Corpus Based Studies on Language Expression of Opinions

Zygmunt Vetulani (1), Marta Witkowska (1), Suleyman Menken (2)
(1) Adam Mickiewicz University in Poznań, Poland
(2) University of Kocaeli, Turkey

Abstract

Opinion processing is, since recently, in focus of interest for computational linguists, public relation experts, marketing companies and politicians. Studies of natural language expression of opinions, desires, emotions and related phenomena require tools and methodologies. Within the empirical approach to these studies, we propose tools for collection of the empirical data in form of a corpus. We limit our research field to the customers' written opinions concerning widely used on-line booking services in the area of hotel reservation (via Booking.com). In the paper, we present the corpus acquisition procedure and our data acquisition tool. Decision concerning selection of source data will be discussed. We also present some limitations of our proposal and propose validation methodology for the resulting corpora.

Keywords: text language resources, opinion processing, corpus-based methods, multilingual corpora, customers' opinions

Introduction

Why is the opinion processing so important? One of possible answers to this crucial question results from the analysis of the social role of opinions, and in particular of their contribution to the evolutionary success of the human kind. This success is based on the aptitude of taking right decisions especially in the situation of a limited access to facts. Opinion, according to a popular definition "is a judgment, viewpoint, or statement about matters commonly considered to be subjective" (Wikipedia, "Opinion"), in opposition to facts considered as verifiable. Three points worth noticing are:

- decisions are being taken on the basis of premises that may be opinions, in particular when facts are not available,
- quality of decision are in obvious relations with the quality of opinions used as decision premises,
- opinions are often emotionally biased, and this bias has an impact on their quality.

Analysis of the emotional layer of opinions used as decision premises may help evaluating the appropriateness of the decision (risk analysis).

Some application fields of opinion processing are:

- political decisions, election campaigns, business PR campaigns,
- questionnaire-based customer profiling, forensic profiling.

The concept of opinion

The sense that we give in this paper to the term *opinion* is close the one that is defined in the popular Collins English Dictionary (Collins English Dictionary - Complete & Unabridged 2012) as "judgment or belief not founded on certainty or proof". We propose to make it more precise (after Charaudeau and Maingueneau 2002) saying that *an opinion expresses a subject's evaluative opinion in favor or against facts*.

Project objectives

The medium-term aim of the project is to create a repository of comparable (size, domain, acquisition mode, nature of texts) opinion text corpora for different languages as a tool for opinion studies (fundamental – descriptive and/or comparative – research on opinion expression in natural language for various languages).[1] The domain (set of domains) should be defined precisely, and corpora should be large enough to allow qualitative observations in order to give insight in the distribution of language phenomena. The resources should be free of legal flaws related to the acquisition procedures and should not have usage restriction for linguistic research purposes.[2] Instead of proposing a closed set of texts for an *a priori* defined list of languages, we decided to design and implement a software tool to compile a corpora of customers' opinions in the area of hotel services. As a source of data, we selected the popular service Booking.com.

Why did we choose hotels?

The choice of hotel opinions as corpora acquisition domain was motivated by several factors, of which the most important was the common (at the world scale) custom of acquiring and publically displaying via the Internet the client opinions (both positive and critical) by the booking portals. We may therefore expect a large volume of accessible data (over 50M opinions accessible

[1] Not many opinion corpora exist. One of the best known is the MPQA Opinion Corpus of English texts (University of Pittsburg, PA, USA), http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/; Stoyanov et al. 2004). Cf. also the a five billion word Corpus of Japanese blogs annotated for affective features (Ptaszyński et al. 2012).

[2] This is the case of Booking.com guests' comments which are not copyright protected elements of the content but just publically presented opinion recordings.

through Booking.com). Opinions are accessible for quasi all languages used by customers.
Although the basic function and structure of hotel services is practically the same in almost all cases, realization modalities and customer expectations may be strong culturally biased and therefore one may *a priori* expect large variety of the content. A positive thing about this choice is the ideology- and world-view-neutral nature of the field of hotel opinion expression. This makes that we do not touch the taboo areas[3] that could disturb acquisition of well-balanced data.

Why did we choose Booking.com?

Nowadays, providing customers' opinions is a common custom in the internet commerce, in particular in the sector of services. Acquisition of such opinions for research purposes is easy. What is more questionable is the data quality and honesty, i.e. authenticity and representativeness. The latter feature means equal access to all opinions regardless their meaning. As long as the opinions are presented by the concerned subject, the risk of manipulation (lack of honesty) should not be ignored. Such a risk is minimal, if the opinions are being gathered by a big, possibly global, hotel reservation broker. It is easy to list a number of such operators. The (alphabetic) list of corresponding portals is given below:

- www.booking.com
- www.hotels.com
- www.hrs.com
- www.tripadvisor.com
- www.trivago.com
- www.worldhotel.com

We selected the popular hotel reservation operator Booking.com. Arguments in favor of this choice are as follows:

- Booking.com provides a huge amount of opinion texts (59,020,000 of verified opinions)
- Booking.com is a global operator covering practically all countries,
- Booking.com allows opinions in all languages,
- Opinion texts are easily accessible,
- Opinions are pre-classified into positives and negatives by the opinion providers (i.e. hotel guests) (see Fig. 1),
- Basic information concerning hotels is available (address, category, price, facilities),
- Availability of comparable data for practically all languages spoken by the users (possibility to create comparable corpora for a large variety of language pairs).

Pre-classification of opinion texts is highly useful when introducing pragmatic information to formalized lexicons as WordNet-like lexical databases (e.g. SentiWordNet (Esuli and Sebastiani 2006); cf. also (Vetulani et al., in preparation) for the discussion of wordnet granulation issues in PolNet). Pak and Paroubek (2010) used emoticons (following similar procedures as in (Read 2005)) as pre-classifiers used in their Twitter-extracted corpus for sentiment analysis and opinion mining.

Of course, Booking.com is not the only possible source of a good corpora and it is not our intention to prove that it is the best one. To check the completeness (representativeness) of the corpus it is useful to apply the chosen acquisition method to other thematic domains (e.g. other kind of services). A good candidate for this research could be, e.g. the Tripadvisor, known for offering a wider spectrum of services addressed to tourists and travelers (including also flights and restaurants).

General characteristics of the available data

The main data of interest for us available from Booking.com are the opinion texts. According to the declaration of the Booking.com service, over 59,020,000 of verified opinions[4] are presented to the clients. Verification means here first of all the authenticity tests (in order to stop provision of abusive opinions it is assumed (and checked) that real hotel guests are opinion authors). The negative about verification practiced by the Booking.com staff is censorship in order to eliminate the use of naughty words, discriminatory remarks, swearing or offensive language. Because of that, the validity of the resulting corpus is limited to the neutral (and high) language register.[5] This rule seems systematic, proper to all subjects publically presenting opinions of other subjects and difficult to avoid. Therefore, it must be taken into account when using the corpus for the study of language expression of emotions. The way to present opinions by Booking.com has interesting properties. It presupposes opinion pre-classification at the acquisition time, as the guests are requested to provide separately positive and negative observations. This feature is precious because it reduces the necessity of intention analysis which would be necessary in order to identify and classify the opinions in the unstructured text. Another positive aspect of the collected text resources is the possibility of partial reconstruction of the context of the situation inspiring the opinion. Although the opinion authors are practically anonymous (they are presented by (first) name (or nickname) and the (declared) provenance country, some relevant information to help interpreting the opinions are supplied (and easy to acquire). These are: the hotel location, the Booking.com ranking (stars and scores calculated on the basis on guests' opinions), information concerning the offer and extras.

[3] Such are often political, religious or custom-related opinions.

[4] All data presented in the paper about Booking.com were acquired in November, 2015).

[5] To get a more precise idea on the nature of these limitations we suggest the Reader to consult *Booking.com Guest Review Guidelines*. To find it, open Booking.com and follow the path: home → … → any hotel → … → hotel reviews; then click *read more*

Figure 1. An example of the Booking.com opinion record completed by the guest

Opinion Corpora Acquisition Software (OCAS)
Specialized software for acquisition of opinions and building corpora assembling opinion texts has been created in 2014/2015 at Adam Mickiewicz University in Poznań (AMU, Faculty of Mathematics and Computer Science). The system OCAS permits to collect opinion texts from Booking.com in a few steps which are implemented by a team composed of visiting Eramus students of computer science (Süleyman Menken, Emre Çelikörs, Veysi Ozan Dağlayan from Turkey and Arcaeli Martinez, Adrian Barreiro Vilalustre from Spain) in collaboration with the Polish students of linguistics Marta Witkowska and Urszula Morzyk, under the supervision of Zygmunt Vetulani (AMU).

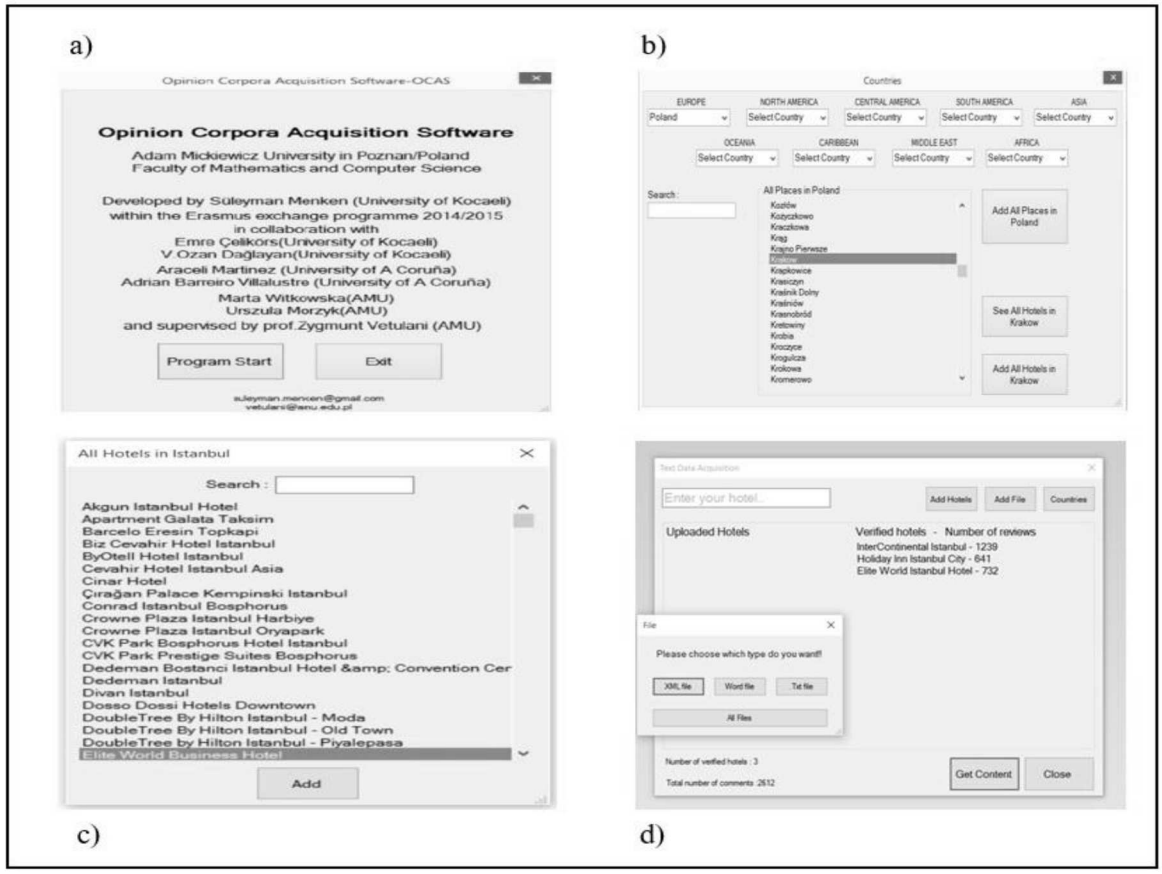

Figure 2. The OCAS screenshots.

presented on Figure 2. First, the user chooses the country and city of interest, then the hotels are being selected. Finally, the user decides on the file format in which the opinion data are to be downloaded (XML, Word .doc file or raw .txt format). OCAS was designed and

The Figure 2 above presents the OCAS screenshots (selection) illustrating data acquisition a) start screen, b) selection of the city (in Istanbul), c) all hotels in Istanbul, d) three selected hotels in Istanbul; using the 'Get

Content' button we can get the opinion texts in any off the following 3 formats: XML, Word, and .txt.

The final data in the XML format looks like presented in Figure3.

6.Evaluation method: saturation tests

For both development and evaluation purposes we selected Turkish and Polish hotels. The first evaluation trials were performed for Turkish. Turkey is a large Euro-Asiatic country with high number of places interesting for tourists and businessmen looking for hotel services at different standards and price categories. Booking.com presents opinions for more than 11,000 hotels or other properties around the country[6], with a major part being written in the local language (usually Turkish) or English. The test corpus was created for opinions of about 300 semi-randomly selected hotels from Booking.com. The first step consisted in selection of the most frequented (for business or tourism) places. When selecting hotels, special attention was paid in order to have a balanced representation of hotels of different categories (from 1 to 5 stars). For each hotel, 1 to 5 opinions of ca. 19 words or more were selected.

Differentiation in hotel locations and categories was intended to make the test-corpus representative for a rich variety of phenomena.

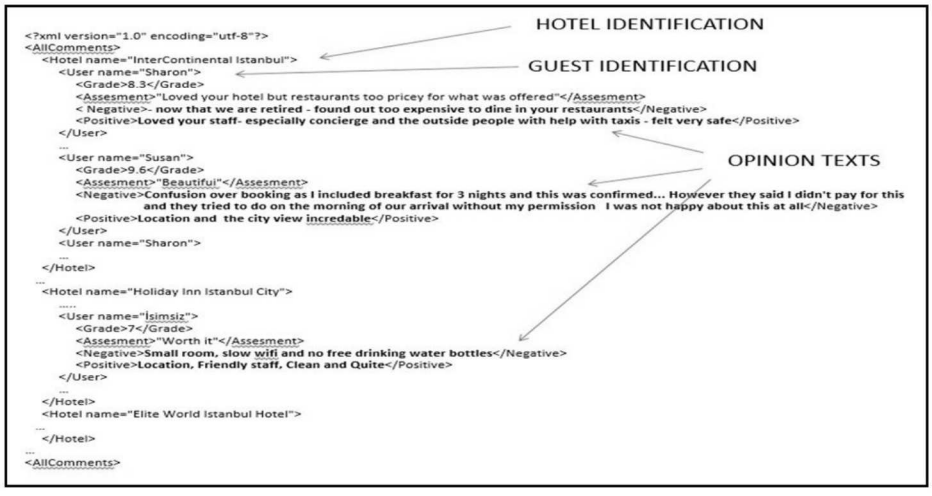

Figure 3. A fragment of the XML formatted corpus presenting three different opinions about hotels in Istanbul.

The fundamental question was when to stop constructing the corpus, i.e. adding new opinions and enlarging the test corpus.

This problem is to be formulated with respect to some pre-defined class of phenomena. Our approach consists in exploring the concept of corpus saturation (Vetulani 1989). This concept was inspired by the research on the evaluation of the size of virtual vocabulary of sublanguages[7] in the 1980s (e.g. in the context of machine translation) and was used to study lexical saturation of corpora (ibid.). Informally, we say that corpus is lexically saturated when "new lexemes appear only sporadically as a result of the extension of the corpus in a natural way" (ibid.). In order to study the lexical saturation of the corpus it is useful to observe the increase of the number of new words corresponding to the increase of the observed fragments of the corpus. In order to evaluate the minimum size of a balanced corpus which could pretend to be considered as representative for an *a priori* given class of phenomena, we did an experiment involving adjectives. Observing how adjectives occur in the corpus was motivated by their importance for opinions wording. For the corpus of 1185 opinions for 300 hotels we observed 131 adjectives ("good" and "bad") and we drew a vocabulary growth rate curve for these adjectives

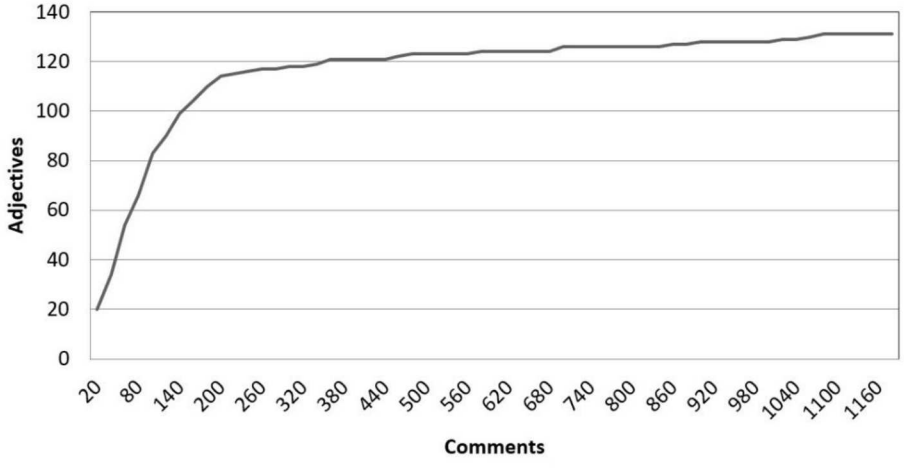

Figure 4. Vocabulary growth rate curve for adjectives

(Fig. 4). What we observe is a very slow, linear increase of the number of observed adjectives after having scanned ca. 20% of the corpus (60 hotels).

[6] 6340 hotels in 28 the most visited cities.

[7] See e.g. (Kittredge 1983).

Of course, this observation does not need to be valid for other categories or phenomena. The only legitimate conclusion is that for some interesting ones, the representative corpus may be surprisingly small.

7.Final remarks

Booking.com is not the only possible source of a multilingual opinion corpus and it is not our intention to prove that it is the best such a source. To check the completeness (representativeness) of a corpus it is useful to apply the same acquisition method to another thematic domains (e.g. other kind of services) and compare the results. We intend to proceed to such evaluation exercise. A good candidate for a texts source may be, e.g., the Tripadvisor information service, known for offering a very large spectrum of services for travelers (including also flights and restaurants).

Among the research tasks related to the concept of *opinion,* the corpora obtained using the tools presented in this paper are intended to serve in the studies of the research areas like:

- description of possible ways people express their opinions (in language),
- description of the emotional content of the opinion texts,
- study of various socio-cultural factors and impact they have on the cross-language and cross-ethnic comparability of opinions.

These studies are to be carried out in the future and are not covered in the present paper.

References

Charaudeau, P. and Maingueneau, D. (2002): *Dictionnaire d'Analyse du Discours*, Seuil

Esuli, A. and Sebastiani,F. (2006): SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining, In: *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06),* 417-422.

Collins English Dictionary - Complete & Unabridged 2012 Digital Edition; © William Collins Sons & Co. Ltd. 1979, 1986 © HarperCollins Publishers 1998, 2000, 2003, 2005, 2006, 2007, 2009, 2012.

Kittredge, R. (1983): Semantic processing of texts in restricted sublanguage, in: *Computers & Mathematics with Applications*, Vol. 9, Issue 1, 45–58Pak, A.,

Paroubek, P. (2010): Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: Nicoletta Calzolari et al. (Eds.). *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. European Language Resources Association 2010, 1320-1326.

Ptaszynski, M., Rzepka, R., Araki, K., Momouchi, Y. (2012): "Automatically Annotating a Five-Billion-Word Corpus of Japanese Blogs for Affect and Sentiment Analysis." *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Jeju, Republic of Korea, 12 July 2012.* Association for Computational Linguistics, 89–98,

Read, J. (2005): Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In: Kevin Knight et al. (Eds.) *43rd Annual Meeting of the Association of Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan.* The Association for Computer Linguistics.

Stoyanov, V., Cardie,C., Litman, D., and Wiebe, J. (2004): Evaluating an Opinion Annotation Scheme Using a New Multi-Perspective Question and Answer Corpus. Working Notes of the 2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications

Vetulani, Z. (1989): *Linguistic Problems in the Theory of Man-Machine Communication in Natural Language*, Universitätsverlag Dr, N. Brockmeyer, Bochum

Vetulani, Z., Vetulani G. and Kochanowski, B.: Recent Advances in Development of a Lexicon-Grammar of Polish: PolNet 3.0. (in preparation).