Polarity Lexicon Building: to what Extent Is the Manual Effort Worth?

Iñaki San Vicente, Xabier Saralegi
Elhuyar Fundazioa
Osinalde Industrialdea 3,
20170 Usurbil
{i.sanvicente,x.saralegi}@elhuyar.com

Abstract

Polarity lexicons are a basic resource for analyzing the sentiments and opinions expressed in texts in an automated way. This paper explores three methods to construct polarity lexicons: translating existing lexicons from other languages, extracting polarity lexicons from corpora, and annotating sentiments Lexical Knowledge Bases. Each of these methods require a different degree of human effort. We evaluate how much manual effort is needed and to what extent that effort pays in terms of performance improvement. Experiment setup includes generating lexicons for Basque, and evaluating them against gold standard datasets in different domains. Results show that extracting polarity lexicons from corpora is the best solution for achieving a good performance with reasonable human effort.

Keywords: Sentiment Analysis, Polarity Lexicons, Evaluation

1. Introduction

Research effort on the sentiment analysis field has seen exponentially increased in the last years, due to its applicability in areas such as VTIC (Technological Surveillance / Competitive Intelligence), marketing or reputation management. One of the main resources of sentiment analysis systems are the polarity lexicons, list of words with prior polarities. Much research has been done on methods for building such methods automatically due to the high cost of manually created lexicons. Then again, automatic methods often produce noisy resources.

Very little work has been done on polarity lexicons for Basque, as is the case for other less resourced languages. Thus, when facing the task of creating such a resource, the doubt arised. Is it worth to make a great manual annotation effort? How much is the gain we obtain by manually annotating polarity words over automatically built polarity lexicons?

This paper compares three strategies for building a polarity lexicon for a less-resourced language. We assumed that for languages of this type the availability of parallel corpora, MT systems and polarity-annotated data is very limited, and we avoided using such resources. We measured the time cost of the manual effort and the gain it brings in terms of accuracy in an extrinsic evaluation. This experiment was carried out for Basque.

2. State of the Art

Polarity lexicons are key resource on sentiment analysis systems. We can group the methods for polarity lexicon building proposed in the literature into three main approaches: manually constructed lexicons (Stone et al., 1966), corpus-based methods (Hatzivassiloglou and McKeown, 1997; Mihalcea et al., 2007) and methods that rely on Lexical Knowledge Bases (LKB) (Kamps et al., 2004; Liu and Singh, 2004; Kim and Hovy, 2004) For major languages there are well known manually constructed lexicons, such as, General Inquirer(Stone et al., 1966), OpinionFinder(Wilson et al., 2005), or SO-CAL(Taboada et al., 2011). Due to the fact that a great human effort is needed to build such resources, some of them are semi-automatically constructed, and manually corrected afterwards. In this line of work, some researchers explore the possibility of using resources already existing in another language (e.g., lexicons, and/or annotated corpora corpora). (Mihalcea et al., 2007) and (Perez-Rosas et al., 2012) analyze the approach of translating English resources into Romanian and Spanish, respectively. However, only a small portion of the translated lexicon entries maintain the correct polarity. The need to treat ambiguous translations becomes clear.

Corpus-based methods require some sort of polarity annotation to construct the lexicons. We can find two main approaches in this group: i) starting from a small list of words with known polarity, find words in a corpus that are semantically close by means of distributional methods (Turney and Littman, 2003), and ii) Based on a corpus that has polarity annotations at document or sentence level, create list of words most related to either positive or negative annotations (Saralegi and San Vicente, 2012).

Finally, the main idea behind LKB-based methods is to propagate to new words the polarity of a small list of seed words with known polarities, by making use of relations between concepts the LKB offers. Propagating polarity through graphs representing the semantic relations existing in WordNet (WN) (Fellbaum, 1998) is a well known strategy (Esuli and Sebastiani, 2006; San Vicente et al., 2014).

With respect to the specific case of Basque, we have found two polarity lexicons in the literature. The NRC Word-Emotion association lexicon, constructed in a crowdsourcing annotation effort, was translated using Google Translate to Basque (NRC_{eu}) (Mohammad and Turney, 2013). The second lexicon is MLSenticon (Cruz et al., 2014), which is an LKB-based lexicon generated in a similar way to SentiWordNet (Esuli and Sebastiani, 2006).

3. Lexicon Building methods.

Our aim is to compare three methods for polarity lexicon building which require a different degree of human edition:

(i) Translating lexicons in other language into our language; (ii) extracting automatically polarity words from corpora; eta (iii) annotating the polarity of the words in an LKB.

3.1. Projection

Projecting polarity lexicons from other languages by means of bilingual dictionaries seems like a direct way to create a lexicon in our language. However, this approach has to deal with the problems derived from the translation process: ambiguous translations and changes in the polarity of the target words.

Spanish lexicon $ElhPolar_{es}$ (Saralegi and San Vicente, 2013) has been translated by means of the Elhuyar Spanish-Basque dictionary[1] (173,931 translation pairs). For each Spanish entry in the lexicon, the first 5 translations are included in the translated lexicon Lex_{pr}.

Lex_{pr} has been initially reviewed by a native speaker correcting the polarity of each word. 3.4. offers details on the cost of this correction effort. Furthermore, a second reference by another annotator was later carried out on part of Lex_{pr}. Details about this second annotation effort are given in section 4..

The corrected lexicon contains 5,335 entries, 1,938 positive and 3,397 negative, very similar numbers to its original Spanish version.

	#entry	#positive	#negative
$ElhPolar_{es}$	5.195	1.892	3.303
Lex_{pr}	11.413	4.934	6.479

Table 1: ElhPolar source and translated lexicons' statistics.

3.2. Corpus-based lexicons

The second approach is based on the idea that words that tend to appear in texts with a certain polarity (positive or negative) are good representatives of that polarity. Usually association measures (AM) are used to find salient words in corpora (Kilgarriff, 2001).

Ideally, we would use a corpus with polarity annotations, which we could divide into positive and negative subparts. Unfortunately, no such resource exists for Basque and many other Less-resourced languages. As a solution, we adopted a semi-automatic approach relying on a corpus including subjective and objective documents (Saralegi et al., 2013). Such a corpus can be built in an easy way from a newspaper corpus taking as subjective documents opinion articles and as objective event news.

Using the Loglikelihood ratio (LLR) (Dunning, 1993) we obtained the ranking of the most salient words in the subjective part with respect to the rest of the corpus. The top 5,000 subjective words were manually checked by a single annotator. The corrected lexicon (Lex_c) contains 1.659 entries (959 negative and 691 positive). This method ranks a lot polar word candidates among the first positions because subjectivity highly correlates to polar words.

[1] http://hiztegiak.elhuyar.eus

3.3. LKB-based lexicons

Using the semantic relations represented in LKBs in order to construct polarity lexicons is a widespread strategy in the literature. In our case, we apply the method presented in (San Vicente et al., 2014) for generating basque polarity lexicons. *Q-WordNet as Personalized PageRanking Vector* (QWN-PPV) represents the concepts and the semantic relations between them stored in a WN like LKB over a graph. The method propagates the polarity of an initial set of words by applying the so-called Personalized PageRank algorithm on a LKB. We use the UKB (Agirre and Soroa, 2009) implementation of the algorithm.

The Basque WN (Pociello et al., 2010) is small compared to others. Thus, we chose to use MCR (Agirre et al., 2012) as LKB. Because it connects WNs for several languages including Basque, we can take advantage of a number of semantic relations existing in larger WNs which offer a bigger chance to propagate polarity information. Two graph representations are used, one including synonymy relations and another antonymy relations. We chose this graph representation because it creates higher quality propagations, although the limited number of relations results on smaller lexicons.

The lexicon produced with this approach ($Lex_{qwn-ppv}$) contains 1.132 entries, 565 positive and 567 negative.

The settings used in this work for QWN-PPV are derived from the experiments carried out in (San Vicente et al., 2014).

3.4. Correction effort

Usually, the main problem of the manual effort is its high cost. In this work we have measured the annotation effort required to correct the lexicons. As an indicator of that effort we have used what we call production rate. We understand production rate as the number of words added to our lexicon per minute.

Projection Altogether, a single annotator needed 36 hours to correct the Basque projected lexicon Lex_{pr}. That means that the correction rate was 5,3 word/minute. As general remark, we can say that the correction requires a great manual effort, because the dictionary-based translation selects many unusual translations (rarely used words), which leads annotators to consult frequently dictionaries and corpora.

Corpus-based lexicon In contrast to the translation approach, the annotator must decide the polarity of a word without any prior information on this regard, but, on the other hand, since the words are extracted from a corpus by means of LLR, the list contains more frequently used words. Hence, it is easier to annotate the polarity of common words as dictionaries and corpora are not so frequently needed. Overall, 10 hours were needed to annotate the polarity of the 5,000 candidate list. This means a correction rate of 8,3 word/minute.

Figure 1 shows the average production rates of the annotation process, for the various candidate ranking intervals. The higher production rates achieved for the first ranked candidates in the corpus-based method

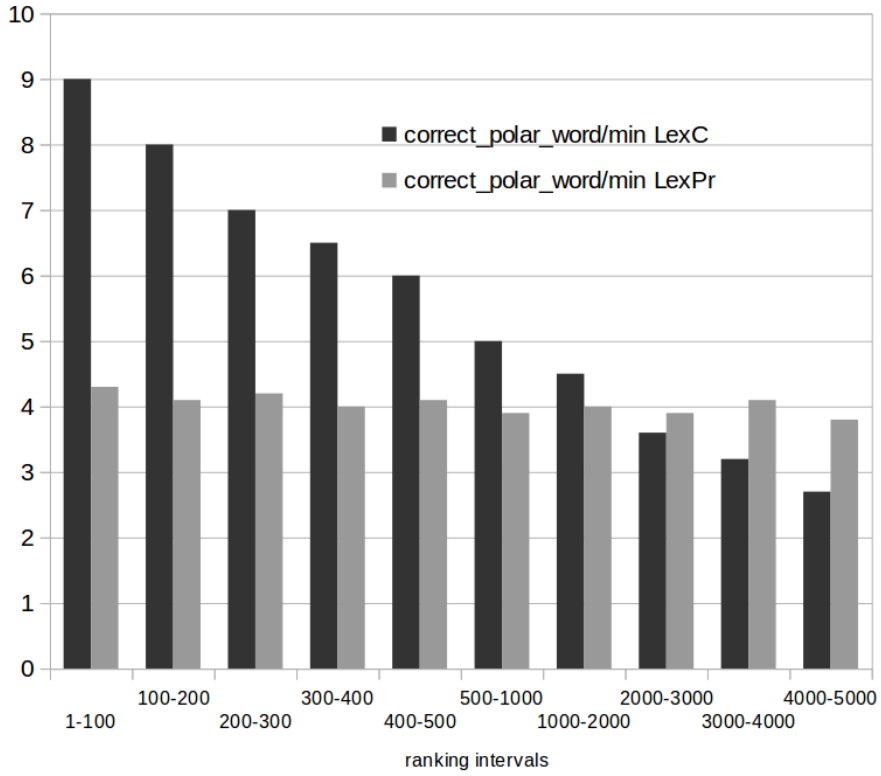

Figure 1: Correction speed and productivity data for Lex_{pr} and Lex_c.

(correct_polar_word/min Lex_C), is due to those candidates being most frequent words. The deeper we go into the ranking, the more unusual words appear, and hence the correction speed is reduced. Also, the higher production rate observed for the corpus-based method indicates that indeed LLR surfaces polar words, having a higher density in the first positions of the ranking. The top-ranking words are those which have most association degree with subjective corpus. For comparison between projection and corpus-based methods, only data for the first 5,000 candidates is shown.

3.5. Second reference

A single reference may not be fully trustworthy, and so we introduced a second reference for both projected and corpus-based lexicons. Due to the time constraints, we asked the second annotator only to review those words in the intersections between the lexicons and the datasets evaluated. Disagreements were resolved by discussion. This second annotation allows us to measure the improvement we can gain with that extra effort, as will be explained in section 4.. Table 2 shows the number of lemmas annotated on this second annotation, inter-annotator agreement (Cohen's Kappa κ value) data for positive (+) and negative (-) words and the time spent on discussing the disagreement cases.

Lexicon	**#Lemmas annotated**	κ +	κ -	**Disagreements**	**Discussion time (min)**
Lex_{pr}	599	0.624	0.765	80	65
Lex_C	542	0.747	0.835	56	40

Table 2: Statistics for the second annotation effort.

4. Evaluation

In order to evaluate the adequacy of the generated lexicons we set up a binary polarity classification task (positive vs. negative). As there is no corpus with gold annotations, we have generated two small datasets manually annotated at sentence level. Section 4. give details about those datasets.

Classifier We implement a simple average polarity ratio classifier. There are two reasons to choose such classifier: on the one hand, the lack of and annotated corpus prevents us from using supervised classifiers, and on the other, our aim is to minimize the role other aspects play in the evaluation and focus on how, other things being equal, polarity lexicons perform in a Sentiment Analysis task. The *average ratio classifier* computes the average ratio of the polarity words found in document d:

$$Pol(d) = \frac{\sum_{w \in d} pol(w)}{\#w} \quad (1)$$

where, for each word w, *pol(w)* is the polarity of the word in the lexicon ($1 = positive, -1 = negative$) or 0 if the word is missing. If $Pol(d) > 0$ d is classified as positive, and otherwise as negative.

Evaluated lexicons Our aim is to evaluate to what extent manual effort brings improvement. Overall we include

11 lexicons in the evaluation. For both Projection and Corpus-based lexicons 3 lexicons are evaluated, one for each of the annotators (Rows starting "AnnotX" in table 4) and a third generated from the consensus of those annotations (Rows starting "Consens" in table 4). In addition, the projected lexicon before manual annotation is included as a baseline. The LKB based lexicon provides comparison with a fully automatic method. The combination of both corpus-based and projected lexicons annotated represents what the greatest manual effort can achieve. Lastly, for the sake of comparison, although we didn't build them, we include the two publicly available polarity lexicons for Basque found in the literature: NRC_{eu} and MLSenticon.

Test datasets Two test-sets were compiled from different sources: One from the news domain, composed of newspaper articles, and another one from music and film reviews. Overall 224 sentences were gathered and manually annotated as positive and negative (see table 3). Neutral polarity sentences were discarded.

Domain	Positive	Negative	Overall
Music&Film reviews	%75.58	%24.42	86
News	%25.36	%74.64	138
Overall	%44.64	%55.36	224

Table 3: Test datasets estatistics.

4.1. Results

Table 4 presents the results obtained by the various lexicons over the test datasets. Accuracy (Acc.) and F-score values per category (Fpos/Fneg) are reported. Corpus-based lexicon achieves the best results across all datasets. As expected manually corrected lexicons perform better that the automatically generated lexicon.

Overall, results show corpus-based lexicons obtain very similar results to those of the translated lexicons, with much less human effort. Furthermore corpus-based lexicons' performance is far better in the Music&Film review domain.

Also, results show that a second annotation and the following discussion does indeed improve the quality of the lexicon in terms of accuracy. This of course means a greater annotation effort.

As an upper bound, the combination of the translated and corpus-based lexicons obtains the best results overall, although it also means the greatest annotation effort.

The performance of the automatically built LKB-based lexicon is far from the manually corrected lexicons, although its performance is similar to that of Lex_{pr}, the other completely automatic lexicon in the evaluation. Moreover, Basque WN suffers from a severe lack of information on adjectives. As adjectives are important for polarity detection, a better coverage would improve the lexicons generated with this strategy.

With respect to external lexicons, NRC_{eu} obtains modest results. There are to main reasons that led to its poor performance. The lexicon contains some incorrect entries and many of the entries correspond to word forms instead of lemmas. This is probably a side effect of the automatic translation. MLSenticon's results are very close to our own automatic Lexicon $Lex_{qwn-ppv}$. This is not surprising, since they both rely in a similar method and use MCR to obtain Basque lemmas.

5. Discussion and Conclusions

This paper explores three methods to build polarity lexicons from scratch. The adequacy of those methods has been evaluated on a polarity classification task over data from two different domains.

Semi-automatic corpus-based generation of polarity lexicons would be an adequate approach for scenarios where time for manual effort is limited. The manual effort required in this strategy is not very costly (10 hours). Even if the lexicon is not very large, the fact that it is corpus-based guaranties that most used polar words will be present.

For the scenarios where the accuracy is critical the combination of both projection and corpus-based strategies with at least two annotators would be desirable for building the polarity lexicon.

We plan to extend this research by constructing new polarity annotated datasets. This will allow us, on the one hand, to evaluate our resources using a machine-learning approach, which would be the first ML sentiment analysis system for Basque; and, on the other, new datasets would provide resources to generate new lexicons. Finally, repeating the experiments with other languages would add robustness to the contribution of this paper.

6. Acknowledgments

This work has been supported by Basque Government Elkartek program, in the framework of the Elkarola project (grants no. KK-2015/00098, KK-2016/00087).

7. Bibliographical References

Agirre, E. and Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, Athens, Greece.

Agirre, A. G., Laparra, E., Rigau, G., and Donostia, B. C. (2012). Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. In *GWC 2012 6th International Global Wordnet Conference*, page 118.

Cruz, F. L., Troyano, J. A., Pontes, B., and Ortega, F. J. (2014). Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994, October.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Esuli, A. and Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language*

Lexicon	News			Music&Films			Overall		
	Acc.	Fpos	Fneg	Acc.	Fpos	Fneg	Acc.	Fpos	Fneg
Projection									
Lex_{pr}	0.63	0.41	0.73	0.63	0.69	0.53	0.63	0.57	0.68
Annot1-Lex_{pr}	0.80	0.61	0.87	0.67	0.75	0.55	0.75	0.69	0.80
Annot2-Lex_{pr}	0.78	0.61	0.85	**0.76**	0.81	0.67	0.77	0.72	0.81
ConsensLex_{pr}	**0.86**	0.68	0.91	0.70	0.75	0.62	**0.79**	0.72	0.84
Corpus-based									
Annot1-Lex_c	0.77	0.56	0.84	0.79	0.85	0.64	**0.78**	0.74	0.80
Annot2-Lex_c	0.75	0.48	0.84	0.74	0.81	0.61	0.75	0.69	0.79
Consens-Lex_c	**0.78**	0.56	0.86	**0.80**	0.86	0.67	**0.79**	0.75	0.82
Automatic									
$Lex_{qwn-ppv}$	0.67	0.21	0.79	0.55	0.68	0.20	0.63	0.53	0.69
Combination									
ConsensLex_{c+pr}	**0.88**	0.74	0.92	**0.83**	0.87	0.73	**0.86**	0.82	0.88
External									
NRC_{eu}	0.62	0.29	0.74	0.47	0.51	0.41	0.56	0.41	0.65
$MLSenticon$	0.65	0.37	0.76	0.55	0.60	0.48	0.61	0.50	0.68

Table 4: Evaluation results for the various lexicons on the test datasets.

Resources and Evaluation (LREC 2006), pages 417–422, Genoa, Italy, May.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

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.

Kamps, J., Marx, M., Mokken, R. J., and Rijke, M. D. (2004). Using wordnet to measure semantic orientation of adjectives. In *Proceedings of LREC 2004*, Lisbon, Portugal.

Kilgarriff, A. (2001). Comparing corpora. *International journal of corpus linguistics*, 6(1):97133.

Kim, S.-M. and Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of Coling 2004*, pages 1367–1373, Geneva, Switzerland, August. COLING.

Liu, H. and Singh, P. (2004). ConceptNet: a practical commonsense reasoning toolkit. *BT Technology Journal*, 22:211226.

Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Annual Meeting of the Association for Computational Linguistics*, volume 45, page 976.

Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Perez-Rosas, V., Banea, C., and Mihalcea, R. (2012). Learning sentiment lexicons in spanish. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May.

Pociello, E., Agirre, E., and Aldezabal, I. (2010). Methodology and construction of the basque WordNet. *Language Resources and Evaluation*, page 122.

San Vicente, I., Agerri, R., and Rigau, G. (2014). Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden*, pages 88–97.

Saralegi, X. and San Vicente, I. (2012). Tass: Detecting sentiments in spanish tweets. In *Proceedings of the TASS Workshop at SEPLN*.

Saralegi, X. and San Vicente, I. (2013). Elhuyar at TASS2013. In *Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS2013)*, pages 143–150, Madrid.

Saralegi, X., naki San Vicente, I., and Ugarteburu, I. (2013). Cross-lingual projections vs. corpora extracted subjectivity lexicons for less-resourced languages. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7817 of *Lecture Notes in Computer Science*, pages 96–108.

Stone, P., Dunphy, D., Smith, M., and Ogilvie, D. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge (MA): MIT Press.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Turney, P. and Littman, M. (2003). Measuring praise and criticism: Inference of semantic oreintation from association. *ACM Transaction on Information Systems*, 21(4):315–346.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). OpinionFinder. In *Proceedings of HLT/EMNLP on Interactive Demonstrations* -, pages 34–35, Vancouver, Canada.