Car Forums: a new Russian Language Dataset Annotated with Keyphrases

Svetlana Popova
Saint-Petersburg State University, Saint-Petersburg, Russia
svp@list.ru

Gabriella Skitalinskaya
Institute of Technology Tallaght, Dublin, Ireland
gabriellasky@icloud.com

***Abstract*—In the paper we present a new, annotated with keyphrases dataset of posts in the Russian language obtained from car forums. The article describes the methodology of building the dataset, as well as its main characteristics.**

I. Introduction

The keyphrases extraction problem has been widely studied in English literature and primarily addressed the keyphrase extraction from texts written in the English language (e.g. [1-14] and other). Keyphrases are sequences of words, that reflect the main topics of texts, their extraction differs from the task of extracting keywords (single words), terminology or collocations. A keyphrase can be a rare phrase and also not a stable expression. For the English language, special test datasets for keyphrase extraction problem have been developed, specialized competitions and seminars have been organized (e.g [1-4]). For the Russian language the extraction of keywords and collocations are widely studied, but keyphrase extraction requires further development, including the development of open datasets.

II. Dataset Description

We obtained a collection of messages from car forums (about 20 different websites). At the next stage, the 3-5 messages of each forum thread were selected. From the selected pool of messages, six non-overlapping collections containing 60 random texts were created. In our studies, such a separation was required to check that the improvement in performance of an algorithm for keyphrase extraction is not due to randomness. Each of the six collections includes texts of different lengths with positive and negative user feedback. Two of the collections contain exactly 30 positive and 30 negative texts per collection. Other collections contain positive and negative posts in random proportions. Keyphrases were assigned to each text in the collections. A number of main annotation strategies were used to define the keyphrases of interest. These strategies can be summarized as follows: which aspects are the most discussed in the reviews, what do the users pay attention to first of all and what statements are significant for the post. In the next section we will describe this strategy in detail. The Dataset is available upon request by email. Tables 1-3 presents the main characteristics of the developed collection. Examples of texts and proposed annotations for them are provided in Table 4. Notice, texts and annotations are written in a colloquial style with Russian-speaking stylistic features, misspellings, expressions, which make it difficult to properly translate the texts to English in presented examples.

TABLE I. Dataset description I. Texts: Min - minimum number of words per document, Max - maximum number of words per document, Avg - average number of words per document, Words - total number of words in the Dataset, Voc - Dataset vocabulary size, Docs - total number of documents

Texts	Min	Max	Avg	Words	Voc	Docs
Dataset	3	461	51	18456	5847	360

TABLE II. Dataset description II. Phrases: Min - minimum number of phrases per document, Max - maximum number of phrases per document, Avg - average number of phrases per document, Num - total number of phrases in all documents, Words - total number of words in all phrases, Voc - vocabulary size for phrases

Phrases	Min	Max	Avg	Num	Words	Voc
Dataset	1	16	5	2068	5275	2553

Table III. Dataset description III. Phrase length: L1, L2,..., L9, L10+ - total number of phrases with the corresponding length: 1,2,..., 9,10+ extracted from the dataset.

Length	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
Num	648	596	370	225	113	51	28	17	7	13

III. Main Annotation Strategies

When compiling the collection the following rules for selecting keyphrases were considered: phrases are formed by sequences of words, and are extracted from the text as is without any changes. We assume that the stage of lemmatization or stemming, if required, is performed by the researcher himself. If the text contains the following information the selected phrases should reflect it:

- the car brand;
- whether the text is about repairing or buying a car;
- emotions;
- information on what was good / bad, does the user recommend (or not) the service / car dealership;
- was it a repair or vehicle inspection, if the latter - what number of the inspection, if there was a repair, then what was being repaired;
- information on price;
- whether a discount was made;
- whether there were queues, waiting time;

TABLE IV. EXAMPLES OF TEXTS AND KEYPHRASES OF THE "GOLD STANDARD"

Examples of texts	Phrases of the "gold standard"(manually assigned phrases)
repaired my AT at the ***** service, good guys, 1 year warranty, did everything that could be done, the transmission is now a fairy tale, no complaints. I recommend. (for those who have a focus, guys the transmission can jerk because of the brain, get it checked, and then check the transmission itself itself). p.s. in Saint Petersburg there is no focus as fast as mine))) p.s.s. with factory settings ...	repaired; ******; service; good guys; 1 year warranty; did everything; the transmission is now a fairy tale; no complaints; recommend
ремонтировал в ***** сервисе свою акпп, ребятки молодцы, 1 год гарантии, сделали всё что тока можно, коробка теперь сказка, нареканий нету. советую. (тем у кого фокус, мужики коробка может дёргаться из-за мозгов, проверяйте их, а потом уже лезьте в саму коробку). p.s. в Питере больше нету такого быстрого фокуса как у меня))) p.s.s. с заводскими настройками . . .	ремонтировал; ******; сервисе; ребятки молодцы; 1 год гарантии; сделали всё; коробка теперь сказка; нареканий нету; советую
I bought my accent in ***** on savushkino. and i can say that I am quite satisfied. didnt impose anything, didn't impose any add-ons, got the car without any problems. another matter - car service in ****** is a total nightmare....	accent; *****; savushkino; bought; quite satisfied; didnt impose anything; didn't impose any add-ons; without any problems; car service in ****** is a total nightmare
я покупал свой акцент в ***** на савушкино. и так скажу, что вполне доволен. Ни чего не навязывали, допы они мне не впаривали, машину забрал без всяких проблем. другой вопрос - обслуживание машины в ****** это полная жуть	акцент; *****; савушкино; покупал; вполне доволен; ни чего не навязывали; допы не впаривали; без всяких проблем; обслуживание машины в ***** полная жуть
changed the bumper covered by insurance at *****. i'm satisfied by the work done .... giving in the car was a pain... ... got it back seems okay and without dirt stains.	changed the bumper; covered by insurance; *****; satisfied by the work done; giving in the car was a pain; without dirt stains
менял бампер по страховке на *****. работой доволен. . . . сдавать машину был целый гемор. . . ... получил кажется нормальную и без пятен грязных.	менял бампер; страховке; *****; работой доволен; сдавать машину был целый гемор; без пятен грязных

- quality of work, quality of service; the name of the service / car dealership, the name of the street and the city where the car service is located; official/non service;
- insurance service or not, under warranty repair or not;
- access to the repair area - allowed / prohibited;
- was the car washed or not;
- diagnostics done or not;
- whether additional equipment was imposed during purchase of the car;
- names / nicknames of managers, chief, mechanics, etc.;
- warranty for the results of work;
- whether there were new scratches or other damages after repair;
- whether the car was recognized as “totaled”.

Prepositions, pronouns and conjunctions within phrases are preserved, including the words “это”, “еще”, “все”(“this”, “still”, “all”), etc. Examples: “ремонтируют мазда и форд”, “пропало на моем фокусе 2 сцепление”, “признавать это не хотели”, “остался я в шоке”, “делают вроде бы нормально”, “устал уже с ними бороться”, “долго все делали” (“repaired Mazda and Ford”, “lost clutch on my focus 2”, “did not want to admit it”, “didn't care at all”, “I'm left in shock”, “the work they do seems to be fine”, “I'm tired already of fight with them”, “it took them a long time to do everything”). “Очень” (“Very”) is usually omitted if it is first word in the phrase. Example: “очень понравился менеджер” (“I really liked the manager”) will be transformed to: “понравился менеджер” (“I liked the manager”). Prepositions, pronouns, and conjunctions at the beginning of the phrases are left in phrases if after lemmatization they remain necessary and removed if they cease to play a binding role (we assume that the results obtained by the annotation algorithm will be compared with the result of manual annotation after the lematization of both annotations). Examples: “в пилот сервисе” will be transformed to “пилот сервисе” (пилот сервис); “из ремонта после дтп” will transformed to “ремонта после дтп” (ремонт после дтп) , “при замене оцарапали стойку” will not be transformed. The numbers are not removed. Examples: “дилер мажор 47” (“dealer major 47”), “2-3 месяца ожидания” (“2-3 months waiting”). The words “dealer”, “car service”, etc. are separated from the phrases, unless it this violates the integrity of the phrase.

When constructing phrases, the phrases are made as short as possible (or divided into several), unless this distorts the phrase or makes it unclear to what object the phrase refers to in the text. Those parts of the phrases that are responsible for the emotions and quality of the process were not removed. “И” (“and”) divides the phrase into two separate phrases, unless it leads to a loss of meaning in both received phrases. Example: “нет очередей и высокий уровень обслуживания” (“there are no queues and high level of service”), these are two phrases: “нет очередей” и “высокий уровень обслуживания” (“there are no queues” and “high level of service”). “Хорошее и быстрое обслуживание” (“Good and fast service”) is one phrase.

IV. DISCUSSION OF AMBIGUOUS CASES

The most complex cases in context of annotations can be divided into two groups. The first group consists of texts with reviews of several car dealerships, usually containing opposite opinions. Normally, they reviews mention something done in one dealership, which did not satisfy the author, and a better experience in solving the same problem at another dealership. In the case of multiple objects mentioned in a review, usually the object with the most descriptions is selected as the main object of the post. And for this object the phrases are chosen following

the general strategy of key phrases selection. The second object is usually described in a small part of the text (several words or 1-2 sentences at most). In this case, for the second object, the phrases are allocated in such a way to include the mention of the object itself, as well as emotions and information on this object. Thus, when reading the extracted phrases it is clear whether they refer to the main object, or to the additional object. Example is presented in Table 5.

TABLE V. EXAMPLES OF TEXTS AND KEYPHRASES OF THE "GOLD STANDARD"

Examples of texts	Phrases of the "gold standard"(manually assigned phrases)
went for car inspection last spring after 10,000, to *****, also got my tires changed at a discount ..... and now in the summer I decided to balance the wheels, though at another place, turned out that the wheels were screwed on in such way, that one bolt had to be torn off. changed the stud at ******, of course not for free, since you can not prove that they were the ones who screwed them on wrong. for 20,000 did the inspection at @@@@@, turned out to be cheaper and better, I saw what they were doing and how. but in ***** they find 1000 reasons for you to not stand there and look. won't go there for inspection anymore, couldn't care less about their warranty!	car inspection; *****; won't go there; tires changed; screwed on in such way, that one bolt had to be torn off; for 20,000 did the inspection at @@@@@, turned out to be cheaper and better
делал то той весной на 10000, в *****, заодно по акции переобувал колеса..... и вот летом решил сделать балансировку, правда в другом месте, оказалось прикрутили колеса так, что один болт пришлось срывать.шпильку менял в ******, разумеется за деньги, им не докажешь, что они прикручивали так. на 20000 тыс. то делал в @@@@@,по деньгам вышло дешевле и лучше,видел ,что делали и как.а в ***** находят 1000 причин, чтобы там не стоять и не смотреть.больше туда не поеду на то, их гарантия мне что шла, что ехала!	то; *****; больше туда не поеду; переобувал колеса; прикрутили колеса так, что один болт пришлось срывать; на 20000 тыс. то делал в @@@@@,по деньгам вышло дешевле и лучше

The second group consists messages written in a conversational style with inconsistencies, torn phrases, and a large number of interjections. There is a workaround to this problem, since the extracted phrases tend to be rather short (mostly phrases consisting of one-two words). Due to this fact, a very small part of the phrases was extracted not as a single sequence of consecutive words from the text, but as a sequence of words after the remove of large pieces of unnecessary information between the main information-bearing phrase words.

V. CONCLUSION

The paper introduces a new dataset for extracting key phrases for the Russian language, describes the basic rules followed when selecting phrases. These rules were developed while working with a significant number of texts and allowed us to resolve ambiguity in the allocation of phrases, as well as helped to identify the main topic units discussed in the texts from the forums. The phrases containing these topic units were considered to be the most important and were included in the annotations.

VI. ACKNOWLEDGMENT

This work was supported by the Committee on Science and the Higher School of the Government of St. Petersburg.

REFERENCES

[1] Hulth, A.: "Improved automatic keyword extraction given more linguistic knowledge". In: Conference on Empirical Methods in Natural Language Processing, pp. 216–223, 2003

[2] Kim, S.N., Medelyan, O., Kan, M.-Y. and Baldwin, T.: "Automatic keyphrase extraction from scientific articles. Language Resources and Evaluation", Springer Kan Timothy Baldwin, 2012

[3] Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: "SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications", SemEval@ACL, 2017

[4] Nguyen, T.D., Kan, MY.: "Keyphrase Extraction in Scientific Publications. In: Goh D.HL., Cao T.H., Sølvberg I.T., Rasmussen E. (eds) Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers. ICADL 2007. Lecture Notes in Computer Science, vol 4822. Springer, Berlin, Heidelberg", 2007

[5] Proceedings of the ACL 2015 Workshop on Novel Computational Approaches to Keyphrase Extraction. The 53rd Annual Meeting of the Association for Computational Linguistics 2015 (ACL 2015), 30-31 July 2015

[6] Popova, S., Kovriguina, L., Mouromtsev, D., Khodyrev, I.: "Stop-words in keyphrase extraction problem", *In: Conference of Open Innovation Association, FRUCT*, pp. 113-121, 2013

[7] Tsatsaronis, G., Varlamis, I., Norvag, K.: SemanticRank: "Ranking Keywords and Sentences Using Semantic Graphs". In: Proc. of the 23rd International Conference on Computational Linguistics, pp. 1074–1082, 2010

[8] Wan, X. and Xiao, J.: "Single document keyphrase extraction using neighborhood knowledge". In: proc. of the 23rd AAAI Conference on Artificial Intelligence, pp. 855–860, 2008

[9] Mihalcea R., Tarau P. TextRank: : "Bringing order into texts". In: proc. of the Conference on Empirical Methods in Natural Language Processing, pp. 404–411, 2004

[10] You, W., Fontaine, D., Barthès, J.P.: "An automatic keyphrase extraction system for scientific documents". *Knowledge and Information Systems* 34, 2013, pp. 691-724

[11] Kazi Saidul Hasan and Vincent Ng: "Automatic Keyphrase Extraction: A Survey of the State of the Art". In: proc. of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1262–1273, Baltimore, Maryland, USA, June 23-25, 2014

[12] Frank, E., Paynter, G.W., Witten, H.I., Gutwin, C., Nevill-Manning, C.G.: "Domain specific keyphrase extraction". In: proc. of the 16th International Joint Conference on Artificial Intelligence, pp. 668–673, 1999

[13] Turney, P.D.: "Learning to extract keyphrases from text". Technical Report ERB-1057, National Research Council, Institute for Information Technology, 1999

[14] Su Nam Kim, Baldwin T., Min-Yen Kan. Evaluating N-gram based Evaluation Metrics for Automatic Keyphrase Extraction // Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Beijing, pp. 572–580, 2010