A Normalizer for UGC in Brazilian Portuguese

Magali Sanches Duran
NILC - Center for Computational Linguistics
São Paulo University (USP)
São Carlos-SP, Brazil
magali.duran@uol.com.br

Lucas Avanço
NILC - Center for Computational Linguistics
São Paulo University (USP)
São Carlos-SP, Brazil
avanco89@gmail.com

M. Graças Volpe Nunes
NILC- Center for Computational Linguistics
São Paulo University (USP)
São Carlos-SP, Brazil
gracan@icmc.usp.br

Abstract

User-generated contents (UGC) represent an important source of information for governments, companies, political candidates and consumers. However, most of the Natural Language Processing tools and techniques are developed from and for texts of standard language, and UGC is a type of text especially full of creativity and idiosyncrasies, which represents noise for NLP purposes. This paper presents UGCNormal, a lexicon-based tool for UGC normalization. It encompasses a tokenizer, a sentence segmentation tool, a phonetic-based speller and some lexicons, which were originated from a deep analysis of a corpus of product reviews in Brazilian Portuguese. The normalizer was evaluated in two different data sets and carried out from 31% to 89% of the appropriate corrections, depending on the type of text noise. The use of UGCNormal was also validated in a task of POS tagging, which improved from 91.35% to 93.15% in accuracy and in a task of opinion classification, which improved the average of F1-score measures (F1-score positive and F1-score negative) from 0.736 to 0.758.

1. Introduction

The increasing volume of text posted by users on the web is regarded as an extremely useful opportunity to reveal public opinion on many issues. For a variety of reasons, governments, companies, political candidates, and consumers want to explore such web content. This type of text is referred to in the literature as UGC (user-generated content) or EWoM (electronic word-of-mouth). However, due to the large amount of data available, it is impossible for humans to analyze all available UGC for most issues. As a result, processing and analyzing UGC became a task of NLP (Natural Language Processing). The problem is that, until now, almost all NLP tools and techniques were developed from, and for, standard language text, but UGC displays a range of creative and idiosyncratic differences, which represent noise for NLP purposes. In order to reuse the NLP tools to process UGC, the normalization or standardization of this genre of text became an essential preprocessing step, aiming to make UGC as close as possible to standard language.

The level of noise in UGC varies depending on the social media in which it is posted. Short messages (SMS and microblogs, such as Twitter) tend to be much noisier than texts posted in blogs and sites of reviews, as users need to be creative to deal with character limitations (140 characters for Twitter and 160 for SMS). The challenge for NLP is to determine the aspects in which UGC deviates from standard language and develop strategies to deal with the normalization of these aspects.

Many of UGC's deviations from standard language are motivated by wordplay (U=you, 4=for), by the need to save space (short messages have a limited length), by the influence of pronunciation, or even by a low level of literacy. Regardless of the causes of UGC deviations from standard language, if they are recurrent, they need to be addressed by normalization processes.

Some characteristics of UGC are language-independent, as the long vowels used to express emphasis (Gooooooooooooood) and the unconventional use of lower and upper cases (proper names in lowercase and common words in uppercase). Other characteristics are language-dependent, such as the apostrophe suppression in English (wont=won't) and the omission of

Proceedings of the ACL 2015 Workshop on Noisy User-generated Text, pages 38–47,
Beijing, China, July 31, 2015.

diacritics and cedilla under "c" in Portuguese (eleicao=eleição).

UGC differs from the standard language mainly in the lexical level. For this reason, the normalization problem is approached by strategies of word correction (the lexical items of the UGC are treated as "errors") and strategies for machine translation (the UGC is treated as source language and the standard language as target language).

We address herein the normalization process as a set of procedures that deal with different types of deviation. The input consists of consumer reviews on electronic products. The main purpose is to convert such texts, as closely as possible, into the form expected by NLP tools trained on corpora of standard language.

This work was preceded by the detection and analysis of out-of-vocabulary[1] (OOV) words in a corpus of product reviews (Hartmann et al. 2014). In another preliminary investigation, we have found other different types of deviations and their impact on a tagging task (Duran et al., 2014). Such diagnosis has resulted in the procedures that integrate the normalization system proposed here.

The remainder of this paper is organized as follows. Section 2 presents related works. Section 3 describes the characteristics of the product review corpus and the problems they pose to normalization. Section 4 reports the methodology used to construct the normalization tool. Section 5 describes and discusses the evaluation and validation results. Finally, in Section 6, we make some final remarks and outline future work.

2. Related works

Text normalization is a term used to convey the idea of converting the format of a text to meet the requirements of a given purpose. There are many text normalization processes reported in the NLP literature and they vary in: i) the genre of the input text; ii) the desired output format; iii) the purpose of the normalization, and iv) the method used to perform the task. It is important to take into account such characteristics to clearly define what "text normalization" means in each context.

The input text may or may not be well-written. The task of normalizing text from a newspaper (as in Schlippe et al., 2012) is quite different from normalizing texts produced by non-professional internet users, i.e. UGC. In addition, the normalization of UGC may depend on the social media used. For example, there are substantial differences between short message texts (SMS and microblogs), on-line chats and users' reviews. Short messages and chats deviate much more from the standard language than users' reviews and are commonly regarded as "noisy texts". The normalization processes of short messages, such as SMS and Twitter messages (Contractor et. al. 2010; Liu et al. 2011; Han et al., 2013; Bali, 2013; Chrupała, 2014) and longer UGC texts, such as reviews and blogs, have much in common, but the differences are sufficiently significant to justify addressing them separately.

Different normalization purposes may require the use of substantially different normalization procedures For example, converting text-to-speech requires the expansion of acronyms and abbreviations, as well as the conversion of numeric or mathematical expressions into words (Boros et al., 2012, Schlippe et al. 2012); conversely, normalization for purpose of storing data may perform the reduction of word forms into their stems. Even a "noisy text" of UGC may be normalized for different purposes. For example, while Mosquera et al. (2012) use normalization to improve the accessibility of web content, Aw et al. (2006) and Contractor et al. (2010) see the normalization as a prerequisite for other automatic processing tasks.

Approaches to text normalization may be roughly divided into two groups: those that "translate" non-standard language into standard language using contextual information (based on language models), and those that replace OOV words (lexical-based) by suitable forms in the standard language. For the latter, lexical information is essential; for the former, parallel corpora of non-standard and standard language are required. Lexical-based approaches are commonly used to normalize general texts, whereas machine-translation approaches are usually an option to tackle SMS normalization.

Aw et al. (2006) first proposed to regard SMS normalization as a machine translation problem. Many other studies have followed this approach

[1] "Out-of-vocabulary (OOV) words are unknown words that appear in the testing speech but not in the recognition vocabulary. They are usually important content words such as names and locations, which contain information crucial to the success of many speech recognition tasks. However, most speech recognition systems are closed-vocabulary recognizers that only recognize words in a fixed finite vocabulary." IN: Long Qin. 2013. Learning Out-of-Vocabulary Words in Automatic Speech Recognition. Phd Thesis. Carnegie Mellon University. 2013.

(Contractor et al., 2010; Schlippe et al., 2012; Bali, 2013, to cite just a few). They differ in the machine translation technique adopted or in the method used to obtain the parallel corpus for training and evaluation. Aw et al. (2006) constructed a parallel corpus with 5,000 SMS, Contractor et al. (2010) generated artificial "clean" sentences in a statistical machine translation approach, and Schlippe et al. (2012) constructed a web interface to receive suggestions of clean versions of noisy sentences.

Many studies have adopted a lexical approach to normalization. For example, Liu et al., 2011, aiming to tackle SMS normalization, proposed the generation of nonstandard tokens by performing letter transformation on the dictionary words. Han et al. (2013) observed that most ill-formed tokens in Twitter are morphophonemically similar to the respective correct forms. Based on this evidence, they proposed an automatic approach to constructing a set of word variants by using edit distance and phonemic transcription; finally, they ranked the candidates using a trigram language model. Mosquera et al. (2012) developed a multilingual lexical-based approach (English and Spanish) to normalize general text from a news corpus. The approaches of Ringlstetter et al. (2006), Clark and Araki (2011), and Bildhauer and Schäfer (2013) are similar to ours, as they regard normalization as a number of subproblems to be solved in sequence. In lexical-based approaches to normalization of web content, lexicons play an important role and require constant updating to keep pace with UGC innovations.

3. Characteristics of User-Generated Content in product reviews

The characteristics we describe in this Section have been observed in the corpus of product reviews Buscapé, built by Hartmann et al. (2014). The corpus is the result of crawling an e-commerce search engine of same name, where users can post comments about several products. This corpus consists of 85,910 reviews, 4,097,905 tokens and 90,513 types. After removing stop words, numbers and punctuation, it has 63,917 types, from which 34,774 are OOV words. To find OOV words, we used Unitex-PB, a Brazilian Portuguese lexicon (Muniz et. al. 2005). Words that miss a diacritic (3,652 or 10.2%) were automatically corrected. From the remaining 31,123 OOV words, we analyzed 5,775, which correspond to words with more than two occurrences in the corpus. Such OOV words were classified in a double-blind annotation task, which obtained 0.752 of inter-annotator agreement (Kappa statistics, Carletta, 1996). The analysis showed that such OOV words encompass misspellings, named entities written in lowercase, foreign loan words and recurrent non-standard words in UGC (Internet slang), for which an equivalent exists in the standard language. The normalization of OOV words, therefore, depends on distinguishing these categories, as they require different procedures: misspellings require spelling correction, named entities require conversion to uppercase, foreign loan words need to be incorporated to the lexicon, and non-standard words require substitution for words from the standard language.

An in-depth analysis of the 1,323 cases classified as misspellings by both annotators (100% of inter-annotator agreement) revealed that 791 were typos, 451 were phonetically-motivated errors, 64 were misused diacritics and 14 were problems related to the recent Portuguese orthographical rules, mostly associated with the use of hyphen in compounds. As open-source Portuguese spellers do not tackle phonetically-motivated misspellings, we undertook the development of a phonetic-based speller (Avanço et al., 2014), which achieved 65.46% of first hit accuracy, against 46.94% of the open-source speller Aspell[2].

Further analysis of the corpus led us to verify that many words that require normalization were not included among the OOV words, a phenomenon known as "real-word errors". In Portuguese there are around 25,000 pairs of words that are distinguished only by diacritics and, due to the systematic absence of diacritics in UGC, such pairs of words remain indistinguishable without contextual information, as the homographs (eg: "varias" (=to vary in the second person singular in the present tense) and "várias" (=several)). There are also some non-conventional words from Internet slang (eg. "vai testa"="vai testar"=will test)) and named entities (eg. the companies Oi, Claro and Sadia), which match existing words ("testa"=forehead; "oi"=hi; "claro"=light, clear; "sadia"=healthy). Therefore, if such words are identical to other words that belong to the lexicon, they are not identified as OOV words. For this reason, the identification of

[2] http://aspell.net/

tokens that require normalization is more complex in UGC than in the standard language.

The unconventional use of case is another characteristic of UGC observed in product reviews. Frequently, capital letters are not used after punctuation as well as for proper nouns. Conversely, common words are written in capital letters to emphasize an opinion (eg. "MUITO BOM" = VERY GOOD). There are also whole reviews written in uppercase or in lowercase or even a mix as: "Fiz Contato com o Vendedor, no qual ele De forma Discarada informa ser um produto ORIGINAL!" (literally: Make Contact with a Seller and he informs In a Shameless manner to be an ORIGINAL product!"). These phenomena cause problems for the recognition of named entities and for the segmentation of sentences since both tasks use capital letters as a clue. Lexical-based strategies can help to identify named entities written in lowercase. However, as proper names and acronyms are in open classes, it is infeasible to construct a comprehensive lexicon for them. Fortunately, the product reviews have metadata that contain most of the named entities found in the respective texts, which help to construct a domain-dependent lexicon of named entities. The opposite problem also exists, that is, to decide whether a word written in uppercase is a named entity or not.

Missing punctuation is another common characteristic of product reviews, which jeopardize sentence and clause segmentations. Some reviews reproduce a kind of uninterrupted stream of consciousness, making it difficult to punctuate the text, even for a human. In addition, most product reviews consist of three sections: Pros, Cons, and General Opinion. General Opinion usually is a plain text, but Pros and Cons may present single words (Pros: inexpensive), noun phrases (Pros: battery life), bulleted lists of words and noun phrases, or complete sentences. For this reason, it is challenging to punctuate the Pros and Cons sections, and the solutions sometimes require arbitrary decisions.

In the corpus of product reviews, unlike in short messages, word abbreviations, agglutination of several tokens into a single one, and suppression of grammatical words rarely occur .

4. A lexicon-based approach to UGC normalization

The nature of the deviations described in Section 3 have motivated us to develop a normalization tool tailored for product reviews. The goal is to normalize the deviations due to: 1) the case use, in what concerns the use of lowercase instead of uppercase; 2) the correction of misspellings, except for those cases that depend on contextual clues to disambiguate two existing words in Portuguese; 3) the substitution of Internet slang by standard language words, and 4) the insertion of missing periods (other punctuation marks will be addressed in future work).

One of the challenges of building a normalization tool refers to how to combine different normalization procedures in such a way that the effect of a procedure does not jeopardize the subsequent ones. For example, there are non-standard words from Internet slang as well as named entities written in lowercase among the OOV words. They need to be identified and protected from spelling correction.

The proposed pipeline architecture of the UGC Normalizer Tool (UCGNormal) is presented in Fig. 1.

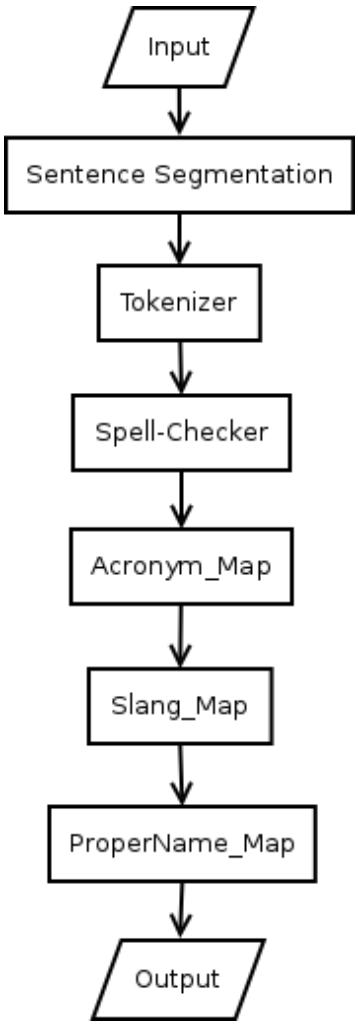

Figure 1: Architecture of UGCNormal

The input is a UGC text written in Brazilian Portuguese. The first step consists in applying the sentence segmentation tool proposed in Condori and Pardo (2015), which is a machine learning-based system trained in a journalistic corpus. It allows us to insert periods where they are missing and, consequently, to properly convert the initial words to uppercase. When evaluated in the Buscapé corpus, it achieved 0.953 for precision; 0.895 for recall; and 0.921 for F-Measure.

Subsequently, the sentences are tokenized, specifically accounting for the nature of UGC texts. Usually, tokenizers consider only blank spaces, punctuation, and few special symbols. However, when processing UGC, it is necessary to consider the occurrence of more complex tokens, like emoticons (‘ :) ’, ‘ :-) ’, ‘ :(‘, etc.), units of measurement (‘1GB’, ‘100Kb’, ‘2mb’, etc.), and URL’s. In order to properly identify and split tokens like those, we have developed a tokenizer using GNU-Flex lexical analyzer tool.

The lexicon-based Spell-Checker developed by Avanço et al. (2014) does the major part of the normalization process. It was specially developed to tackle phonetically-motivated misspellings, i.e. words written as they are pronounced. Another important characteristic of this speller is the automatic correction, as it does not presuppose user interaction. Therefore, instead of suggesting some candidates for correction, it automatically replaces the misspelled word with the best-ranked candidate. In such a scenario, the accuracy of the first hit is essential.

In short, the algorithm consists of (a) identifying misspelt words, using the UNITEX-PB[3] lexicon; b) generating candidates for the substitute word by using the edit distance (Levenshtein, 1966); (c) ranking the candidates by considering corpus-based frequency information; (d) looking for phonetic similarities by using several specific rules for Portuguese and using a variation of the Soundex[4] algorithm.

For UGCNormal, we made major improvements to the original algorithm of the speller, as well as adapting it to fit in the pipeline. As many misspellings are related to the omission of diacritics and cedilla under “c”, we have incorporated some heuristics to correct this kind of error before the generation of candidates.

As the correction of real-word errors is a hard context-dependent problem, this phonetic-based speller cannot handle them well. In order to overcome this limitation, we applied a simple strategy that enables the correction of some real-word errors without contextual information. For this, we have compiled, from the lexicon Unitex-PB, a list of 25,722 pairs of words that differ from each other by a single diacritic. From this list, we analyzed the pairs that differ in morphological tags (2,877), and selected 561 pairs of a highly frequent word and a highly infrequent word (eg. “óbvio” (=obvious) and “obvio” (an inflection of “obviar”=to obviate). The infrequent word was then excluded from the lexicon in order to enable the speller to eventually correct the more frequent one.

The remaining pairs are not addressed by the tool since the frequency of the words is similar. The most serious problem is related to pairs of frequent words, like “e” (=and) and “é” (=is); “da” (=of the) and “dá” (third person of the verb “dar”=to give).

Another modification was made in the speller to prevent the correction of acronyms and Internet slang. Foreign loan words and proper nouns have been incorporated to the lexicon, which is used to identify misspelled words and to generate candidates for misspelling correction. This decision was motivated by the high frequency of misspelled technology jargon in the domain of product reviews (eg. “desing” instead of “design” and “Blutoth” instead of “Bluetooth”).

The lexical resources, created especially for this, comprise: Internet slang (420 items), foreign loan words (248 items), proper nouns (20,730 items), and acronyms (156 items). These sets of items were partially compiled by Hartmann et al., (2014) and further complemented during the analysis of the corpus.

The module Acronym_Map sets all letters to uppercase whenever it detects an acronym (the detection of acronyms is based on the lexicon). The module Slang_Map substitutes some frequent slang words by their equivalent in standard language and normalizes long vowels by using regular expressions. There are two types of Internet slangs: 1) those that can be identified in a lexical-based approach (eg. “vc”=“você”; “tb”= “também”), and 2) those that have a homonym in the standard language, as “fala” in “vo fala” (=“vou falar”=I will speak) and “fala” (=he/she speaks; speech). Here we deal only with the correction of the first kind, as the second kind requires context knowledge to be identified and

[3] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/dicionarios.html

[4] http://www.archives.gov/research/census/soundex.html

corrected. All these modules use their own lexicons as well as a set of regular expressions for recognizing the items.

The last module, ProperName_Map, uses a lexicon of named entities, which consists of 8,465 proper nouns from the NILC Lexicon (Nunes et al., 1996). We have also added a further 12,265 proper nouns, consisting of product names including brands and models. These were extracted from the metadata available in the Buscapé corpus, and the addition of these resulted in 20,730 lexical items. When a proper noun is recognized, this module capitalizes it. However, detection of proper nouns written in lowercase is far from a simple task, because many proper nouns are also common words in the language lexicon, as mentioned in Section 3. Although there are some named entity recognizer (NER) systems for Portuguese, they do not perform well for UGC, since they heavily rely on the occurrence of a capital letter starting the proper nouns, and the problem is in discovering proper nouns that are not capitalized. That is why we have adopted a domain and lexical-based approach.

5. UGCNormal Evaluation

We evaluated the normalization tool intrinsically, in two corpus, and extrinsically, in a POS tag task and in an Opinion Classifier.

5.1. Intrinsic Evaluation

In the intrinsic evaluation we used two samples, one from the Buscapé corpus, and one from another corpus of the same genre, extracted from the e-commerce website Mercado Livre, which constitutes unseen data. In both cases, a sample of 60 product reviews was manually annotated with respect to punctuation errors, case use, and misspellings.

Our two samples (random selection from both corpora) are described in Table 1.

Table 1: Samples' statistics

	Buscapé Sample	**Mercado Livre Sample**
reviews	60	60
tokens	3,179	3,897
tokens without stop-words	2,061	2,732
tokens without stop-words and punctuation marks	1,563	1,967
types	887	1,096

Table 2 shows the recall figures of UGCNormal in both samples. The second and third columns contain X/Y=Z, where X shows the number of items to be normalized, Y shows the number of correctly normalized items, and Z shows the corresponding accuracy rate. As expected, the results in the Buscapé corpus (used for diagnosis) are better than in Mercado Livre, because some lexical resources were constructed from analysis of OOV words in Buscapé. In spite of both samples having the same number of reviews, the Mercado Livre sample contains proportionally more items to be normalized than the Buscapé sample, that is, the reviews from Mercado Livre deviate more from standard language than those from Buscapé.

For the misspellings whose corrections are context-free, UGCNormal achieved a recall of 89% in Buscapé corpus and 80% in Mercado Livre corpus. This difference may be due to the small size of both samples and the number of misspellings (in Mercado Livre there are almost twice as many misspellings as in Buscapé).

Table 2: Distribution of errors and corrections for each UGC sample, and the recall values for each error type.

Error type	**Buscapé**	**Mercado Livre**	**Average**
common misspellings	**50/56 = 0.89**	**87/108 = 0.80**	**0.84**
real-word misspellings	**15/39 = 0.38**	**24/76 = 0.31**	**0.34**
internet slang	**4/6 = 0.67**	**15/25 = 0.60**	**0.61**
case use (proper names and acronyms)	**11/12 = 0.92**	**13/19 = 0.68**	**0.77**
case use (start of sentence)	**14/14 = 1.00**	**7/12 = 0.58**	**0.81**
glued words	**0/2 = 0**	**2/6 = 0.33**	**0.25**
punctuation	**44/47 = 0.94**	**58/79 = 0.73**	**0.81**

We evaluated the task noise removal in a single pass, identifying and correcting errors simultaneously. Therefore, cases where errors were identified but not corrected were taken to be failures just like unidentified errors.

However, it is worth mentioning that the normalizer failed to correct 6 true errors identified in the Buscapé sample and 14 true errors identified in the Mercado Livre sample. The other non-corrected errors were not even identified.

The normalization tool corrected 66% (138 of 209) of the manually annotated errors in the Buscapé sample, and 63% (206 of 325) in the Mercado Livre sample.

Misspellings whose correction depends on contextual information were not expected to be corrected, as the speller is based only on lexical information. However, thanks to the strategy of excluding highly infrequent words that are homographs of frequent words without diacritics, some such errors were corrected (38% of the annotated errors of such category in Buscapé and 31% in Mercado Livre).

The case use in the start of sentences and the punctuation are treated by the sentence segmentation tool. These procedures are simultaneous: if a punctuation mark is not inserted, the initial word after a period is consequently not converted into uppercase. In the Mercado Livre corpus, the use of uppercase and lowercase is far more unconventional than in the Buscapé corpus and this explains the deterioration of results in case use and punctuation. For example, in Mercado Livre, unlike in Buscapé, we found reviews completely written in uppercase.

The conversion of proper nouns and acronyms to uppercase, as well as the conversion of Internet slangs to the standard language, are two issues that depend on the respective lexicons. As such, lexicons resulting from the analysis of the Buscapé corpus are not sufficient to identify all the proper nouns, acronyms and Internet slangs from the Mercado Livre corpus.

Finally, the glued words are normalized by the tokenizer only in cases where numbers are followed by units of measurement. Glued words are rare in both evaluated corpora, but we need to tackle them in the future if we want to address other categories of UGC, such as chats and short messages.

UGCNormal made 149 corrections in the Buscapé sample, of which 138 were true positives and 11 were false positives (well-formed words that were incorrectly modified), representing a precision of 93%. In the Mercado Livre sample, UGCNormal made 220 corrections, of which 206 were true positives and 14 were false positives, also representing a precision of 93%.

From the 82 OOV words in the Buscapé sample, UGCNormal corrected 65 (79%), and the remaining 17 words are constituted of 6 (7.3%) true errors and 11 (13.4%) real words.

In the Mercado Livre sample, UGCNormal identified 145 OOV words and appropriately corrected 117 (80.6%). From the remaining 28 OOV words, 14 (9.6%) are true errors and 14 (9.6%) are real words.

The false positives (real words identified as errors) are mainly foreign loan words, proper nouns, acronyms and Internet slang absent from the UGCNormal's lexicons.

5.2. Extrinsic Evaluation

To validate the normalization tool, we evaluated its impact as a preprocessing step in two NLP tasks: POS tagging and opinion classification.

For the first task, we used the tagger MXPOST (Ratnaparkhi, 1996), trained in the MAC-Morpho corpus (1.2 million tokens, Aluisio et al., 2003). The better reported results of MXPOST are around 97%, for journalistic texts, the same genre used to train the tagger.

For this experiment, we first randomly selected a sample of ten reviews from the Buscapé corpus. Then we tagged the sample with MXPOST and performed a linguistic revision of the POS tags, in order to create a gold-standard POS-tagged version of the sample. Subsequently, we POS-tagged three different versions of the same sample: 1) the original one; 2) a version manually normalized, and 3) a version automatically normalized by UGCNormal. The results of the three versions evaluated against the gold-standard version are presented in the Table 3.

Table 3: The number of correct tags produced by the tagger, for each sample version.

	Without Normaliz.	After Human Normaliz.	After Automatic Normaliz.
Correct tags	1120	1145	1142
Accuracy - MXPOST	91.35%	93.39%	93.15%

The accuracy values are the ratio between the number of correct tags and the total number of tags (1226). The result achieved by the automatically normalized version (UGCNormal) is almost the same as that achieved by the human normalized version.

We have also made a test of statistical significance to evaluate the probability that such improvement in the tagger precision could have been obtained by chance. Given the sample size and some relevant considerations while evaluating NLP tasks (Sogaard et al., 2014), we opted for the non-parametric test Wilcoxon Signed-Rank. We observed a significance of 0.05, the p-value being equal to 0.02249.

The other extrinsic evaluation is based on a lexicon-based opinion classifier (Avanço and Nunes, 2014), which assigns polarity to texts (positive, negative or neutral). We applied the classifier on a sample of 13,685 reviews (6,812 positives and 6,873 negatives) extracted from the Buscapé corpus, before and after normalization by UGCNormal. The average of F1-score measures (F1-score positive and F1-score negative) was 0.736 for non-normalized texts, and 0.758 for normalized texts.

The performance of a lexicon-based opinion classifier is highly dependent of the recognition of sentiment words in the text. As errors like "exelente" (excelente=excellent) and "otimo" (ótimo=great) are very frequent, such improvement in the precision, after normalization, was expected.

5.3 Some limitations of the normalization tool

The UGCNormal corrects a few real-word misspellings thanks to the strategy of extracting from UNITEX-PB those infrequent words that are homographs (except by the diacritics) of frequent words. However, many real-word misspellings remain unsolved, as those corrections would require contextual information. This problem is more serious when the homographs are very frequent words, such as "esta" (=this) and "está" (=is). Besides homographs, we also have to deal with the homophone words (those with identical pronunciation), which also frequently cause real-word misspellings, such as "segmento" (=segment) and "seguimento" (=follow up).

The normalization of acronyms, Internet slang, and proper names is dependent on their respective lexicons, which are not only domain-dependent, but also corpus-dependent, as we observed in the evaluation. The lexicons have been constructed with data from the Buscapé corpus and this justifies the best performance of the normalizer in such corpus.

The normalization of punctuation presupposes a plain text. For this reason, some product reviews that consist of simple items or noun phrases are difficult to normalize. If each item starts with uppercase, the sentence segmentation tool inserts a period after each item. Conversely, if an item starts in lower case and there is another item in the sequence, the sentence segmentation tool does not insert periods.

Another problem that remains unsolved is related to common words written in uppercase. We only convert uppercase to lowercase when the whole review is in uppercase. Otherwise, we maintain the uppercase, because it may indicate an acronym or a proper noun.

6. Final remarks and future work

The UGCNormal performance ranges from an average of 25% (for glued words) to 84% (for common misspellings). The validation of the tool shows that the results of both POS tagging and opinion classification tasks improved around by two percentage points after normalization.

Although there is no all-purpose normalization process, it is possible to reuse some modules of a normalization pipeline, assembling them differently in order to suit another purpose. The proposed normalization tool will certainly be useful for the development of UGC normalization tools that encompass short messages normalization. In order to be suitable for short messages normalization, this tool needs to address some problems related to word agglutination and informal abbreviations of nouns with stem preservation.

This normalizer evolved from a phonetic-based speller aimed at tackling common errors in UGC (words written as they are pronounced). Our approach is largerly dependent on lexical resources, incurring a high maintenance cost. In addition, this normalizer does not perform well with real-word errors. We believe that machine learning approaches will enable us to overcome these shortcomings. We have, indeed, made some preliminary experiments with language models, but the high occurrence of false positives (well-written words wrongly corrected) remains as a challenge.

Acknowledgements

Part of the results presented in this paper were obtained through research activity in the project entitled "Semantic Processing of Texts in Brazilian Portuguese", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law number 8.248/91.

References

Aluísio, S. M.; Pelizzoni, J. M.; Marchi, A. R.; Oliveira, L. H.; Manenti, R.; Marquivafável, V. (2003). An account of the challenge of tagging a reference corpus of Brazilian Portuguese. In: *Proceedings of PROPOR´2003*. Springer Verlag, 2003, pp. 110-117.

Avanço, L. V., Duran, M. S.; Nunes, M. G. V. (2014) Towards a Phonetic Brazilian Portuguese Spell Checker. *TorPorEsp - Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish.* Available at: http://www.lbd.dcc.ufmg.br/bdbcomp/servlet/Evento?id=755.).

Avanço, L. V.; Nunes, M. G. V. (2014). Lexicon-based sentiment analysis for reviews of products in Brazilian Portuguese. *Proceedings of the Brazilian Conference on Intelligent Systems* (BRACIS) - 2014, October 18-23, 2014, in São Carlos, SP, Brazil, pp. 277–281.

Aw, A.; Zhang, M.; Xiao, J.;Su, J. (2006) A phrase-based statistical model for SMS text normalization. In *Proceedings of COLING 2006*. ACL.

Bali, R. (2013) A Theoretical Review on SMS Normalization using Hidden Markov Models (HMMs). *International Journal of Computer Trends and Technology* (IJCTT), V.4 (7):2388-2387 July Issue 2013 .ISSN 2231-2803.www.ijcttjournal.org. Published by Seventh Sense Research Group.

Bildhauer, F.; Schäfer, R. (2013) 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).

Boros, T.; Stefănescu, D.; Ion, R. (2012) Bermuda, a data-driven tool for phonetic transcription of words info. *Proceedings of Natural Language Processing for Improving Textual Accessibility* (NLP4ITA) Workshop.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, vol. 22, n. 2, pp. 249--254.

Chrupała, G. (2014). Normalizing tweets with edit scripts and recurrent neural embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), pp. 680-686

Clark, E., & Araki, K. (2011). Text normalization in social media: progress, problems and applications for a pre-processing system of casual English. *Procedia-Social and Behavioral Sciences*, 27, 2-11.

Condori, R. E. L.; Pardo, T. A. S. (2015) Experiments on Sentence Boundary Detection in User-Generated Web Content. In: 16th International Conference on Intelligent Text Processing and Computational Linguistics, 2015, Cairo. *Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics*, 2015. v. 9041. p. 227-237.

Contractor, D.; Faruquie, T. A.; Subramaniam, V. (2010) Unsupervised cleansing of noisy text. *Coling 2010: Poster Volume*, pages 189–196, Beijing, August 2010.

Duran, M. S.; Avanço, L. V., Pardo, T. A. S.; Aluísio, S. M.; Nunes, M. G. V. Some issues on the normalization of a corpus of products reviews in Portuguese. In: Felix Bildhauer & Roland Schäfer (eds.), *Proceedings of the 9th Web as Corpus Workshop* (WaC-9) EACL 2014, pages 22–28, Gothenburg, Sweden, April 26 2014. 2014 Association for Computational Linguistics.

Han, B.;Cook, P.; Baldwin, T. (2013) Lexical Normalisation of Short Text Messages. *ACM Transactions on Intelligent Systems and Technology* 4(1), pp. 5:15:27.

Hartmann, N. S.; Avanço. L.; Balage, P. P.; Duran, M. S.; Nunes, M. G. V.; Pardo, T.; Aluísio, S. (2014). A Large Opinion Corpus in Portuguese - Tackling Out-Of-Vocabulary Words. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC 2014).

Levenshtein, V. I. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 1966.

Liu, F.; Weng, F.; Wang, B.; Liu, Y. (2011) Insertion, Deletion, or Substitution? Normalizing Text Messages without Pre-categorization nor Supervision. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*:shortpapers, pages 71–76, Portland, Oregon, June 19-24, 2011. Association for Computational Linguistics.

Mosquera, A.;Lloret, E.; Moreda, P. (2012) Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalisation. *Proceedings of Natural Language Processing for Improving Textual Accessibility* (NLP4ITA) Workshop.

Muniz, M.C.M.; Nunes, M.G.V.; Laporte. E. (2005) "UNITEX-PB, a set of flexible language resources for Brazilian Portuguese", *Proceedings of the Workshop on Technology of Information and Human Language* (TIL), São Leopoldo (Brazil): Unisinos.

Nunes, M.G.V.; Vieira, F. M. C.; Zavaglia, C.; Sossolote, C. R. C.; Hernandez, J. (1996) (In Portuguese) The design of a Lexicon for Brazilian Portuguese: Lessons learned and Perspectives. *Proceedings of the II Workshop on Computational Processing of Written and Speak Portuguese*. CEFET-PR, Curitiba, October 23-25, p. 61-70.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In: *Proceedings of the conference on empirical methods in natural language processing* (Vol. 1, pp. 133-142).

Ringlstetter, C.; Schulz, K. U.; Mihov, S. (2006). Orthographic Errors in Web Pages: Toward Cleaner Web Corpora. In: *Computational Linguistics* Volume 32, Number 3, p. 295-340.

Schlippe, T.; Zhu, C.; Gebhardt, J.; Schultz, T. (2010). Text normalization based on statistical machine translation and internet user support. *Interspeech*, 2010, pp. 1816-1819.

Søgaard, A., Johannsen, A., Plank, B., Hovy, D., & Martinez, H. (2014). What's in a p-value in NLP? In: *Proceedings of the eighteenth conference on computational natural language learning* (CONLL'14), pp. 1-10.