Some issues on the normalization of a corpus of products reviews in Portuguese

Magali S. Duran
NILC-ICMC
University of São Paulo
Brazil
magali.duran@gmail.com

Lucas V. Avanço
NILC-ICMC
University of São Paulo
Brazil
avanco89@gmail.com

Sandra M. Aluísio
NILC-ICMC
University of São Paulo
Brazil
sandra@icmc.usp.br

Thiago A. S. Pardo
NILC-ICMC
University of São Paulo
Brazil
taspardo@icmc.usp.br

Maria G. V. Nunes
NILC-ICMC
University of São Paulo
Brazil
gracan@icmc.usp.br

Abstract

This paper describes the analysis of different kinds of noises in a corpus of products reviews in Brazilian Portuguese. Case folding, punctuation, spelling and the use of internet slang are the major kinds of noise we face. After noting the effect of these noises on the POS tagging task, we propose some procedures to minimize them.

1. Introduction

Corpus normalization has become a common challenge for everyone interested in processing a web corpus. Some normalization tasks are language and genre independent, like boilerplate removal and deduplication of texts. Others, like orthographic errors correction and internet slang handling, are not.

Two approaches to web corpus normalization have been discussed in Web as a Corpus (WAC) literature. One of them is to tackle the task as a translation problem, being the web texts the source language and the normalized texts the target language (Aw et al., 2006; Contractor et al., 2010; Schlippe et al., 2013). Such approach requires a parallel corpus of original and normalized texts of reasonable size for training a system with acceptable accuracy. The other approach is to tackle the problem as a number of sub problems to be solved in sequence (Ringlstetter et al., 2006; Bildhauer & Schäfer, 2013; Schäfer et al., 2013).

The discussion we engage herein adopts the second approach and is motivated by the demand of preprocessing a Brazilian Portuguese web corpus constituted of products reviews for the specific purpose of building an opinion mining classifier and summarizer. Our project also includes the task of adding a layer of semantic role labeling to the corpus. The roles will be assigned to nodes of the syntactic trees and, therefore, SRL subsumes the existence of layers of morphosyntactic and syntactic annotations. The annotated corpus will be used as training corpus for a SRL classifier. The aim of SRL classifier, on its turn, is to provide deep semantic information that may be used as features by the opinion miner. If the text is not normalized, the POS tagger does not perform well and compromise the parsing result, which, as consequence, may generate defective trees, compromising the assignment of role labels to their nodes.

In fact, mining opinions from a web corpus is a non-trivial NLP task which often requires some language processing, such as POS tagging and parsing. Most of taggers and parsers are made to handle error-free texts; therefore they may jeopardize the application results when they face major noises. What constitutes a major noise and which noise may be removed or corrected in such a corpus is the challenge we are facing in this project.

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.

2. Related Work

Depending on the point of view, there are several studies that face problems similar to those faced by us. The general issue is: how to convert a non-standard text into a standard one? By non-standard text we mean a text produced by people that have low literacy level or by foreign language learners or by speech-to-text converters, machine translators or even by digitization process. Also included in this class are the texts produced in special and informal environments such as the web. Each one of these non-standard texts has its own characteristics. They may differ in what concerns spelling, non-canonical use of case, hyphen, apostrophe, punctuation, etc. Such characteristics are seen as "noise" by NLP tools trained in well written texts that represent what is commonly known as standard language. Furthermore, with the widespread use of web as corpus, other types of noise need to be eliminated, as for example duplication of texts and boilerplates.

The procedures that aim to adapt texts to render them more similar to standard texts are called normalization. Some normalization procedures like deduplication and boilerplate removal are less likely to cause destruction of relevant material. The problem arises when the noise category contains some forms that are ambiguous to other forms of the standard language. For example, the words "Oi" and "Claro" are the names of two Brazilian mobile network operators, but they are also common words ("oi" = hi; "claro" = clear). Cases like these led Lita et al. (2003) to consider case normalization as a problem of word sense disambiguation. Proper nouns which are derived from common nouns (hence, distinguished only by case) are one of the challenges for case normalization reported by Manning et al. (2008). Similar problem is reported by Bildhauer and Schäfer (2013) regarding dehyphenation, that is, the removal of hyphens used in typeset texts and commonly found in digitized texts. In German, there are many hyphenated words and the challenge is to remove noisy hyphens without affecting the correct ones. There are situations, however, in which both the corrected and the original text are desired. For example, social media corpora are plain of noises that express emotions, a rich material for sentiment analysis. For these cases, the non-destructive strategy proposed by Bildhauer and Schäfer (2013), keeping the corrected form as an additional annotation layer, may be the best solution.

3. Corpus of Products Reviews

To build the corpus of products reviews, we have crawled a products reviews database of one of the most traditional online services in Brazil, called Buscapé, where customers post their comments about several products. The comments are written in a free format within a template with three sections: Pros, Cons, and Opinion. We gathered 85,910 reviews, totaling 4,088,718 tokens and 90,513 types. After removing stop words, numbers and punctuation, the frequency list totaled 63,917 types.

Customers have different levels of literacy and some reviews are very well written whereas others present several types of errors. In addition, some reviewers adopt a standard language style, whereas others incorporate features that are typical of the internet informality, like abusive use of abbreviations, missing or inadequate punctuation; a high percentage of named entities (many of which are misspelled); a high percentage of foreign words; the use of internet slang; non-conventional use of uppercase; spelling errors and missing of diacritic signals.

A previous work (Hartmann et al. 2014) investigated the nature and the distribution of the 34,774 words of the corpus Buscapé not recognized by Unitex, a Brazilian Portuguese lexicon (Muniz et. al. 2005). The words for which only the diacritic signals were missing (3,652 or 10.2%) have been automatically corrected. Then, all the remaining words with more than 2 occurrences (5775) were classified in a double-blind annotation task, which obtained 0,752 of inter-annotator agreement (Kappa statistics, Carletta, 1996). The results obtained are shown in Table 1.

Table 1. Non-Recognized Words with more than 2 occurrences in the corpus

Common Portuguese misspelled words	44%
Acronyms	5%
Proper Nouns	24%
Abbreviations	2%
Internet Slang	4%
Foreign words used in Portuguese	8%
Units of Measurement	0%
Other problems	13%
Total	100%

The study reported herein aims to investigate how some of these problems occur in the corpus and to what extent they may affect POS tagging. Future improvements remain to be done in the specific tools that individually tackle these problems.

4. Methodology

As the same corpus is to be used for different subtasks – semantic role labeling, opinion detection, classification and summarization – the challenge is to normalize the corpus but also keep some original occurrences that may be relevant for such tasks. Maintaining two or more versions of the corpus is also being considered.

To enable a semi-automatic qualitative and quantitative investigation, a random 10-reviews sample (1226 tokens) of the original corpus was selected and POS tagged by the MXPOST tagger which was trained on MAC-Morpho, a 1.2 million tokens corpus of Brazilian Portuguese newspaper articles (Aluísio et al., 2003).

It is worthwhile to say that the sampling did not follow statistical principles. In fact, we randomly selected 10 texts (1226 tokens from a corpus of 4,088,718 tokens), which we considered a reasonable portion of text to undertake the manual tasks required by the first diagnosis experiments. Our aim was to explore tendencies and not to have a precise statistical description of the percentage of types of errors in the corpus. Therefore, the probabilities of each type of error may not reflect those of the entire corpus.

We manually corrected the POS tagged version to evaluate how many tags were correctly assigned. The precision of MXPOST in our sample is 88.74%, while its better precision, of 96.98%, has been obtained in its training corpus. As one may see, there was a decrease of 8.49% in performance, which is expected in such change of text genre.

In the sequence, we created four manually corrected versions of the sample, regarding each of the following normalization categories: spelling (including foreign words and named entities); case use; punctuation; and use of internet slang. This step produced four golden corpus samples which were used for separate evaluations. The calculation of the difference between the original corpus sample and each of the golden ones led us to the following conclusions.

The manual corrections of the sample were made by a linguist who followed some rules established in accordance with the project goals and the MXPOST annotation guidelines[1]. As a result, only the punctuation correction allowed some subjective decisions; the other kinds of correction were very objective.

5. Results of diagnosing experiments

Regarding to spelling, 2 foreign words, 3 named entities and 19 common words were detected as misspelled. A total of 24 (1.96%) words have been corrected. There are 35 words (2.90%) for which the case have been changed (6 upper to lower and 29 in the reverse direction).

Punctuation has showed to be a relevant issue: 48 interventions (deletions, insertions or substitutions) have been made to turn the texts correct, representing 3.92% of the sample. Regarding internet slang, only 3 occurrences (0.24%) were detected in the sample, what contradicted our expectation that such lexicon would have a huge impact in our corpus. However due to the size of our sample, this may have occurred by chance.

The precision of the POS tagged sample has been compared with the ones of the POS tagged versions of golden samples. The results showed us the impact of the above four normalization categories on the tagger performance.

We have verified that there was improvement after the correction of each category, reducing the POS tagger errors as shown in Table 2. When we combine all the categories of correction before tagging the sample, the cumulative result is an error reduction of 19.56%.

Table 2. Improvement of the tagger precision in the sample

Case Correction	+ 15.94%
Punctuation Correction	+ 4.34%
Spelling	+ 2.90%
Internet Slang Convertion	+ 1.45%
Cumulative Error Reduction	19.56%

These first experiments revealed that case correction has major relevance in the process of normalizing our corpus of products reviews. It is important to note that case information is largely

[1] Available at http://www.nilc.icmc.usp.br/lacioweb/manuais.htm

used as feature by Named Entities Recognizers (NER), POS taggers and parsers.

To evaluate whether the case use distribution is different from that of a corpus of well written texts, we compared the statistics of case use in our corpus with those of a newspaper corpus (http://www.linguateca.pt/CETENFolha/), as shown in Table 3.

Table 3. Percentage of case use in newspaper and products reviews corpus genres

CORPUS	Newspaper	Products Reviews
Uppercase words	6.41%	5.30%
Initial uppercase words	20.86%	7.30%
Lowercase words	70.79%	85.37%

The differences observed led us to conclude that the tendency observed in our sample (proper names and acronyms written in lower case) is probably a problem for the whole corpus.

To confirm such conclusion, we searched in the corpus the 1,339 proper nouns identified in our previous annotation task. They occurred 40,009 times with the case distribution shown in Table 4.

Table 4. Case distribution of Proper Nouns

Initial uppercase words	15,148	38%
Uppercase words	7,392	18%
Lower case words	17,469	44%
Total	40,009	100%

The main result of these experiments is the evidence that the four kind of errors investigated do affect POS tagging. In the next section we will detail the procedures envisaged to provide normalization for each one of the four categories of errors.

6. Towards automatic normalization procedures

After diagnosing the needs of text normalization of our corpus, we started to test automatic procedures to meet them. The processing of a new genre always poses a question: should we normalize the new genre to make it similar to the input expected by available automatic tools or should we adapt the existing tools to process the new genre? This is not a question of choice, indeed. We argue that both movements are needed. Furthermore, the processing of a new genre is an opportunity not only to make genre-adaptation, but also to improve general purpose features of NLP tools.

6.1 Case normalization: truecasing

In NLP the problem of case normalization is usually called "truecasing" (Lita et al, 2003, Manning et al., 2008). The challenge is to decide when uppercase should be changed into lower case and when lower case should be changed into upper case. In brief, truecasing is the process of correcting case use in badly-cased or non-cased text.

The problem is particularly relevant in two scenarios; speech recognition and informal web texts.

We prioritized the case normalization for two reasons: first, badly-cased text seems to be a generalized problem in the genre of products reviews and, second, it is important to make case normalization *before* using a spell checker. This is crucial to "protect" Named Entities from spelling corrections because when non-recognized lowercase words are checked by spellers, there is the risk of wrong correction. Indeed, the more extensive is the speller lexicon, the greater is the risk of miscorrection.

The genre under inspection presents a widespread misuse of case. By one side, lower case is used in place of uppercase in the initial letter of proper names. On the other side, upper case is used to emphasize any kind of word.

Our first tentative to tackle the problem of capitalization was to submit the samples to a Named Entity Recognizer. We chose Rembrandt[2] (Cardoso, 2012), a Portuguese NER that enhances both lexical knowledge extracted from Wikipedia and statistical knowledge.

The procedure was: 1) to submit the sample to Rembrandt; 2) to capitalize the recognized entities written in lower case; 3) to change all the words capitalized, except the named entities, to lower case. Then we tagged the sample with MXPOST to evaluate the effect on POS tagging accuracy.

The number of errors of POS tagging increased (149) when compared to the one of the sample without preprocessing (138). The

[2] The Portuguese named entity recognition is made by system Rembrandt (http://xldb.di.fc.ul.pt/Rembrandt/)

explanation for this is that among the words not recognized as named entities there were capitalized named entities which were lost by this strategy.

Next we tried a new version of this same experiment: we only changed into lower case the words not recognized as named entities that were simultaneously recognized by Unitex. The results were slightly better (143 errors) compared to the first version of the experiment, but still worse than those of the sample without preprocessing.

Our expectation was to automatically capitalize the recognized entities written in lower case. In both experiments, however, no word was changed from lower to upper case because all the entities recognized by the NER were already capitalized.

The sample contains 57 tokens of named entities (corresponding to proper nouns and acronyms) from which 24 were written in lower case. The NER recognized 22 of the 57 or 18 of the 38 types of named entities (a performance of 47.4%). Unfortunately the NER is strongly based on the presence of capitalized initial letters and was of no aid in the procedure we tested.

We argue that a finite list of known proper nouns and acronyms, although useful for improving evaluation figures, is of limited use for an application such as an opinion miner. In real scenarios this constitutes an open class and new entities shall be recognized as well.

We observed that many of the named entities found in the reviews relate to the product being reviewed and to the company that produces it. Then we realized an advantage of the source from which we have crawled the reviews: the customers are only allowed to review products that have been previously registered in the site database. The register of the name of the product is kept in our corpus as metadata for each review. This situation gave us the opportunity to experiment another strategy: to identify named entities of each review in its respective metadata file. We first gathered all the words annotated as Proper Nouns and Acronyms in our previous annotation task[3]. Then we search for the matches. The result is promising: from 1,334 proper nouns and from 271 acronyms, respectively 676 (50.67%) and 44 (16.23%) were found in the metadata. Adding both types of named entities, we have a match of 44.85% (720 of 1605). This is pretty good mainly because the named entities recognized are precisely the names of products for which opinions will be mined.

However, we still need to solve the recognition of the other named entities in order to support the truecasing strategies.

Following Lita et al. (2003) and Beaufays and Strope (2013), we are considering using a language model. Lita et al. developed a truecaser for news articles, a genre more "stable" than products reviews. Beaufays and Strope, on their turn, developed a truecaser to tackle texts generated from speech recognition. Language modeling may be a good approach to our problem because many named entities of products domain do not sound as Portuguese words. For example, they frequently have the consonants k, y and w, which are only used in proper names in Portuguese. Other approaches to truecasing reported in the literature include finite state transducers automatically built from language models and maximum entropy models (Batista et al. 2008).

6.2 Punctuation problems

Many reviews have no punctuation at all. This prevents processing the text by most of NLP tools which processes sentences. Some grammatical rules may be used to correct the use of comma, but the problem is more complex in what concerns full stop. We are now training a machine learning based program with a corpus of well written texts by using features related to n-grams. We aim at building a sentence segmentation tool which does not depend on the presence of punctuation or case folding, since these are major noises in the corpus.

6.3 Spelling correction

The common Portuguese words in the corpus which were not recognized by Unitex have been spell checked. Manual analysis is being undertaken to determine whether the word has been accurately corrected or not. Early results evidenced opportunity to extend Unitex and to improve our spellers with more phonetic rules in order to suggest more adequate alternatives. As we have already mentioned, product reviewers have several levels of literacy and those of lower level frequently swap the consonant letters that

[3] Confusion matrix of our double annotated data show that annotators diverged in what concerns Proper Nouns and Acronyms. For our purposes, however, all of them are named entities and need to be capitalized, so that this kind of disagreement did not affect the use we have made of the annotated words.

conveys the same phonetic value. For example, in Portuguese the letters “s”, “c”, “xc” “ss” and “ç” can have the same sound: /s/. Therefore, it is a common mistake to employ one instead of the other. These rules shall be incorporated in spell checker. In addition, there are many words which were correctly spelled, but were not part of Unitex or of the speller’s dictionary or both. Both lexicons will be extended with the missing words.

In the same way, the foreign words of current use in Brazilian Portuguese shall be incorporated in the spell checkers in order to improve their suggestions of correction. As a matter of fact, foreign words are frequently misspelled. For example, “touchscreen” appeared as 10 different spelling forms in our corpus with more than 2 occurrences (“toch escreen”, “touch screem”, “touch sreen”, “touche”, “touch scream”, “touchscream”, “touchscreem”, “touch-screen”, “touchsren”, “touch screen").

6.4 Internet slang normalization

Internet slang is a class that combines: 1) words written in a different way and abbreviations of recurrent expressions, for which there is an equivalent in the standard language (in this case the procedure is to substitute one for another); 2) repeated letters and punctuation (e.g. !!!!!!!!!!!!, and ameiiiiiiiiiiiiiiiiiiiii, in which the word "amei" = “love” is being emphasized), which may be normalized by eliminating repetitions; and 3) sequences of letters related to emotion expression, like emoticons (e.g. “:)”, “:=(”), laughing (e.g. rsrsrsrs, heheheh, kkkkkkkk), which for some purposes shall be eliminated and for others shall not. The procedures relating to internet slang will be implemented carefully to allow the user to activate each one of the three procedures separately, depending on his/her interest in preserving emotion expression or not.

7. Final Remarks

This preliminary investigation about the needs of text normalization for the genre of products reviews led us to deep understand our challenges and to envisage some solutions.

We have opened some avenues for future works and established an agenda for the next steps towards corpus normalization.

Acknowledgments

This research work is being carried on as part of an academic agreement between University of São Paulo and Samsung Eletrônica da Amazônia Ltda.

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.

Aw, A.; Zhang, M.; Xiao, J.; Su, J. (2006). A Phrase-based Statistical Model for SMS Text Normalization. In: *Proceedings of the COLING-2006* .ACL, Sydney, 2006, pp. 33–40.

Batista, F.; Caseiro, D. A.; Mamede, N. J.; Trancoso, I. (2008). Recovering Capitalization and Punctuation Marks for Automatic Speech Recognition: Case Study for the Portuguese Broadcast News, *Speech Communication*, vol. 50, n. 10, pages 847-862, doi: 10.1016/j.specom.2008.05.008, October 2008

Beaufays, F.; Strope, B. (2013) Language Model Capitalization. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), p. 6749 – 6752.

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*) .

Cardoso, N. (2012). Rembrandt - a named-entity recognition framework. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation* (LREC'12). May, 23-25, Istanbul, Turkey.

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

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

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). Forthcoming.

Lita, L., Ittycheriah, A., Roukos, S. & Kambhatla,N. (2003), Truecasing, In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Japan.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge: Cambridge university press.

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.

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

Schäfer, R.; Barbaresi, A.; Bildhauer, F. (2013) The Good, the Bad, and the Hazy: Design Decisions in Web Corpus Construction. In: *Proceedings of the 8th Web as Corpus Workshop* (WAC-8).

Schlippe, T.; Zhu, C.; Gebhardt J.;, Schultz, T.(2013). Text Normalization based on Statistical Machine Translation and Internet User Support. In: *Proceedings of The 38th International Conference on Acoustics, Speech, and Signal Processing* (ICASSP-2013) p. 8406 – 841.