Construction of a Portuguese Opinion Lexicon from multiple resources

Marlo Souza[1], Renata Vieira[1], Débora Busetti[2], Rove Chishman[2], Isa Mara Alves[2]

[1]Faculdade de Informática – PUCRS
Porto Alegre – RS – Brazil

[2]Faculdade de Letras – Unisinos
Vale dos Sinos, Brasil

marlo.souza@acad.pucrs.br, renata.vieira@pucrs.br,

deborabusetti@hotmail.com, {rove,ialves}@unisinos.br

Abstract. *Opinion Lexicons are linguistic resources annotated with semantic orientation of terms (positive and negative) and are important for opinion mining tasks. In the literature we see a variety of proposals for the construction of opinion lexicons using different linguistic resources and techniques. In this work, we propose and evaluate the integration of such linguistic resources to create a single lexicon for the Portuguese language.*

1. Introduction

Sentiment Analysis, or Opinion Mining, corresponds to the problem of identifying or extracting emotions, opinions or points of view expressed in text. This area has received a great deal of attention in the last years due to its potential applicability, according to [Wilson et al. 2006, Liu 2010], among others. The solutions provided to such problem have been applied to several tasks in the literature that encompasses the area of study [Liu 2010].

Opinion Lexicons are commonly used within various techniques for sentiment analysis in the literature [Liu 2010]. Their importance lies on easily improve the recall on identifying opinion-bearing expressions and providing clues for identification of new-ones when associated with linguistic rules. Also, it can be used both to determine the polarity of an expression - using what is called previous polarity - or in aiding the determination of its polarity within a context.

Important to this problem is the identification and the determination of polarity (or semantic orientation) of individual terms and words. In such a task, an opinion lexicon has an important role in documenting already known terms and their semantic orientation within a certain context.

Work on word or term orientation detection usually fall on three approaches: a corpus-based approach, a lexicon or dictionary-based approach or a multilingual/translation approach.

The first uses the relations encountered in large-corpora between words and expressions to determinate their polarity. Works as [Hatzivassiloglou and McKeown 1997, Turney 2002, Riloff and Shepherd 1997] fall in this category. Their advantage is the pos-

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 59–66,
Cuiabá, MT, Brazil, October 24–26, 2011.

sibility to identify multi-word opinion-bearing expressions such as ”pain in the ass”[1] and identify neutral expressions such as ”great deal of” which do not have any evaluative connotation, expressions with evaluative connotations based on social usage, but not directly accessible through their lexicographic sense, and not always reported in lexicons or dictionaries, such as ”fantastic”[2]. The results directly reflect the nature of the corpus, so different senses and connotations of a word might not be captured. These methods require a great amount of data to be processed.

The second approach explores the semantic relations annotated in resources such as thesauri and dictionaries. Representatives of such methods are the work of Kamps et al who makes use of the WordNet [Fellbaum 1998] relation of synonymy to determine polarity; or [Esuli and Sebastiani 2005] that uses an online dictionary and the WordNet relations. The advantage is the possibility to explore well-defined, formally code and validated semantic relations between the words and a vast lexical base. There are restrictions imposed to such methods since multi-word expressions, slang and social attributed connotations not contemplated in the thesaurus or dictionary are not accessible.

Finally, the multi-lingual and translation-based methods explore available resources some languages, as in English, to be used in different ones. Those methods have advantages, since in some languages, linguistic resources are not available. They must deal, nevertheless, with the great challenge of translating a word or expression to another language maintaining its original sense.

Based on this, we propose the integration of different methods in the literature [Turney 2002, Kamps et al. 2004] and different linguistic resources to identify opinion-bearing terms to create a opinion lexicon for the Portuguese language, exploiting their individual qualities to overcome their individual shortcomings.

This paper is organized as follows. First, we present related works, focusing on three classical methods in the literature (Section 2). We then explain how each technique was implemented to create a lexicon of words and expressions for the Portuguese language (Section 3) and the linguistic resources used in such a work. The results achieved are described (Section 3) and analyzed (Section 4) and we conclude with our opinions about the proposed technique and the final product.

2. Related Work

The work of Hatzivassiloglou and Mckeown [Hatzivassiloglou and McKeown 1997] aims to identify the previous polarity - or polarity outside any particular context - of adjectives exploring conjunctions. The work hypothesize that two different adjectives are usually involved in a additive conjunction when they have same semantic orientation and in adversative conjunction when their polarities are opposite.

Turney [Turney 2002] used a distance-based approach to determine the polarity of the expressions. Differently than the previous work, the authors treated not only adjectives but also modifiers and adjectival-phrases - extracting polarized bi-grams instead of single words. The semantic orientation is determined by the algebraic sum of distances of the bi-grams to the words of a seed. The authors used three different distances in their work

[1]Connotes a negative evaluation, similar to ”annoying”.

[2]Which connotes a positive evaluation, similar to ”awesome”.

based in PMI statistics, Latent Semantic Analysis and Association Rules.

Many works use the semantic relations of the WordNet [Fellbaum 1998] to identify the polarity of adjectives [Kamps et al. 2004, Riloff and Shepherd 1997]. Kamps et al [Kamps et al. 2004] use an initial set of polar words - a seed set - that is expanded through the exploration of synonymy relations. The hypothesis of these works bring the idea that synonyms share the same semantic orientation.

Mihalcea et al [Mihalcea et al. 2007] use translation-based approach to explore existing opinion lexicons for languages as English to others in which linguistic resources such as WordNet are not available. They translate the opinion lexicon using bilingual dictionaries. The use of such resources imposes great restrictions to the work, as opinion lexicons usually contain multi-word expressions that are not contemplated in the dictionaries. JijKoun and Hofmann [Jijkoun and Hofmann 2009] explore this methodology further by applying an online automatic translation system and the WordNet to improve the results.

An opinion lexicon for Portuguese language has been recently developed for the domain of social judgment [Silva et al. 2010].Their work, however, focuses in the domain-specific characteristics of the lexicon and the choice of listing only adjectives, while we crafted a domain-free lexicon composed by polar words (adjectives, verbs and nouns) and expressions. It has been argued that the usage of domain-independent lexicons is unsatisfactory and domain-specific lexicons should be constructed. We agree that a domain-specific lexicon is potentially more useful than a domain-independent one, when available, but domain-independent lexicons have their importance when dealing with non-specified domains or when resources are not available for this construction and can be satisfactorily used when complemented with domain-specific data - as a classifier trained in such a domain.

Despite their advantages and shortcomings, we believe that all described methods can be implemented satisfactorily when the resources are available for it in any given language - for example the Portuguese language. It is our belief, however, that each different method tackles a slightly different problem when generating an opinion lexicon and thus are complementary, since each collects the polarity attributed by a different linguistic or social process. We propose the usage of different methods to enrich a Brazilian Portuguese Opinion Lexicon exploring their best qualities.

3. Opinion lexicon construction

The proposed technique consists in applying three methods in the literature: the Turney's corpus-based [Turney 2002], one of the top-performing methods in the literature; a thesaurus-based similar to Kamps et al's [Kamps et al. 2004] and a variation from Mihalcea et al's [Mihalcea et al. 2007] using an online automatic translation system instead of a bilingual dictionary.

Our method, then creates three different opinion lexicons that are conjoined to create a large lexicon for the Portuguese language. The seed sets, which can be seen in Figure1, were identical in all approaches.

$$Positive : \left\{ \begin{array}{c} \text{bom, ótimo, excelente, feliz, brilhante, fenomenal,} \\ \text{fantástico, espetacular, melhor, satisfatório} \end{array} \right\}$$

$$Negative : \left\{ \begin{array}{c} \text{ruim, péssimo, horrível, infeliz,} \\ \text{estúpido, odioso, pior, feio, insatisfatório} \end{array} \right\}$$

Figure 1. Seed Sets

3.1. Corpus-based lexicon

Due to time restrictions and given that we already disposed of processed corpus whose statistics were easily accessible to us, we decided to use a document corpus instead of an Internet search engine as in the original work [Turney 2002].

The corpus used in our experiments is composed by 346 movie reviews written in Brazilian Portuguese and extracted from the sites CinePlayers[3] and Cinema com Rapadura[4] websites and 970 journalistic texts about different themes extracted from the PLN-Br CATEG corpus [Bruckschen et al.]. The resulted corpus contain 1317 documents and around one million words.

Given all expressions extracted from the corpus and annotated using the Pointwise Mutual Information, we selected only the expressions for which its polarity were above the class' medium polarity, to guarantee a greater accuracy.

3.2. Thesaurus-based lexicon

The method implemented in this paper was based on Kamps et al's method [Kamps et al. 2004] using a distance function based synonymy and antonymy defined as the length of the minor path between the minimal-path of synonyms from one word to another or the minimal-path between its antonyms. Given the seed set (Figure 1), the polarity was, then, computed by the difference of minimum-distance to each seed class.

$$EVA(x) = \frac{min\{d(x,p)\} - min\{d(x,n)\}}{min\{d(p,n)\}}, \text{ for each } p \text{ in } \textit{Seed pos} \text{ and } n \text{ in } \textit{Seed neg}$$

As lexical resource, we used the TEP thesaurus [Maziero et al.] containing 44077 words and annotation of synsets and antonymy. As for the WordNet for English language [Kamps et al. 2004] the calculated distance between the words "bom" (good: adjective) and "ruim" (bad: adjective) is 4 with "bom" having 10 senses (25 on the WordNet for English language). For this, we only selected words in which the distance to at least one of seed sets are equal or lesser than 3.

3.3. Translation-based lexicon

We also tried to evaluate the use of existing resources for other languages - most notably the English language - to sentiment analysis in the Portuguese language. The resource used was the Liu's English Opinion Lexicon [Hu and Liu 2004][5] composed by nearly

[3]http://www.cineplayers.com

[4]http://cinemacomrapadura.com.br

[5]Available at http://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html.

6800 entries. To overcome some drawbacks of Mihalcea et al's method - the use of bilingual dictionary - and to decrease the manual work involved in their approach, we decided to use an automated translation system. In this work we used the Google Translate Online translation engine[6]. All translated words and expressions were used. The ones which the translation system could not translate - due to the high presence of linguistic variation, as common misspelling, in the original lexicon - were discarded by manual revision.

4. Results

The application of the three described methods resulted in three different opinion lexicons composed of 359 expressions for the corpus-based lexicon, 2400 words for the thesaurus-based lexicon and 4909 expressions for the translation-based lexicon. The intersection of words and expressions and agreement in their polarities among the lexicons - inside the parenthesis - can be seen in Table 1. We also compared the generated lexicons with the SentLex Lexicon [Silva et al. 2010] of human-related adjectives[7] for Portuguese language, composed of around 6322 annotated adjectives.

Note that the corpus-based lexicon is composed of bi-grams and, for such, doesn't have intersection with the ones composed of single words.

Table 1. Intersection and agreement of the generated lexicons

	SentLex	Corpus	Thesaurus	Translation
Sentlex	-	0	1347 (951)	1553 (1262)
Corpus	0	-	0	4(3)
Thesaurus	1347 (951)	0	-	587 (498)
Translation	1553 (1262)	4 (3)	587 (498)	-

The resulting lexicon, when conjoining the three generated lexicons is composed of 7077 polar words and expressions - the neutral words and expressions were not computed. The cases in which the polarity of a word or expression differed in two different sources were decided using simple heuristics based on the reliability in the sources. Since the translation-based lexicon was created using resources designed for other languages, and given that the translation process is not perfect, when a conflict occurs we always choose the other source.

Applying these lexicons to review classification, in a similar fashion and, thus, comparable to [Turney 2002], we obtained the results of Table 2 by using each lexicon separately, the conjunction of the lexicons constructed with Portuguese-specific resources and the conjunction of all three lexicons. The corpus of movie reviews used as test set is composed of 320 unseen documents from the same sources of the ones used in the corpus-based lexicon construction. The F-measure used was $F = 2pr/(p + r)$.

Aiming to evaluate the lexicon in a more qualitative manner, we selected an extract of 150 terms - 50 for each method - randomly selected to be analyzed. Since

[6]Accessible at http://translate.google.com/.

[7]Human-related adjectives are defined by being characterized as co-occurring with a human subject [Silva et al. 2010]

Table 2. Results of the review classification

Lexicon	Precision	Recall	F-measure	Accuracy
Corpus	0.468	0.275	0.346	0.247
Thesaurus	0.520	0.956	0.674	0.522
Translation	0.656	0.513	0.576	0.613
Corpus + thesaurus	0.526	0.963	0.680	0.528
Conjoined	0.745	0.769	0.757	0.741
SentLex	0.586	0.725	0.648	0.591

the terms are presented outside a given context and polarity is highly dependent of it, we decided to compare the polarity of terms using the SentiWordNet Lexicon [Baccianella and Sebastiani 2010], which differs from the others because it does not list nor annotate words, but their senses - given by the synsets they belong to in WordNet.

This analysis was made in 2 steps. First it was necessary to translate the words from Portuguese to English, since SentiWordNet was crafted for the English language. Then, by using the online interface of the lexicon[8] we searched the polarity of the words. Given the different senses of a lexical unit, sometimes they can be classified into positive, objective and also negative. It happens because some words can have both meanings, it depends on the semantics of words in the sentence. In cases where more than one polarity class was attributed to the lexical unit, the most frequent polarity was selected. The results of this analysis can be visualized in Table 3.

Table 3. Results of the evaluantion of lexicons using SentiWordNet

	Correct	Error	Acurracy
Corpus	21	29	0.42
Thesaurus	18	32	0.36
Translation	25	25	0.5
Total	64	86	0.427

Although this is a modest evaluation, given the small number of terms analyzed, we can see that the best accuracy coincides with the review classification - with the translation-based lexicon presenting higher results.

5. Discussion

The resulting lexicon, composed of 7077 words and expressions, is comparable to many other used in the literature, such as Liu's lexicon [Hu and Liu 2004] for the English language and SentLex [Silva et al. 2010] for the Portuguese language. Given the high intersection rate with SentLex and the overall concordance of polarity we can evaluate both the thesaurus-based and translation-based lexicons as satisfactory, which can be confronted with the results in review classification - F-measures of 0.674 and 0.576, respectively.

In the review classification analysis, it is worth noticing that the SentLex performed slightly worse than the generated lexicon. An explanation for it is that the SentLex

[8] Available at www.sentiwordnet.isti.cnr.it/

is domain specific - for the case of social judgment of people- and it has been, in part, automatically annotated with reported precision of 67%, 45% and 82% for the positive, neutral and negative classes, respectively.

The corpus-based lexicon, due to its particularities - constituted only of bi-grams - cannot be evaluated in a comparative way. Its low performance in the review classification can be understood by the few expressions composing the lexicon. A less rigid restriction in the selection of the expressions may increase such results. Another possible source of error is the relative small corpus used in the creation of this lexicon. For future research, we intend to test the performance of this method in a larger corpus and using a search engine, as in the original work.

The qualitative analysis of the lexicons has shown a low performance of the lexicons. Possible explanations to such performance are the relative small amount of terms analyzed, when compared to the length of each lexicon, the process of manual translation required the use of the SentiWordNet, which may introduce many errors, and the differences in the design of the SentiWordNet and of our lexicon, in which the terms are annotated outside a given context. It is remarkable that for the translation-based lexicon - which terms come originally from the English language - the results of both analysis are similar.

Differently than our initial expectations, the lexicon generated by the translation process has achieved better performance than the others in the review classification evaluation and manual comparing to the SentiWordNet. Possible explanations to such performance are the larger amount of terms in this lexicon compared to the others - exceeding the double of terms than the thesaurus-based one - which affects directly the review classification analysis and, since it was generated from a resource originally in English, it is not impressive that in the comparison to the SentiWordNet yielded higher scores.

The combination of the lexicons has clearly improved the results, both just using only the Portuguese-specific resources and using the translated resource. We consider that the resulting lexicon can yet be extended using the SentLex lexicon, but since it is a domain-specific resource, it is necessary to evaluate the discordance in annotations. Even thought we have applied and analyzed our method with the Portuguese language, we argue that it can be applied in any language. In those in which a thesaurus as WordNet has not yet been constructed a dictionary may be to extract synonyms through lexical and syntactic patterns and the other approaches can be easily implemented resulting in a satisfying lexicon.

References

Baccianella, A. E. S. and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 2200–2204, Valletta, Malta. European Language Resources Association (ELRA).

Bruckschen, M., Muniz, F., de Souza, J. G. C., Fuchs, J. T., Infante, K., Muniz, M., Gonçalves, P. N., Vieira, R., and Aluísio, S. Anotação lingüística em xml do corpus pln-br. Technical report, Universidade de São Paulo, São Paulo, Brazil.

Esuli, A. and Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. In *Proceedings of the International Conference on Information and Knowledge Management*, pages 617–624, Bremen, DE.

Fellbaum, C., editor (1998). *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, US.

Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 174–181, Morristown, US. Association for Computational Linguistics.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, US. ACM.

Jijkoun, V. and Hofmann, K. (2009). Generating a non-english subjectivity lexicon: relations that matter. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 398–405, Stroudsburg,US. Association for Computational Linguistics.

Kamps, J., Marx, M., Mokken, R. J., and de Rijke, M. (2004). Using WordNet to measure semantic orientation of adjectives. In *Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation*, pages 1115–1118, Lisbon, PT.

Liu, B. (2010). Sentiment analysis and subjectivity. In Indurkhya, N. and Damerau, F. J., editors, *Handbook of Natural Language Processing*. CRC Press, Taylor and Francis Group, Boca Raton, US, 2 edition.

Maziero, E. G., Pardo, T. A. S., Di Felippo, A., and Dias-da Silva, B. C. In *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, pages 390–392, Vila Velha, Brazil. ACM.

Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 976–983, Prague, CZ.

Riloff, E. and Shepherd, J. (1997). A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pages 117–124, Prague, CZ.

Silva, M. J., Carvalho, P., Costa, C., and Sarmento, L. (2010). Automatic expansion of a social judgment lexicon for sentiment analysis. *Technical Report TR 1008 University of Lisbon Faculty of Sciences LASIGE*.

Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Morristown,US. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hwa, R. (2006). Recognizing strong and weak opinion clauses. *Computational Intelligence*, 22(2):73–99.