics, Knowledge and Grid – SKG'2005*.

Ye, H. and Young, S. (2006). A Clustering Approach to Semantic Decoding. In *Proceedings of 9th International Conference on Spoken Language Processing – ICSLP*, Pittsburgh, PA, USA.