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.

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 the text is about repairing or buying a car;
- whether the car brand;
- 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 I. 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, VOC - VOCABULARY SIZE FOR PHRASES

<table>
<thead>
<tr>
<th>Phrases</th>
<th>Min</th>
<th>Max</th>
<th>Avg</th>
<th>Num</th>
<th>Words</th>
<th>Voc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>1</td>
<td>16</td>
<td>5</td>
<td>2068</td>
<td>5275</td>
<td>2553</td>
</tr>
</tbody>
</table>

Table II. DATASET DESCRIPTION II. PHRASE LENGTH: L1, L2, ..., L9, L10+ - TOTAL NUMBER OF PHRASES WITH THE CORRESPONDING LENGTH: 1, 2, ..., 9, 10+ EXTRACTED FROM THE DATASET

<table>
<thead>
<tr>
<th>Length</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
<th>L7</th>
<th>L8</th>
<th>L9</th>
<th>L10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num</td>
<td>648</td>
<td>996</td>
<td>301</td>
<td>225</td>
<td>117</td>
<td>310</td>
<td>28</td>
<td>17</td>
<td>7</td>
<td>13</td>
</tr>
</tbody>
</table>

TABLE III. DATASET DESCRIPTION II. VOCABULARY SIZE, DOCS - TOTAL NUMBER OF DOCUMENTS

<table>
<thead>
<tr>
<th>Texts</th>
<th>Min</th>
<th>Max</th>
<th>Avg</th>
<th>Words</th>
<th>Voc</th>
<th>Docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>3</td>
<td>151</td>
<td>51</td>
<td>19456</td>
<td>5847</td>
<td>360</td>
</tr>
</tbody>
</table>
IV. DISCUSSION OF AMBIGUOUS CASES

The most complex cases in context of annotations can be divided into two groups. The first group consists of phrases with an object mentioned in a review, usually the object with the same problem at another dealership. In the case of multiple reviews of texts with reviews of several car dealerships, usually the object with which the placements are chosen following the same problem at another dealership. In the case of multiple reviews, the object with the same problem is usually mentioned in the text.

In the second group, the phrases are chosen following the object with the same problem at another dealership. In the case of multiple reviews, the object with the same problem is usually mentioned in the text.

Whether additional equipment was imposed during the purchase of the car, or the names/nicknames of managers, chief mechanics, etc.;

whether the additional equipment was imposed during the purchase of the car, or the names/nicknames of managers, chief mechanics, etc.;

whether the car was recognized as "totaled";

whether the car was recognized as "totaled";

whether the car was recognized as "totaled";

whether the car was recognized as "totaled";

whether the car was recognized as "totaled";

whether the car was recognized as "totaled";

whether the car was recognized as "totaled";
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"**

<table>
<thead>
<tr>
<th>Examples of texts</th>
<th>Phrases of the &quot;gold standard&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>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. 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! **deval той весной на 10000, в ***<strong>, заодно по акции перебукал колеса.... и вот летом решил сделать балансировку, правда в другом месте, оказалось прикрутили колеса так, что один болт пришлось срывать; на 20000 тыс. то делать в @@@@, но денег вышло дешевле и лучше, идешь и думаешь, что делать и как, в ***** находят 1000 причин, чтобы там не стоять и не смотреть. Больше туда не поеду на то, их гарантия мне что шла, что такая их акция!</strong></td>
<td></td>
</tr>
</tbody>
</table>

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 phases 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**


