Towards Semantic Role Labeling Annotation on Product Reviews in Brazilian Portuguese

Nathan Siegle Hartmann, Marina Coimbra Viviani, and Leandro B. dos Santos

Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
nathanshartmann@gmail.com, rina.coimbra@gmail.com, sborgesleandro@gmail.com

Abstract. Our main focus in this work is tailoring Semantic Role Labeling (SRL) taggers to the scenario of User-Generated Content which represents a challenge for most Natural Language Processing (NLP) tasks. Our object of study is SRL annotation on a Brazilian Portuguese product reviews corpus and the challenges associated with it. We believe that this study will be beneficial for many Natural Language Processing tasks such as question and answering and opinion mining.

Keywords: semantic role label, user-generated content, opinion corpus.

1 Introduction

User-generated content (UGC) [1] websites like social networks, blogs and e-commerce systems are becoming more popular as the Internet becomes more accessible. The information generated by UGC media is extremely valuable since it can be used by companies to improve their understanding about their customers needs. However, the language employed by UGCs represents a challenge for most Natural Language Processing (NLP) tools since it is written in a nonstandard format. While NLP tools are mostly designed for dealing with well written texts [2], UGC is marked by the absence of complex syntactic structures (as subordinate constructions) and the presence of Internet slang [3]. Xue et al. [4] reinforce that especially young users abuse of texting abbreviation style, resulting in a new subcategory of writing that is very different from well-written texts.

Our main focus in this work is tailoring Semantic Role Labeling[1] (SRL) taggers to the scenario of UGC. Semantic role labels are usually annotated over syntactic trees. However, there is a high probability of generating incorrect trees when the parser is not set up to deal with missing diacritics, vowel repetitions, texting abbreviations, emoticons and Internet slang. One of the most challenging aspects of our task is to stop the propagation of lexical errors to the semantic level.

Hartmann et al. [6] proposed a text normalization tool[2] to address the lexical issue on UGC texts. In their work they compiled a corpus of 85,910 Brazilian

[1] SRL is a NLP task which aims to detect predicates and label their semantic arguments in a sentence [5].

[2] http://nilc.icmc.usp.br/semanticnlp/LexicalNormalization

Portuguese (BP) product reviews from the Web, containing 4,097,905 tokens and 68,633 types. They also categorized 5,775 sample tokens in misspellings, acronyms, proper names, abbreviations, Internet slang words, foreign words, units of measure and unrated for doubtful tokens. The categorization sparked the creation of specific dictionaries to each category and a Phonetic Brazilian Portuguese Spell Checker.

In this paper we study the challenges of manual annotation of semantic roles on a corpus of product reviews written in BP based on the previous lexical analysis on the corpus of product reviews [6]. We are certain that this study will be beneficial for many NLP tasks such as question and answering, opinion mining and information retrieval.

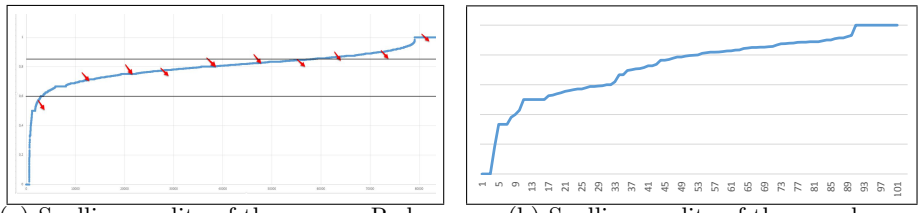
2 Methodology

The semantic role labeling annotation over this product review corpus was divided into three stages: preprocessing, manual annotation and evaluation. The preprocessing stage comprises automatic syntactic parsing by parser Palavras [7] and elimination of ill-formed syntactic trees. The annotation task was performed in double blinded pairs and we have used the kappa statistic [8] for evaluating the annotation. It is important to perform experiments on a sample from the corpus to ensure that the annotation guidelines are well defined and identify preprocessing needs as well as linguistic features. We collected 100 reviews, ensuring that this sample is representative of the entire corpus, in order to evaluate challenges before conducting a larger scale annotation on SRL. The spelling quality of the reviews is a proper feature to describe the corpus since it can distinguish between well-written texts, and texts with misspellings and other problems analysed by Hartmann et al. [6]. Our indicator of spelling quality for each review is the percentage of words that were comprised by Unitex-PB dictionary[3] [9]. The indicator ranged from 0 (no matches) to 1 (fully matched reviews). The curve on Fig.1(a) represents the distribution of the reviews according to our evaluation criteria.

The lexical properties pointed out by Hartmann et. al [6] are shown in the quality curve since less than 10% of the reviews are fully written with words recognized by Unitex-PB. However, since more than 95% of the curve was evaluated as having a spelling quality over 60%, we can assume that there are very few reviews evaluated as having extremely low quality. In fact, more than 55% of the reviews were evaluated as having an overall quality over 80% on the lexical level. Therefore, even though most of the reviews contain unmatched words, over than 90% of them are composed mostly of words matched by Unitex-PB.

A K-Means execution identified 10 centers that are evenly distributed among the horizontal axis (red arrows on Fig.1(a)). This distribution aided us in splitting the data into 3 segments: bad quality reviews (rated between 0 to 0.6), medium quality reviews (rated between 0.6 to 0.85) and good quality reviews

[3] Unitex-PB is a large freely available lexicon dictionary for Brazilian Portuguese. It is available at http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html



(a) Spelling quality of the corpus. Red arrows point to K-Means prototypes. Dark lines split the data by their quality.

(b) Spelling quality of the sample.

Fig. 1. Chart of spelling quality of reviews. X-axis represents the number of reviews and the Y-axis represents their quality in scale ranging from 0 to 1.

(rated between 0.85 to 1). A balanced sampling from each of the segments ensures a representative corpus regarding the lexical level, as shown in Fig.1(b).

After selecting the sample, we parsed all selected reviews using the syntactic parser PALAVRAS [7], generating syntactic trees. However, this parser has limitations on parsing words that are not covered by its dictionary, resulting in ill-formed trees. To address this issue we discarded trees that are not a connected graph. Furthermore, we discarded trees without verb. Once the aim of our SRL annotation is the verb of a sentence, sentences without verb can not be annotated. For the annotation task, the set of labels used is defined in [10] and has already been adapted to the Brazilian Portuguese scenario by Duran and Alusio [11]. The annotator's task consisted of assigning a sense[4] to each verb besides semantic roles for the arguments in the sentences of the review.

3 Results

Many syntactic trees had to be discarded. The parser PALAVRAS generated 311 trees from the 100 reviews selected from the corpus. A total of 86 trees were discarded by did not contain a verb, remaining 225 syntactic trees. However, 19 trees presented some sort of problem and were discarded as well: 6 trees did not contain a verb tagged by the parser and 13 trees was not a connected graph. We also discarded 12 trees which contained more than 30 tokens as terminals. Large sentences are difficult for human comprehension and even more troublesome for the parser to generate a correct syntactic tree. Therefore, only 206 sentences were selected for semantic role labeling annotation.

We generated one annotation instance per verb on each sentence, totalizing 348 annotation instances. These annotation instances have an average length of 10 tokens and their most frequent length is 5 tokens. A double blind annotation of the 348 instances resulted in a Kappa agreement of 0.91, which expresses good agreement [8].

[4] Verbo-Brasil (http://www.nilc.icmc.usp.br/verbobrasil/) is a repository of verb senses and their arguments for BP language [12] that aids in human annotation.

Table 1. Frequency of manual SRL agreement.

Arguments	Occurrence	Agreement (%)
Arg4	1 (00.17%)	100.00
ArgM-Adv	2 (00.41%)	100.00
ArgM-Neg	51 (08.70%)	92.16
ArgM-Cau	10 (01.70%)	90.00
ArgM-Tmp	19 (03.24%)	84.21
Arg1	246 (41.97%)	81.30
Arg0	48 (08.19%)	77.08
ArgM-Mnr	31 (05.29%)	74.19
Arg2	142 (24.23%)	73.24
ArgM-Ext	6 (01.02%)	66.67
ArgM-Loc	9 (01.53%)	44.44
Arg3	5 (00.85%)	40.00
ArgM-Asp	1 (00.17%)	0.00
ArgM-Dis	5 (00.85%)	0.00
ArgM-Nse	1 (00.17%)	0.00
ArgM-Prd	4 (00.68%)	0.00
ArgM-Prp	2 (00.34%)	0.00
ArgM-Pas	2 (00.34%)	0.00
ArgM-Tml	1 (00.17%)	0.00

Table 2. Most frequent verbs of the sample.

Verb	Frequency (%)
Ser - To be	20.1%
Words incorrectly classified as verbs	15.8%
Ter - To have	7.4%
Recomendar - To recommend	2,8%
Comprar - To buy	2,5%

Table 3. Most frequent elements of the top roles.

Arguments	Most frequent element
Arg1	*o produto* - the product
Arg2	*bom* - good
ArgM-Neg	*não* - no
Arg0	*eu* - I

The results shown in Table 1 expose the most frequent arguments of the sample in which there was agreement between the annotators: Arg1 (41.97%), Arg2 (24.23%) and ArgM-Neg (8.70%). As shown in Table 2, the most frequent verb in the sample is the verb *ser* - to be, with 20.1% of occurrence. Instead of expecting an agent (Arg0), the most common sense of this verb expects a topic (Arg1) and a comment about it (Arg2). The most common elements for each role, shown in Table 3, illustrate what sort of information we can extract from each label. These elements reveal that SRL is able to represent well the nature of the corpus. It is also possible to assert that even though ArgM-Neg is mostly associated with words like "no" or "never", it is not associated directly with the polarity of the reviews. Finally, since omitting the subject of the verb is possible in BP language, Arg0 label represents only 8.19% of labels annotated and it is usually related to the reviewer, e.g., "I like Samsung products.".

We also identified that the parser incorrectly assigned a target verb to 55 annotation instances (15.80%). Those errors occur because every time PALAVRAS cannot identify the correct part of speech tag to a word**,** it tags the word as a verb. The reasons for these errors are: 29 words had their diacritic accent missing; 6 words had spelling errors; 3 words were from a foreign language; 2 words were Internet slang; 2 words were proper names and 13 wrong verbs were assigned for spelling errors in other words of the instance. It show us that if a spelling checker is used to correct at least missing diacritic signals, 52.72% of wrong target verbs are solved.

4 Conclusion

The results presented in this paper are very encouraging. At least, 95% of the corpus reviews are composed more by well written words and the annotation Kappa agreement is acceptable for SRL task. However, it is necessary to preprocess the sentences before the parsing. Reduction of redundant punctuation and spelling checking are potential preprocessing steps. The experience aided us in

identifying verbs that need to be included in Verbo-Brasil or have their senses expanded. The analysis of the most frequent arguments reveals SRL's potential to extract information from the reviews. As future work, we intend to extend the annotation to a larger sample of the corpus and train a SRL classifier for this genre of texts. The classifier will enable an evaluation of automatic semantic role labeling in UGC texts.

Questions. (1) Is the usage of syntactic trees the best choice to this scenario? (2) Is it more appropriate to fit the data to the SRL classical task or to do otherwise? (3)Would it be interesting to adapt the labels of SRL in such a way that they would relate to the product in question instead of the verb subject?

Acknowledgements. Part of the results presented in this paper were obtained through research activity in the project titled "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

[1] J. Krumm, N. Davies, and C. Narayanaswami, "User-generated content," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 10–11, 2008.

[2] C. Ringlstetter, K. U. Schulz, and S. Mihov, "Orthographic errors in web pages: Toward cleaner web corpora," *Computational Linguistics*, vol. 32, no. 3, pp. 295–340, 2006.

[3] L. Squires, "Enregistering internet language," *Language in Society*, vol. 39, pp. 457–492, 2010.

[4] Z. Xue, D. Yin, and B. D. Davison, "Normalizing microtext.," *Analyzing Microtext*, vol. 11, p. 05, 2011.

[5] M. Palmer, D. Gildea, and N. Xue, "Semantic role labeling," *Synthesis Lectures on Human Language Technologies*, vol. 3, no. 1, pp. 1–103, 2010.

[6] N. S. Hartmann, L. Avanco, P. P. Balage Filho, M. S. Duran, M. d. G. Volpe Nunes, T. S. Pardo, and S. M. Aluisio, "A large opinion corpus in portuguese – tackling out-of-vocabulary words," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, vol. 2014, (Reykjavik, Iceland), European Language Resources Association (ELRA), 2014.

[7] E. Bick, *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*, vol. 202. Aarhus University Press Aarhus, 2000.

[8] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Computational linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[9] M. C. Muniz, M. D. G. V. Nunes, E. Laporte, *et al.*, "Unitex-pb, a set of flexible language resources for brazilian portuguese," in *Proceedings of the Workshop on Technology on Information and Human Language (TIL)*, pp. 2059–2068, 2005.

[10] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," *Computational Linguistics*, vol. 31, no. 1, pp. 71–106, 2005.

[11] M. S. Duran and S. M. Aluísio, "Propbank-br: a brazilian treebank annotated with semantic role labels.," in *LREC*, (Istanbul, Turkey), pp. 1862–1867, 2012.

[12] M. S. Duran, J. P. Martins, and S. M. Aluisio, "Um repositorio de verbos para a anotação de papéis semânticos disponivel na web," in *Brazilian Symposium in Information and Human Language Technology (STIL)*, (Fortaleza, CE), pp. 168–172, October 2013.