Automatic Speech Recognition Texts Clustering*

Svetlana Popova[1,2], Ivan Khodyrev[2], Irina Ponomareva[3], and Tatiana Krivosheeva[3]

[1] Saint-Petersburg State University, Saint-Petersburg, Russia
[2] ITMO University, Saint-Petersburg, Russia
svp@list.ru, kivan.mih@gmail.com
[3] Speech Technology Center, Saint-Petersburg, Russia
{ponomareva,krivosheeva}@speechpro.com

Abstract. Abstract. This paper deals with the clustering task for Russian texts obtained using automatic speech recognition (ASR). The input for processing are recognition result for phone call recordings and manual text transcripts for these calls. We present a comparative analysis of clustering results for recognition texts and manual text transcripts, make an evaluation of how recognition quality affects clustering and explore approaches to increasing clustering quality by using stop words and Latent Semantic Indexing (LSI).

Keywords: clustering, speech-to-text, recognition result clustering, Latent Semantic Indexing, information retrieval, stop words

1 Introduction

The development of Internet communication and multimedia raises the issue of searching and structuring data not only in textual form, but also in the form of sound recordings, graphics and video. Spoken content retrieval is becoming increasingly important and this fact motivates extensive research on techniques and technologies in this area [1]. There are two approaches to solving this task. The first one involves transforming speech into text and further processing of text data. The second one is based on analyzing the acoustic signal itself without preliminary transformation of speech into text [2]. This paper uses the first approach.

Our aim is to group together documents that are thematically close, and to classify documents on different topics into different groups. We explore possible approaches to clustering Russian data produced by a speech-to-text system.

Our results demonstrate an extremely small influence of recognition word error rate (WER) of about 20–35% for Russian database. For clustering algorithms we also found a substantial improvement in clustering quality if high-frequency words, excepting context words (keywords) are removed.

In natural language processing it is common to use methods for detecting and using latent features of documents, for instance, Latent Semantic Indexing (LSI) [3]. We explored the possibility of improving clustering quality by means of LSI for a database of near-spontaneous speech with a large overlap of high frequency words between

* This work was partially financially supported by the Government of Russian Federation, Grant 074-U01

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 489–498, 2014.

documents. We demonstrate that using LSI leads to an improvement in the results of the EM (Expectation Maximization [4]) clustering algorithm but does not make much difference for k-means [5]. This fact shows that using LSI does help to detect latent features, however the overlap between clusters is retained in the semantic space built using LSI.

2 The Goal of the Research

The goal of our research is to explore and evaluate different approaches to the task of clustering Russian texts produced by speech-to-text conversion (namely, k-means [5] and the EM (Expectation Maximization) [4] algorithms), as well as to estimate the influence of recognition quality (100% versus 80–65%) on clustering results. We used the implementation of these algorithms provided by the WEKA library (`http://www.cs.waikato.ac.nz/ml/weka/`).

Our choice of algorithms was determined by the fact that k-means and EM are both classical iterative optimization algorithms whose results depend on their initialization. EM is less sensitive to mutual cluster overlapping than k-means, while k-means can work more efficiently if clusters are separated well. Next task was to evaluate the possibility of improving the results of each algorithm using LSI and a domain dependent list of stop words.

We speculated that LSI would detect latent features in the documents, which would make it possible to better identify thematically close documents. The stop list was expected to help remove high-frequency words that did not define the topic and could be treated as noise.

3 Experimental Data

A speech dataset was collected for the purposes of the research. The dataset consists of spontaneous speech recordings (8kHz sample rate) recorded in different analogue and digital telephone channels. The recordings were provided by Speech Technology Center Ltd (www.speechpro.ru) and contain customer telephone calls to several large Russian contact centers. All test recordings have manual text transcripts. In order to be able to evaluate clustering quality, the test dataset was manually labeled with the most frequent call topics. Each text document that contained a transcript of a call was analyzed by three experts. In difficult cases the decision of attributing the text to a certain topic was made by vote. The experts could attribute the text to one or several thematic categories out of the list they were provided with. Only the recordings that were attributed to a single topic by the majority of experts were later used for the test dataset.

As a result, we obtained a dataset of manually prepared text transcripts of the speech recordings, which were divided into five thematic clusters.

We used the speaker-independent continuous speech recognition system for Russian developed by Speech Technology Center Ltd [6,7] which is based on a CD-DNN-HMM acoustic model [8]. The ASR system included interpolation of a general language model (LM) trained on a 6GB text corpus of news articles (300k words, 5 million n-grams) with a thematic language model trained on a set of text transcripts of customer calls to the

contact center (70MB of training data, the training and test datasets did not overlap). We used Good-Turing smoothing (cutoff=1 for all orders of n-grams). Recognition accuracy on the test dataset under these conditions was 80–65%. The recognition results were used to create a second experimental dataset, which corresponded to the same sound files as the first one but contained texts that were produced by using an automatic (rather than manual) speech-to-text transformation with recognition accuracy below 100%.

Table 1 contains a description of both datasets. Each text in the dataset is the recognition result for one short phone call, which is a text of small length. This means that we are faced with the short text clustering task and it becomes difficult to gather enough statistics to improve text processing [9,10].

Table 1. Text datasets

	Manual transcripts	Recognition results
Cluster number	5	
Cluster sizes	44, 24, 28, 55, 35	
Cluster topics	Municipal issues, Military service issues, Political issues, Family and maternity issues, Transport issues	
Word count	12641	11784
Dataset lexicon size	3819	3519

4 Text Pre-processing and Stop Words

Text pre-processing included lowering the case of all characters, removing punctuation marks and deleting words that were found in fewer than three documents (collection size 186 documents). We expected that this would remove words specific to a particular speaker, as well as incorrectly recognized words. The experiments confirmed that using this threshold improves clustering quality.

Stop words. Manual text analysis showed that most texts contain the same high-frequency words and phrases, which carry no information about the clusters topic. For instance, all texts contain greetings, goodbyes and thanks, as well as common function words. Although the clusters are well segmented thematically (they do not belong to similar topics), such uninformative words can introduce a lot of noise in the texts that need to be clustered, and lead to high cluster overlapping. Table 2 shows the top 20 most frequent words for each cluster. Words that carry information about the cluster topic are given in bold.

These words are high-frequency expressions that are not informative and do not refer to the contents of the document. We created a frequency lexicon using both the

Table 2. Demonstrates the high rate of cluster overlap for common words

Russian	English
Cluster: Municipal issues	
в, и, не, у, я, нас, за, на, по, а, мы, это, вот, что, с, спасибо, вопрос, нам, город, меня	*v "in", i "and", ne "not", u "by", ya "I", nas "us", za "for", na "on", po "along", a "and", my "we", eto "this", vot "so", chto "that", s "with", spasibo "thanks", vopros "question", nam "to us", gorod "city", menya "me"*
Cluster: Military service issues	
в, и, я, с, не, вот, на, что, по, спасибо, это, город, как, **пенсии**, вопрос, у, а, меня, так	*v "in", i "and", ya "I", s "with", ne "not", vot "so", na "on", chto "that", po "along", spasibo "thanks", eto "this", gorod "city", kak "how",* ***pensii "pensions"****, vopros "question", u "by", a "and", menya "me", tak "this way"*
Cluster: Political issues	
в, и, не, на, вот, по, вопрос, у, бы, как, это, меня, а, за, так, что, я, спасибо, город	*v "in", i "and", ne "not", na "on", vot "so", po "along", vopros "question", u "by", by (particle), kak "how", eto "this", menya "me", a "and", za "for", tak "this way", chto "that", ya "I", spasibo "thanks", gorod "city"*
Cluster: Family and maternity issues	
я, в, и, не, на, у, что, меня, вот, как, спасибо, вопрос, здравствуйте, это, **детей**, по, бы, нас, почему, до	*ya "I", v "in", i "and", ne "not", na "on", u "by", chto "that", menya "me", vot "so", kak "how", spasibo "thanks", vopros "question", zdravstvuyte "hello", eto "this",* ***detey "children"****, po "along", by (particle), nas "us", pochemu "why", do "until"*
Transport issues	
И, в, у, не, нас, на, я, это, вот, с, ни, вопрос, нет, что, нам, **дороги**, мы, меня, как, по	*i "and", v "in", u "by", ne "not", nas "us", na "on", ya "I", eto "this", vot "so", s "with", ni "not", vopros "question", net "no", chto "that", nam "to us",* ***dorogi "roads"****, my "we", menya "me", kak "how", po "along"*

manual transcripts and the recognition results, after which an expert selected high-frequency common function words from the lexicon and added them to the stop word list. Removing stop words was expected to facilitate thematic clustering. We should note that the list of high-frequency words was analyzed by an expert, and some words from it were not added to the stop list. It is important because the call transcripts contain few context words that reflect the topic. Deleting context words (such as "pensija" (pension), "deti" (children), "dorogi" (roads)) deteriorates clustering quality, which cancels out the effect of stop word removal. For this reason, adding all high-frequency words to the stop list does not improve clustering quality (experiment result), and the list needs to be edited by an expert.

5 The Algorithms

We use two algorithms that are well-known in the field of information retrieval: k-means [5] and EM [4], in the implementation provided by the WEKA library. To define the feature space we used the dataset lexicon obtained after text pre-processing. Each document was represented as a vector in the obtained feature space. The weight of the feature (word) in the document was estimated using *tf-idf* [11]. Both clustering algorithms were given input information about the number of clusters, which equaled 5. For k-means, we used the value 1−(cosine of the angle between document vectors [11]) to calculate the distance between documents. For EM, we set the minimum allowable standard deviation 1.0E-6 and the maximum iteration number as 100.

6 Evaluation

We used the classic clustering quality estimate based on combining information about the cluster Precision and Recall [12,13], F-measure, we will sign it as FM:

$$FM = \sum_i \frac{G_i}{|D|} \max_j F_{ij}, \quad \text{where } F_{ij} = \frac{2 \cdot P_{ij} \cdot R_{ij}}{P_{ij} + R_{ij}}, \\ P_{ij} = \frac{|G_i \cap C_j|}{|G_i|}, \quad R_{ij} = \frac{|G_i \cap C_j|}{|C_j|}, \tag{1}$$

$G = \{G_i\}_{i=\overline{1,m}}$ 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 the collection.

7 Experiments and Results

All experiments were performed on two versions of the same data collection: one contained manual transcripts, the other contained recognition results (with 80–65% recognition quality, see the description of the data in Section "Experimental Data" and in Table 1).

7.1 Experiments Group 1

For these experiments, all documents were pre-processed without removing stop words. Then the documents were clustered using k-means and EM algorithm. We also tested the possibility of improving clustering quality by using LSI, due to mapping the feature space to a lower dimensional semantic space.

7.2 Experiments Group 2

In this group of experiments, all documents were pre-processed and stop words were removed. Then the documents were clustered using k-means and EM algorithm. In these experiments we also tested the possibility of improving clustering quality by using LSI.

7.3 Results

For the Group 1 experiments we obtained the following results. Table 3 shows clustering results for k-means and EM for the experimental conditions without removing stop words. Since both algorithms do not have a single constant result and depend on the initial solution, we performed 100 tests for each algorithm. For each test, we estimated clustering quality using the *FM* (1). We then chose the highest (max) and lowest (min) score for each algorithm, and calculated the average score for all 100 experiments (avg). Table 3 demonstrates two conditions: when LSI was not used, and when LSI was used (for the case of the optimal choice of semantic space dimensions).

Table 3. Clustering results for k-means and EM on the manual transcript dataset and recognition results dataset with and without LSI. Stop words were not removed

	Clustering result for text transcripts			Clustering result for ASR results		
	Without LSI					
	avg	max	min	avg	max	min
k-means	0.36	0.46	0.29	0.35	0.47	0.29
EM	0.35	0.43	0.29	0.36	0.48	0.29
	With LSI (for the case of the optimal choice of semantic space dimensions)					
K-means	0.42	0.49	0.34	0.40	0.42	0.37
EM	0.41	0.42	0.36	0.40	0.47	0.35

Figure 1 shows how mapping the initial feature space onto a lower-dimensional semantic space (LSI) influences clustering results. We selected space dimensions of 2 to 49. First diagram shows the dependency of the average clustering result (avg) for both algorithms on the dimension of the semantic space. Next diagram shows the same dependency for the best clustering results (max). Average FM=0.36 and maximum FM=0.46 were chosen as baseline because these values reflect the algorithms results without stop word deletion and LSI. Both diagrams show clustering result both for manual transcripts and for recognized text.

For Group 2 experiments we obtained the following results. Table 4 demonstrates clustering results for k-means and EM when stop words were removed, with and without LSI. As in Table 3, we show the estimate for the best (max), worst (min) and average (avg) clustering result over 100 tests.

Table 4. Clustering results for k-means and EM on the manual transcript dataset and recognition results dataset with and without LSI. Stop words were removed

	Clustering result for text transcripts			Clustering result for ASR results		
	Without LSI					
	avg	max	Min	avg	max	min
k-means	0.44	**0.57**	0.34	0.42	**0.58**	0.35
EM	0.36	0.45	0.30	0.37	0.48	0.29
	With LSI (for the case of the optimal choice of semantic space dimensions)					
K-means	0.41	0.51	0.32	0.40	0.49	0.34
EM	**0.47**	0.56	**0.38**	**0.46**	0.53	**0.39**

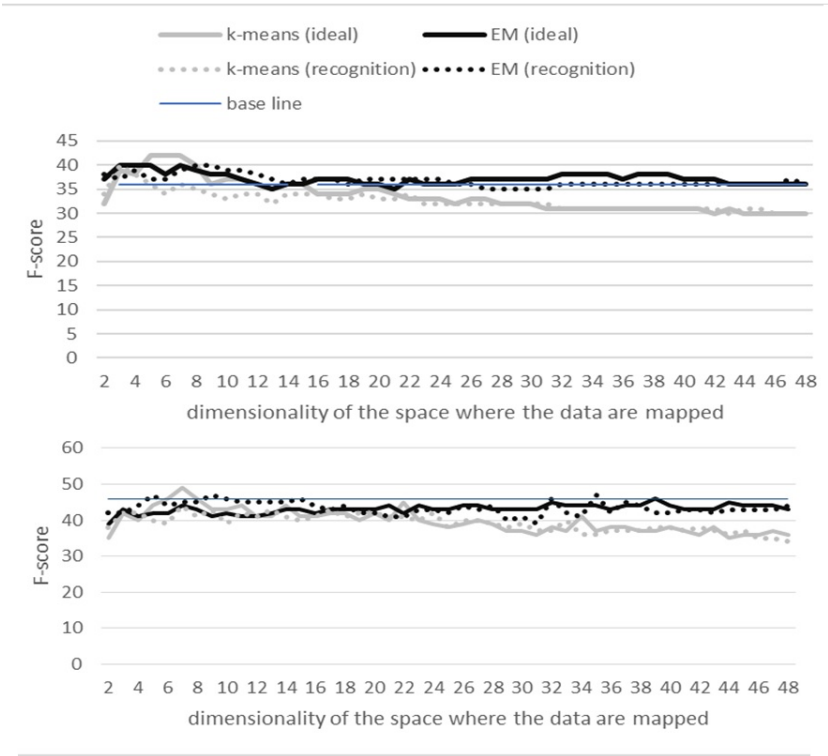

Fig. 1. Dependency of clustering quality on LSI semantic space dimension. First: dependency of the average score for 100 tests (avg). Next: best score (max). Ideal – result for manual text transcripts, recognition – result for ASR results. Stop words were not removed

Figure 2 shows how using LSI and mapping the initial feature space onto a lower-dimensional semantic space influences clustering results when stop words are removed from the texts (2 to 49 dimensions). First diagram shows the dependency of the average clustering result (avg) for both algorithms on the dimension of the semantic space. Next diagram shows the same dependency for the best clustering results (max).

8 Discussion

Our results show that recognition quality has no strong influence on the k-means and EM clustering results when comparing manual text transcripts and recognition results with 80–65% accuracy. This shows that a decrease in recognition quality does not have a strong influence on clustering if the accuracy is about 80–65% (it should be noted that most data in the collection have recognition accuracy close to 80%).

Both algorithms show a similar result if we do not use stop word deletion and LSI. Using LSI improves average clustering quality (avg) when stop words are not

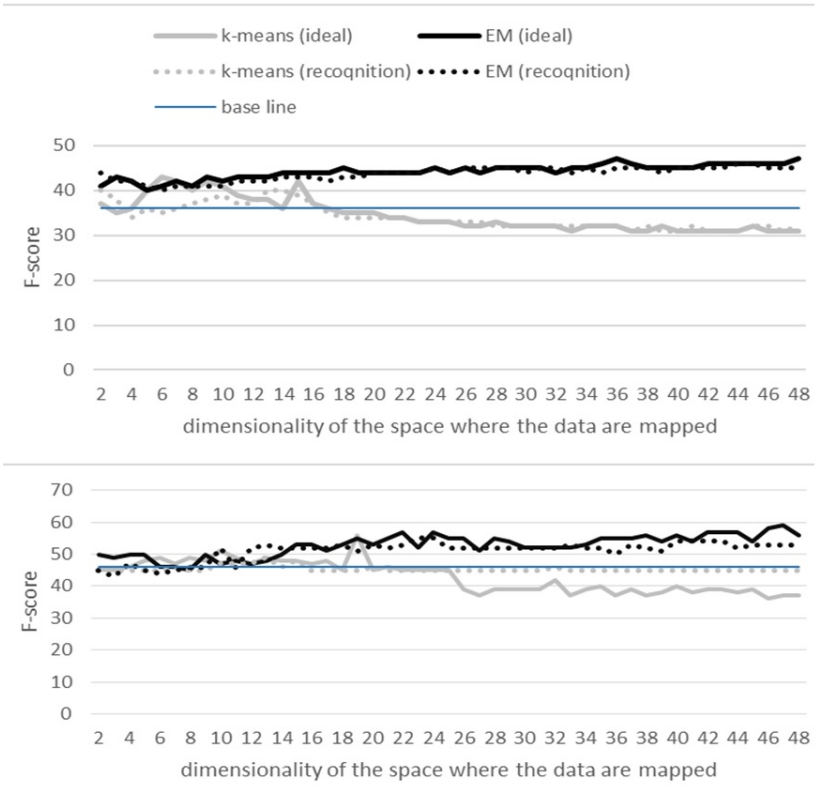

Fig. 2. Dependency of clustering quality on LSI semantic space dimension. First: dependency of the average score for 100 tests (avg). Next: best score (max). Ideal – result for manual text transcripts, recognition – result for ASR results. Stop words were removed

removed. However, for EM this improvement is practically stable on 2 to 49 – dimensional semantic space, while for k-means the improvement is only observed when the dimension is lower than 15; when the dimension increases further, clustering quality begins to decline. This is probably due to the assumption that as semantic space dimension increases, clusters begin to overlap more strongly. That has a negative effect on k-means, which is more sensitive to cluster overlap than EM. K-means provides the highest average score under the optimal semantic space dimension (avg FM=0.42 for manual transcripts), however this value only slightly improves upon the best average EM result (FM=0.41).

If we compare the best results (max) obtained with LSI, we observe a situation similar to the average results (avg), although the difference in maximum results is more pronounced than in average results.

Removing stop words considerably improves k-means results, which illustrates better cluster segmentation when a stop word list is used. EM results remain virtually unchanged compared to baseline if we use stop words.

If we use both the stop word list and LSI we observe a steady increase in EM results when semantic space dimension is increased (from 2 to 49). In case of k-means, clustering results with LSI are worse compared to those without LSI, and the result deteriorates steadily when semantic space dimension is increased. This is probably due to the assumption that clusters overlap more in the semantic space than in the feature space before its dimension decreases.

When we use both the stop word list and LSI, EM average result (avg) exceeds the average result of k-means. The results for the best scores (max) when using both stop words and LSI behave similarly to the average scores (avg). It should be noted that the highest (max) clustering quality is reached when using k-means, stop word deletion and no LSI (FM=0.58), which improves the EM result under the same conditions but with LSI by 0.02.

To sum up, on average, the best results are demonstrated by EM when stop word deletion and LSI are used.

9 Conclusion

Our research demonstrates that average clustering results of k-means and EM on manual transcripts and recognition results with 80-65% accuracy do not show a large difference. Using a stop word list of high-frequency common function words (without context words) leads to a substantial increase in k-means clustering quality if LSI is not used. Removing stop words improves EM results if we use feature space mapping onto a semantic space by means of LSI. In the latter case we observe stable improvements in quality as the semantic space dimension increases from 2 to 49. The following conclusions can be drawn from experiments. Using a list of stop words improves cluster segmentation, which influences the performance of k-means and LSI. The use of the latter improves the performance of the EM algorithm.

On average, the best results are achieved when using a domain dependent stop list, LSI and the EM algorithm.

References

1. Larson, M. and Jones, G.J.F. Spoken content retrieval: A survey of techniques and technologies. Foundations and Trends in Information Retrieval, 5 (4-5). pp. 235-422. ISSN 1554-0669, (2012)
2. Park,A. and Glass, J.R. Unsupervised pattern discovery in speech. IEEE Trans. Acoustics, Speech and Language Processing, 8(1):186197, 2008)
3. Deerwester, S., et al, Improving Information Retrieval with Latent Semantic Indexing. Proceedings of the 51st Annual Meeting of the American Society for Information Science 25, pp. 3640, (1988)
4. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B, (1977)

5. MacQueen, J.B. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297, (1967).
6. Chernykh, G., Korenevsky, M., Levin, K., Ponomareva, I. and Tomashenko, N. Cross-Validation State Control in Acoustic Model Training of Automatic Speech Recognition System. Scientific and Technical Journal Priborostroenie, ITMO University, Vol57 2, p.23-28, (2014)
7. Kudashev, J., Kozlov, A. The Diarization System for an Unknown Number of Speakers. Lecture Notes in Artificial Intelligence (LNAI), vol. 8113, P. 340344, 2013.
8. Dahl, G.E. Dong Yu, Li Deng and Acero, A., "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition", IEEE Trans. Audio, Speech and Language Proc., 20 (1): 30-42, (2012)
9. Pinto, D. Analysis of narrow-domain short texts clustering. In: Research report for Diploma de Estudios Avanzados (DEA), Department of Information Systems and Computation, UPV, (2007)
10. Pinto, D., Rosso, P., Jimenez, H.: A Self-Enriching Methodology for Clustering Narrow Domain Short Texts. In: Comput. J. 54(7). P. 1148-1165, (2011)
11. Manning, C., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. In: Cambridge University Press, (2009)
12. Eissen, S., Stein, B.: Analysis of Clustering Algorithms for Web-based Search. In: Practical Aspects of Knowledge Management, LNAI N 2569. Springer, P.168178, (2002)
13. Stein, B., Meyer zu Eissen, S., Wibbrock, F. On Cluster Validity and the Information Need of Users. In 3rd IASTED Int. Conference on Artificial Intelligence and Applications (AIA 03). Benalmadena, Spain. Edited by M. H. Hanza. P. 216-221, ISBN 0-88986-390-3, ACTA Press, IASTED, (2003)