Development of a Brazilian Portuguese Hotel's Reviews Corpus

Joana Gabriela Ribeiro de Souza, Alcione de Paiva Oliveira(✉), and Alexandra Moreira

Universidade Federal de Viçosa, Viçosa, MG 36570-900, Brazil
joana.souza@ufv.br, alcione@gmail.com

Abstract. The provision of voluntary textual information mediated by the Internet, and particularly by Web 2.0, provided an opportunity for the creation of large linguistic corpora. These corpora can serve as a fundamental resource for the development of applications focused on natural language, especially those using deep learning techniques that require big datasets. One type of application that benefits from these resources is the ones that perform sentiment analysis. This article describes the creation of corpus aimed to support sentiment analysis applications. It consists of reviews hotels located in the Brazilian capitals and the Federal District, written in Brazilian Portuguese language. The reviews that make up the corpus have been taken from TripAdvisor and have undergone normalization and POS tagging. The primary goal is to make it available to the community to be used in machine learning tasks geared toward natural language.

Keywords: Linguistic corpus · Portuguese corpus · Hotel's reviews Sentiment analysis

1 Introduction

The Web allows users to interact, collaborate with each other and share information on a variety of subjects. There are digital spaces such as social networks, forums, e-commerce sites among others where people usually express their opinions. In these environments the users usually make use of an informal language, being very common the use of slang, abbreviations, different spellings of words existing in the grammar, besides the creation of new terms. These are some of the challenges encountered by researchers in compiling corpus with data scraped from the Internet.

Corpora are key resources for the training and testing of applications focused on natural language processing. Nevertheless the creation of these features can be quite complex and time-consuming. There are several compromise decisions and processing steps, such as normalization and annotation. According to [4] the Web has been widely used as a corpus, due to the amount of data available in several languages and textual genres, free of charge and easy to access. [6] presented the

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 353–361, 2018.
https://doi.org/10.1007/978-3-319-99722-3_36

difficulties and effort involved in extracting information from the Web. In order to extract and store textual data from the Internet, in order to become useful for NLP tools, it is necessary the adoption of Web scraper tools and to apply several pre-process steps to the data to remove undesirable "noises". Occurrences such as the use of acronyms, terms in different languages, emojis, non-formal use of the Portuguese language among other situations are common in texts written by people on the Internet. Patil [6] also points out the challenges for extracting Web texts due to the variety of content and formats available that vary from one genre to another (from a social network to a government website, for instance). Meyer et al. [5], listed important details to consider when deciding the use of the Web as a corpus, such as selecting the appropriate search tool, the necessary pre-processing, among others. [2,3] also point out interesting normalization that were taken in account in our development of a corpus composed of reviews in Brazilian Portuguese.

This work was motivated due to the need for Brazilian Portuguese corpora for performing sentiment analysis. For this task, we choose hotel reviews to make up our corpus considering its availability on the Web. The textual information was collected from a well-known tourist attraction review site and went through several stages of processing, from the adjustments to the text to the addition of syntactic annotation, which will be described in this article. The corpus developed was made available on the Internet in order to help meet the need for corpus in Brazilian Portuguese.

This article describes the steps performed in the development of a corpus of hotel reviews in Brazilian Portuguese for performing sentiment analysis. Section 2 presents some related work. Section 3 describes the compilation of the corpus composed of texts written by people who have stayed in hotels in the 26 Brazilian capitals and the Federal District. Section 4 presents the process of analysis and annotation of the types of occurrences that originate terms outside the vocabulary in this type of text. The procedures involved in pre-processing and normalization are described in Sect. 4, and finally we present a brief conclusion along with the indication of future work in Sect. 5.

2 Related Works

The construction of corpus is one of the fundamental tasks of the natural language processing area. Here we will present some work related to the development corpus written in the Portuguese language.

[2] points out some questions about the standardization of a product reviews corpus in Portuguese. On this corpus the authors performed the semantic role labeling, sentiment analysis, classification and summarization. They used the MXPOST tagger for the part-of-speech (POS) annotation, and used a small portion of the corpus for measuring the accuracy. Afterwards, they manually created 4 golden corpora following the spelling normalization (including foreign words and named entities); case; punctuation; and use the of Internet slang. They observed that case information got the highest correction rate, albeit they

expected that slang would have greater impact (only 0.24% of the corpus sample). They selected four tasks to be performed: (1) true casing normalization using Named-entity recognition (NER) as one of the main strategies; (2) punctuation problems correction; (3) spelling correction using Unitex and then manually checking common words to evaluate if the word had been accurately corrected, and; (4) Internet slang normalization, where the words were categorized as written different from cultured norm and abbreviations, repetition of signs and letters, and sequences related to emotions (such as emoticons).

[3] described how the corpus used in [2] has been compiled and the strategies used to normalize it. They defined 8 categories based on the types of noise found in the corpus: misspellings, acronym, proper name, abbreviation, Internet slang, foreign word, unit of measure, and unrated or doubtful tokens. As a result, they made lists with the most common terms for each category and concatenated it in a tool where it is possible to carry out the normalization of a text (especially in the context of product reviews). In the list they identified and corrected acronym, proper name, abbreviation, Internet slang, as well as the spelling according to the spell checker Aspell.

[1] presented two types of normalization (destructive and non-destructive) and an architecture developed by them in order to normalize a corpus in German, without losing information that, for a POS tagger, can be considered "noises", but may give important clues about non-standard language. The architecture was based on the use of two normalization techniques. First a destructive normalization was performed, using HyDRA that unite a rule pattern with n-grams frequency to define when a word actually contains a hyphen, correcting it. Subsequently a non-destructive normalization that aims to maintain the "noise" of the corpus by rewriting words that are emphatically written ("looooooove") or with typing errors. It uses an annotation layer, so it doesn't lose information that may be useful for sentiment analysis or for discovering new aspects of the language, for instance.

In [7] it is described how the corpus CETEM/Público was created (Corpus of Extracts of Electronic Texts MCT/Public). A corpus developed with the support of the Ministry of Science and Technology (MCT). The CETEM/Público was created from journalistic texts from newspaper "O público" founded in 1990. The newspaper is written in Portuguese, being almost exclusively from European Portuguese texts, with exceptions of texts written by Brazilians and Africans. The creation steps involved cleaning of texts of images, subject classification, and sentence separation using the program library developed in the AC/DC project. The result was a corpus of 180 million words. The main difference is from the present work is that the corpus is not aimed to sentiment analysis.

3 The Corpus Development

As the source of textual information to build up our corpus of hotel reviews we have chosen the site TripAdvisor[1], since it is one of the most used sites,

[1] https://www.tripadvisor.com.br.

by travelers, to evaluate not only hotels, but also several other types of tourist attractions and related services. Many tourist related establishment presents at their front door or reception desk the seal of recommendation and/or the quality certificate giving by TripAdvisor. Our corpus is only composed of hotel reviews, so we will always talk about this evaluation context in this paper. On TripAdvisor when making a review, users should enter the number of circles (meaning similar to stars) corresponding to the overall evaluation of the hotel, give a title to the evaluation (which can be understood as a summary of the review), the evaluation (with at least 200 characters where one can give more details about the stay), choose the month and year of the visit, as well as other non-mandatory information. We collected four information from the evaluations: the number of circles, the title, the evaluation and instead of the date of the stay, we collected the date on which the evaluation was performed.

The data were collected only from accommodations classified as hotels. The reviews were collected from February to March 2018, so the most recent review dates from March 20, 2018. Reviews were taken from hotels of the 26 capitals of the Brazilian states and the Federal District as well. We chose to collect reviews of hotels in the capitals in order to have a clear criteria of delimitation of the number of cities and to cover all Brazilian states as well. We gathered a total of 730,069 reviews. Until this moment the corpus had not gone through any linguistic pre-processing, being just removed the HTML tags. Using the NLTK (Natural Language Toolkit)[2], a Python library for NLP, it was verified that the corpus contained 55,950,007 tokens and 457,337 types. Among the most common words are: hotel (hotel), não (no), bem (well or good), manhã (morning), bom (good), quarto (bedroom), café (coffee), localização (location), atendimento (attendance or service) and excelente (excellent). What is expected given the context of the evaluations. In contrast, the less used terms refer to writing errors, very common due to the different levels of literacy of the users.

The Fig. 1 presents in summary form the steps followed by us for the development of the corpus. Firstly, the capture of hotel reviews yielded four files (dates, notes, titles and comments) for each hotel of that capital. After this process, we gathered all the files by city, by region and later in only four files containing all the data collected row by row sequentially, in such way that the i-th line of the dates file corresponds to the evaluation date that is on i-th line in the comments file and so on. After this junction we performed the first normalization where we remove all the HTML tags and in the case of the dates we converted from the format “2 de janeiro de 2018” (January 2, 2018) to “2/01/2018” and for grades we converted the internal code used by TripAdvisor (from “bubble_10”, to “bubble_50”) to numbers from “1” to“5”. Later we did the second normalization where we tried to make several non-destructive corrections according to [1] resulting two versions of the comments: one with the normalizations for tokenization described in the later section and another also without the stopwords.

[2] https://www.nltk.org.

We provided four files (comments normalized, dates, grades and titles) with free access for the community[3].

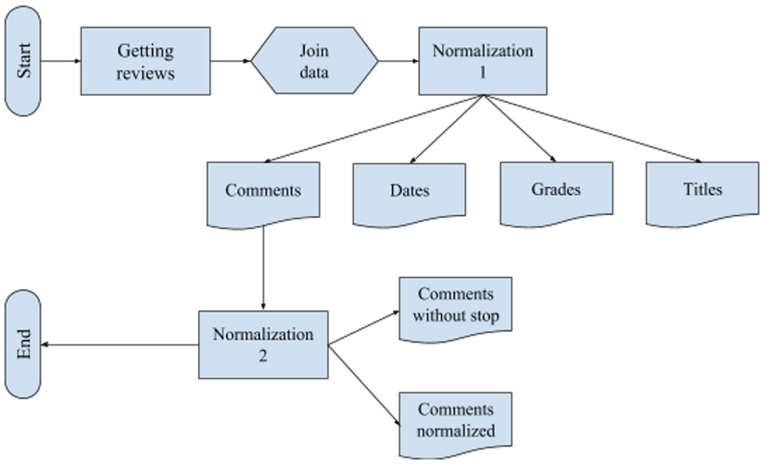

Fig. 1. The steps of the process of the corpus creation and normalization.

4 The Corpus Analysis and Normalization

We used NLTK to get a general idea of the size of the corpus and we studied the related works in order to establish a basis for normalizations that could be applied. The first normalization we did in the corpus was the removal of the excesses of punctuation and sequence of repeated letters that did not form words. As TripAdvisor requires the user to write a comment of at least 200 characters, in several occasions the users completed the comments with punctuation sequences and random characters that did not form words, so we removed several of these occurrences. We kept reticence and sequences of up to three exclamations or questions marks (which may have some meaning when it comes to the sentiment analysis). We have developed a lexical dataset to help us to reduce the number of words that were linked to each other (e.g. "Bomcafédamanhã"). We separated terms such as numbers or hyphens (preceded and followed by spaces) linked to words (e.g. "8 Limpeza" to "8 Limpeza" and "-Gostei" to "- Gostei"). In addition, we kept the words the way they were written, even if incorrectly, due to typing errors or intentionally. So we kept terms with "adoreeeeei" and "lliiiixxxoooo" (similar to "I loveeeeed it" and "trrraaaaaassshh", respectively). On the Internet it is common to write uppercase terms as a way of emphasizing something either positively or negatively, for this reason we also kept the texts capitalization intact.

[3] The corpus files are available at: https://bit.ly/2JVRJbI.

We observed that the corpus had several types of errors in the formatting of the words that prevented them from being tokenized correctly, and those errors did not fit the types of errors or "noises" mentioned previously. Table 1 indicates examples of occurrence that were very common in the corpus and the correction made the tokenizer more efficient. After this normalization the number of tokens increased (even with previously destructive normalization) and the number of types was considerably reduced since the corrections of these occurrences produced more words that could be recognized and counted correctly. By doing some empirical tests we noticed that NLTK tokenizer fails in some common cases, as in the examples shown in Table 1. We also tested the Spacy[4] tokenizer which is an API also developed in Python for NLP which according to developers is the fastest tool and provides the most features up to date. In addition, Spacy supports deep learning, which is a hot topic these days. However, considering the cases in Table 1 Spacy does not separate all the tokens, even though it separates a few more cases that the NLTK is not able to handle ("calçada.." tokenized for ["calçada",".","."] by Spacy and ["calçada.."] by the NLTK), but we have corrected many of these problems with the standardization process. The reason why we continue to use NLTK as a NLP tool in this work is due to its approach of treating words that contain hyphens. NLTK keeps the term as a single token ("wi-fi" tokenized to ["wi-fi"]) while Spacy treats it as distinct tokens ("wi-fi" tokenized to ["wi", "-","fi"]). As one of the objectives of the work is to make the corpus available to the community, we opted to keep these terms in this way. After picking the tool and carrying out the normalizations described here, we have obtained a count of 56,743,114 tokens and 246,307 types. Considering stopwords and punctuation signs, we can see that each review, on average, consists of about 77 tokens, with the largest review having 2,857 tokens and the lowest having only 2 tokens.

Table 1. Some occurrences found in the corpus

Occurrence	Correction
Muito.Porém	Muito.Porém
Residência..	Residência...
*apartamentos	*apartamentos
Custo/benefício	Custo/benefício
2 km	2 km

As them main purpose of the corpus is to be used in sentiment analysis applications, words such as "não" and"sem" (no and without, respectively) can change the meaning of the phrase by inverting its polarity, for example (in sentiment analysis the polarity of a term is basically its classification between the classes: positive, negative or neutral). Figure 2 contains examples of phrases found in the

[4] https://spacy.io.

corpus that when removing these words have their polarity altered. Due to this type of event, stopwords that could indicate change of polarity, intensity or clues to the next sentence classification were maintained in the normalized corpus.

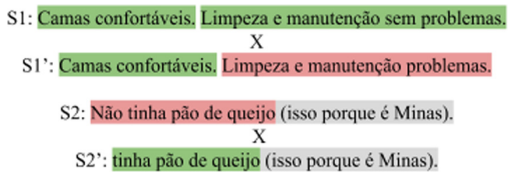

Fig. 2. Examples of polarity changes by removing "sem" and"não". Where green, red and gray colors mean positive, negative and neutral polarity respectively. (Color figure online)

Thus, by manipulating the set of Portuguese stopwords incorporated in NLTK we kept the terms "não","mais", "mas","muito", "sem" and"nem" (no, more/plus, but, much/very, without and nor/neither, respectively). After this normalization we created a corpus of reviews without stopwords with 39,165,169 tokens. After these normalizations in the reviews, we produced a set of four files that can be used by the community for various purposes (we do not provide the stopwords file since we do not remove all of them from our corpus), as well as serve as a resource for our own research in sentiment analysis.

After some standardization, some analysis was done on the content of the corpus. The reviews are divided into five classes: horrible, bad, reasonable, very good and excellent. The Table 2 presents the distribution of the reviews in the five available classes, showing that there is a considerable imbalance between the negative, neutral and positive classes, whereas the negative classes (horrible and bad) together correspond to only 7.1% of the corpus, 16.6% of the reviews are neutral (reasonable) and 76.2% are positive (very good and excellent). Depending on the application, it would be interesting to balance these classes or make some kind of compensation in the machine learning algorithm.

Table 2. The table shows the unbalanced distribution of the reviews among the 5 classes

Class distribution	
Class	Percentage
Horrible	2.8%
Bad	4.3%
Reasonable	16.6%
Very good	40.2%
Excellent	36.0%

Although we have selected only Portuguese comments on the TripAdvisor site, the resulting corpus still contains emojis and many words in other languages, mainly terms in English as well as Chinese and Spanish. Several of these reviews date back to the 2014 World Cup, which brought tourists from all over the world to host cities. We emphasize that several of these occurrences mix Portuguese with other languages, and we chose to keep these terms as they were written, with no translations.

The Fig. 3 displays a graph with the corpus most common words, after eliminating the punctuation marks. Not surprisingly, prepositions are among the most common words. Furthermore, other frequent words are highly related to the

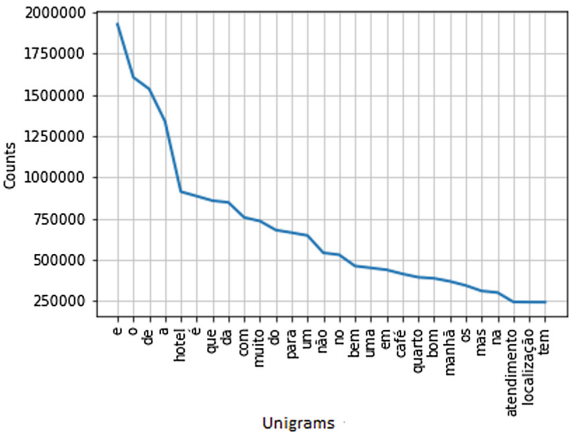

Fig. 3. Graph with the most common words/unigrams in the corpus.

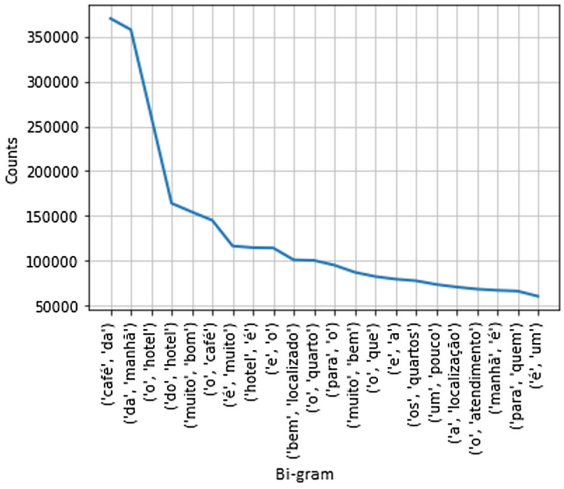

Fig. 4. Graph with the bi-gram distribution in the corpus.

hotel context, such as: hotel, coffee, room, service and location. Figure 4 shows the most common bi-grams in the corpus pointing out that terms like "caf/'e da manhã" (breakfast) and manhã (morning) are the most common terms, words that refers directly to the reviews' context.

5 Conclusions

The main contribution of this work is the production of a hotel review corpus with considerable size that will serve as a dataset for future work in sentiment analysis. We are making it free available to the community. We carried out a semi-automatic normalizations to reduce noise present in the corpus, but with the intention of keeping it as accurate as possible, considering that it is a corpus made up of texts extracted from the web. The corpus also can be used to aid in the tasks of extracting information, identifying patterns and new aspects present in the context of hotel evaluations among others. As future work, techniques such as [1] can be applied to spell checking the "noisy" terms while still keeping its original meaning, as well as normalizing the titles archive. Moreover, it is possible to adopt the use of multilingual methodologies to address cases of reviews written partially or entirely in languages other than Portuguese.

Acknowledgments. This research is supported in part by the funding agencies CAPES, FAPEMIG, and CNPq.

References

1. Bildhauer, F., and Schfer, R.: Token-level noise in large Web corpora and non-destructive normalization for linguistic applications. In: Proceedings of Corpus Analysis with Noise in the Signal (CANS 2013) (2013)
2. Duran, M.S. et al.: Some issues on the normalization of a corpus of products reviews in Portuguese. In: Proceedings of the 9th Web as Corpus Workshop (WaC-9), pp. 22–28 (2014)
3. Hartmann, N.S., et al.: A large corpus of product reviews in Portuguese: tackling out-of-vocabulary words. In: 9th European Language Resources Association-ELRA International Conference on Language Resources and Evaluation, pp. 3865–3871 (2014)
4. Kilgarriff, A., Grefenstette, G.: Introduction to the special issue on the web as corpus. Comput. linguist. **29**(3), 333–347 (2003)
5. Meyer, C.: The world wide web as linguistic corpus. Lang. Comput. **46**, 241–254 (2003)
6. Patil, P.: Application for data mining and web data mining challenges. Int. J. Comput. Sci. Mob. Comput. **6**(3), 39–44 (2017)
7. Rocha, P.A., Santos D.: CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In: Proceedings of V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000) (2000)