Using lexical level information in discourse structures for Basque sentiment analysis

Jon Alkorta
IXA research group
UPV-EHU
jon.alkorta@ehu.eus

Koldo Gojenola
IXA research group
UPV-EHU
koldo.gojenola@ehu.eus

Mikel Iruskieta
IXA research group
UPV-EHU
mikel.iruskieta@ehu.eus

Maite Taboada
Discourse Processing Lab
Simon Fraser University
mtaboada@sfu.ca

Abstract

Systems for opinion and sentiment analysis rely on different resources: a lexicon, annotated corpora and constraints (morphological, syntactic or discursive), depending on the nature of the language or text type. In this respect, Basque is a language with fewer linguistic resources and tools than other languages, like English or Spanish. The aim of this work is to study whether some kinds of discourse structures based on nuclearity are sufficient to correctly assign positive and negative polarity with a lexicon-based approach for sentiment analysis. The evaluation is performed in two phases: i) Text extraction following some constraints on discourse structure from manually annotated trees. ii) Automatic annotation of semantic orientation (or polarity). Results show that the method is useful to detect all positive cases, but fails with the negative ones. An error analysis shows that negative cases have to be addressed in a different way. The immediate results of this work include an evaluation on how discourse structure can be exploited in Basque. In the future, we will also publish a manually created Basque dictionary to use in sentiment analysis tasks.

1 Introduction

Sentiment analysis is "the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" (Liu, 2012, p. 7).

Automatic sentiment analysis is an area in continuous development. It first started with the identification of subjectivity (Wiebe, 2000) and, after that, polarity identification and measurement of strength have become the center of new developments (Turney, 2002). The objectives of sentiment analysis are evolving as well, as different types of information are used. For instance, initially, entity- and aspect-based information was used (Hu and Liu, 2004) but, later, new types of information, such as discourse structure information, have been used (Polanyi and Zaenen, 2006).[1]

This study is the first work that examines lexical and discourse structure information for sentiment analysis of Basque. The main aim is to evaluate which discourse structures can help in polarity detection following a lexicon-based approach. Our hypothesis is that some discourse structures are more related to opinions than others and we want to identify and study how they can help in a sentiment analysis task.

The paper is organized as follows: Section 2 discusses related works. Section 3 explains the methodology of the study and Section 4 presents the results and error analysis. Finally, conclusions and future work are given in Section 5.

2 Related Work

Various studies from different theoretical approaches analyze the influence of nuclearity and some rhetorical relations in sentiment analysis tasks. For example, Zhou et al. (2011) use discursive in-

[1] See a detailed review of sentiment analysis in Taboada (2016).

Proceedings of the 6th Workshop Recent Advances in RST and Related Formalisms, pages 39–47,
Santiago de Compostela, Spain, September 4 2017.

formation in Chinese to eliminate noise at the intra-sentence level, improving not only polarity classification but also the labeling of rhetorical relations at sentence level.

Wu and Qiu (2012) analyze sentiment analysis based on Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) in Chinese texts. They split texts in segments and, then, they train weights taking into account relations and nuclearity, showing that CONTRAST, CAUSE, CONDITION and GENERALIZATION have a more important role in this task than other discourse relations. Bhatia et al. (2015) use a simpler classification of relations into CONTRAST or NON-CONTRAST, and they show that the distinction improves the results of bag-of-words classifiers using Rhetorical Recursive Neural Networks.

Chardon et al. (2013) rate documents using three approaches: i) bag-of-words, ii) partial discourse information and iii) full discourse information. The discursive approach gives the best result in the framework of Segmented Discursive Representation Theory (SDRT).

Trnavac et al. (2016) propose that a few rhetorical relations have a significant effect on polarity: CONCESSION, CONTRAST, EVALUATION and RESULT. They also conclude that nuclei tend to contain more evaluative words than satellites.

Alkorta et al. (2015) analyze which features perform better in order to detect the polarity of texts using machine learning techniques on Basque texts. Their results show that discourse structure is needed to improve results along with other types of features. They use a dictionary created by automatic means with an unsupervised method (Vicente et al., 2017). The dictionary values of their work are binary (-1 for negative polarity and $+1$ for a positive one).

In this work, we analyze which coherence relations could help to improve lexicon-based sentiment analysis, so that we can assign different weights to discourse structures following Bhatia et al. (2015) when calculating sentiment analysis for a whole text. For this task, we use the RST framework.

The main contributions of this work are: i) A fine-grained dictionary, manually created for Basque with 5 different negative values and 5 different positive ones, ranging from -5 to $+5$. ii) A study of how discourse structure interacts with this polarity lexicon.

3 Methodology

The subsections below detail the main steps followed in the present study.

3.1 Extraction of discourse structures

In the first phase, different discourse structures were compared. They will be used to determine which ones can be helpful in sentiment analysis. To extract as many discourse structures as possible, we use the corpus described in Alkorta et al. (2016), annotated for discourse relations according to RST.

The corpus contains 29 book reviews. Regarding polarity, it is a balanced corpus, with 14 positive reviews and 15 negative ones. The majority of reviews were collected from a website specialized in Basque literary reviews (Kritiken Hemeroteka).[2]

The following subcorpora were created, following some discourse constraints:

- Full text, containing all the RS-tree of the text.
- Texts extracted from central units (CU)[3] of the text.
- Text spans extracted from the CU of the text and from the central subconstituent (CS)[4] of some rhetorical relations (see Table 1).

Relation	CS	Relation	CS
ELABORATION	34	CONCESSION	2
EVALUATION	32	RESTATEMENT	2
PREPARATION	32	SUMMARY	2
BACKGROUND	13	ANTITHESIS	1
CIRCUMSTANCE	8	PURPOSE	1
INTERPRETATION	6	MOTIVATION	1
CAUSE	4	JUSTIFY	1

Table 1: Number of central subconstituents (CS) in the corpus per relation type linked to the CU.

We extracted 139 instances of rhetorical relations from our corpus. For some relations, such as ELABORATION and PREPARATION (66 of 139), we do

[2] http://kritikak.armiarma.eus/.

[3] Central units are defined as the most important EDU (Elementary Discourse Unit), and it is the main nucleus when tree structure is constructed (Iruskieta, 2014).

[4] Central subconstituents are "the most important unit of the modifier span that is the most important unit of the satellite span" (Iruskieta et al., 2015, p. 5).

not expect them to contain important polarity information, because these relations only add extra information to the central unit. In fact, Mann and Thompson (1988, p. 273) mention that in the case of ELABORATION "R(eader) recognized the situation presented in S(atellite) as providing additional detail for N(uclei). R(eader) identifies the element of subject matter for which detail is provided". Similarly, in PREPARATION "R(eader) is more ready, interested or oriented for reading N(uclei)". We did not take into account relations with low frequency (a single instance), such as MOTIVATION, JUSTIFICATION, ANTITHESIS and PURPOSE. Consequently, we will work with a subcorpus containing 69 relations, where almost half of them are central subconstituents of EVALUATION.[5]

3.2 Polarity extraction and evaluation

Polarity was extracted from all the discourse structures using a dictionary (v1.0) of words annotated with their semantic orientation: polarity (positive or negative) and strength (from 1 to 5). To do so, the Spanish SO-CAL dictionary (Taboada et al., 2011) was translated using the Elhuyar (Zerbitzuak, 2013) and Zehazki (Sarasola, 2005) bilingual Spanish-Basque dictionaries. Our dictionary contains information about grammatical categories: nouns, adjectives, verbs and adverbs.

Dictionary	Words	SO(-)	SO(+)
Nouns	2,882	1,635	1,247
Adjectives	3,162	1,733	1,429
Adverbs	652	225	427
Verbs	1,657	1,006	651
Total	**8,353**	**4,599**	**3,754**

Table 2: Characteristics of the Basque dictionary.

As Table (2) shows, the dictionary contains a total of 8,353 words. The majority of words are nouns and adjectives. In terms of polarity, there are more negative words (almost one thousand more).

We created a polarity tagger, based on this dictionary. The polarity tagger used the output of Eustagger (Aduriz et al., 2003), which is a robust and wide-coverage morphological analyzer and a Part-of-Speech tagger (POS) for Basque, to enrich the text with a POS analysis information and to assign polarity to every lemma of the dictionary that matches with the lemma and category of the text. With the aim of comparing the results of the system, a linguist annotated the polarity (positive, negative or neutral) of all the discourse structures described in Section (3.1).

Figure 1 shows a portion of the RST tree of one text (LIB28).[6] After the full RST analysis was performed for each text, we extracted the following discourse structures: i) the text of the central unit (EDU_2), as shown in Example (1), and ii) the central subconstituent of the EVALUATION ($EDU_{21,22,23,25}$), in Example (2).

(1) XIX. mendean Gasteiz inguruak izutu$_{(-3)}$ zituen Juan Diaz de Garaio Sacamantecas pertsonaia hartu$_{(+2)}$ du Aitor Aranak (Legazpi, 1963) bere azken eleberrian$_{(+2)}$. (LIB28_CU)
English: Aitor Arana (Legazpi, 1963) has taken$_{(+2)}$ in his last novel$_{(+2)}$ the character Juan Diaz Garaio Sacamantecas who scared$_{(-3)}$ the surroundings of Gasteiz in the 19th century.

(2) Hala ere, nahiko$_{(+2)}$ planoa da nobela$_{(+2)}$, erritmoa falta$_{(-1)}$ zaio eta bortxaketen kontaketak aspergarriak$_{(-3)}$ ere bihurtzen$_{(-2)}$ dira, Bestalde, alabaren ikuspuntua$_{(+2)}$ ez da batere argi geratzen$_{(-2)}$, (...). (LIB28_EVA)
English: However, the novel$_{(+2)}$ is fairly$_{(+2)}$ flat, it lacks$_{(-1)}$ rhythm, and the stories of rapes also become$_{(-2)}$ boring$_{(-3)}$. On the other hand, the point of view$_{(+2)}$ of the daughter is not clear$_{(-2)}$ (...)

The classifier then assigns polarity to each word in the dictionary, as shown in Table 3 and in examples (1) and (2). The table shows that the semantic

[5] All the reviews of the corpus were coded, assigning the domain LIB (for literature review) and a number, and each discourse structure extracted from them was also coded: CU stands for text that only contains the central unit of the text, CAUS for texts that contain CAUSE relation, INT for INTERPRETATION, ELAB for ELABORATION, CIR for CIRCUMSTANCE, BACK for BACKGROUND and finally, EVA for EVALUATION. In addition, if the same relation appears more than once in each text, we added letters (e.g., a, b, c) to each relation, to indicate their order of appearance.

[6] Size constraints prevent us from showing the entire tree.

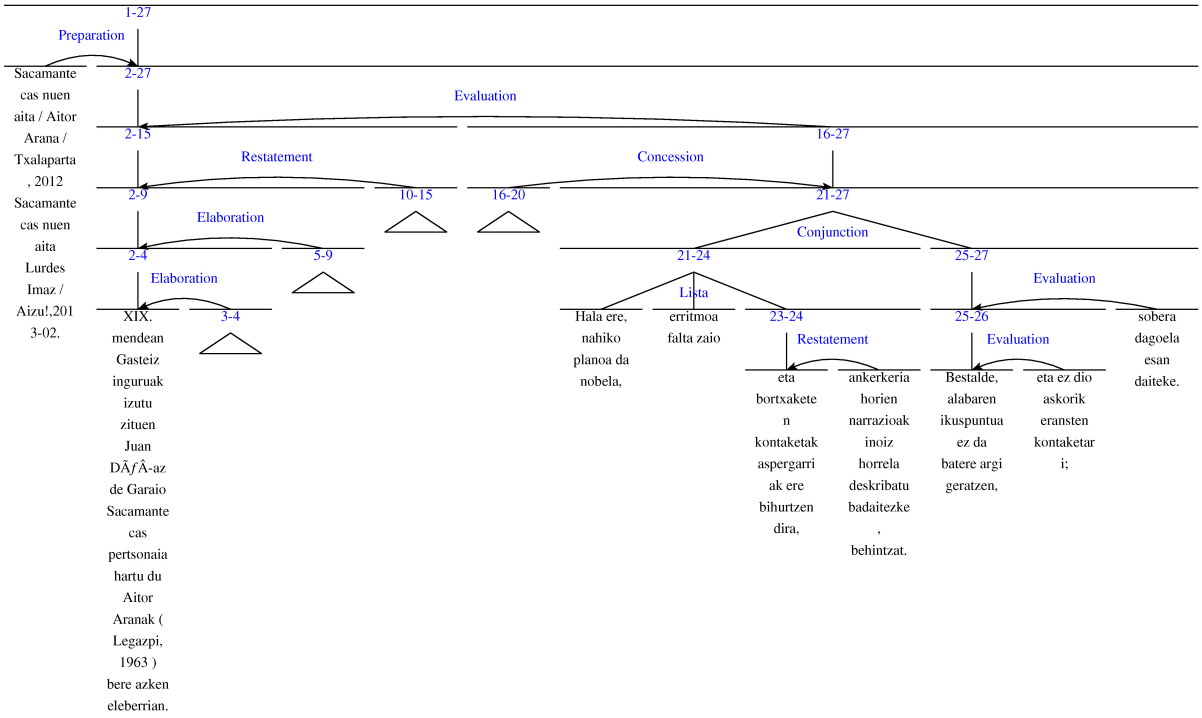

Figure 1: Central unit and the central subconstituent of EVALUATION in text LIB28

orientation of the central unit (LIB28_cu) is positive, while the semantic orientation of the central subconstituent (LIB28_EVA) is negative.

Ex.	CS ID	Classifier	SO	Manual
1	LIB28_cu	$-3+2+2$	$+1$	Neutral
2	LIB28_EVA	$+2+2-1-3-2$	-2	Negative

Table 3: Semantic orientation of LIB28_cu and LIB28_EVA: results of the classifier and of the manual annotation.

3.3 Normalization of semantic orientation results

We normalized the results obtained with the classifier to compare the different discourse structures, as in the following examples:

(3) Gure izaeraz$_{(+3)}$ hausnartzeko$_{(+1)}$ manual gisa eta, etxetik ibiltzeko$_{(+2)}$ dosi psikoanalitiko ttipi$_{(-1)}$ moduan$_{(+1)}$ hautematen$_{(+4)}$ dut nik. (LIB26_INT)
English: I consider$_{(+4)}$ it is like a manual with a small$_{(-1)}$ dose of psychoanalysis, a domestic$_{(+2)}$ consideration$_{(+1)}$ to reflect about$_{(+1)}$ our being$_{(+3)}$.

(4) Nolanahi$_{(-2)}$ den dela, saihestezina da gatazka$_{(-4)}$. (LIB13_CIR)
English: In any case$_{(-2)}$, the conflict$_{(-4)}$ is inevitable.

The results obtained by the classifier are $+112$ (LIB10),[7] $+10$ (LIB26_INT) and -6 (LIB23_CIR), as shown in Table 4. To compare those results among them, we normalized the frequencies dividing these results by the number of the words in each discourse structure. We show the normalized frequencies in Table 4.

Ex.	CS ID	SO	Words	NV
	LIB10	$+112$	418	$+0.27$
3	LIB26_INT	$+10$	17	$+0.59$
4	LIB13_CIR	-6	8	-0.75

Table 4: Examples of semantic orientation results after normalization (NV = Normalized Value).

Table 4 shows how normalization helps to better adjust the weight of the automatically assigned polarities. As a matter of fact, the values are adjusted

[7] Remember that this notation, LIB10, represents the entire text.

to a smaller range and, therefore, they are more easily comparable.

4 Results and error analysis

The results show that using a simple classifier with a manually built dictionary, along with different rhetorical structures, helps to identify the strength of such structures. For example, the result obtained in the central subconstituent of EVALUATION is strong.

(5) Guztiz$_{(+3)}$ gomendagarria$_{(+3)}$. (LIB26_EVA)
English: Highly$_{(+3)}$ recommended$_{(+3)}$.

(6) Liburu$_{(+5)}$ sano gomendagarria$_{(+3)}$ da, (LIB23d_EVA)
English: It's a very recommendable$_{(+3)}$ book$_{(+5)}$,

In Examples (5) and (6) the strength is higher than 1: $+2$ ($+6$ / 3 = 2) and $+1.6$ ($+8$ / 5 = $+1.6$), respectively, while the strength in other relations is lower.

(7) Izugarri$_{(+5)}$ gustura irakurri dut Bertol Arrietaren Alter ero narrazio bilduma. (LIB26_CAUS)
English: I have read very$_{(+5)}$ comfortable the Alter ero narration collection of Bertol Arrieta.

(8) Udako giro$_{(-2)}$ sapa horretan gertatzen diren kontakizun xumeak$_{(+3)}$ ekarriko dizkigu idazleak. (LIB15_CIR)
English: The writer will bring us the common$_{(+3)}$ stories that happen in that sticky atmosphere$_{(-2)}$ of summer.

The strength of CAUSE shows in Example (7) a value lower than 1 ($+5$ / 11 = $+0.45$). In Example (8) the central subconstituent of INTERPRETATION shows a value lower that 1 with a value of $+0.08$ ($+1$ / 12 = $+0.08$) and lower value than in Example (7).

We have analyzed the discourse structure with the aim of determining the strongest discourse structures of our corpus and therefore the structures that contribute most to improving sentiment labeling. Most of the values are between -1 and $+1$, but in 11.59% of the relations (8 of 69 relations), the values are higher than one (see Table 5).

RR	**Total**	**Total (<1)**	**%**
EVALUATION	32	6	18.75
INTERPRETATION	6	1	16.67
BACKGROUND	13	1	7.69
Others	18	0	0.00
Total	69	8	11.59

Table 5: Polarity strength ($< +1$ and > -1) of central subconstituents.

The most frequent and strongest value is obtained in EVALUATION (18.75%, 6 of 32). After that, the second strongest relation is INTERPRETATION with 16.67% (1 of 6). And, finally, BACKGROUND is once above one (7.69%, 1 of 13).

As examples (9, 10, 11, 12, 13) show, these relations have similar characteristics: short central subconstituents with many and strong evaluative words.

(9) berriz, zuzenean$_{(+3)}$ egin$_{(+2)}$ dut. (LIB14a_EVA)
English: whereas, I have done$_{(+2)}$ it directly$_{(+3)}$.

(10) Abentura$_{(+2)}$ liburu$_{(+5)}$ ederra$_{(+3)}$ iruditu$_{(+1)}$ zait, eta erremate paregabea$_{(+4)}$ trilogiarentzat. (LIB14b_EVA)
English: It seemed$_{(+2)}$ to me a beautiful$_{(+3)}$ adventure$_{(+2)}$ book$_{(+5)}$, and extraordinary$_{(+4)}$ finish for the trilogy.

(11) izenburua zuzen$_{(+3)}$ jarrita$_{(+1)}$, (LIB29a_EVA)
English: the title set$_{(+1)}$ correctly$_{(+3)}$,

(12) Intrigazko$_{(+2)}$ argumentua garatu$_{(+1)}$ nahi$_{(+3)}$ da. (LIB01b_EVA)
English: You want$_{(+3)}$ to develop$_{(+1)}$ an argument of intrigue$_{(+2)}$.

(13) Folklorean ikusi$_{(+4)}$ nahi$_{(+3)}$ ditu idazleak komunitate$_{(+1)}$ baten bizi$_{(+2)}$ nahi$_{(+3)}$ eta indarra$_{(+3)}$. (LIB35_INT)
English: The author wants$_{(+3)}$ to see$_{(+4)}$ in the folklore the strength$_{(+3)}$ and the desire$_{(+3)}$ to live$_{(+2)}$ of one community$_{(+1)}$.

Consequently, their value is higher than one, as shown in Table (6).

Ex.	**CS ID**	**NV**
9	LIB14a_EVA	1
10	LIB14b_EVA	1.36
11	LIB29a_EVA	1
12	LIB01b_EVA	1
13	LIB35_INT	1.33

Table 6: Central subconstituents and their value ($<+1$).

In contrast, we did not see any case of other central subconstituents with a value higher than one. If we compare partial discourse structures with the results obtained with all words of a text, the strength is lower in all cases. This is because polarity words do not have the same frequency in other rhetorical relations and, as a consequence, the concentration of words with semantic orientation is smaller. The highest value across the texts is $+0.50$ (LIB35), and the lowest value is -0.1 (LIB28).

These results suggest that opinions and, consequently, words with semantic orientation, are mainly found in the central subconstituent of EVALUATION, INTERPRETATION and BACKGROUND.

Apart from helping to identify the strongest central subconstituents, we have observed that the dictionary together with some central subconstituents can help in sentiment analysis. In fact, assigning a weight to some CSs could help to improve sentiment analysis results, as in text LIB34.

(14) "Behi eroak$_{(-3)}$" bilduman, ordea, egileak aurrekoan izan zituen arazoak$_{(-1)}$ konpondu$_{(+3)}$ ditu. Zoritxarrez$_{(-4)}$ bilduma honek batzuetan xelebrekeria$_{(-1)}$ merketik$_{(+3)}$ badu nahiko$_{(-2)}$. (LIB34b_EVA)
English: However, in "Behi eroak$_{(-3)}$" collection, the author has solved$_{(+3)}$ the problems$_{(-1)}$ that he had before. Unfortunately$_{(-4)}$, this collection has enough$_{(-2)}$ cheap$_{(+3)}$ eccentricity$_{(-1)}$.

The human annotator marked LIB34 as a negative review and the system assigns a value of $+0.15$ for the entire text, but a negative value of -0.2 ($-5/25=-0.2$) for LIB34b_EVA, Example (14). If the proper weight was assigned to this CS (LIB34b_EVA), the semantic positive orientation of the entire text (LIB34) would be corrected and tagged as negative.

We analyzed the previous finding in all the CSs of EVALUATION, but taking the results of the human annotator, instead of the classifier. In total, in 29 texts, there are 32 CSs of EVALUATION and in 24 of them, the human annotation of polarity of CSs and texts agree. So, the agreement happens in 75% of CSs and 86.20% of texts (25 texts).

Even though most of the times there is agreement between the annotated polarity of CSs and texts, this does not happen in all cases. For example, in other cases, the same text has one positive central subconstituent and another negative central subconstituent of EVALUATION. These cases are 12.50% of central subconstituents and 6.89% of texts (LIB03ab and LIB12ab).

Finally, there are two cases in which the polarity of the central subconstituent of EVALUATION and the polarity of all text are the opposite (LIB02ab and LIB19ab).

(15) eta apustu ausarta$_{(+3)}$ egin$_{(+2)}$ du bertan. (LIB19a_EVA)
English: and has made$_{(+2)}$ a strong$_{(+3)}$ bet there.

(16) Batetik, idazleak goi-literaturaren jokalekua hautatu duelako —liburuaren$_{(+5)}$ erlazio estratestualak eta baliatutako$_{(+1)}$ errekurtso andana$_{(-1)}$ lekuko—. Bestetik, borgestarretik asko duen jokoa$_{(-4)}$ delako liburuan$_{(+5)}$ dagoena. (LIB19b_EVA)
English: On the one hand, because the writer has chosen a scene from high literature —extratextual relations and a lot$_{(-1)}$ of resources used$_{(+1)}$ in the book$_{(+5)}$ as proof—. On the other hand, because there is a game$_{(-4)}$ that has a lot of Borges in the book$_{(+5)}$.

In this case, the text LIB19 is negative, whereas examples (15) and (16) are positive. We observe that the change of polarity happens in the EVALUATION situated inside an ELABORATION coherence relation.

(17) Baina, horiek horrela izanik ere, emaitza$_{(+1)}$ zalantzagarria$_{(-1)}$ da. Izan ere, liter-

aturan, baliabide$_{(+2)}$ orok medio izan behar$_{(-1)}$ du, eta irakurleak ikusi$_{(+4)}$ behar$_{(-1)}$ du errekurtsoak literaturaren mesedetan$_{(+3)}$ daudela "baita metaliteraturaz ari$_{(+2)}$ garenean ere". Hemen, ordea, medioak emaitza$_{(+1)}$ estaltzen$_{(-2)}$ du maiz$_{(+1)}$: literaturaren mekanismoekin egindako$_{(+2)}$ jokoek$_{(-4)}$ ipuinetan$_{(+2)}$ dauden istorioak$_{(-1)}$ indartu$_{(+1)}$ beharrean$_{(-1)}$, higatu$_{(-2)}$ egiten$_{(+2)}$ dituzte. Aldamioa oso$_{(+1)}$ nabarmena$_{(+4)}$ da, idazle askok beretzat nahi$_{(+3)}$ lukeen ahalmenez$_{(+2)}$ jasoa$_{(+2)}$. Haatik, hartatik sortzen$_{(+2)}$ den literatura ez da hain ikusgarria$_{(+4)}$. (LIB19_ELAB)

English: But, they being so, the result$_{(+1)}$ is doubtful$_{(-1)}$. In fact, in the literature, all resources$_{(+2)}$ need$_{(-1)}$ to be the medium, and the reader needs$_{(-1)}$ to see$_{(+4)}$ that resources are in favor$_{(+3)}$ of literary, "also when we are talking$_{(+2)}$ about metaliterature." But here, the medium hides$_{(-2)}$ the result$_{(+1)}$ in many times$_{(+1)}$: games$_{(-4)}$ made$_{(+2)}$ by literary devices wear away$_{(+2)(-2)}$ the tales$_{(-1)}$ of the stories$_{(+2)}$ instead$_{(-1)}$ of strengthening$_{(+1)}$ them. The scaffolding is very$_{(+1)}$ evident$_{(+4)}$, built$_{(+2)}$ with capacity$_{(+2)}$ as many writers would like$_{(+3)}$. However, the literature created$_{(+2)}$ is not very impressive$_{(+4)}$.

In Example (17), there are some discourse markers (*but, however*) and words (*doubtful, wear away, not very impressive*) that suggest a change of polarity that affects all text. Consequently, this example shows that, apart from central constituents of EVALUATION, a deeper analysis of nuclearity assigning different weighs could be necessary in order to improve sentiment analysis.

4.1 Error analysis

In this section, we will analyze the errors that can affect accurate detection of sentiment analysis, and specially the ones that were relevant in this study: i) errors in negative reviews, and ii) errors related to syntax.

4.1.1 Errors in negative reviews

Brooke et al. (2009) mention that lexicon-based sentiment classifiers show a positive bias because humans tend to use positive language (see also Taboada et al. (2017)). We also found this problem by examining the results of the classifier.

As Table (2) shows, the majority of the words in the dictionary are negative. Therefore, it is expected that we will detect more negative words in the texts. However, the results of the classifier with our dictionary show a tendency to classify texts as positive in different discourse structures of the texts.

For example, this tendency is observed in results of the CS of EVALUATION[8] (see Table 7).

CS of EVALUATION	Total	Guess	%
Positive	20	19	95.00
Negative	11	4	36.36
Neutral	1	0	0.00
Total	32	23	71.88

Table 7: Positive polarity tendency in central subconstituents of EVALUATION.

Table 7 demonstrates that the classifier tends to consider as positive the majority of central subconstituents of this rhetorical relation. In fact, 26 of 32 central subconstituents have been classified as positive. Consequently, the correct guess rate in CSs is higher in positive (95%) versus negative (36.36%).

A tendency to positive semantic orientation is higher if we analyze the results of all texts instead of just central subconstituents of EVALUATION as shown in Table 8.

Texts	Total	Guess	%
Positive	14	14	100
Negative	15	1	6.67
Total	29	15	51.72

Table 8: Positive polarity tendency in texts of the corpus.

As a consequence of this positive bias, our classifier guesses easily the texts with positive polarity and the correct guess rate is 100%. In contrast, the rate is very low in negative texts, as a matter of fact, there is only one right guess in text LIB28 (-0.1) and consequently, the correct guess rate is 6.67%.

[8]We have analyzed this relation and not others because it accounts for almost half of all the studied rhetorical relations.

However, if we compare the results of central subconstituents and texts, we can observe another tendency. The rate of correct assignments in positive texts is higher (95% vs. 100%) on the full texts (long text), while for negatives it is higher (36.36% vs. 6.67%) in central subconstituents (short text). This suggests that the tendency to positive semantic orientation is stronger using our dictionary as a bag-of-words approach as the text is longer.

In summary, the dictionary classifier shows the same problem already described in previous research, as there is a strong tendency towards positive semantic orientation, which increases as the text is longer.

4.1.2 Errors related to syntax

As we mentioned in Section 4.1.1, there is a tendency towards positive polarity caused by the use of positive language and, for that reason, the correct guess rate is lower in negative texts. However, it is not the only reason, and information at the syntactic level also affects the results. As an example, we will discuss one particular problem, negation. Due to negation, the polarity of a sentence is changed and it is necessary to take this characteristic into account in sentiment analysis.

(18) (...) narrazioak ere ez du arretarik bereganatzen$_{(+4)}$ (...) (LIB18_EVA).
English: (...) the narration also does not get attention$_{(+4)}$ (...)

In Example (18), the semantic orientation of the sentence would be negative but our classifier regards it as positive. The classifier has detected *bereganatu* 'to get hold of' as a positive word (+4/7=+0.57). But, in this case, a correct analysis should assign it a negative value.

In a first study of our subcorpus of CSs of different rhetorical relations, we estimate that this affects to 11.43% of the constituents, since 8 of 70 CSs have some type of negation.

5 Conclusions and future work

This study has analyzed whether combining a semantic oriented dictionary with some discourse structure constraints is helpful in sentiment analysis of Basque.

The results show that i) the central subconstituents (CS) of EVALUATION, INTERPRETATION and BACKGROUND are the units with the strongest semantic orientation, and ii) the CSs of EVALUATION could help in improving semantic orientation of the texts, given that the results of the human annotation of polarity of CSs and the full text text agree in 75% of the cases.

On the other hand, error analysis has shown that there are some aspects that should be addressed: i) a tendency to positive semantic orientation, and ii) sentence and more discourse level constraints are needed.

In the near future, we plan to pursue the following aspects:

i) Do reviews have a specific discourse structure? We hypothesize that reviews have a specific structure and, consequently, the same discourse relations will be repeated with high frequency, and they will appear in the same place.

ii) How we can weigh properly the central subconstituents of EVALUATION and INTERPRETATION, and neutralize the positive tendency, to improve the results for negative reviews?

iii) Are other CSs not linked to the CU important for sentiment analysis?

Acknowledgments

We thank Arantxa Otegi for assistance with the lexicon-based polarity tagger. Jon Alkorta's work is funded by a PhD grant (PRE_2016_2_0153) from the Basque Government and Mikel Iruskieta's work is funded by the TUNER project (TIN2015-65308-C5-1-R) funded by the Spanish *Ministerio de Economia, Comercio y Competitividad*.

References

[Aduriz et al.2003] Itziar Aduriz, Izaskun Aldezabal, Inaki Alegria, J Arriola, Arantza Dıaz de Ilarraza, Nerea Ezeiza, and Koldo Gojenola. 2003. Finite state applications for basque. In *EACL2003 Workshop on Finite-State Methods in Natural Language Processing*, pages 3–11.

[Alkorta et al.2015] Jon Alkorta, Koldo Gojenola, Mikel Iruskieta, and Alicia Prez. 2015. Using rela-

tional discourse structure information in basque sentiment analysis. In *SEPLN 5th Workshop RST and Discourse Studies. ISBN: 978-84-608-1989-9. https://gplsi.dlsi.ua.es/sepln15/en/node/63*.

[Alkorta et al.2016] Jon Alkorta, Koldo Gojenola, and MIkel Iruskieta. 2016. Creating and evaluating a polarity - balanced corpus for basque sentiment analysis. In *IWoDA16 Fourth International Workshop on Discourse Analysis. Santiago de Compostela , September 29 th - 30 th . Extended Abstracts. ISBN: 978 - 84 - 608 - 9305 - 9*.

[Bhatia et al.2015] Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. *arXiv preprint arXiv:1509.01599*.

[Brooke et al.2009] Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *RANLP*, pages 50–54.

[Chardon et al.2013] Baptiste Chardon, Farah Benamara, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2013. Measuring the effect of discourse structure on sentiment analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 25–37. Springer.

[Hu and Liu2004] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

[Iruskieta et al.2015] Mikel Iruskieta, Iria Da Cunha, and Maite Taboada. 2015. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation*, 49(2):263–309.

[Iruskieta2014] Mikel Iruskieta. 2014. Pragmatikako erlaziozko diskurtso-egitura: deskribapena eta bere ebaluazioa hizkuntzalaritza konputazionalean (a description of pragmatics rhetorical structure and its evaluation in computational linguistic). *Doktore-tesia. EHU, informatika Fakultatea*.

[Liu2012] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

[Mann and Thompson1988] William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

[Polanyi and Zaenen2006] Livia Polanyi and Annie Zaenen. 2006. Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*, pages 1–10. Springer.

[Sarasola2005] Ibon Sarasola. 2005. *Zehazki: gaztelania-euskara hiztegia*. Alberdania.

[Taboada et al.2011] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

[Taboada et al.2017] Maite Taboada, Radoslava Trnavac, and Cliff Goddard. 2017. On being negative. *Corpus Pragmatics*, 1(1):57–76.

[Taboada2016] Maite Taboada. 2016. Sentiment analysis: an overview from linguistics. *Annual Review of Linguistics*, 2:325–347.

[Trnavac et al.2016] Radoslava Trnavac, Debopam Das, and Maite Taboada. 2016. Discourse relations and evaluation. *Corpora*, 11(2):169–190.

[Turney2002] Peter D Turney. 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*, pages 417–424. Association for Computational Linguistics.

[Vicente et al.2017] Inaki San Vicente, Rodrigo Agerri, and German Rigau. 2017. Q-wordnet ppv: Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. *arXiv preprint arXiv:1702.01711*.

[Wiebe2000] Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *AAAI/IAAI*, pages 735–740.

[Wu and Qiu2012] Fei Wang1 Yunfang Wu and Likun Qiu. 2012. Exploiting discourse relations for sentiment analysis. In *24th International Conference on Computational Linguistics*, page 1311.

[Zerbitzuak2013] Elhuyar Hizkuntza Zerbitzuak. 2013. Elhuyar hiztegia: euskara-gaztelania, castellano-vasco. usurbil: Elhuyar.

[Zhou et al.2011] Lanjun Zhou, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 162–171. Association for Computational Linguistics.