ISSN 2007-9737

Construction of Paraphrase Graphs as a Means of News Clusters Extraction

Elena Yagunova[1], Ekaterina Pronoza[1], Nataliya Kochetkova[2]

[1] St.-Petersburg State University,
Department of Informational Systems in Arts and Humanities, St.-Petersburg,
Russian Federation

[2] National Research University Higher School of Economics,
School of Computer Engineering, St.-Petersburg,
Russian Federation

{iagounova.elena, katpronoza}@gmail.com, natalia_k_11@mail.ru

Abstract. In this paper, we construct paraphrase graphs for news text collections (clusters). Our aims are, first, to prove that paraphrase graph construction method can be used for news clusters identification and, second, to analyze and compare stylistically different news collections. Our news collections include dynamic, static and combined (dynamic and static) texts. Their respective paraphrase graphs reflect their main characteristics. We also automatically extract the most informationally important linked fragments of news texts, and these fragments characterize news texts as either informative, conveying some information, or publicistic ones, trying to affect the readers emotionally.

Keywords. News cluster, paraphrase graph, paraphrase extraction, linked text segments, text analysis.

1 Introduction

Paraphrase extraction, identification and generation are increasingly popular topics of research in natural language processing nowadays. Paraphrase corpora can be helpful for various tasks like text summarization, text entailment recognition, information extraction, sentiment analysis, and many more. However, potential application of paraphrases is not limited to the mentioned tasks. We believe that paraphrases can help us to analyze the structure and other characteristics of text collections.

In this paper we work with thematically homogeneous news text collections (clusters). News clusters differ in style, themes and structure: they can be static or dynamic, informational or publicistic, conventional or unconventional, with different hierarchy and number of subtopics. We believe that news cluster can be considered a recognized object of text linguistics today. Indeed, modern media audience is often interested in a series of news on a given topic rather than in a single news report.

In this paper we test a hypothesis that paraphrase construction method allows us to identify thematically homogeneous news clusters. A paraphrase graph is a graph where news headlines are vertices, and two vertices are connected by an edge if they are paraphrases [4]. Such graph reflects the structure of the corresponding news cluster: for example, similar headlines tend to group into subgraphs which refer to the subtopics in the news cluster.

There is no doubt that paraphrase graph construction approach can only be used if we are sure that the headlines do reflect the main topics of the news reports. However, it is not true for every news report. Apart from the purely informative headlines, there are also publicistic headlines which aim to affect the reader emotionally. That is why we also conduct experiments to identify clusters with informative and publicistic headlines. These experiments involve extraction of the most important linked text segments from headlines and bodies of the news reports, and their comparison.

doi: 10.13053/CyS-22-4-3065

2 Related Work

Clustering is a widely discussed problem in natural language processing, and there is a large number of papers on clustering news articles. But to the best of our knowledge, most papers dealing with news clusters usually solve the problem of clustering news articles, however, in our case, the articles are already clustered and we work with the obtained clusters.

Some researchers propose methods of clustering news headlines, without considering the bodies of news reports. For example, in [2] news headlines are clustered using heuristic frequent term-based and frequent noun-based methods. These methods are reported to be as good as the traditional clustering methods, however, tested on scientific abstracts, the results are worse.

In [7] the authors solve the problem of selecting appropriate labels for the clusters of news headlines.

Some researchers work at the problem of online incremental clustering of news articles [1].

Note that the problem of automatic extraction of the text of a news is not so trivial due to the amount of advertisements, non-standard structure of the news, etc. [8].

In our study, we focus on the analysis of the existing news clusters, and show that it can be conducted via the construction of paraphrase graphs based on the headlines of the clusters.

3 Data

Our data include thematically homogeneous news collections from Galaktika-Zoom news aggregator [11]. We work with five news clusters. They are extracted from various Russian media sources; each cluster consists of several news reports about one event. The clusters are about:

– The arrival of A. Schwarzenegger in Moscow (360 texts, about 110 thousand tokens);
– The appointment of S. Sobyanin as the Mayor of Moscow (660 texts, 170 thousand tokens);
– The predictions and the death of Paul the Octopus (310 texts, about 1 million tokens);
– The protests in Ecuador (569 texts, about 1 million tokens);
– The release of toxins in Hungary (346 texts, about 100 thousand tokens).

All the five clusters describe a single macroevent and are more or less thematically homogeneous. The “Sobyanin” cluster is the least homogeneous cluster because it consists of texts describing several events inside one main macroevent.

Our choice of the clusters is motivated by several purposes:

– text should have more or less clear and simple syntactic and semantic structure (e.g., compared with fiction);
– clusters should be informationally important and of a large volume, with a clear plot, with only one main character;
– at least one of the clusters should be more dynamic (i.e., its texts are dynamic, with a chain of different situations) – see the cluster about A. Schwarzenegger;
– at least one of the clusters should be more static (i.e., the texts are about one main event rather than a chain of situations) – see the cluster about S. Sobyanin;
– at least one of the clusters should be mixed, when a macroevent is described, with complex causal and temporal relationships inside – see the cluster about Paul the Octopus.

4 Method

4.1 Paraphrases Extraction

We test a hypothesis that paraphrase graph construction method can be applied towards the problem of news texts (or rather headlines) clustering. Paraphrase graph is a graph where sentences are vertices, and edges connect headlines if they are para-phrases.

To test our hypothesis, we consider gold standard news clusters and construct paraphrase graphs for each of the clusters. The constructed graphs are supposed to be represented by single connected components with minimal number of outlying vertices.

To decide whether two headlines are paraphrases, we use our paraphrase identification model trained on the ParaPhraser corpus [6].

ParaPhraser is a project dedicated to the collection of the Russian paraphrase corpus. It consists of about 11 thousand of sentences pairs annotated as precise, loose or non-paraphrases. The sentences are derived from news headlines.

Our paraphrase identification model is actually an SVM with two types of features: shallow string-overlap-based features and se-mantic features (covering synonymy and word-building relations). Detailed description of the features can be found in [5]. This is our best paraphrase identification model so far (it obtained accuracy of 0.7448 and F1 of 0.8078 at Russian Paraphrase Detection Shared Task'2016 in binary classification subtask).

Thus, for each of the five news clusters in question, we extract their headlines and compare them pairwise inside each cluster. Each pair of news headlines inside a cluster is assigned a paraphrase class (1 – for paraphrases, either precise or loose, and 0 – for non-paraphrases) which is the output of our paraphrase identification model. The resulting set of sentences together with their paraphrase relations form a paraphrase graph for each news cluster.

It is assumed that since a news headline is a convolution of a news report and texts inside each cluster in question describing the same event, each cluster or at least the greater part of each cluster it will be represented by a single connected component in paraphrase graphs. The scope of our method can be evaluated by the number of the outlying vertices (headlines). In other words, to evaluate the appropriateness of our paraphrase graph construction method for the problem of news clustering, we calculate the portion of the vertices inside the largest connected component for each of the considered news clusters: the closer the portion value is to 1, the better our method is.

4.2 Extraction of Linked Text Segments

As a baseline we analyze texts inside their respective news clusters via their key-words.

When analyzing a text using keywords method, a focus can be on different units of analysis:

– A keyword in the context of other keywords (e.g., when we analyze the role of this keyword).

– A keyword in the context of the whole text (e.g., when we analyze the distribution of the word in the text).

– A text (or the list of keywords as text compression) in the context of the whole cluster [10].

Table 1. Top-3 central nodes of the paraphrase graphs

Cluster	Top-3 central nodes
Ecuador	There is unrest in Ecuador A coup in Ecuador: police attacked the president A coup attempt in Ecuador
Paul the Octopus	Paul the Octopus died Died Paul the Octopus Died Paul the Octopus…
A. Schwarzenegger	Schwarzenegger in Moscow Medvedev took Schwarzenegger to Skolkovo in "Tchaika" Medvedev drove with Schwarzenegger to Skolkovo in "Tchaika"
S. Sobyanin	Sobyanin is the Mayor of Moscow Moscow City Duma approved Sobyanin as the new Mayor of Moscow Sergey Sobyanin is the new Mayor of Moscow
Toxins	In Hungary toxic waste fell into the Danube Toxic waste from Hungary fell into the Danube Hungary: toxic waste fell into the Danube river

Table 2. Percentage of vertices in the largest connected component

Cluster	Percentage %
Ecuador	93
Paul the Octopus	93
A. Schwarzenegger	92
S. Sobyanin	91
Toxins	88

In this part of the research we analyze keywords in the context of the text, and the text as a part of the cluster.

The computational experiment in question is based on the calculation of the Dice coefficient. The experiment is conducted in two stages:

1. We calculate Dice coefficient for each pair of tokens for all text collections.

2. We link pairs of tokens together into text segments according to their Dice coefficient values and their context. As we move along the text, we decide whether cur-rent word should join the current text segment. Namely, we do not link the word to the preceding word if Dice coefficient for this pair is lower than the threshold or if it is lower than the average between the corresponding values for the left and the right pairs. As a result of such procedure, we obtain a list of linked segments, calculated for each text independently and then united into a frequency list. In other words, each text becomes marked up with linked segments (word sequences).

As a result of such experiment, we obtain:

– A list of linked segments describing a cluster;
– A list of linked segments describing each text of a cluster;
– A list of linked segments describing headlines inside a cluster.

At this stage of the research we prove the following hypotheses:

– Headlines can be used as a traditional baseline for the analysis of the news clusters.
– The most frequent words from the headlines can be used for the express analysis of the news clusters.
– The comparison of the linked text segments obtained from the headlines and from the news reports themselves reveals the publicistic and the informational components in the headlines.

5 Results

5.1 Results of the Experiment with Paraphrase Graphs

According to the algorithm described in Section 4.1, we have built paraphrase graphs for the news clusters in question. In the current section we provide a brief description of the constructed graphs and error analysis, i.e., analysis of the headlines which are missing from the largest connected components of each paraphrase graph.

We extracted nodes with the largest node degrees from each paraphrase graph. They are presented in Table 1 (top 3 nodes per graph).

It turned out that for each graph central nodes correspond to the shortest headlines. For all the clusters except for the Schwarzenegger cluster, central nodes are brief and informative and represent the compression of the news reports of the clusters. As for the central nodes of the Schwarzenegger cluster, the corresponding headlines are also laconic, but not purely informative, in fact they are publicistic rather than in-formative.

For each constructed graph, we extracted its connected components. We assume that since headline is a convolution of text and the texts inside each cluster refer to the same event, the whole cluster or at least the greater part of it will be one connected component. The outlying headlines, in their turn, reveal the scope of the method.

For each of the five constructed graphs we calculated percentage of vertices covered by their largest connected component (see Table 2). The closer these values are to 100%, the better performance our paraphrase graph construction method demonstrates.

Quite high percentage scores allow us to speak about the validity of the proposed method for the task of extracting thematically homogeneous news clusters. It can be seen that our method shows best results (93%) against the “Schwarzenegger” and the “Paul the Octopus” clusters, while the lowest result is achieved against the “Sobyanin” cluster. According to the results of our second experiment with linked text segments, publicistic headlines are abundant in the former two clusters, while the latter one (the “Sobyanin” cluster) being purely informative, includes several sub-events inside one macroevent, i.e. is the least homogeneous cluster of the five considered ones.

In the five constructed clusters, the outlying vertices are of the main interest to us because they reveal the limits of our clusters extraction method. Let us consider such vertices further in this section.

Firstly, news headlines are often represented by citations or snippets of dialogues – vivid and memorable phrases aimed to attract readers' attention, which are not the convolutions of the news report (e.g., "У нас есть вакантное место" / "We have a vacancy"; "Медведев выразил сожаление, что у Шварценеггера нет российского паспорта, а то вакантное место есть" / "Medvedev sorry that Schwarzenegger has no Russian passport because there is a vacancy for him"; "Д.А.Медведев: Опыт Сколково надо тиражировать по всей стране" / "D. Medvedev: Skolkovo experi-ence should be …")

In the "Schwarzenegger" cluster the headlines are often references to the "Terminator" movie (which is not surprising since A. Schwarzenegger is mostly known to the Russians thanks to this movie), e.g., "Восстание машин" / 'Rise of the machines", "Иногда они возвращаются" / "Sometimes they do come back".

Short headlines (consisting of 2–3 words) especially without proper names, receive low value of the closeness measure with other headlines, and such low values are insufficient to get into the main connected component, e.g., "Военный мятеж" / "Military mutiny".

Similarly, very long headlines consisting of a whole paragraph also appear to be outliers in paraphrase graphs: they receive low similarity value with other headlines because of the abundance of mismatched vocabulary and complex syntax, e.g., "На ликвидацию последствий разлива химикатов на алюминиевом заводе в Венгрии потребуется как минимум год работы и несколько десятков миллионов долла-ров" / "It will take at least a year of work and several tens of millions of dollars to eliminate the consequences of chemical spills at an aluminum plant in Hungary".

Despite the abundance of emotional and attention attracting (purely publicistic

) headlines, the "Schwarzenegger" cluster shows the maximal thematic homogeneity while the "Sobyanin" cluster consists of a multitude of events related to the inauguration of Sobyanin as the mayor of Moscow. However, our paraphrase identification model could not divide it into sub-events. Only a part of the sub-events formed separate components, but the larger part of sub-events were stucked to the main one.

It is worth noting that some errors were also initially contained in the news corpora, for example, "Тема недели"/ "Topic of the week", "На идут уже 331 человек!" / "331 people being sent already!". However, their percentage is very small (about 0.2%).

Thus, the result of news headlines clustering is influenced by the usage of specific expressions, especially references or emotional words, in different headlines and in different contexts.

General phrases, analogies, opinions and evaluations of events are not amenable to clusters. They have lack of proper names (or contain names of experts which further reduces the value of the closeness measure) and they have very tiny intersection with words that describe the event.

5.2 Results of the Experiment with Linked Text Segments

Results of the experiment with linked text segments have proved our ideas about the differences in the style of the considered news clusters.

At this part of our research we only worked with the three news clusters out of fixe: about the arrival of A. Schwarzenegger in Moscow, about the appointment of S. Sobyanin as the Mayor of Moscow and about the death of Paul the Octopus as they cover all types of news clusters we consider in this paper, and are the most typical representatives of different style.

In the cluster about A. Schwarzenegger the theme of investment and innovations turned to be less important (according to headlines analysis) than the ride of A. Shwarzenegger on the Moscow subway and his posting about it in Twitter: «шварценеггер_прокатился в московском_метро» /Schwarzenegger took a ride in the Moscow metro/ (8 occurrences), «медведев_прокатил_шварценеггера_на "_чайке_"» /Medvedev took Schwarzenegger for a ride in "Tchaika"/ (6 occurrenc-es), «шварценеггер_гуляет в московском_метро и переписывается с_медведевым через_Twitter» /Schwarzenegger takes a ride in metro and corresponds with Medvedev in Twitter/ (3 occurrences), «шварценеггер_хочет_вернуться в_кино» /Schwarzenegger wants to return to acting/ (3 occurrences).

Express analysis of the second cluster has shown that it has one main theme: the appointment of S. Sobyanin as the Mayor of Moscow.

The second most important theme is about the retirement of Luzhkov. Such conclusions are proved by the frequency list of linked text segments obtained from the corresponding news cluster: «новым_мэром москвы стал сергей_собянин» /Sergey Sobyanin has become a new Mayor of Moscow/ (7 occurrences), «мосгордума _утвердила сергея_собянина на_пост мэра_ москвы» /Moscow City Duma has approved Sobyanin as the Mayor of Moscow/ (4 occurrences), «сергей_собянин - новый_мэр_ москвы» /Sergey Sobyanin as a new Mayor of Moscow/ (4 occurrences), «мосгордума_ «мосгордума_проголосует по_кандидатуре собянина на_пост мэра_москвы» /Moscow City Duma will vote for Sobyanin as the Mayor of Moscow/ (3 occurrences), «в_москве назначили но-вого_мэра» /A new Mayor appointed in Moscow/ (3 occurrences), «москва_получила нового_мэра» /Moscow got a new Mayor/ (3 occurrences) и «москва_получит нового_мэра» /Moscow will get a new Mayor/ (3 occurrences), etc. There are much fewer news reports with headlines like «лужков_не пойдет на инаугурацию_преемника» /Luzhkov will not attend the inauguration of his successor/ (2 occurrences).

The main theme of the third cluster concerns the death of Paul the Octopus. Other less important themes are his predictions and other moments of his biography as well as the appointment of his successor. This is also confirmed by the list of linked text segments derived from the corresponding news cluster. Thus, most frequent text seg-ments are «умер_осьминог_пауль» /Paul the Octopus died/ (30 occurrences), «умер_осьминог-предсказатель_пауль» /Paul the Octopus, the predictor, died/ (20 occurrences), «в_германии_умер осьминог_пауль» /Paul the Octopus died in Germany/ (19 occurrences), others are quite similar: «скончался_осьминог_пауль» /Paul the Octopus deceased/ (9 occurrences), «знаменитый_ осьминог-предсказатель пауль_умер» /Famous Octopus-Predictor died"/ (6 occurrences), «в_германии_умер осьминог- предсказатель_ пауль» /The Octopus-Predictor died in Germany/ (5 occurrences). Only some of them convey some additional information, for example, «мемориал_ осьминога_пауля установят в_его родном_ немецком_океанариуме» /Paul the Octopus Memorial will be installed in his native German oceanarium/ (5 occurrences).

All the text segments were extracted using a program by Daudaravičius [3]. The evaluation of the corresponding algorithm was conducted by its author and it was shown that in the task of keywords extraction it helps to increase F1 by 17-27% de-pending on the data.

Some of the segments we obtained during express analysis of the clusters consist of 3 and more words, and such segments reflect the distribution of levels of importance of different structural components of the text. It correlates with the idea that the addressee of the texts tends to operate the largest operational units (see [11]). The above supposition is part of the hypothesis that heterogeneous structure of a news text is reflected in the markup of the segments which are most important for the addressee's perception (see [8, 10]):

– In most cases linked text segments of 3 and more words have the largest weights (i.e., the highest importance);
– Sometimes such trend may be broken, which reflects the structure of the text:
– In this case either the structural components are short (which corresponds to the syntax of the text)
– Or the structural components are split to intensify the effect of the text on the addressee.

Dynamic sketchy style of the most texts about A. Schwarzenegger is similar to that of Twitter. It is characterized by short structures and thus has short linked segments. There is no strict compositional structure of a plot although the text is certainly close to the narrative. We believe that the cluster about A. Schwarzenegger most clearly reflects the differences between the data obtained from the headlines and from the bodies of the news reports. Even on the baseline level of the linked segments there is a clear tendency: headlines are publicistic (trying to affect the addressee rather than share the information) while the bodies are informational and sound “more quiet”. Thus, the list of linked text segments gives us a useful insight into the news cluster in question.

In the cluster about S. Sobyanin there is no such striking contrast between the segments obtained from the headlines and the bodies of the news reports.

Texts are static, they tell about a single event and the main theme stands out very clearly. Linked segments derived from the headlines consist of informationally important fragments as well as the publicistic ones. Linked segments derived from the bodies of the news reports allow us to get an idea about important additional subtopics (e.g., the party path of Sobyanin, voting circumstances, etc.).

In the cluster about Paul the Octopus the intersection between the segments de-rived from the headlines and those from the bodies is minor. The texts in this cluster are static, they narrate about a single event and at the same time are full of various details which form the event and their compositional structure is not conventional.

There could be two reasons for such disconformity: firstly, a bright publicistic style both in the headlines and the bodies of the texts (which forces to abandon clichés and use a variety of design options for the messages), and secondly, the abundance of details, i.e., topics and subtopics, which are difficult to attribute weights of informational importance. Thus, if we consider lists of linked segments as the compressed text of the cluster, then the cluster about Paul the Octopus is the most complex one among the three considered clusters.

6 Conclusion

This paper presents our results of the analysis of news clusters different in style, themes and structure. News clusters can be static or dynamic, informational or publicistic, conventional or unconventional, with different hierarchy and number of sub-topics. We present a method of news cluster analysis via its paraphrase graph structure.

Our main idea is based on the idea of news cluster compression: we analyze news headlines as specific forms of compression. The specialty of headlines is that they are given by the journalists or editors and may be oriented mainly on information trans-mission function, or rather the function of the impact on the addressee.

We conduct two experiments. During the first experiment, we work with headlines inside each news cluster and extract paraphrase headlines.

These paraphrase head-lines form a paraphrase graph which reflects the structure and the characteristics of the news cluster. As the news clusters are more or less thematically homogeneous, most vertices inside each paraphrase graph are supposed to form a single connected components – and they do, as was shown in our experiments.

It also appears that such clusters extraction method as paraphrase graphs construction can be applied for both publicistic and informative headlines. A lot of errors (see the outlying headlines in paraphrase graphs of the news clusters) occur on headlines containing emotional, anaphoric and attention attracting headlines, however, the real problem is the hetero-generous structure of a news cluster – in such cases our clusters extraction method shows the lowest headlines coverage score.

During the second experiment, we work with linked text segments extracted from the headlines and the bodies of the news reports. It turns out that the characteristics of the derived lists of linked segments and the differences between the segments de-rived from headlines and bodies are closely related to the type of considered news cluster.

As a result of the experiments, we were able to confirm the following hypotheses:

– Paraphrase graphs can be successfully used for the extraction of thematically homogeneous news clusters.
– The most frequent words from the headlines can be used for the express analysis of the news clusters.
– The comparison of the linked text segments obtained from the headlines and from the news reports themselves reveals the publicistic and the informative components in the headlines.

References

1. **Azzopardi, J. & Staff, Ch. (2012).** Incremental Clustering of News Reports. *Algorithms*, Vol. 5, No. 3, pp. 364–378. DOI: 10.3390/a5030364.
2. **Bora, N. N., Mishra, B. S. P., & Dehuri, S. (2012).** Heuristic Frequent Term-Based Clustering of News Headlines. *Procedia Technology*, Vol. 6, pp. 436–443. DOI: 10.1016/j.protcy.2012.10.052.
3. **Daudaravičius, V. & Marcinkevičienė, R. (2004).** Gravity Counts for the Boundaries of Collocations. *International Journal of Corpus Linguistics*, Vol. 9 No. 2, pp. 321–348. DOI: 10.1075/ijcl.9.2.08dau.
4. **Pronoza, E., Yagunova, E. & Kochetkova, N. (2016).** Sentence Paraphrase Graphs: Classification Based on Predictive Models or Annotators' Decisions?. In: Sidorov G., Herrera-Alcántara O. (eds.) *Advances in Computational Intelligence*, Lecture Notes in Computer Science, 10061, Springer, pp. 41–52. DOI: 10.1007/978-3-319-62434-1_4.
5. **Pronoza, E. & Yagunova, E. (2015).** Comparison of sentence similarity measures for Russian paraphrase identification. *Proceedings of the Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT)*, pp. 74–82. DOI: 10.1109/AINL-ISMW-FRUCT. 2015.7382973.
6. **Pronoza, E., Yagunova, E., & Pronoza, A. (2015).** Construction of a Russian Paraphrase Corpus: Unsupervised Paraphrase Extraction. *Communications in Computer and Information Science*, pp. 146–157. DOI: 10.1007/978-3-319-41718-9_8.
7. **Thirunarayan, K., Immaneni, T., & Shaik, M. V. (2007).** Selecting Labels for News Document Clusters. *Lecture Notes in Computer Science*, pp. 119–130. DOI: 10.1007/978-3-540-73351-5_11.
8. **Yagunova, E. & Pivovarova, L. (2010).** The nature of collocations in the Russian language. The experience of automatic extraction and classification of the material of news texts. *Automatic Documentation and Mathematical Linguistics*, Vol. 44, No. 3, pp. 164–175. DOI: 10.3103/S0005105510030088.
9. **Viveros-Jiménez, F., Sánchez-Perez, M.A., Gómez-Adorno, H., Posadas-Durán, J.P., Sidorov, G., & Gelbukh, A. (2018).** Improving the Boilerpipe Algorithm for Boilerplate Removal in News Articles Using HTML Tree Structure. Computación y Sistemas, Vol. 22, No. 2, pp. 483–489.
10. **Yagunova, E., Pivovarova, L., & Volskaya, S. (2015).** News Text Segmentation in Human Perception. *Proceedings of Natural Language Processing and Cognitive Science*. Eds. Sharp B., Delmonte R. de Gruyter, pp. 63–74.
11. **Antonov, A. V. & Yagunova, E. V. (2010).** Procedure of working with text information collections via information portraits analysis. *Proceedings of RCDL'10*, pp. 79–84.

Article received on 18/12/2017; accepted on 15/02/2018.
Corresponding author is Elena Yagunova.