

SentiTegi: building a semantic oriented Basque lexicon

Jon Alkorta, Koldo Gojenola, and Mikel Iruskieta

IXA NLP Group. University of the Basque Country.
Manuel Lardizabal Pasealekua, 1 - E20018 Donostia
{jon.alkorta,koldo.gojenola,mikel.iruskieta}@ehu.eus

Abstract. The creation of a semantic oriented lexicon of positive and negative words is often the first step to analyze the sentiment of a corpus. Various methods can be employed to create a lexicon: supervised and unsupervised. Until now, methods employed to create Basque polarity lexicons were unsupervised. The aim of this paper is to present the construction and evaluation of the first semantic oriented supervised Basque lexicon ranging from -5 to +5. Due to the lack of resources, the Basque lexicon was created translating the SO-CAL Spanish dictionary by means of two bilingual dictionaries following specific criteria and then slightly corrected with the SO-CAL English dictionary and frequency data obtained from the Basque Opinion Corpus. Evaluation results show that the correlation between human annotators is slightly better than between a gold standard lexicon (obtained from human annotation) and the translated dictionary. This shows that the quality of the translated dictionary is satisfactory, although there is a space to improve it.

1 Introduction

The aim of this paper is to present a semantic oriented lexicon for Basque. We will emphasize the process of creating this lexicon, and particularly the solutions adopted to solve the problems encountered during the process.

Sentiment analysis is a task that classifies documents according to their polarity. This research area has had a big development in the last years due to social networks and Internet, which have increased the quantity of opinions and other types of text with emotion, and is in demand of methods for automatic processing.

There are many resources for sentiment analysis for the most used languages such as English [1], Chinese [2] and Spanish [3]. Additionally, competitions like SemEval [4] have greatly contributed to the development of resources and tools for sentiment analysis. However, the development is not symmetric on lesser used languages or languages in normalization process like Basque.

The creation of a semantic oriented Basque lexicon is a part of a bigger study with the objective of analyzing sentiment analysis in Basque. Its aim is to study how different phenomena affect to linguistic structures where opinions or sentiments are situated in. The study of sentiment analysis takes into account

three levels of the language: i) word level, where this work is located, ii) sentence level and iii) document level.

The semantic oriented lexicons are related to the lexical level and so, they are useful and important in sentiment analysis. If the semantic orientation of the words is known, the opportunities open up to calculate the semantic orientation of sentences and, therefore, the semantic orientation of texts taking into account syntax and discourse constraints.

The creation of the semantic oriented Basque lexicon has been semi-manual translating from the SO-CAL Spanish dictionary and then, enriching it with corpus analysis and the English SO-CAL dictionary. In the translation process, different bilingual dictionaries have been used. We have decided to use a semi-manual procedure to create our lexicon because it takes the language characteristics of Basque into account.

The main contributions of this work are: i) the creation of a domain-specific semantic oriented Basque lexicon, ii) the use of a semi-manual technique for the lexicon creation and iii) a thorough evaluation.

This paper has been organized as follows: after presenting related work in Section 2, Section 3 describes the methodology of the translation process. Then, Section 4 discusses the design decisions, while Section 5 gives the main results. In Section 6 the quality of the lexicon is evaluated and, finally, Section 6 concludes the paper, also proposing directions for future work.

2 Related Work

There are various approaches for the creation of polarity lexicons, based on knowledge or on automatic methods. Each of the approaches has its advantages and drawbacks.

SO-CAL [5] is a dictionary-based tool to extract sentiment from texts. The dictionary was created manually, where words are annotated with polarity and strength (semantic orientation). In addition, it has incorporated a treatment for intensifiers and negation. In total, the tool contains five dictionaries, each corresponding to one grammatical category: nouns, adjectives, adverbs, verbs and intensifiers. The English and Spanish dictionaries (V1.11) contain 6,610 and 4,880 words, respectively. A disadvantage of manually-created lexicons is the hard-work to make modifications. In contrast, they can be tailored to be domain-specific and, depending on the linguistic information used, they can treat a variety of different linguistic phenomena.

ML-SentiCon [6] is a multilingual polarity lexicon, where the lexicons have been automatically generated from an improved version of SentiWordNet, layered according to their precision. It contains a Basque lexicon that contains 4,323 lemmas. The polarity values are situated between -1 and $+1$, in a continuous scale. Additionally, the QWN-PPV tool [7] is able to generate multilingual polarity lexicons, including Basque. This unsupervised tool makes use of a corpus and WordNet. The main disadvantage of these lexicons is that they are not domain-

specific, so their results could vary from one domain to another. In contrast, their main advantage lies on the facility to create them.

The methods to evaluate lexicons are different depending on each technique. Some works [8] use intrinsic method where the result of the system is compared to answers, predefined by evaluators. In contrast, there are other systems [9] which use extrinsic methods where the system is evaluated in an applied setting. Finally, some works [10] use both extrinsic and intrinsic methods.

The lexicon presented in this work differs from previous ones in several respects. SO-CAL dictionaries have also been manually-created but, until now, they have dealt with languages which are not morphologically rich (Spanish and English) in contrast with Basque. Another relevant difference of this study has been the evaluation. We will measure, using Pearson correlation, to describe the reliability between two human annotator agreement and, the reliability between the gold standard (based on human annotation) and the translated dictionary. In other other words, we will apply an intrinsic evaluation.

3 Methodology

In order to create a semantic oriented lexicon for Basque, we have adopted some decisions taking some factors into account:

- i) Time. The creation of semantic oriented lexicon for Basque is related to the project of linguistic-based Basque sentiment analysis and, for that reason, the time to create the lexicon is limited.
- ii) Resources. The Basque language is still in a normalization process and this has some limitations to create corpora and to reuse computational resources. On the one hand, it is difficult to create a large opinion corpus of different topics. This situation could affect to the quality of the lexicon if the corpus is used for that. The collaboration of lexicographers would be ideal but it is a costly resource, not available. This situation adds a difficulty to create a semantic oriented Basque lexicon.
- iii) Quality. We want to develop the lexicon with the best possible quality (and in the less time possible) and with that aim we will first evaluate and then take steps for improvement of our semantic oriented lexicon.

3.1 Resources for translation

We have used mainly four resources in translation process.

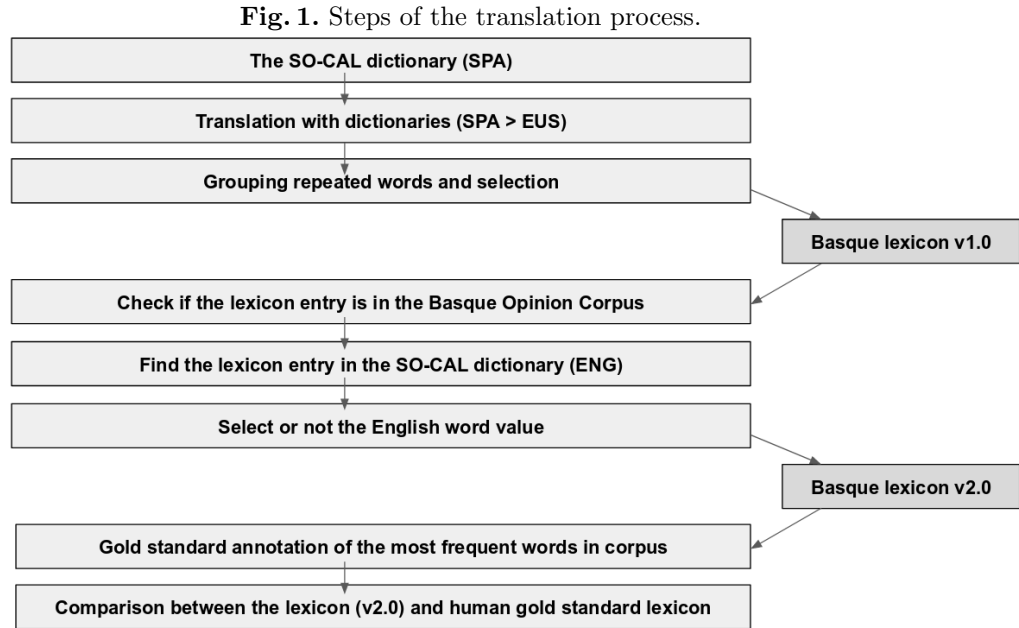
- i) **The SO-CAL Spanish dictionary [5]**. It contains 4,880 words of five grammatical categories (noun, adjectives, adverbs, verbs and intensifiers). It is the dictionary which has been translated.
- ii) **Elhuyar dictionary [12]**. It has been one of the dictionaries to translate the Spanish dictionary. Moreover, it has been used to check if the translated word is an entry of that dictionary since some translated words can be an entry while other can not be because they are collocations or expressions. We will work only with words which are entry of the dictionary.

- iii) **The Basque Opinion Corpus [14].** After getting the first version of the lexicon, the words which contains it have been checked in the corpus to create a domain-based lexicon. The corpus contains 240 texts of six different domains.
- iv) **The SO-CAL English dictionary [5].** This version which contains 6,610 words has been used to enrich the already created domain-based lexicon.

Taking all the factors explained above into account and using the mentioned resources, we have decided to translate the SO-CAL Spanish dictionary with a specified methodology (see Figure 1).

3.2 Translation steps

Figure 1 shows the steps followed in the translation process. Firstly, the first version of semantic oriented Basque lexicon has been created from the Spanish version of the SO-CAL dictionary. After that, the second version has been created enriching it with the English lexicon version (V1.11) and limiting to the domains of Basque Opinion Corpus.



In the translation process of SO-CAL dictionaries from Spanish and English versions (V1.11), several phenomena, presented in Table (1) have been detected.

In total, five phenomena have been treated:

Table 1. Different options in the lexicon creation process.

| Phenomenon | SPA | SPA grouping | EUS | ENG | Value |
|------------|-----------------------------|---|-------------------------------|------------------------|-------|
| 1 | desacreditar “discredit” | desacreditar -2 “discredit” | ospea_kendu -2 “discredit” | - | - |
| 2 | atrofiar “atrophy” | atrofiar -1 “atrophy” | atrofiatu -1 “atrophy” | - | - |
| 3 | amago “feint” | amago “feint” -1 cicatriz “scar” -2 | seinale “signal” -1 | - | - |
| 4 | franquismo “Francoism” | franquismo -2 “francoism” | frankismo -2 “francoism” | - | -2 |
| 5 | correcto “correct” | acertado “correct” +3 correcto “correct” +3 decente “decent” -2 | zuzen +3 “correct” | right +1 correct +3 | +3 |

- 1) The Spanish word is translated but the translation is not an entry of Elhuyar dictionary [12], so we do not take it into account.
- 2) The Spanish word is translated, it is an entry of Elhuyar but the translation does not appear in the Basque Opinion Corpus. Consequently, it will appear in the first version but not in the second one.
- 3) The Spanish word is translated, it is an entry, it appears in the corpus but it is not in the SO-CAL English dictionary. So, it will appear in the first version of the dictionary, but not in the second one.
- 4) The Spanish word is translated, it is an entry, it appears in the corpus and it is not present in the SO-CAL English dictionary. Then, it will be included in the (first and) second version.
- 5) The Spanish word is translated, it is an entry, it appears in the corpus and it also a word of the SO-CAL English dictionary. It will appear in the first and second versions.

The previous phenomena are the result of the translation process which is shown below.

- i) **Translation.** The Spanish version of the dictionaries has been translated using Elhuyar [12] and Zehazki [13] dictionaries. When one word of the dictionary has more than one entry, all the entries have been taken into account. The value of the Spanish word has been transferred to all the translations in Basque.
For example, the Spanish word *desacreditar* -2 “discredit” has been translated to Basque in different forms: *izena_kendu*, *ospea_kendu* and *sona_kendu* “discredit”. This example shows how one Spanish word could be translated in different forms to Basque. The word of this example the same that appear in Table 1. In the next step, the Basque words will be grouped.
- ii) **Grouping.** After translating all the words and transferring their values, the repeated words in Basque have been grouped.

Table 1 shows how the Basque words (fourth column) can have one or more translations in Spanish (third column). The phenomena numbered 1, 2 and 4 have one translation in Spanish whereas 3 and 5 have more than one.

This phenomenon occurred because those words are polysemic. There are cases where two or more words in Spanish correspond to the same word in Basque and vice versa. Consequently, each word in Basque has several meanings and values (related to Spanish words) in Spanish.

- iii) **Check if the Basque translation is an entry in the Elhuyar dictionary.** We have only accepted the translations which are entries of the Elhuyar dictionary. Consequently, the phenomenon 1 in Table 1 has occurred: *ospea_kendu* “discredit” is a collocation and not an entry, so we will not take it into account. In contrast, other words in the table are entries in the dictionary and they are maintained.
- iv) **Selection.** The value (and meaning in Spanish) of each word in Basque will be selected. In order to choose the value, we have followed the following criteria:
 - If the word in Basque has one translation (and value) in Spanish and if that translation is correct, the translation is selected. This is the case of phenomena 2 and 4 in Table 1. Sometimes the translation is not “correct” or “direct” as we will observe in Section 4.
 - If the word in Basque has many translations (and values) in Spanish, the translation has been selected according to what of the translations is the best to use in the Basque Opinion Corpus [14]. We have analyzed the context of the words in the corpus using Key Word In Context (KWIC) format for concordance. This is the case of phenomena 3 and 5 in Table (1).
 - There have also been cases where the word in Basque has not examples in the corpus. In these cases, the meanings which are used more frequently has been selected.

After these three steps, the first version of the Basque lexicon (V1.0) has been created. However, we detected some inconsistencies and we have felt the necessity to feed more information and, for that reason, we followed new steps to create the Basque lexicon (V2.0).

- v) **New lexicon based on the Basque Opinion Corpus [14].** We have created a new lexicon based on the first one. The new lexicon has been created with the words of the lexicon that appear in the corpus. The effects of this step are showed in phenomenon 2 in Table 1. The word *atrofiatu* “to atrophy” does not appear in the corpus, so it is not related to the domains of the corpus and, consequently, we do not take it into account. In Table 1, phenomena 3, 4 and 5 are not affected by this limitation while phenomenon 2 is. With this procedure, the number of entries in the lexicon was reduced from 8,140 to 1,813 words.
- vi) **Find the English translations of lexicon words in the SO-CAL English dictionary.** Using the Elhuyar dictionary, we have translated the

words in Basque to English and, after that, we have checked if the words are in the SO-CAL English dictionary. If the word is in the dictionary, we have added this translation and its value to the word in Basque. If the word is not in the dictionary, we have left empty the space.

In Table 1, phenomena 3 and 4 do not have any translation in the English dictionary and, consequently, their (English) column in Table 1 is empty. In contrast, phenomenon 5 has two translations according to the English dictionary: *right* and *correct*.

- vii) **Compare and choose the best translation and value.** In this step, each word in Basque has one translation and value in Spanish and sometimes, the English word and its value are linked to Basque one. There are 3 different cases in this situation. The number of each case is linked to the phenomenon number in Table (1):

- 3) There is not word in the English version corresponding to the Basque word and the previous Spanish one is not accepted. In phenomenon 3, the word *seinale* “sign” has been assigned the value -1 (Table 1, fourth column) but there is not a corresponding value in the English version and, consequently, we have removed that value.
- 4) There is not a corresponding word in the English version for Basque and the previous Spanish translation and value are accepted. The word *frankismo* “francoism” is related to Spain and, for that reason, it appears in the Spanish version and not in English. In this case, we have maintained the assigned value.
- 5) The English translation and value are the same or better quality than the Spanish ones. Phenomenon 5 shows that the Spanish and English values agree, so we have assigned the value $+3$ to the word. In other cases, the English and Spanish values, which the Basque words contain from the first version, differ and, in that case, the English value prevails to the Spanish one. The reasons are explained in Section 4.

Phenomena 3 and 5 show how we have decided to give more relevance to English version, although sometimes there is not a corresponding word in the English dictionary. The reasons for that decision are explained in Section (4).

4 Discussion

In the translation process, some problems have been detected and we have taken some decisions in order to solve them.

- i) **Source language and preferred language.** English and Spanish could be the source language but we have chosen Spanish due to several reasons. The overall accuracy of the English SO-CAL is 76.62% while in the Spanish version is 71.81% [11]. In other words, the difference between them is not big enough. On the other hand, there are many more resources to translate the dictionary from Spanish to Basque than to translate from English to Basque. So, the translation from Spanish is more reliable and extended as shown in

Table 1, where the phenomenon numbered 4 (*frankismo* "francoism") shows that although the English dictionary contains more items, there are some words in the Spanish dictionary that are not present in the English one.

In contrast, the English version has helped to check if the assigned value to the Basque word in the first version from Spanish is correct. In the cases where the value of the Spanish and English versions are different, we have preferred the English one as phenomenon 3 (*señale* "signal") shows. Due to this decision, the number of words of the lexicon has decreased from 1,813 to 1,237.

- ii) **Taking into account all the translations in Basque.** Another problem was presented when, in the translations, the Spanish word could be translated to Basque in different forms. We have decided to use all the translated words in Basque so as to get the higher recall possible. The first step (3.2) in the process shows that one or more entries have been taken in Basque.
- iii) **Polysemic words.** There are some words that have opposite meanings according to their context. The best solution would be to create two entries but then it would be difficult to implement it in a system that does not distinguish between word senses. In this situation, we have decided to take only one meaning and we have used the Basque Opinion Corpus to choose the meaning. When the word did not appear in the corpus, we took the meaning used with more frequency.

For example, the Basque word *deigarri* "flashy" can be in Spanish *aparatoso* -3 "spectacular" or *llamativo* +3 "showy". Taking the corpus into account, we have chosen the value +3 for the word.

- iv) **Coherence consistency.** In the process of choosing the value, we have to try (when the values match) to maintain the coherence of the values taking these criteria into account. Examples of the criteria are shown in Table 2.
 - A) Sometimes, the same word appears in different forms. For example, it is usual that one word appears sometimes with genitive *-ko* "with" (genitive) and other times with an elided genitive. In these cases, we decided to assign the same value.
 - B) We have to try (when the values match) to assign the same value to words with similar meaning.
 - C) We have also assigned the value of the same intensity to antonymic words when the values coincide in Basque words. In Basque, some prefixes (*des-*, *ez-* "dis-") and suffixes (*-ezin*, *-gabe* "without") are used to invert the meaning of the words and we have put special attention on these ones.

Table 2. Examples of translations applying the coherence criteria.

| Criteria | EUS | Value | EUS | Value |
|----------|-------------------------|-------|--------------------------------|-------|
| A | errukigabe "ruthless" | -4 | errukigabeko "(with) ruthless" | -4 |
| B | tonto "stupid" | -3 | tuntun "stupid" | -3 |
| C | arduradun "responsible" | +2 | arduragabe "irresponsible" | -2 |

- v) **“Incorrect” translations.** There have been some translations which are incorrect because of different factors. The Spanish word *provinciano* “backward” (−1) is employed to refer to people of Bizkaia and Gipuzkoa provinces. The Elhuyar dictionary [12] has defined the word as “inhabitant of Bizkaia or Gipuzkoa”, a translation which is not useful for our purpose.
- vi) **“Indirect” translations.** There have been some translations that we have considered as indirect. They are correct translations but since they have an extensive meaning and they are used in limited situations, they are not useful for us. We have not used them in the translation process.
- For example, the word *beltz* “black” could have two meanings: *i*) a color *ii*) “black, sad; gloomy, depressing” (figurative meaning). The figurative use of that word is less usual, there are other words with the same meaning and, taking into account that the word could create problems, we have decided not to assign any value.

5 Results

As a result of this translation process, two versions of the semantic oriented Basque lexicon have been created. Table 3 shows the characteristics of these two versions.

Table 3. The semantic oriented Basque lexicons (V1.0 and V2.0).

| Grammatical category | V1 | | V2 | |
|----------------------|--------------|------------|--------------|------------|
| | Words | % | Words | % |
| Noun | 2,282 | 28.06 | 461 | 37.27 |
| Adjectives | 3,162 | 38.85 | 446 | 36.05 |
| Adverbs | 652 | 7.98 | 54 | 4.36 |
| Verbs | 1,657 | 20.36 | 276 | 22.32 |
| Intensifiers | 387 | 4.75 | | |
| Total | 8,140 | 100 | 1,237 | 100 |

The first version is the result of the first fourth steps in the translation process (Figure 1). It is translated directly from the Spanish SO-CAL dictionary with a strict criteria. But unlike the second version, it is not subject to the restrictions of the English SO-CAL dictionary, Basque Opinion Corpus and syntactic constraints (for example, with intensifiers).

As a result of these considerations, the first and second versions contain 8,140 and 1,237 words, respectively. In both cases, adjectives and nouns are the grammatical categories with more words. After that, there are verbs and adverbs. The intensifiers have not been taken into account in the second version because they affect to other words, so we think that it is better to analyze them in other levels.

6 Evaluation

In this section, we want to evaluate two aspects of the task. On the one hand, we want to evaluate the difficulty of the task. We think that the annotation of sentiment polarity is a difficult task because subjective perceptions are evaluated, which are completely different from grammatical categories. This aspect has been evaluated with two human annotators. On the other hand, we also want to evaluate the quality of the translated lexicon. With this objective, a gold standard annotation has been created from previous annotation by two annotators.

In order to evaluate these two aspects, we have extracted the most frequent 400 words (100 per each grammatical category) using Anallhitza [15] from the Basque Opinion Corpus [14]. We have used the Pearson correlation [16] to evaluate them. This method has been used in two different ways: i) Pearson 1: when the correlation is only calculated on the annotated words and ii) Pearson 2: the correlation of all words (which means that there are cases where one word has been annotated by one annotator or by nobody).

6.1 Correlation between annotators

We have decided to measure the correlation of two annotators to create the gold standard, taking into account the results achieved in the correlation coefficient. Table 4 shows the coefficient for each grammatical category and overall, together with a contingency table.

Table 4. Pearson correlation measurement and contingency table between two annotators.

| Grammatical category | | | Total categories | | | |
|----------------------|-------------|-------------|------------------|------------|------------|----|
| | Pearson 1 | Pearson 2 | 0 | NEG | POS | |
| Noun | 0.87 | 0.59 | 0 | 187 | 12 | 27 |
| Adjectives | 0.71 | 0.60 | NEG | 14 | 42 | 5 |
| Adverbs | 0.93 | 0.82 | POS | 39 | 5 | 69 |
| Verbs | 0.87 | 0.76 | | | | |
| Total | 0.79 | 0.73 | | | | |

The Pearson 1 value shows that the correlation coefficient is high (0.79). That means that the value assigned is similar in a big percentage of the annotated words. The coefficients for different grammatical categories are situated between 0.71 and 0.93. In a similar way, Pearson 2 also shows high correlation, although it is slightly lower (0.73), with values between 0.59 and 0.82.

Our interpretation of these results is that the assigned value in annotated words is similar. However, while one annotator has assigned a value, the other annotator has not assigned the value and that is reason why the coefficient drops in Pearson 2.

The contingency table (see Table 4) of 400 words shows that the biggest difference comes from the 0-POS/POS-0 pairs. That means that while one annotator has assigned a value to one word (39 cases), the other one did not assign any value and vice versa (27 cases).

6.2 Correlation between the lexicon and gold standard

The correlation between the human gold standard lexicon and the translated lexicon shows some differences compared to the correlation between two annotators as presented in Table (5).

Table 5. Pearson correlation measurement and contingency table between the gold standard and the Basque semantic oriented lexicon (V2.0).

| Grammatical category | Pearson 1 | Pearson 2 | Total categories | | | |
|----------------------|-------------|-------------|------------------|--|--|--|
| Noun | 0.96 | 0.59 | | | | |
| Adjectives | 0.78 | 0.56 | | | | |
| Adverbs | 0.75 | 0.47 | | | | |
| Verbs | 0.69 | 0.54 | | | | |
| Total | 0.76 | 0.54 | | | | |

| | 0 | NEG | POS |
|------------|-----|-----|-----|
| 0 | 195 | 2 | 15 |
| NEG | 30 | 34 | 8 |
| POS | 59 | 4 | 53 |

With Pearson 1, the cases in which the dictionary and gold standard contain an annotation for the word show similar correlation when compared to the results of two annotators (0.79). The correlation is high since the coefficients for the different grammatical categories are situated between 0.69 and 0.96. In contrast, Pearson 2 shows a lower correlation (0.54) and the coefficients of grammatical categories are situated between 0.47 and 0.59.

The interpretation of this results is that the values assigned to the dictionary and gold standard are similar (Pearson 1). The difference is mainly in the words which have been annotated only by one annotator. That is the reason of the decline of the correlation from Pearson 1 to Pearson 2.

The contingency table shows us how the gold standard and the created dictionary differ. The biggest differences are between the 0-NEG (30 cases) and 0-POS (59 cases) pairs which means that while the gold standard includes an annotation for the word, the lexicon does not include it.

7 Conclusion and Avenues for Future Work

In this paper we presented the first semi-manually created semantic orientation lexicon for Basque. Time factor, few resources and quality (the resources does not permit us) pushed us to translate the SO-CAL Spanish dictionary to Basque.

The translation process has followed several steps: firstly, translation from Spanish to expand the recall and, then, enrichment with the English version.

Moreover, in contrast with the first version, the second one is specific to a domain, based on the Basque Opinion Corpus. In the translation process, polysemy and figurative meaning have been the limitations of this work.

The Pearson correlation shows that coefficient is high between both annotators with respect to the following two factors: *i*) assigning a value and *ii*) deciding if a word has any value. In contrast, in the case of the comparison between human gold standard and lexicon, the correlation coefficient is high when the value is assigned but not in the case of deciding if the word has a value or not, which has been lower. This lower coefficient is because there are less words annotated in our lexicon.

In a foreseeable future, our aim is to implement this lexicon in a Basque sentiment analysis system. This lexicon will be the basis of this system and we will consider how to enrich the system with sentence level and text level information.

Acknowledgements

Jon Alkorta's work is financed by a Basque Government scholarship (code: PRE_2017_2.0041). Koldo Gojenola's work is partially financed by the PROSA-MED (TIN2016-77820-C3-1-R) funded by the Spanish Ministerio de Economía, Comercio y Competitividad) and UPV/EHU IXA Group (University of the Basque Country). Mikel Iruskieteta's work is partially financed by the TUNER project (TIN2015-65308-C5-1-R) funded by the Spanish Ministerio de Economía, Comercio y Competitividad) and UPV/EHU IXA Group (University of the Basque Country). We would like to offer our special thanks to Maite Taboada, Arantxa Otegi and Oier Lopez de Lacalle.

References

1. Pak, A., & Paroubek, P. "Twitter as a corpus for sentiment analysis and opinion mining." *LREc*. Vol. 10. No. (2010)
2. Tan, S. , & Zhang, J. "An empirical study of sentiment analysis for chinese documents" *Expert Systems with Applications*. (2007). doi:10.1016/j.eswa.2007.05.028
3. Cruz, F., Troyano, J.A., Enrquez de Salamanca, F., Ortega & F.J. "Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español." *Procesamiento del lenguaje natural* 4: 73-80. (2008). ISSN 1135-5948
4. Rosenthal, S., Farra, N., & Nakov, P.. "SemEval-2017 task 4: Sentiment analysis in Twitter." *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada. (2017)
5. Taboada, M., J. Brooke, M. Tofiloski, K. Voll & M. Stede. "Lexicon-Based Methods for Sentiment Analysis". *Computational Linguistics* 37 (2): 267-307. (2011)
6. Cruz, F. L., Troyan, J. A., Pontes, B., & Ortega, F. J. "ML-SentiCon: Un lexicón multilingüe de polaridades semánticas a nivel de lemas." *Procesamiento del Lenguaje Natural* 53: 113-120. (2014)

7. San Vicente, I., Agerri, R., Rigau, G., & Sebastin, D. S. "Simple, Robust and (almost) Unsupervised Generation of Polarity Lexicons for Multiple Languages." EACL. (2014)
8. Chetviorkin, I., & Loukachevitch, N. V., "Extraction of Russian Sentiment Lexicon for Product Meta-Domain." COLING, pp. 593610. (2012)
9. Chetviorkin, Ilia, & Loukachevitch, N. "Two-step model for sentiment lexicon extraction from Twitter streams." Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. (2014).
10. Goyal, A., & Daumé, H. III. "Generating semantic orientation lexicon using large data and thesaurus." Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 9096. Association for Computational Linguistics. (2014)
11. Brooke, J., Tofiloski, M., & Taboada, M. "Cross-Linguistic Sentiment Analysis: From English to Spanish." In Proceedings of RANLP 2009, Recent Advances in Natural Language Processing, pp. 50-54. (2009)
12. Zerbitzuak, Elhuyar Hizkuntza. "Elhuyar hiztegia: euskara-gaztelania, castellano-vasco. Usurbil: Elhuyar." (2013). ISBN: 9788495338709
13. Sarasola, Ibon. "Zehazki: gaztelania-euskara hiztegia." Alberdania. (2005)
14. Alkorta, J., Gojenola K., & Iruskieta M. "Creating and evaluating a polarity - balanced corpus for Basque sentiment analysis." IWoDA16. (2016)
15. Otegi, A., Imaz, O., Díaz de Ilarraza, A., Iruskieta, M., & Uria, L. ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. Procesamiento del Lenguaje Natural 58: p.p. 77-84. (2017)
16. Zou, K. H., Tuncali, K., & Silverman, S. G. Correlation and simple linear regression. Radiology, 227(3): p.p. 617-628. (2003)