KnCr: A Short-Text Narrow-Domain Sub-Corpus of Medline*

[1,2]David Pinto & [1]Paolo Rosso
[1]Department of Information Systems and Computation
Polytechnic University of Valencia
Camino de Vera s/n, 46022, Valencia, Spain
{dpinto, prosso}@dsic.upv.es

[2]Faculty of Computer Science, BUAP
14 Sur & Av. San Claudio (CU), 72570, Puebla, Mexico
dpinto@cs.buap.mx

Abstract

Clustering of short texts in narrow domains is one of the most difficult tasks due to the high overlapping of vocabularies among the texts and also to the specific terminology used by researchers. Here, we are presenting a new corpus of scientific texts in medicine domain, specifically about "Cancer" topics. This corpus is a subset of the last MEDLINE sample, made up of 900 abstracts of 16 different categories. This compilation is provided as a dataset for the evaluation of algorithms in this area. Preliminary experiments carried out with this corpus highlight its difficulty and reinforce the hypothesis of using it in this challenging new task.

1. Introduction

Clustering is the most important unsupervised learning problem, due to its wide real possible applications. The goal of this task is to determine the intrinsic grouping in a set of unlabeled data. Nowadays there exist datasets widely used in classification, a clustering close-related task, like Reuters [1] and 20 Newsgroups [1].

Currently, Reuters is the most widely used test collection for text categorization research. The data was originally collected and labeled by Carnegie Group, Inc. and Reuters, Ltd. in the course of developing the CONSTRUE text categorization system. The most known version of Reuters used nowadays is Reuters-21578 [1], which contains four set of categories (EXCHANGES, ORGS, PEOPLE, and PLACES), where each one corresponds to named entities of the specified type. Differents subsets of Reuters-21578 have been constructed (ModApte, R10, R90, etc), but no one can be considered suitable for experiments in clustering short texts of narrow domain. On the other hand, 20 Newsgroups is a collection of 20.000 messages, collected from UseNet postings over a period of several months in 1993. The data are divided almost evenly among 20 different UseNet discussion groups and, therefore, is quite far to be a narrow domain corpus.

Clustering short texts of narrow domain task is a new challenging task that has been attended in just a few papers. For instance, in [2] Makaganov et al. presented simple procedures for clustering feature selection using two narrow-domain corpora. The first collection was made up of 48 abstracts from the computational linguistics and intelligent text processing conference (CICLing 2002)[2], whereas the second was composed by 200 abstracts from the international federation of classificaton societies conference (IFCS-2000)[3]. The first corpus was also used by Alexandrov et al. [3], Jiménez et al. [4], and Pinto et al. [5] for their experiments in clustering abstracts in a narrow domain. However, this collection is very small and, therefore, the results obtained may be imprecise when a cross-validation evaluation is not used [6]. Recently, in [5] another short-text corpus in the particles physics domain was used for experiments in clustering; the size of it was approximately 2.000 abstracts, but the distribution of the topics was very unbalanced. The clustering of these kind of corpora implies a very big challenge if a correct performance measure is applied, because identifying a class with one element is more difficult than identifying another one

*This work was partially supported by the R2D2 (CICYT TIC2003-07158-C04-03) and ICT EU-India (ALA/95/23/2003/077-054) research projects, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant.

[1] http://people.csail.mit.edu/jrennie/20Newsgroups/

[2] http://www.cicling.org

[3] http://www.Classification-Society.org

with many elements. Moreover, in real situations this kind of corpora are very difficult to be found. From this viewpoint, we are interested in a moderate-sized and balanced corpus and, therefore, the aim of this work consists in gathering abstracts from a high quality source for constructing a balanced corpus suitable for experiments in clustering short texts of the narrow cancer domain. We have selected MEDLINE for extracting those documents that are related with the cancer topics. In this way, we have structured this paper for explaining the characteristics of this new corpus and the hardness of clustering the documents inside it. Section 2 presents a brief introduction of the MEDLINE repository. In Section 3 we describe the composition of the KnCr corpus. Moreover, a set of experiments carried out in order to determine the hardness of clustering the content this new corpus are shown. Finally, a discussion is presented.

2. The MEDLINE repository

The National Library of Medicine (NLM) collects materials in all areas of biomedicine and health care, as well as works on biomedical aspects of technology, the humanities, and the physical, life, and social sciences. The collections stand at more than 8 million items–books, journals, technical reports, manuscripts, microfilms, photographs and images. NLM is a national resource for all U.S. health science libraries through a National Network of Libraries of Medicine.

Althought the last annual statistical profile of NLM, given in September 2005, stands this collection in 606.000 articles indexed from 4.900 journals for MEDLINE, the access to the complete collection is not free available for all people; MEDLINE data is licensed by the NLM at low cost to anyone who wants to make the information available to a user group. Moreover, a sample data for experiments is provided[4]; for instance, the last sample file "medsamp2006f.xml" is about 20,5MB.

The use of MEDLINE in literature is wide extended. Several works use this collection for different tasks (see http://www.nlm.nih.gov/bsd/licensee/reports/name.html). The last sample provided by NLM contains abstracts, texts, and sometimes just the title and authors from the medicine domain investigations and, therefore, in order to construct a short text narrow domain corpus, an analysis of such documents have to be done for selecting those that have both, abstract and keywords. The process for the construction of this new corpus is described in the next section.

3. The KnCr corpus

The absence of a specific forum for the evaluation of systems for the clustering short text narrow-domain task, has not allowed to create a good corpus for using it as a standard evaluation. We have done several experiments on constructing new narrow domains corpora, specifically in the medicine domain. Currently, we have constructed one, by downloading the last sample of documents provided by MEDLINE[5], which contains approximately 30.000 abstracts, and selecting those related with the "Cancer" domain. In the following subsections we will explain how we have created the gold standard for this new corpus.

3.1. Automatic gold standard generation

In order to correctly evaluate results of clustering, a corpus must be provided with a gold standard of the possible clustering classes distribution. Although the gold standard is normally constructed by humans, we tried to create it automatically.

Due to the fact that each retrieved abstract of our document set contains "keywords" provided by each author, we used them for constructing the gold standard for this collection. We selected three clustering methods for this experiment, two are already implemented in the Weka machine learning software [7]: Expectation Maximization (EM) and K-Means. The third clustering method is KStar [8]. We used the F-Measure [9] for comparing each pair of clustering methods. The formula used is described as follows:

Given a set of clusters $\{G_1, \ldots, G_m\}$ and a set of classes $\{C_1, \ldots, C_n\}$, the F-measure between a cluster i and a class j is given by the following formula.

$$F_{ij} = \frac{2 \cdot P_{ij} \cdot R_{ij}}{P_{ij} + R_{ij}}, \qquad (1)$$

where $1 \leq i \leq m$, $1 \leq j \leq n$. P_{ij} and R_{ij} are defined as follows:

$$P_{ij} = \frac{\text{Number of texts of cluster } i \text{ in class } j}{\text{Number of texts from cluster } i}, \qquad (2)$$

and

$$R_{ij} = \frac{\text{Number of texts of cluster } i \text{ in class } j}{\text{Number of texts in class } j}. \qquad (3)$$

The global performance of a clustering method is calculated by using the values of F_{ij}, the cardinality of the set of clusters obtained, and normalizing by the total number of

[4] http://www.nlm.nih.gov/bsd/sample_records_avail.html

[5] ftp://ftp.nlm.nih.gov/nlmdata/sample/medline/

documents in the collection ($|D|$). The obtained measure is named F-measure and it is shown in equation 4.

$$F = \sum_{1 \leq i \leq m} \frac{|G_i|}{|D|} \max_{1 \leq j \leq n} F_{ij}. \qquad (4)$$

The results obtained are presented in Table 1. None pair combination of clustering methods obtained more than 0,51 of F-Measure and it was not possible to determine a winner clustering method for constructing the gold standard. This first experiment has shown that clustering narrow-domain corpora is really a difficult task, eventhought we have available the keywords of each abstract.

Table 1. Results obtained by clustering abstract keywords (without gold standard)

	EM	KMeans	KStar
EM	–	0,51	0,45
KMeans	0,31	–	0,36
KStar	0,36	0,33	–

3.2. Manual gold standard generation

Once obtained the previous results, we had to do manual inspection for classifying every document in its correct class for constructing the gold standard. We used the ontology made available by the National Cancer Institute (NCI)[6], in order to construct the gold standard categories. This ontology describes a hierarchy of cancer terms based in the anatomy kind and specifies the fine grain categories of this domain (the current owl version of the NCI thesaurus can be found in http://www.mindswap.org/2003/CancerOntology/). Table 2 and 3 show the complete characteristics of this new cancer corpus. As can be seen, only 900 from 30.000 abstracts are related with the cancer topic, and the average length of each of them is about 126 words which makes it suitable for experiments in the task described before.

Once constructed the gold standard, we carried out some experiments to compare different methods of clustering against it, in order to investigate the hardness of clustering the texts that made up this corpus. We implemented two hierarchical clustering methods, namely Single and Complete Link Clustering (SLC, CLC) [10], and three agglomerative clustering methods (K-NN [11], KStar [8], NN1 [4]). The results obtained by clustering the abstracts instead of the keywords, and by using two well known vocabulary reduction techniques (Document Frequency-DF and Term Strength-TS) [12], are presented in Table 4. We can observe low F-measure values for each clustering method, which highlights again the hardness of this task.

Table 2. Distribution of *KnCr*

Category	# of abstracts
blood	64
bone	8
brain	14
breast	119
colon	51
genetic studies	66
genitals	160
liver	29
lung	99
lymphoma	30
renal	6
skin	31
stomach	12
therapy	169
thyroid	20
Other (XXX)	22
Total	**900**

Table 3. Other features of *KnCr*

Feature	Value
Size of the corpus (bytes)	834.212
Number of categories	16
Number of abstracts	900
Total number of terms	113.822
Vocabulary size (terms)	11.958
Terms average per abstract	126,47

Table 4. Results obtained by clustering abstracts: evaluation with the gold standard

	DF	TS
KStar	0,39	0,39
SLC	0,52	0,51
CLC	0,36	0,36
NN1	0,42	0,41
KNN	0,38	0,37

In order to verify whether the clustering by keywords, provided by abstract authors, behaves better than using the vocabulary reduction techniques presented above, we carried out a third experiment: in this case we compared the results obtained by clustering those keywords with EM, KMeans and KStar methods with the gold standard built manually. The results are presented in Table 5. We can see

[6] http://ncimeta.nci.nih.gov/

that using keywords instead of abstracts can lead to more confusion in the clustering short texts narrow-domain task. This may be due to the different viewpoints of scientific text author, and the few words added as keywords. That is, a little variation in the keyword set leads to classify similar documents as different. We consider that more investigation must be done in order to clearly determine this behaviour.

Table 5. Comparison against the gold standard of clustering abstracts keywords

	F-Measure
EM	0,20
KMeans	0,22
KStar	0,22

4. Discussion

Up to now, clustering very short texts of narrow domains has not received too much attention by the computational linguistic community and only few are the related works which can be found in literature. This could be derived from the high challenge that this problem implies, since the obtained results are very unstable or imprecise when clustering abstracts of scientific papers, technical reports, patents, etc. As a consequence, there exist a lackness of this type of corpora that led us to compile scientific abstracts from high quality sources. We have selected MEDLINE as a repository source for the construction of a new corpus in the cancer domain. Our corpus is a moderate sized one, with 900 abstracts and 16 different balanced categories.

In order to investigate the possible hardness of clustering this corpus, we have carried out a set of experiments. First we tried to construct automatically the gold standard by comparing three different clustering methods upon the use of the keywords of each abstract. Due to the difficulty to evaluate the goodness of the automatically obtained gold standard, we decided to obtain it manually. Moreover, we compared the results of clustering keywords against clustering abstracts (using a vocabulary reduction), and in this particular case we found that author keywords may confuse the clustering process. Further analysis should investigate this behaviour.

We have made free available this new corpus by email request to authors considering that this corpus, together with its gold standard, will allow to test algorithms for clustering very short texts of the cancer narrow domain.

References

[1] D. D. Lewis, Reuters-21578 text categorization test collection, Tech. rep., Distribution 1.0 (1997).

[2] P. Makagonov, M. Alexandrov, A. Gelbukh, Clustering Abstracts instead of Full Texts, in: Proceedings of the Seventh International Conference on Text, Speech and Dialogue (TSD 2004), Vol. 3206 of Lecture Notes in Artificial Intelligence, Springer-Verlag, Brno, Czech Republic, 2004, pp. 129–135.

[3] M. Alexandrov, A. Gelbukh, P. Rosso, An Approach to Clustering Abstracts, in: Proceedings of the 10th International Conference NLDB-05, Lecture Notes in Computer Science, Springer-Verlag, Alicante, Spain, 2005, pp. 8–13.

[4] H. Jiménez, D. Pinto, P. Rosso, Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos, Procesamiento del Lenguaje Natural 35 (1) (2005) 114–118.

[5] D. Pinto, H. Jiménez-Salazar, P. Rosso, Clustering abstracts of scientific texts using the transition point technique., in: A. F. Gelbukh (Ed.), CICLing, Vol. 3878 of Lecture Notes in Computer Science, Springer-Verlang, 2006, pp. 536–546.

[6] D. Pinto, P. Rosso, A. Juan, H. Jiménez-Salazar, A comparative study of clustering algorithms on narrow-domain abstracts, Procesamiento del Lenguaje Natural(To appear in 2006).

[7] I. H. Witten, E. Frank, Data mining: Practical machine learning tools and techniques with Java implementations, Morgan Kaufmann, 2000.

[8] K. Shin, S. Y. Han, Fast clustering algorithm for information organization., in: A. F. Gelbukh (Ed.), CICLing, Vol. 2588 of Lecture Notes in Computer Science, Springer-Verlang, 2003, pp. 619–622.

[9] C. J. V. Rijsbergen, Information Retrieval, 2nd edition, Dept. of Computer Science, University of Glasgow, 1979.

[10] S. C. Johnson, Hierarchical clustering schemes, Psychometrika (2) (1967) 241–254.

[11] E. Fix, J. L. Hodges, Discriminatory analysis: nonparametric discrimination: small sample performance, Tech. Rep. 11, USAF School of Aviation Medicine, Randolph Field, Texas, project No. 21-49-004 (1952).

[12] T. Liu, S. Liu, Z. Chen, W. Ma, An evaluation on feature selection for text clustering., in: T. Fawcett, N. Mishra (Eds.), ICML, AAAI Press, 2003, pp. 488–495.