Local Theme Detection and Annotation with Keywords for Narrow and Wide Domain Short Text Collections

Svetlana V. Popova
Saint-Petersburg State University
Faculty of applied mathematics and control processing
Saint-Petersburg, Russia
spbu@bk.ru

Ivan A. Khodyrev
Saint-Petersburg State Electro-technical University
Faculty of computer science and informatics
Saint-Petersburg, Russia
kivan.mih @gmail.com

***Abstract*—This paper presents a clustering approach for text collections and automatic detection of topic and keywords for clusters. Present research focuses on narrow domain short texts such as short news and scientific paper abstracts. We propose a term selection method, which helps to significantly improve hierarchic clustering quality, and also the automatic algorithm to annotate clusters with keywords and topic names. The results of clustering are good comparing with the results of other approaches and our algorithm also allows extracting keywords for each cluster, using the information about the size of a cluster and word frequencies in documents.**

Keywords-narrow domain short text clustering; automatic annotation; hierarchical clustering; Pearson correlation.

I. INTRODUCTION

In the presented paper, we are solving two main tasks: clustering and annotation tasks with keywords for small collections of short texts. We have chosen two types of collections for our tasks: first type collections contain texts from one narrow domain and second type collections contain texts from different domains. In our experiments, we are using collections, which are used for clustering in other papers [2][8][9][12]. We also observe that there is not much attention paid in literature in respect to annotation of narrow domain short texts for small collections.

Topics/trends detection and annotation is a popular theme today. Annotations help user to understand if a document or a group of documents is useful in respect to his goals or not without reading the full source. Annotations also help in a search process when user tries to find documents similar to some target document. New keywords appearance in sets of scientific articles could signify emerge of a new research domain or a new trend in present domains. The task of novelty detection is highly demanded today, but it is also a hard task to deal with. Main themes detection in news collections is related to topic detection and tracking domain (TDT) [4][5][15]. Keyword detection and annotation for document collections could be used in automated ontology's creation task.

The task of short text processing and analysis is emerged with the development of social networks. Today, the practical interest to analyze messages in blogs, forums, e-mails, sms is constantly growing [3][16]. There is a wide variation of tasks in this field: social analysis, opinion mining and sentiment analysis, searching for useful and redundant branches on forums, social network search engines etc. Electronic libraries also benefit from the research in the field of short texts, because it could help automating searching and sorting documents by using abstracts.

The importance to separate small collections could be defined as follows. Consider an analysis of text documents' collection with clustering goal. It leads to situation where from big collections small subgroups of texts are extracted, which need further processing. Analysis of these subgroups needs changes in text processing. Small sizes of texts and collections which contain them make word evaluation a hard task, because amount of data is very limited

We are basing annotation results of preceding clustering. So our first task was clustering. Short texts clustering is a task with high complexity [2][8][12]. In present paper, we propose clustering approach based on Pearson correlation coefficient [19] and special term selection technique.

As a clustering algorithm we are using one of the hierarchical clustering algorithms [7][18] and Pearson correlation as measure between texts. On term selection step not more than 10% of a collection's vocabulary left. Our research showed that quality of clustering is increased if words with high value of document frequency are used, with exception to some words with the highest document frequency. Obtained clustering results are relatively good comparing with the other methods [2][8][12]. Approach based on Pearson correlation measure seems productive and we are planning to test it with different clustering algorithms in the future. There is still unsolved question: how to determine the right number of clusters for hierarchical clustering algorithm.

Second task is annotation of given type of collections. In this paper, we consider only keyword annotation. Word's overlapping between clusters makes this task difficult. Choosing frequent words in some cluster as a keyword usually lead to situation where common word for the whole

 ISBN: 978-1-61208-175-5

collection is choosing which is not informative for cluster. From the other hand, setting a threshold for a words which appear outside of cluster, could lead to loss of semantically significant words. In present paper we propose novel algorithm which helps to deal with these problems.

The rest of the paper organized as follows. In section 2, we describe related work. In section 3, we present test collections and the measure, with which we could compare the results automatic and manual clustering. In section 4, proposed clustering algorithm, term selection method and keywords detection algorithm are described. Section 5 contains experimental results, and we make a conclusion in section 6.

II. RELATED WORK

Clustering of narrow domain short text collections was addressed in David Pinto's PhD and in [12]. Pinto tested a number of algorithms, similarity measures to compare documents and term selection techniques. Pinto suggests that it is possible to increase the clustering quality using self-term expansion before term selection. Idea of self-term expansion was further developed in [13]. In [11], weblog clustering task is solving using different topics detection inside documents with preceding self-term expansion. The best clustering results for narrow domain short texts were obtained in [2][8][9]. In [2], algorithm CLUDIPSO is introduced; it is based on discrete particle swarm optimization. It needs precise information about the number of clusters and some other parameters, which were calculated in [2] during experiments. However even for fixed parameters on the same date, the quality of CLUDIPSO's clustering result could vary. In [8], Ant-Tree-Silhouette-Attraction algorithm (*AntSA*) was introduced, which is based on AntTree algorithm and use some initial data partitions by using CLUDIPSO (*AntSA-CLU*). *AntSA-CLU* gives better results comparing to CLUDIPSO, but it also needs input parameters to be set and the result may vary from experiment to experiment as well. In [9], iterative method for short text clustering tasks (ITSA) was proposed. This method does not make clustering itself, but it integrates and refines results of arbitrary clustering algorithms and based on them generates final result.

In [2][8][9][12], authors show clustering results on narrow domain short texts using different algorithms: Single Link Clustering, Complete Link Clustering, K-Nearest Neighbour, K-Star and a modified version of the K-Star method (NN1), K-means, MajorClust, CHAMELEON, DBSCAN. Obtained results are relatively low for these algorithms. Algorithms which show the best results (CLUDIPSO, *AntSA-CLU*) do not show these results constantly on narrow domain collections with low topics differentiation. Clustering quality changes on each independent run for these algorithms and it could vary: it could be very good or it could be relatively low on different runs on the same data with the same input parameters. In practice such situation is usually does not satisfy user because when user receives bad results from some algorithm a number of times, he will most likely stop using it. So for presented work we have chosen hierarchical clustering algorithms, which give the same result for fixed number of clusters. We defined the term selection method and similarity measure between documents to reach results comparable with best clustering result of other algorithms. Also, to obtain stable results; we have made universal definition of input parameters for all test collections, which leads us to the problem of universal term selection.

III. TEST COLLECTIONS AND QUALITY VALUE

A. Collections

In present research, we used three collections with narrow domain short texts: CICling_2002 (this collection is recognized as one of the hardest for analysis), SEPLN_CICling and EasyAbstracts; one wide domain collection: Micro4News. All collections with "gold standards" and descriptions may be found [17]. Table I contains information about gold standard and vocabulary sizes of test collections. EasyAbstracts collection contains scientific abstracts on well differentiated topics. It could be considered as medium complexity. Collection for clustering CICling_2002 and SEPLN_CICling both contain narrow domain short abstracts and their complexity for analysis is relatively high. Micro4News contains short news and its documents are longer than in other collections, also its topics are well differentiated, so the complexity is relatively low. For each collection a golden standard exists, which is a result of classification by experts and it contains 4 groups for each

TABLE I. TEST COLLECTIONS INFORMATION

Test collections	**Collection's information**		
	Cluster's topics	***Vocabulary size***	***Vocabulary size after stop words filtering***
CICLing 2002	Linguistic, Ambiguity, Lexicon, Text Processing	953	942
SEPLN CICLing	Morphological – syntactic analysis, Categorization of documents, Corpus linguistics, Machine translation	1 169	1 159
Easy Abstracts	Machine Learning, Heuristics in Optimization, Automated reasoning, Autonomous intelligent agents	2 169	1 985
Micro 4News	Sci.med, soc.religion.christian, rec.autos, comp.os.ms-windows.misc	12 785	12 286

 ISBN: 978-1-61208-175-5

collection. Collections contain 48 texts each. For our experiments test collections were additionally parsed to remove stop words.

B. Quality Values

To test quality of clustering, we use measure based on F -measure [4], we will sign it as FM :

$$FM = \sum_i \frac{G_i}{|D|} \max_j F_{ij}, \text{ where } F_{ij} = \frac{2 \cdot P_{ij} \cdot R_{ij}}{P_{ij} + R_{ij}},$$

$$P_{ij} = \frac{|G_i \cap C_j|}{|G_i|}, \; R_{ij} = \frac{|G_i \cap C_j|}{|C_j|},$$

$G = \{G_i\}_{i=\overline{1,m}}$ is an obtained set of clusters, $C = \{C_j\}_{j=\overline{1,n}}$ is set of classes, defined by experts, D - number of documents in taken collection. We use FM as quality value in this paper.

IV. ALGORITHM DESCRIPTION

A. Pearson Correlation as a Metric for Clustering

We assumed that texts in the same subject have several features that could be measured.

- There exists a group of words which always occur together in texts of one thematic group.
- Some of these words occur often in each text of a subject, some words occur rarely in each text, but all these words could be found in significant number of texts.

These assumptions lead us to the idea that if two texts have words with the same frequency characteristics, then they are semantically close to each other. Relation between texts based on the mutual word frequencies could be expressed using correlation coefficient. In our research, we present texts as N - dimension vectors, where N is the number of selected words for text representation. In our research we used Pearson correlation coefficient between two texts as a similarity function. It is calculated using formula:

$$p_{x,y} = \frac{\sum_{i=1}^{N}(x_i - M_x)(y_i - M_y)}{(N-1)\sigma_x \sigma_y},$$

where N – is a number of clustering space dimensions; x_i, y_i are values of paired variables: frequencies of a word i in document x and in document y ; M_x, M_y are values for x and y which represent average frequencies of all words in document x and y ; σ_x, σ_y - standard deviation for documents x and y .

Consider two texts test_1 and text_2 and let these texts be represented by the same set of 20 words. Consider a 2-dimension plot where horizontal and vertical axis contain frequencies of words occurrence in each of two texts. Each

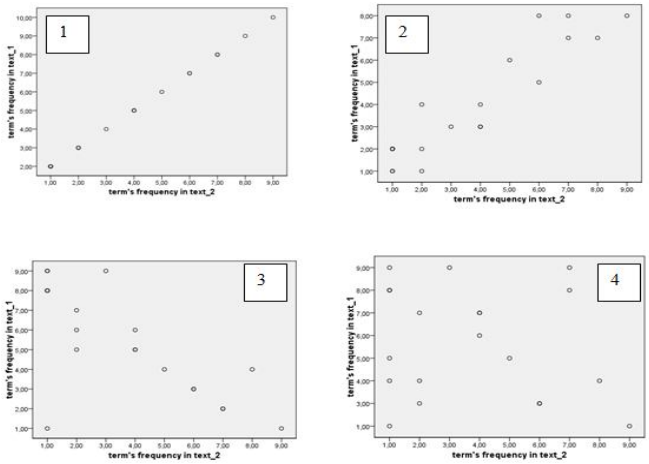

Figure 1. Pearson correlation (1: +1; 2: +0,926; 3: -0,722; 4: -0,192).

dot on such plot represents concrete word and it is placed according to frequencies in first and second texts. Four such plots are depicted in "Fig. 1". On the first plot each word of the first text occurs one more time than in the second text. In this case correlation coefficient between two texts is equals to 1. However in reality such relation is almost impossible. Second plot represents the positive relation between words: frequency characteristics of words for both texts are almost the same. But difference between frequencies of words in two texts is defined empirically and it couldn't be expressed as a function. In this case correlation coefficient is between 0 and 1. If the value of the correlation coefficient is close to 1 then more positive relation between frequencies of words in two texts is found. In the third plot, an example of negative relation is presented: if in the first text some word occurs often, then in another text this word occur rarely and vise versa. Value of correlation coefficient in this case will be from -1 to 0. On the fourth plot an example of a near zero correlation coefficient value is depicted: the relation between frequencies of words does not have significant ordered behavior.

Our research is based on the heuristic that the closer correlation coefficient between two texts is to 1, the semantically closer these texts are to each other.

Our usage of vectors as a representation for texts does not take into account the size of texts. We assume that average frequency to meet a word in text is proportional to the text size. If so, the size of text does not have much influence on correlation metric between two texts. Let we have two very similar documents d_1 and d_2 , where document d_2 is four times longer than d_1 . Let d_1 be represented by a vector (4,3,5) and document d_2 with vector (16,12,20). In this case, Pearson correlation between texts will be 1 anyway, which we interpret as semantic equivalence.

 ISBN: 978-1-61208-175-5

B. Hierarchical Clustering

We tried algorithms of hierarchic clustering such as *Between Groups Linkage* (UPGMA) [18], *Single Linkage and*, *Complete Linkage* [7]. Working scheme is the same for all of them. In the beginning each clustering object becomes a cluster. Then, on each step, two clusters with the most value of similarity between them are linked into one cluster. These steps are made until the given number of clusters is not reached. The difference between methods is in the choice of similarity function. In Single Linkage similarity between clusters is calculated as a similarity between two most similar objects in clusters. In Complete Linkage similarity between clusters is defined as a similarity between less similar objects in clusters. In Between Groups Linkage method, a mean value of similarity is calculated between each pair of objects from both clusters. Two clusters are linked if average distance between their objects is less than average distance between objects of other clusters.

Number of clusters for hierarchic clustering should be predefined and it seems like a significant disadvantage. We investigated if the result of clustering is relatively good in case the number of clusters was determined wrong. Our goal was to check which method suits the clustering task best, if the number of clusters differs from a golden standard. We calculated clustering quality with each method as an mean value of clustering results for 3-8 clusters. Experiments showed that single linkage gives bad results on all collections. We investigated if it's possible to increase clustering quality by additionally using term selection technique.

C. Terms Selection

In our research, a simple term selection method to reduce clustering space is used. Experiments showed that for Between Groups Linkage method, term selection technique, which filters words with low value of document frequency, increases the quality of clustering. Improvement of quality is observed until the number of selected terms reaches a value about 10% of initial collection vocabulary. If the number of selected words exceeds 10% limit, then clustering quality becomes worse. Our experiments also showed that filtering words with the highest values of document frequency improves clustering quality. So, we first selected about 10% of initial vocabulary terms and then from the obtained set we removed a small number of terms with the highest document frequency values. Combination of this technique with the Between Groups Linkage clustering gives best results. For Complete Linkage such term selection method could lead to further quality reduction. Based on our experiments we conclude that for narrow domain short text clustering a Between Groups Linkage method enhanced with the given term selection method is the most suitable.

D. Detection of Keywords

In our research, a simple term selection method to reduce clustering space is proposed.

After clustering was done the problem of keyword detection should be solved. We used an algorithm presented in listing in "Fig. 2" to deal with keywords. We are using three main assumptions to deal with keywords.

- If the word is semantically significant, then its occur frequency is low in most documents, but in some documents its occur frequency is high.
- If the word is significant for cluster, then it occurs in most documents of a cluster.
- If the word is significant for cluster, then the number of documents in which this word occurs, does not exceed much the size of a cluster.

```
Let D is a set of all collection's documents;
Let C is a set of clusters for annotation with keywords;
Let W is set of all words from vocabulary of collection after term selection step;
For every cluster c from C do {
    For every word w from W do {
        Let Q is a set of documents from D where occurrence number of word w
        is less then four;
        If |Q| < |c| {
            If more then |c|-α documents from c contain word w {
                Select w as a key word for cluster c;
            }
        }
    }
}
```

Figure 2. Listing of algorithm for keywords detection.

First and third rule allow filtering the commonly used words for a given collection. Second rule allows detecting words which are typical for a cluster. We defined α parameter to regulate the minimal number of documents in cluster in which a word should occur in order to be chosen as a keyword. Increasing α will reduce the number of clusters documents in which a word should be found and thus we obtain more keywords which less reflect clusters features.

V. PRESENTATION OF RESULTS

Results of our experiments are shown in Table II. For each collection we present such information: clustering quality evaluation using different number of predefined clusters (3-8); best and worst quality measure for each clustering method. This information is given for 3 cases: 1 – without initial term selection, 2 – 10% term selection, 3 – 10% term selection with filtering 3-4 terms with the highest document frequency. BGL stands for Between Groups Linkage and CL for Complete Linkage. In most cases best results are obtained for test collections with the number of clusters equal to 4, and sometimes with 3 or 5,

Using proposed algorithm we have reached good results of clustering for mentioned collections. We link this fact with the proposed combination of chosen similarity measure and term selection approach. We remove words that occur in a small number of texts and act as a noise. The description is as follows: let a word be occurring in a small number of documents. When texts are presented as N -

 ISBN: 978-1-61208-175-5

dimensional vectors, the part of vector representing a word will be like "0" in most cases and it does not affect much the correlation between texts. From the other hand there is plenty of words, which occur in a small number of texts. To leave about 10% of a collection's terms, it was enough to remove words, which occur only in 2-4 documents, most of which occur only in 1 or 2 documents. These words act as noise and they make clustering results worse. Whenever we remove 3-4 words with highest document frequencies, the actual removed words occur in half of documents, but their frequency is usually 1 (such words as: paper or based). These words act as noise and have negative influence on the result of clustering. *Between Groups Linkage* gives better results, than *Complete Linkage*, and we think it happen because test collection includes texts, which are not near the main clusters. Single linkage method tries to build one big cluster, because clusters are placed near each other and their borders are not precise.

In Table III, results of automatic topic and keywords' set detection for each cluster are presented. We also give the value of α parameter which leads to the given results. If the cluster contains small number of texts then the annotation becomes impossible. Information is given for two cases: 1) clusters from golden standard were used 2) clusters, obtained with Between Groups Linkage clustering enhanced with 10% term selection with filtering 3-4 terms with the highest document frequency were used.

Let, $w_i \in W$, $d_j \in D$, $c_k \in C$, $d_l \in D$ correspond to definitions from "Fig. 2". For the annotation process from the "Fig. 2", value of α parameter is important. This parameter is used to determine keywords: the word w_i is a keyword if it occurs at least in $|c_k| - \alpha$ documents of cluster c_k. Words, found with a small value of α, occur often in cluster and they reflect its contents. However, sometimes with the small value of α, words included in the keyword set are specific not only for concrete cluster c_k, but also for the documents of the whole collection. This problem could be solved, with introduction of limitations for w_i : w_i reflects the topic of cluster only if the number of documents, containing w_i , is less than some threshold value. For example as threshold $|c_k|$ could be taken. In this case, common words for the whole collection will not be included in resulting set (such words as: word or corpora). From the other hand, with such approach, we can loose words, which are frequent for some concrete cluster but also are in documents, outside that cluster (words like: translation or linguistic). However we found that words, which are related to topic of cluster, occur frequently in some documents, but for collection specific and common words this is not the case. We have made an assumption that for each word w_i if it relates to the topic of cluster, measures of following two points are almost equal.

- Number of documents d_j of a cluster c_k , which were not included in the set Q because w_i occurred in document d_j more than 3 times.
- Number of documents d_j, which are not included in cluster, but in the same time contain word w_i .

First and second points are balancing each other and allow finding a topic defining word despite the threshold for occurrence, even if this word occur in more than $|c_k|$ documents. Collection specific and common words do not have significant frequencies in single documents so the first point for them will not balancing with the second point. So the introduced thresholds and limitations in the annotation algorithm allow filtering most of the collection specific words without loosing the important keywords for clusters. However as the results in Table III shows us, some collection specific words still persist in the resulting keyword set, giving more challenges for future work.

VI. CONCLUSION AND FURTURE WORK

Research presented in this paper shows that for short text narrow domain collections usage of hierarchical clustering enhanced with special term selection technique could lead to good results. Comparing with other methods discussed in [2][8][12] our approach shows results which are near best and sometimes exceed them. Proposed algorithm of keywords and topic detection allows to detect words which reflect specific of each cluster. Our algorithm gives better results on well differentiated collections, but to process collections like CICling_2002 it needs improvement and this will be the subject for future work.

REFERENCES

[1] M. Alexandrov, A. Gelbukh, and P. Rosso, "An Approach to Clustering Abstracts," In Proc. 10th Int. NLDB-05 Conf., volume 3513 of Lecture Notes in Computer Science, 2005, pp. 8–13. Springer-Verlag.

[2] L. Cagnina, M. Errecalde, D. Ingaramo, and P. Rosso, "A discrete particle swarm optimizer for clustering short text corpora," BIOMA08, 2008, pp. 93–103.

[3] G. Cselle, K. Albrecht, and R. Wattenhofer "BuzzTrack: topic detection and tracking in email," IUI'07, doi:10.2011/www.arnetminer.org/viewpub.do?pid=459847

[4] A. Feng and J. Allan "Incident threading for news passages," CIKM 2009, pp. 1307-1316, doi:10.2011/www.arnetminer.org/viewpub.do?pid=1239503

[5] J, Makkonen., "Semantic Classes in Topic Detection and Tracking," 2009, Helsinki University Print, doi:10. 2011/www.doria.fi/bitstream/handle/10024/48180/semantic.pdf

[6] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering,". In Proc. 5th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 1999.

[7] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to Information Retrieval," Cambridge University Press, doi:10.2011/nlp.stanford.edu/IR-book/information-retrieval-book.html

[8] D. Ingaramo, M. Errecalde, and P. Rosso, "A new anttree-based algorithm for clustering Short-text corpora," Journal of CS&T, 2010.

[9] M. Errecalde., D. Ingaramo, and P. Rosso, "ITSA*: An Effective Iterative Method for Short-Text Clustering Tasks," In Proc. 23rd Int. Conf. on Industrial, Engineering & Other Applications of Applied

 ISBN: 978-1-61208-175-5

Intelligent Systems , IEA-AIE-2010, Springer-Verlag, LNAI(6096), 2011, pp. 550-559.
[10] R. Ortiz, D. Pinto, M. Tovar, and H. Jimenez-Salazar, "BUAP: An Unsupervised Approach to Automatic Keyphrase Extraction from Scientific Articles," In Proc. 5th Int. Workshop on Semantic Evaluation, ACL 2010, pp. 174–177.
[11] F. Perez-Tellez, D. Pinto, J. Cardiff, and P. Rosso "Clustering weblogs on basis of topic detection method," doi: 10.2011/users.dsic.upv.es/~prosso/resources/PerezEtAl_MCPR10.pdf
[12] D. Pinto, "Analysis of narrow-domain short texts clustering. Research report for «Diploma de Estudios Avanzados (DEA)»," Department of Information Systems and Computation, UPV, 2007, doi:10.2011/users.dsic.upv.es/~prosso/resources/PintoDEA.pdf
[13] D. Pinto, P. Rosso, and H. Jiménez, "A Self-Enriching Methodology for Clustering Narrow Domain Short Texts," Comput. J. 54(7): 1148-1165, 2011.
[14] D. Pinto, H. Jimenez-Salazar, and P. Rosso, "Clustering abstracts of scientific texts using the transition point technique," In Proc. CICLing 2006 Conf., volume 3878 of Lecture Notes in Computer Science, pp. 536–546. Springer-Verlag.
[15] S. Smith and M. Rodr´ıguez, "Clustering-based Searching and Navigation in an Online News Source," doi:10.2011/www.inf.udec.cl/~andrea/papers/ECIR06.pdf
[16] Y. Tian, W. Wang, X. Wang, J. Rao, and C. Chen, "Topic detection and organization of mobile text messages," CIKM'10, doi:10.2011/arnetminer.org/viewpub.do?pid=2898431
[17] doi:10.2011/sites.google.com/site/merrecalde/resources
[18] doi: 10.2011/www.adelaide.edu.au/acad/events/workshop/LockhartUPGMA&NJ_calculation.pdf
[19] doi:10.2011/sjsu.edu/faculty/gerstman/StatPrimer/correlation.pdf.

TABLE II. RESULTS OF CLUSTERING

Test collections	**Results of 3cases of testing**								
	1: without initial term selection			2: 10% term selection			3: 10% term selection with filtering 3-4 terms with the highest document frequency		
CICLing 2002	FM_{avg}	FM_{max}	FM_{min}	FM_{avg}	FM_{max}	FM_{min}	FM_{avg}	FM_{max}	FM_{min}
BGL	0,482	0,53	0,42	0,635	0,68	0,54	**0,645**	**0,73**	**0,59**
CL	0,508	0,54	0,48	0,503	0,56	0,45	0,5312	0,58	0,49
SEPLIN CICLing	1			2			3		
	FM_{avg}	FM_{max}	FM_{min}	FM_{avg}	FM_{max}	FM_{min}	FM_{avg}	FM_{max}	FM_{min}
BGL	0,598	0,66	0,42	0,665	0,73	0,56	**0,722**	**0,84**	**0,65**
CL	0,625	0,74	0,54	0,598	0,67	0,55	0,703	0,84	0,58
Easy Abstracts	1			2			3		
	FM_{avg}	FM_{max}	FM_{min}	FM_{avg}	FM_{max}	FM_{min}	FM_{avg}	FM_{max}	FM_{min}
BGL	0,640	0,83	0,48	0,748	0,81	0,72	**0,788**	**0,82**	**0,72**
CL	0,787	0,9	0,72	0,713	0,75	0,63	0,680	0,71	0,61
Micro4 News	1			2			3		
	FM_{avg}	FM_{max}	FM_{min}	FM_{avg}	FM_{max}	FM_{min}	FM_{avg}	FM_{max}	FM_{min}
BGL	0,832	0.89	0,75	0,868	0,96	0,79	**0,873**	0,96	0,79
CL	0,753	0,81	0,67	0,843	0,94	0,8	0,840	0,94	0,78

TABLE III. RESULTS OF OF KEYWORDS DETECTION

CICling 2002	**Clusters**			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Gold Standard	Document α=3	Natur α=7	Word α=6	Relat α=6
	tradit, perform, select, order, rule, document, need, larg, techniqu, automat, compar, identifi, obtain// α = 9	natur, linguist, corpu, kind, work, develop, larg, main, known, translat, obtain, provid // α=9	lexic, word, speech, part, tag, knowledg, sens, english, compar, ambigu, algorithm, disambigu, accuraci, approach, context, method // α=11	type, rule, defin, analysi, sentenc, structur, context, relat // α=9
Automati-cally clustering	Document α=5	Word α=4	No	Represent α=7
	natur, tradit, perform, select, order, rule, document, need, techniqu, experi, automat, compar, identifi, propos, algorithm, gener, discuss, evalu, represent, obtain, provid // α=11	lexic, word, corpu, inform, speech, text, on, part, differ, describ, spanish, sens, automat, compar, disambigu, accuraci, approach, dictionari, method// α = 14	No	atur, lexic, type, mean, analysi, propos, structur, context, translat, represent, relat // α=11

 ISBN: 978-1-61208-175-5

SEPLN CICLing	Clusters			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Gold Standard	Translation α=1	Syntactic α = 8	Clustering α=4	Linguistic α=6
	systems, task, automatic, order, experiments, smt, english, target, spanish, model, translation, statistical // α =8	languages, describe, grammar, parser, parsing, information, syntactic // α =11	obtained, domain, kind, short, performance, clustering, text, measures, propose, work, clusters, cluster, narrow // α =8	presents, order, resources, level, work, time, linguistic, computational, grammar, process, spanish, considered, architecture // α =9
Automati-cally clustering	Syntactic α = 11	Translation α=4	Clustering α=3	No
	grammar, parser, corpus, formalism, information, describe, syntactic // α =14	system, translation, word, machine // α =9	measure, domain, determine, kind, short, method, algorithms, clustering, propose, clusters, cluster// α =7	No

Easy Abstracts	Clusters			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Gold Standard	Objective, search α = 6	Theorem, proof, based, words, key α = 7	Agents α=6	Learning α=6
	tabu, heuristic, computational, order, optimisation, function, constraints, heuristics, objective, scheduling, multi, quality, time, search // α=8	automated, terms, theorem, system, proof, order, implemented, proving, based, words, key// α=8	communication, system, modeling, applications, semantics, flexible, independent, model, agents, information, framework, high, agent, present, work, engineering // α=9	general, classification, set, data, real, model, algorithms, function, analysis, problems, training, methods, learning, results, method, machine // α=11
Automati-cally clustering	Solution α=3	Theorem, proof α = 4	Learning α=8	Agents α=4
	heuristic, computational, algorithm, problem, solution, problems, objective, multi, quality, time, search // α=8	automated, theorem, proof, order, complete, implemented, proving, based, design, describe, words, key// α=6	general, form, class, classification, set, algorithm, support, data, real, space, model, problem, algorithms, function, analysis, problems, number, training, methods, linear, learning, results, method, machine // α =16	communication, variety, context, importance, modeling, world, semantics, flexible, independent, level, complexity, agents, models, information, notion, high, agent, effective, dynamic, formal, work, engineering // α =7

Micro 4News	Clusters			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Gold Standard	Car α=1	Windows α=1	Jesus α=1	Medical α=1
	performance, transmission, ford, road, car, sounds, suspension, tires, driving, cars, buy, mph, engine, honda, parts, bought// α = 5	software, ms, dos, running, windows, version, microsoft, user, files // α = 5	man, god, desire, spirit, acts, words, jesus, biblical, law, christians, sins, church, bible, sin, lord, christ, christian, moral // α=5	dr, study, american, news, patient, health, disease, treatment, control, national, number, related, human, year, patients, medical // α=4
Automati-cally clustering	Car α=1	Windows α=1	Jesus α=1	Medical α=1
	performance, transmission, ford, road, car, sounds, suspension, tires, driving, cars, buy, mph, engine, honda, parts, bought// α=5	software, dos, running, windows, file, version, user// α=5	man, god, desire, spirit, acts, words, jesus, biblical, law, christians, sins, church, bible, sin, lord, christ, christian, moral// α=5	dr, fax, news, patient, women, hiv, health, drug, disease, treatment, data, states, national, research, prevention, public, clinical, david, year, patients, medical, university, medicine // α=6

 ISBN: 978-1-61208-175-5