Some Experiments on Clustering Similar Sentences of Texts in Portuguese

Eloize Rossi Marques Seno and Maria das Graças Volpe Nunes

NILC-ICMC University of São Paulo
CP 668P, 13560-970 São Carlos – SP, Brazil
{eloize,gracan}@icmc.usp.br

Abstract. Identifying similar text passages plays an important role in many applications in NLP, such as paraphrase generation, automatic summarization, etc. This paper presents some experiments on detecting and clustering similar sentences of texts in Brazilian Portuguese. We propose an evalution framework based on an incremental and unsupervised clustering method which is combined with statistical similarity metrics to measure the semantic distance between sentences. Experiments show that this method is robust even to treat small data sets. It has achieved 86% and 93% of F-measure and Purity, respectively, and 0.037 of Entropy for the best case.

Keywords: Sentence Similarity, Sentence Clustering, Statistical Metrics.

1 Introduction

Identifying similar text passages plays an important role in many Natural Language Processing (NLP) applications, such as paraphrase generation [1], automatic summarization [4] [5] [6], ontology building [11], digital library systems [13], dialogue systems [15], etc. In this paper, we present experiments on identifying and clustering similar sentences from one or multiple documents written in Brazilian Portuguese. Sentence clustering is performed as a primary step towards aligning and fusing common information (e.g., paraphrases and synonyms) among semantically similar sentences.

We propose an evaluation framework named SiSPI – *Similar Short Passages Identifier*, which is based on an incremental and unsupervised clustering method. The incremental method is particularly appealing since it is not based on learning and, therefore, it does not require a great training data set.

In order to compute semantic distance between a sentence and a cluster, SiSPI implements three different statistical similarity measures. The first measure, called Word Overlap [16], is based on the total of words in common between a sentence and a cluster. The two latter are the well-known TF-IDF (Term Frequency Inverse *Document* Frequency) measure from Information Retrieval [10] and the TF-ISF (Term Frequency Inverse *Sentence* Frequency) measure [3], which is an adaptation of the TF-IDF (see Section 3).

Aiming at identifying sets of highly semantically-related sentences from a collection of documents, a key concept to SiSPI is the notion of similarity. In this study, we

A. Teixeira et al. (Eds.): PROPOR 2008, LNAI 5190, pp. 133–142, 2008.

follow Hatzivassiloglou et al.'s similarity definition [5], which has been proposed for the same task of detecting similar sentences. Thus, we regard two sentences as similar if they refer to the same object or event and i) the object either accomplishes the same action in both units, or ii) is the subject of the same description. Next, we present three sentences on the same event, the domestic bomb explosion, extracted from the experimental corpus (see Section 4)[1]. Despite all sentences refer to the same fact, sentences (a) and (b) focus on the explosion in Ministério Público, while sentence (c) focuses on the explosion in Secretaria de Estado da Fazenda. Therefore, only sentences (a) and (b) are considered similar.

(a) Uma bomba caseira foi atirada contra a sede do Ministério Público (MP).
(b) Uma bomba caseira foi jogada contra o prédio do Ministério Público, na capital do estado.
(c) Uma bomba caseira atingiu o prédio da Secretaria de Estado da Fazenda, localizado na avenida Rangel Pestana, ao lado do Poupatempo Sé.

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

2 Related Work

Various methods for detecting similar short passages (e.g. sentences and paragraphs) have been proposed in the literature recently. Most of them are based on machine learning techniques and rely on statistics of words in common [11] [15]. In general, they make use of the Salton et al.'s vector space model [10] and of some statistical similarity measure to identify similar passages. In [11], for example, the TF-IDF model, which is widely used for document clustering (e.g., [3] [8]) is combined with a non-hierarquical clustering algorithm in order to cluster sentences and paragraphs for ontology enhancement. No evaluation result for the clustering process in specific is presented by the authors.

Despite those works treat short passages, our concept of similarity is more restrict than the one used in those works. The concept of similarity used in this work is similar to the one used in Hatzivassiloglou et al. [5] (see Section 1). The differences rely on the fact that they utilize a supervised approach based on linguistic knowledge to classify paragraph pairs of documents written in English as similar or non-similar. More specifically, those authors make use of a rule induction method, called RIPPER, which combines 43 linguistics features. Such features include morphological, syntactic and semantic information. RIPPER has been trained with a corpus of 10.345 manually-classified paragraph pairs and obtained 45.6% F-measure. In a subsequent experiment, reported by [6], a log-linear regression model was 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 with RIPPER. In [6] an

[1] The sentences have been kept in Brazilian Portuguese in order to avoid noise in the translation.

experiment using a variation of the TF-IDF model which treats paragraphs rather than documents is also presented. By using the same data set used by RIPPER and by the regression model, such model has obtained 36.7% F-measure on average.

In spite of machine learning techniques being widely used, they usually require a great data set of similar passage instances, which is hard to obtain. Trying to solve this, we employ an incremental clustering method which does not require training. Our hypothesis is that with an incremental clustering approach it is possible to achieve satisfactory results even using statistical similarity metrics only.

3 The Clustering Framework

SiSPI is composed by two main processing modules named Sentence Splitting and Sentence Clustering (Figure 1). The former splits each document of a collection into sentences. The latter identifies and clusters similar sentences. During this process, SiSPI makes use of a stemmer [2] and a stoplist. The output is a set of sentence cluster files.

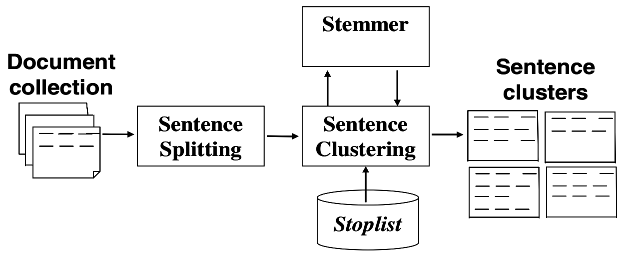

Fig. 1. SiSPI architecture

SiSPI is domain independent, for it is based only on lexical information. It is also weakly language-dependent, for it does not use any deeper linguistic knowledge (e.g., syntactic and semantic information).

The Sentence Splitting is performed by a textual-segmentation tool called SENTER [7], which is based on a list of abbreviations and some sentence delimiters. SiSPI could manage longer passages, as paragraphs, by just substituting this tool.

The Sentence Clustering module uses the incremental clustering method Single-pass [14], an effective and widely used algorithm for document clustering ([8]).

Single-pass requires a single sequential pass over the set of sentences to be clustered. The first cluster is created by selecting the first sentence of the first document. At each iteration, the algorithm decides on whether a new input sentence should be inserted in an existing cluster or should originate a new one. This decision is based on a condition specified by the similarity function employed, that is, a similarity threshold.

In this study, two different similarity functions are evaluated. The first one is based on the Word Overlap metric [16], which calculates the number of common words between a sentence S and a cluster C, normalized by the total of words of S plus C (Formula 1). According to (1), the similarity threshold is a value that ranges from 0 to 0.5, which is derived experimentally (see Section 4). The larger the similarity value, the more similar the sentence and that cluster are. Notice that in SiSPI each sentence belongs to a single cluster.

$$Wol\ (S,C) = \#\text{CommonWords}(S,C) \,/\, (|S| + |C|). \tag{1}$$

The second similarity function is the cosine coefficient [10], which is applied to the term frequency vector of a sentence and to the vector that represents the most important terms of a cluster, named centroid. According to this function, the similarity threshold is a value in the range of 0 to 1. The larger the similarity value between the vectors, the more similar the sentence and the cluster are.

The determination of a cluster centroid is based on the relevance of the corresponding words of that cluster, computed by two different metrics. The first metric is a slightly modified version of TF-IDF (Term Frequency Inverse Document Frequency) [10]. The TF-IDF value of a word *w* of a cluster *c*, denoted *TF-IDF(w,c),* is given by Formula 2.

$$TF\text{-}IDF(w,c) = TF(w,c) * IDF(w). \tag{2}$$

where *TF(w,c)* depictes the number of times the word *w* occurs in cluster *c, i.e.,* the frequency of *w* in *c*. The higher the *TF value*, the more representative the word *w* is of cluster *c*. The inverse document frequency of a word *w*, denoted *IDF(w)*, is given by Formula 3, where *C* is the total of sentences of the collection and *DF(w)* is the sentence frequency of the collection in which *w* occurs.

$$IDF(w) = 1 + \log\,(|C| \,/\, DF(w)). \tag{3}$$

According to (3), the *IDF* value is high if the word *w* occurs in few sentences of a collection, meaning that *w* has a great document-discriminating power. On the other hand, the *IDF* value is low if the word *w* occurs in many sentences of the collection, indicating that *w* has a little document-discriminating power.

The second metric used is TF-ISF (Term Frequency Inverse Sentence Frequency) [3]. The TF-ISF measure is similar to (1), but we compute the inverse sentence frequency for a specific cluster rather than for the document collection. The inverse sentence frequency of a word *w*, denoted *ISF(w)*, is given by Formula 4, where *C* is the total number of sentences in the current cluster, and *SF(w)* is the sentence frequency of the cluster in which *w* occurs.

$$ISF(w) = 1 + \log\,(|C| \,/\, SF(w)). \tag{4}$$

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

4 Experimental Evaluation

External or internal quality measures can be used to assess the quality of a clustering solution [12]. External quality measures evaluate how good the clusters are when compared with reference clusters (often manually classified clusters). So, this kind of evaluation can be carried out only if the class of each sentence is determined a priori. On the other hand, internal quality measures do not use any kind of 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 appropriate. In this study, we use three external quality measures that are described in Section 4.2. Next, we describe the corpus used for the evaluation.

4.1 The Corpus

The corpus is composed by 20 collections of news articles, with 3.6 documents on average on the same topic per collection (one example of topic is the Virginia Tech massacre). This corpus has been manually collected from several web news agencies and totalizes 1.153 sentences in 71 documents.

Aiming at creating a reference clustering corpus, each sentence of each document collection has been manually classified (i.e. associated with a cluster name) by the first author of this work, according to the similarity definition presented in Section 1. In cases when there were more than one possible cluster for a single sentence, only one has been chosen. Decisions about the best cluster 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 as *classes* and automatic clustering as *clusters*.

4.2 The Evaluation Measures

The accuracy of the produced clustering solution has been assessed by using the well-known Precision and Recall metrics, redefined in the cluster domain (see [4] and [12]).

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 and Recall for k_i and c_j, denoted $P(k_i,c_j)$ and $R(k_i,c_j)$, respectively, are computed by formulas 5 and 6. Precision is given by the number of sentences of cluster c_j that belong to the class k_i, thus measuring the homogeneity of cluster c_j with respect to class k_i. Similarly, Recall is given by the number of sentences of class k_i that are present in cluster c_j, thus measuring how complete cluster c_j is with respect to class k_i. We also measure 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 (Formula 7). This is also known as F-measure.

$$P(k_i,c_j) = \quad n_{ij} / |c_j|. \tag{5}$$

$$R(k_i,c_j) = n_{ij} / |k_j|. \tag{6}$$

$$F(k_i,c_j) = (2*R(k_i,c_j)* P(k_i,c_j)) / (R(k_i,c_j)+P(k_i,c_j)). \tag{7}$$

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 8. $F(S)$ values range from 0 (worse) to 1 (best).

$$F(S) = \sum_{k_i \in K} \frac{|k_i|}{N} \max_{c_j \in C} \{F(k_i,c_j)\}. \tag{8}$$

The second metric employed is Entropy [12]. 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 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 in each cluster. This is exactly what is done by Precision metric. In fact, Precision represents the probability of a sentence randomly chosen 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 9.

$$E(c_j) = -\sum_{k_i} P(k_i,c_j) \log P(k_i,c_j). \tag{9}$$

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, (Formula 10). $E(S)$ values are always positive. The smaller the $E(S)$, the better the clustering solution is.

$$E(S) = \sum_{c_j} \frac{|c_j|}{N} E(c_i). \tag{10}$$

The third metric used is Purity [9], which is given by the percentual of the most frequent class of a given cluster. Thus, the Purity of a cluster c_j, denoted $P(c_j)$, is defined by the class k_i that maximizes the Precision of that cluster (Formula 11).

$$P(c_j) = \max_{k_i} \{P(k_i,c_j)\}. \tag{11}$$

The overall Purity of a clustering solution, denoted $P(S)$, is given by a weighted sum of the individual cluster purities (Formula 12). $P(S)$ values range from 0 (worse) to 1 (best).

$$P(S) = \sum_{c_j \in C} \frac{|c_i|}{N} P(c_j). \tag{12}$$

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

4.3 Experimental Results

Regarding TF-IDF and TF-ISF models, two parameters are relevant for evaluating a clustering solution: the centroid size and the similarity threshold. The first one is used to measure the similarity between a cluster and a candidate sentence to be added to it. The second one plays the role of a similarity limit, indicating when a sentence originates a new cluster.

The first experiment was carried out 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 obtained for each assessment measure for all collections are depicted in Table 1. The purpose of this experiment was to identify the centroid configuration that best describes our data set for each similarity measure.

Table 1. Average results obtained for TF-IDF and TF-ISF with 4 different centroid sizes

Centroid size in words	TF-IDF			TF-ISF		
	Entropy	**F-measure**	**Purity**	**Entropy**	**F-measure**	**Purity**
5	**0.035**	0.860	**0.941**	**0.101**	0.860	**0.917**
10	0.037	0.860	0.939	0.106	0.863	0.912
15	0.036	**0.862**	0.940	**0.101**	**0.864**	0.913
20	0.042	**0.862**	0.938	0.106	0.863	0.913

In general, the difference between the results of all configurations for both models is little. Regarding effectiveness (i.e. F-measure), the TF-IDF best performance was achieved using a 15 and a 20-word centroid, while the TF-ISF best performance was achieved using a 15-word centroid. However, regarding cluster goodness (measure in terms of Entropy and Purity), a 5-word centroid was the best configuration for both cases (except for TF-ISF whose Entropy values were the same for both configurations).

As F-measure is more complete than Entropy and Purity (those do not address the question of whether all elements of a given class are present in a single cluster), we preferred to use the configuration with the highest F-measure instead of the highest Entropy and Purity values. So, in the following experiments we have used a 15-word centroid. This value is close to the one used in document clustering, whose experiments show a 10-word centroid is enough to give a clear idea of what each cluster is about [8].

To identify the best similarity threshold, each similarity model has been assessed with several different threshold configurations that range from 0.1 to 1 (except Word Overlap that ranges from 0.1 to 0.5). The average values for all collections are shown in Table 2.

According to Table 2, in all cases, the Entropy values improve in a considerably way as the threshold increases. This also happens with F-measure and Purity values, but up to a given point, from which those values decrease smoothly. F-measure achieves its maximum at a threshold of 0.2, 0.3 and 0.4 for Word Overlap, TF-IDF and TF-ISF, respectively. Regarding Purity, the values increase until a similarity of 0.3 for Word Overlap, and of 0.5 for TF-IDF and TF-ISF models.

Table 2. Average results obtained for each similarity measure with different thresholds

	Similarity	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
TF-IDF	Entropy	0.843	0.287	0.096	0.037	0.016	0.005	0.004	0.003	0.002	0.001
	F-measure	0.603	0.814	**0.886**	**0.860**	0.841	0.828	0.812	0.799	0.775	0.736
	Purity	0.549	0.808	0.907	**0.934**	**0.945**	0.945	0.942	0.940	0.941	0.938
TF-ISF	Entropy	1.759	0.900	0.319	0.101	0.043	0.013	0.004	0.003	0.002	0.002
	F-measure	0.348	0.603	0.805	**0.864**	0.856	0.843	0.828	0.813	0.798	0.786
	Purity	0.315	0.564	0.804	0.913	**1.000**	0.950	0.954	0.953	0.952	0.951
Word Overlap	Entropy	0.572	0.079	0.010	0.000	0.001	-	-	-	-	-
	F-measure	0.695	**0.860**	0.838	0.809	0.786	-	-	-	-	-
	Purity	0.654	0.908	**0.946**	0.943	0.941	-	-	-	-	-

Specifically regarding Entropy and Purity values, they can be explained by the fact that whereas the threshold increases, the number of clusters also grows in a way that they become more homogeneous, i.e., the variety of classes in each cluster tend to decrease. Moreover, since the corpus contains many non-similar sentences, it is expected that those values increase even more, once many clusters contain only one sentence. With respect to F-measure, in spite of the cluster tendency to become more homogenous (increasing the precision), as the threshold increases, it becomes harder to identify those sentences that are semantically equivalent but lexically different (e.g. paraphrases). Hence, the recall values tend to decrease, damaging the model performance.

In terms of providing both good performance and cluster goodness, the TF-IDF model with a similarity of 0.4[2] (here TF-IDF-0.4), performed as the most appropriate for our purpose. Besides TF-IDF-0.4 has achieved a F-measure of 86.0% (the best F-measure was 88.6% (TF-IDF-0.3)), its Entropy and Purity values are good, mainly if they were compared with those obtained for TF-IDF-0.3, TF-ISF-0.4 and Word-Overlap-0.2. Moreover, the standard deviation obtained for TF-ISF-0.4 (0.07 for F-measure, 0.06 for Purity and 0.05 for Entropy) was smaller than that obtained for TF-IDF-0.3 (0.08 for F-measure, 0.07 for Purity and 0.10 for Entropy), TF-ISF-0.4 (i.e. 0.09 for F-measure, 0.08 for Purity e 0.09 Entropy) and Word Overlap (0.08 for F-measure, 0.06 for Purity and 0.07 Entropy). Figure 2 shows an example of sentence cluster built by using TF-IDF-0.4. According to the human classification, this cluster consists of 4 sentences and SiSPI found 3 of them (therefore, 85% F-measure, 100% Purity and 0 Entropy for this specific cluster).

[1] A polícia informou que o grupo já desviou R$ 70 milhões, desde 2004.
[2] O grupo é acusado de lesar os cofres públicos em cerca de R$ 70 milhões.
[3] Segundo divulgado pela PF, o grupo criminoso desviou desde 2004 cerca de R$ 70 milhões dos cofres públicos, por meio do pagamento de serviços, compras e obras superfaturadas.

Fig. 2. Example of a cluster generated by SiSPI with TF-IDF-0.4 version

[2] Coincidentally, this value is equal to the empirical value used in the first experiment.

5 Conclusions

We presented experiments using SiSPI, a sentence clustering framework which, for our best knowledge, is the first one proposed for Portuguese. SiSPI is domain independent and may be easily customized to other languages. Moreover, it can treat other similarity definitions just by adjusting the similarity threshold.

SiSPI's incremental clustering approach makes it robust even to treat small data sets. We believe that such approach will allow SiSPI to manage larger corpora with similar performance to that achieved using small corpus. Performance gains should be obtained by making use of, for instance, a synonym and/or paraphrase set, what may be useful to identify sentences with a lot of paraphrases.

Acknowledgements. We thank CNPq for financial support.

References

1. Barzilay, R., McKeown, K.: Sentence Fusion for Multi-document News Summarization. Computational Linguistics 31(3), 297–327 (2005)
2. Caldas Junior, J., Imamura, C.Y.M., Rezende, S.O.: Avaliação de um Algoritmo de Stemming para a Língua Portuguesa. In: 2nd Congress of Logic Applied to Technology, vol. 2, pp. 267–274 (2001)
3. Larocca Neto, J., Santos, A.D., Kaestner, C.A.A., Freitas, A.A.: Document Clustering and Text Summarization. In: 4th International Conference Practical Applications of Knowledge Discovery and Data Mining – PAAD 2000, pp. 41–55 (2000)
4. Fung, B.C.M., Wang, K., Ester, M.: Hierarchical Document Clustering using Frequent Itemsets. In: Barbará, D., Kamath, C. (eds.) 3rd SIAM International Conference on Data Mining, pp. 59–70 (2003)
5. Hatzivassiloglou, V., Klavans, J.L., Eskin, E.: Detecting Text Similarity over Short Passages: Exploring Linguistic Feature Combinations via Machine Learning. In: Empirical Methods in Natural Language Processing and Very Large Corpora – EMNL 1999, pp. 203–212 (1999)
6. Hatzivassiloglou, V., Klavans, J.L., Holcombe, M.L., Barzilay, R., Kan, M., McKeown, K.R.: SimFinder: A Flexible Clustering Tool for Summarization. In: Workshop on Automatic Summarization at NAACL 2001, pp. 41–49 (2001)
7. Pardo, T.A.S.: SENTER: Um Segmentador Sentencial Automático para o Português do Brasil. Technical Report NILC-TR-06-01, São Carlos-SP, Brazil, 6p (2006)
8. Radev, D.R., Hatzivassiloglou, V., McKeown, K.R.: A Description of the CIDR System as Used for TDT-2. In: DARPA Broadcast News Workshop (1999)
9. Rosell, M., Kann, V., Litton, J.: Comparing Comparisons: Document Clustering Evaluation Using Two Manual Classifications. In: Sangal, R., Bendre, S.M. (eds.) International Conference on Natural Language Processing, pp. 207–216. Allied Publishers Private Limited (2004)
10. Salton, G., Allan, J.: Text Retrieval Using the Vector Processing Model. In: 3rd Symposium on Document Analysis and Information Retrieval. In: 3rd Symposium on Document Analysis and Information Retrieval. University of Nevada, Las Vegas (1994)
11. Schaal, M., Müller, R.M., Brunzel, M., Spiliopoulou, M.: RELFIN - Topic Discovery for Ontology Enhancement and Annotation. In: The Semantic Web: Research and Applications. LNCS, pp. 608–622. Springer, Berlin (2005)

12. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: International Conference on Knowledge Discovery & Data Mining - KDD 2000 (2000)
13. Tombros, A., Jose, J.M., Ruthven, I.: Clustering Top-Ranking Sentences for Information Access. In: Research and Advanced Technology for Digital Libraries. LNCS, pp. 523–528. Springer, Berlin (2004)
14. Van Rijsbergen, C.J.: Information Retrieval, 2nd edn. Butterworths, Massachusetts (1979)
15. Ye, H., Young, S.: A Clustering Approach to Semantic Decoding. In: 9th International Conference on Spoken Language Processing – ICSLP, Pittsburgh, PA, USA (2006)
16. Radev, D., Otterbacher, J.: Zhang, Zhu.: Cross-document Relationship Classification for Text Summarization. In: Computational Linguistics (to appear, 2008)