Computers in Industry 78 (2016) 16–28

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.elsevier.com/locate/compind

A distributional approach to open questions in market research

Stefan Evert [a], Paul Greiner [a,*], João Filipe Baigger [b], Bastian Lang [b]

[a] Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Department Germanistik und Komparatistik, Professur für Korpuslinguistik (Corpus Linguistics Group), Bismarckstr. 6, 91054 Erlangen, Germany
[b] Rogator AG, Emmericher Str. 17, 90411 Nürnberg, Germany

ARTICLE INFO

Article history:
Received 13 December 2014
Received in revised form 8 October 2015
Accepted 13 October 2015
Available online 29 November 2015

Keywords:
Topic clustering
Distributional semantics
Sentiment analysis
Market research

ABSTRACT

Free-text responses to open questions are a rich and valuable resource in modern-day market research, but often pose problems for a traditional analysis, which requires prohibitively expensive manual coding of topic categories. The Klugator Engine (TKE) is a system for semi-automatic identification, exploration and visualization of topics and sentiment in large collections of such free-text responses or other short text fragments. The system utilizes state-of-the-art techniques of natural language processing and machine learning to transform textual input into a structured corpus, complemented by automatically determined polarity scores for individual responses. Statistical and distributional methods are then applied in order to identify semantic clusters of responses, label each topic cluster with a set of salient keywords, and evaluate the sentiment associated with the topic. This process can run in fully automated fashion, but it also offers the opportunity of interactive parameter tuning and refinement guided by the end user. Results are presented in a concise graphical visualization supported by detailed tables with numerical information. Embedded in RogTCS, the Rogator Text Clustering Solution, TKE enables customers to obtain a good overview of the main topics in a text collection comprising thousands of responses within 20 min of interactive exploration. An evaluation study based on a data set of more than 60,000 word tokens has shown good agreement with the topics identified by manual coding, rendering TKE a powerful tool for the analysis of unstructured textual data.

1. Overview

This article describes the purpose, design and implementation of The Klugator Engine (TKE), a specialized text clustering software for online surveys developed by the Corpus Linguistics Group at Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) in co-operation with local market research company Rogator AG. TKE has already been applied successfully to the analysis of data collected in online surveys, but it could also be used for a variety of other tasks on similar kinds of text data.

1.1. Motivation

A large proportion of the data collected in market research are strictly quantitative in nature, usually obtained from opinion polls with predefined answer options. While such data are easy to analyze with standard techniques of descriptive and inferential statistics, there are two important drawbacks: (i) usually the level of measurement is restricted to categorical and numerical scales, which cannot account for qualitative aspects and complex relationships; (ii) they do not adequately reflect the true variety of respondents' opinions, discarding any information that the researcher designing a poll did not have in mind a priori.

Open questions permitting unrestricted free-text answers give much more fine-grained insights, but they are often shunned as an alternative (or complement) to quantitative polls because their analysis is time-consuming and expensive. All responses have to be examined manually and assigned to topic categories, based on an ad-hoc code plan specially designed for each question. In this way, the qualitative data are transferred into a quantitative representation, which can then be analyzed with the usual statistical techniques. Code plans can rarely be reused without discarding important aspects of the unrestricted answers, and they often need to be revised and extended during the coding process.

With the recent trend towards online polls and other electronic forms of surveys, large amounts of free-text responses can easily be collected in machine-readable form. It is not uncommon for an online poll to produce several tens of thousands of responses. Open

* Corresponding author.
E-mail addresses: stefan.evert@fau.de (S. Evert), paul.greiner@fau.de (P. Greiner), f.baigger@rogator.de (J.F. Baigger), b.lang@rogator.de (B. Lang).

http://dx.doi.org/10.1016/j.compind.2015.10.008
0166-3615/

questions are sometimes included merely to improve the user experience, and are then discarded when analyzing the survey.

TKE was developed in order to enable market researchers to exploit this wealth of qualitative information. It is able to carry out a mostly automatic identification of key topics within a large set of unstructured free-text answers to open questions, to determine the general sentiment expressed towards each topic, and to visualize the quantitative distribution of topics among the answers. The software is already in commercial use as the core engine of the Rogator Text Clustering Solution (RogTCS), distributed by German market research company Rogator AG.

1.2. Goals

Key requirements for TKE include (i) fast automatic analysis (so that the open questions of a typical poll can be explored interactively by customers), (ii) ease of use (as RogTCS is operated directly by customers), (iii) domain independence (so that all open questions can be analyzed) and (iv) support for multiple languages (for use with international surveys). Therefore, the system has been designed to use as few language- and domain-specific resources as possible, which also helps to avoid steep licensing costs for ontologies, sentiment lexica, etc. Instead, the main text clustering component of TKE builds on unsupervised statistical procedures that have been carefully tuned in order to give a faithful representation of the content of the input data, condensing the main topics and sentiments into a manageable amount of graphs, numerical indicators and tables.

In most cases, TKE is able to generate satisfactory results without any manual intervention, so that a rough analysis of a text collection can be carried out within a few seconds. If desired, this initial analysis can afterwards be refined by users in an interactive, semi-automatic procedure.

1.3. Scope of application

TKE is commercially available, providing the core engine of RogTCS, a tool for fast and cost-effective analysis of open questions in market research.[1] Therefore, the typical text material processed by TKE at this point consists of short textual answers and comments from online surveys concerned with a range of different products and services. Texts may be in different languages, with a main focus on German and English. RogTCS offers market researchers an alternative to the time-consuming and expensive manual coding of answers to open questions, which involves the creation and application of a specialized code plan for each question.

While optimized for online surveys, TKE can also be applied to other kinds of digital collections of natural language data, provided that they consist of short, focused texts and contain enough material for a statistical topic analysis. Possible use cases are, among others, messages (tweets) sent via Twitter or similar micro-blogging services, online discussion boards, customer product reviews, as well as other material obtained from various social media platforms.

1.4. An example session

In order to give readers a better idea of the functionality of TKE and the quality of its results, this section presents an example analysis based on a real-life data set. The data comprise 4627 responses to an open question posed in an online survey about the redesign of an e-mail provider's Web site, written in English. The average length of a response in this data set is 13.7 words, resulting in a total size of 63,314 word tokens. The data were analyzed with the most recent TKE version (as of December 2014), coupled with a Web-based GUI that has also been used for development and in-house testing of the engine (see Section 3.4 for more information).

After uploading the text data, TKE carries out a fully automated analysis and presents its results in the form of a semantic map, as shown in Fig. 1. The quality of this initial analysis depends strongly on the quality of the input data – the statistical topic identification works best with focused responses to a clearly phrased question – and might not always be as neat and easily interpretable as the example presented here. A key parameter is the number of topic clusters, which has to be specified by the user in the current TKE implementation. The result in Fig. 1 was achieved after a few minutes of manual experimentation by reducing the number of clusters from its default value of 20. The interactive development GUI, which is similar to the RogTCS user interface, allows users to explore different settings for a wide range of system parameters with ease. Note, however, that only the number of clusters – a simple and intuitive parameter – had to be adjusted in this example.

Each circle in Fig. 1 represents one of the automatically identified topic clusters, labelled with single words or bigrams (pairs of consecutive words) that occur frequently in responses assigned to the corresponding cluster. While labels are not perfect, it is usually easy for human users to infer the main topic of a cluster. For example, the labels *like*, *new*, *service* clearly indicate that respondents enjoyed the new product. Multiword expressions consisting of more than two words (e.g. *good spam filter*) are broken down into multiple labels (*good spam*, *spam filter*) that can easily be pieced together by users. If necessary, users can request additional label suggestions or directly inspect the answers assigned to a topic cluster.

The area of each circle reflects the number of responses assigned to the topic, which we refer to as the *mass* of the topic. The colours of the circles indicate the overall sentiment expressed towards topics, using a colour palette ranging from green (for positive sentiment) to red (for negative sentiment). Neutral or mixed sentiment corresponds to a yellow hue. The grey circle in the centre stands for responses that could not be assigned reliably to one of the topic clusters, e.g. because they are formulated in an unusual way, express a very infrequent opinion, or consist of meaningless text. In this case, the data set contains 3698 well-formed responses (excluding e.g. empty answers), out of which 3156 were automatically assigned to one or more topic clusters by the engine.

Fig. 2 gives an alternative overview of the main topics in tabular format, including detailed quantitative information. Market researchers are particularly interested in the importance of each topic as represented by its mass, i.e. the number of responses in which the topic was addressed. Notice that these counts do not add up to the total number of texts because TKE can assign a response to multiple topic clusters. Here, 829 responses were (at least partially) assigned to the most frequent topic (C11: *ease of use*); 1028 responses contain a text fragment that does not fit any of the topics and are thus assigned to the residual cluster R01 shown as a grey circle in the map; and 1193 responses are either empty or include a fragment that cannot be analyzed by the system at all (because all words have been filtered out by the stopword list and frequency threshold, cf. Section 3.2). Additional columns indicate the quality of each topic cluster (based on the homogeneity of the corresponding text fragments) and the average sentiment towards the topic (cf. Section 3.1).

Some of the responses from topic clusters C09: *web interface, good interface, user interface* and C06: *time, problems, signed* are shown in Figs. 3 and 4. The responses are ranked by their

[1] https://www.rogator.de/software/textanalysesoftware.html.