Exploring the Enrichment of Basque WordNet with a Sentiment Lexicon

Jon Alkorta, Itziar Gonzalez-Dios
Ixa Group, University of the Basque Country (UPV/EHU)
Manuel Lardizabal Ibilbidea, 1, 20018 Donostia
{jon.alkorta,itziar.gonzalezd}@ehu.eus

Abstract

Wordnets are lexical databases where the semantic relations of words and concepts are established. These resources are useful for many NLP tasks, such as automatic text classification, word-sense disambiguation or machine translation. In comparison with other wordnets, the Basque version is smaller and some PoS are underrepresented or missing e.g. adjectives and adverbs. In this work, we explore a novel approach to enrich the Basque WordNet, focusing on the adjectives. We want to prove the use and and effectiveness of sentiment lexicons to enrich the resource without the need of starting from scratch. Using as complementary resources, one dictionary and the sentiment valences of the words, we check if the word of the lexicon matches with the meaning of the synset, and if it matches we add the word as variant to the Basque WordNet. Following this methodology, we describe the most frequent adjectives with positive and negative valence, the matches and the possible solutions for the non-matches.

Keywords: Basque_WordNet, sentiment_lexicon, resource enrichment

1. Introduction and Background

Creating and maintaining language resources is an expensive and costly task. Moreover, in the case of the languages with recent standarisation processes, the update and constant redesign of the resources is mandatory. In the case of Basque, a language whose standarisation process officially began in 1968, the maintenance and updating lexico-semantic resources such as the Basque Wordnet (BWN) or *Euskal Wordnet* (Pociello et al., 2011) requires a big lexicographic effort (Aldezabal et al., 2018).

BWN[1] is a version in Basque language of WordNet (Miller, 1995; Fellbaum, 1998). It was created following the expand approach, but special care was taken for cultural concepts and lexicalization issues. It was developed together with the EuSemCor corpus (Agirre et al., 2006). In the latest distribution (2016), BWN had 30 263 synsets, 40 420 variants for nouns, 9 469 for verbs and 148 for adjectives. There were no variants for adverbs. As far as the size of BWN is concerned, its size is limited in comparison with wordnets in other languages. As updating it from scratch is costly, in this paper we explore multimodal approaches to add new variants, namely based on the sentiment lexicon for Basque *SentiTegi* (Alkorta et al., 2018)

Regarding wordnets and sentiment lexicons, wordnets are usually complemented with polarity and sentiment information. They have been created above all for sentiment analysis and opinion mining. For example, in SentiWordNet 3.0 (Baccianella et al., 2010) each synset is tagged with the notions positivity, negativity, and neutrality and associated to three numerical scores to indicate how positive, negative, and objective/neutral the variants are. Based on SentiWordNet, SentiWord (Gatti et al., 2015) profits from prior built polarity lexica in order to achieve higher precision and coverage. Finally, in WordNet-feelings (Siddharthan et al., 2018) synsets are classified in nine broad feeling categories and as a feeling based on a depth linguistic study of human feelings. This resource is complementary to SentiWordNet. All these resources have been created based on the English WordNet.

In this paper, however, we want to explore the opposite option: the possibility of using the sentimen lexicon *SentiTegi* to increase the size of BWN. We want to focus on the adjectives, whose coverage is limited in BWN. Exactly, as case study we will analyze the most frequent adjectives with positive and negative valence in *SentiTegi*, is a manually-created sentiment lexicon for Basque. The aim is to explore if *SentiTegi*, and to a certain extend sentiment lexicons, can be used as a source to enrich BWN and wordnets. Moreover, we want to examine which linguistic issues arise when comparing both resources.

This paper is structured as follows: in Section 2 we describe the characteristics of the sentiment lexicon *SentiTegi*, which is the source for the enrichment of the BWN. In Section 3, we explain the casuistry regarding the match between meanings of synsets and the sentiment valence of the words by some examples of the enrichment. Finally, in Section 4, we conclude the work and enumerate the future works.

2. *SentiTegi*, the Basque Sentiment Lexicon

SentiTegi (Alkorta et al., 2018) is a manually-created sentiment lexicon for Basque. It is a part of the Basque version of the sentiment classifier called SO-CAL (Taboada et al., 2011). The SO-CAL sentiment classifier is a lexicon-based sentiment classifier. The words in lexicon have a sentiment valence[2] that determines if the words make a positive or negative evaluations or judgements. The sentiment valence of the words is numerical and the numbers rank from -5 to $+5$.

[1]The Basque Wordnet is available in `https://adimen.si.ehu.es/cgi-bin/wei/public/wei.consult.perl` and in the Open Multilingual wordnet `http://compling.hss.ntu.edu.sg/omw/`

[2]Sentiment valence and semantic orientation are used to determine the subjectivity of words. Semantic orientation is a signal ($+$ or $-$) that indicate if the word makes a positive (or good) or negative (or bad) evaluation. In contrast, the sentiment valence indicates the intensity of the evaluation with numbers in addition to type of evaluation (good or bad evaluation).

(1) *Bikain* "excellent" (+5)

(2) *Eskas* "insufficient" (−1)

(3) *Txar* "bad" (−3)

Examples (1), (2) and (3) indicate the scale of the words in the sentiment lexicon regarding their subjective evaluation. In Example (1), the word *bikain* "excellent" expresses the most intense positive evaluation and, consequently, its sentiment valence is +5. On the other hand, in Example (2), the word *eskas* "insufficient" makes a negative evaluation, therefore the sign is negative (−). In contrast with Example (1), its intensity is lower and, for that reason, its sentiment valence is (−1). Finally, in Example (3), the word *txar* "bad" makes a negative evaluation and it has a medium intensity. Therefore, its sentiment valence is (−3).

The sentiment lexicon in Basque has been created by translating the sentiment lexicon in Spanish and enriching with the sentiment lexicon in English of different language versions of the SO-CAL tool. First, the sentiment lexicon from Spanish (Brooke et al., 2009) was translated into Basque with the *Elhuyar* (Elhuyar Hizkuntza Zerbitzuak, 2013) and *Zehazki* (Sarasola, 2005) dictionaries. In the second step, the Basque translations have been grouped with different source in Spanish lexicon. For example, the Spanish words *amago* "feint" (−1) and *cicatriz* "scar" (−2) have been translated into Basque as *seinale* "signal" and they have been grouped in the Basque word *seinale* "signal". In addition, the Basque translations have inherited the sentiment valence from Spanish words. In the case of the words that had various sources, the most adequate for Basque has been chosen. Consequently, the sentiment valence (−1) has been chosen for *seinale* "signal", based on the Spanish *amago* "feint".

In order to choose the sentiment valence of the Basque words, the context where the words appear in the *Basque Opinion Corpus* (Alkorta et al., 2017) has been taken into account. Then, the Basque translations that are not entries of the *Elhuyar* (Elhuyar Hizkuntza Zerbitzuak, 2013) and *Zehazki* (Sarasola, 2005) dictionaries or do not appear in the *Basque Opinion Corpus* (Alkorta et al., 2017) have been removed. As a consequence of that, e.g. the word *atrofiatu* "atrophy" (−1) has been removed from the lexicon. After this step, the size of the sentiment lexicon has been reduced from 8,140 words to 1,237 words. Finally, the translated sentiment lexicon has been enriched with the English sentiment lexicon (Taboada et al., 2011). The sentiment valences of the Basque words and their equivalents in English have been compared: in some cases the sentiment valence has been changed and in other cases, the sentiment valence of the Basque word has been kept.

Table 1 shows the characteristics of the sentiment lexicon in Basque. The first version of the sentiment lexicon has been created following the first two steps (in other words, until grouping the Basque translations and choosing their sentiment valence). In contrast, the second version has been created following all the steps mentioned before. Among the first and second version, the size of the lexicon has been reduced from 8,140 words to 1,237 words due to the constraints of dictionaries and domains of the corpus. However, in both corpora, nouns and adjectives are the most common grammatical category.

	V1.0		V2.0	
Part-Of-Speech	**Words**	**%**	**Words**	**%**
Nouns	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

Table 1: The characteristic of two versions of the sentiment lexicon *SentiTegi*

3. Methodology and Casuistry

In this section, we explain the methodology and the casuistry in the enrichment of the BWN with *SentiTegi*. In order to select the sample for the analysis, we have extracted the most frequent adjectives in *Basque Opinion Corpus* (Alkorta et al., 2017) with *AnalHitza* (Otegi et al., 2017), a tool that extracts basic linguistic information from texts and corpora. We have also filtered the adjectives taking into account their valence. The list we have created has the following information the Basque adjective, frequencies, valences and the respective English equivalents. For example, the Basque adjective *handi*, has a frequency of 101, its valence is +1, and its English equivalent is "big".

Once having the list with frequencies, valences and the respective English equivalents, we have looked up the English in the word in the MCR. Taking also as a reference the *Euskaltzaindiaren Hiztegia* (Euskaltzaindia, 2016) the dictionary of the Academy of Basque Language that includes definitions, we have checked if the meaning of Basque word corresponds to the meaning of the synsets that contained the English variants. In addition to that, we also use the sentiment valence of the Basque adjectives to determine if the variant corresponds to the synsets.

In the following subsections, we explain the casuistry we have found for the adjectives by means of some examples.

3.1. Adjectives with positive semantic orientation

In Table 2, we show the most frequent positive adjectives of the sentiment lexicon *SentiTegi* in the *Basque Opinion Corpus*. Following, we present the analysis for these cases.

Basque	**Instances**	**Sentiment valence**	**English**	**Sentiment valence**
Handi	101	+1	*Big*	+1
Berri	57	+2	*New*	+2

Table 2: Two positive words with their sentiment valence taken from *SentiTegi*

As far as *handi* "big" is concerned, its sentiment valence is (+1). If we want to enrich BWN with this word of the sentiment lexicon, the sentiment valence of the Basque word and the the meaning of synset need to agree. According to

the sentiment lexicon, *handi* "big" means something positive because the sentiment valence is (+1). Moreover, the variant "big"appears in several synsets.

Synset	Meaning	Match
ili-30-01382086-a	above average in size or number or quantity or magnitude or extent	Yes
ili-30-01276872-a	significant	Yes
ili-30-01510444-a	very intense	Yes
ili-30-01453084-a	loud and firm	Yes
ili-30-00579622-a	conspicuous in position or importance	No
ili-30-02402439-a	prodigious	Yes
ili-30-01890752-a	exhibiting self-importance	No
ili-30-01890187-a	feeling self-importance	No
ili-30-01488616-a	(of animals) fully developed	No
ili-30-01191780-a	marked by intense physical force	Yes
ili-30-01114658-a	generous and understanding and tolerant	Yes
ili-30-01111418-a	given or giving freely	Yes
ili-30-00173391-a	in an advanced stage of pregnancy	No

Table 3: Synsets including the variant "big" and its matches with the meaning and valence of *handi*

In Table 3.1., we list the synsets and meanings related to the word "big". Eight of them (ili-30-01382086-a, ili-30-01276872-a, ili-30-01510444-a, ili-30-01453084-a, ili-30-02402439-a, ili-30-01191780-a, ili-30-01114658-a, ili-30-01111418-a) match with sentiment valence and meaning of the word *handi* "big". The match happens with the synsets associated to the intensity in different ways (physical or psychic) but not with self-importance. In these last cases (ili-30-00579622-a, ili-30-01890752-a, ili-30-01890187-a, ili-30-01488616-a, ili-30-00173391-a), "big" makes negative evaluation and it does not match with the sentiment valence of the word *handi* (+1). To express those physiological features, there is another word in Basque: *handinahi* "arrogant" and this word should be included in those synsets. Morphologically, *handinahi* is a compound and includes *handi* "big", but it is an independent word. However, this lead us to think that derivative words and compounds can also play a role towards an automatic candidate proposal for non-matching synsets.
Regarding the word *berri* "new", presented in Table 3.1., it matches with all the synsets (ili-30-01640850-a, ili-30-01687167-a, ili-30-00937186-a, ili-30-00128733-a, ili-30-02070491-a, ili-30-02584699-a, ili-30-01687965-a and ili-30-00818008-a) and their meanings. In addition, the novelty means something positive or good and it goes in line with the sentiment valence (+2) of the word. However, in some meanings like ili-30-00024996-a, the novelty could be positive or negative according to the context.

Synset	Meaning	Match
ili-30-01640850-a	not of long duration; having just (or relatively recently) come into being or been made or acquiredor discovered	Yes
ili-30-01687167-a	original and of a kind not seen before	Yes
ili-30-00937186-a	lacking training or experience	Yes
ili-30-00128733-a	having no previous example or precedent or parallel	Yes
ili-30-02070491-a	other than the former one(s); different	Yes
ili-30-02584699-a	unaffected by use or exposure	Yes
ili-30-01687965-a	(of a new kind or fashion) gratuitously new	Yes
ili-30-00818008-a	(of crops) harvested at an early stage of development; before complete maturity	Yes
ili-30-00024996-a	unfamiliar	Yes(?)

Table 4: Synsets related to word "new" in the BWN

3.2. Adjectives with negative semantic orientation

In this subsection we analyse the most frequent adjectives that have negative connotation. We present these synsets in Table 5.

Basque	Instances	Sentiment valence	English	Sentiment valence
Politiko	33	−1	*Political*	−1
Txiki	30	−1	*Little*	−1

Table 5: Three negative words with their sentiment valence taken from *SentiTegi* Alkorta et al. (2018)

The examples in Table 5 show a different casuistry regarding the match with the meaning of synsets.
In the case of the sentiment word *politiko* "political", its meaning matches with three possible meanings (ili-30-01814385-a, ili-30-02857407-a and ili-30-02857587-a). But, when it comes to semantic orientation, the meanings of "political" in the English *WordNet* are neutral while in the case of the word *politiko* "political" is (−1). Therefore, there is a disagreement from the point of view of sentiment analysis. This suggests us that another synset may be necessary for this variant.
Finally, in the case of the sentiment word *txiki* "little", there are two cases. The word matches with some synsets (ili-

Synset	Meaning	Match
ili-30-01814385-a	involving or characteristic of politics or parties or politicians	Yes(*)
ili-30-02857407-a	of or relating to your views about social relationships involving authority or power	Yes(*)
ili-30-02857587-a	of or relating to the profession of governing	Yes(*)

Table 6: Synsets related to word "political"

Synset	Meaning	Match
ili-30-01391351-a	limited or below average in number or quantity or magnitude or extent	Yes
ili-30-01554510-a	(quantifier used with mass nouns) small in quantity or degree; not much or almost none or (with 'a') at least some	Yes
ili-30-01649031-a	(of children and animals) young, immature	Yes
ili-30-01280908-a	(informal) small and of little importance	Yes
ili-30-01455732-a	(of a voice) faint	No
ili-30-02386612-a	ow in stature; not tall	Yes
ili-30-01467534-a	lowercase	No
ili-30-00855670-a	small in a way that arouses feelings (of tenderness or its opposite depending on the context)	Yes

Table 7: Synsets related to word "little" in the BWN

30-01391351-a, ili-30-01554510-a, ili-30-01649031-a, ili-30-01280908-a, ili-30-02386612-a and ili-30-00855670-a) but, in other cases, there is no match. For the synset ili-30-01455732-a, the word *baxu* "low" is more suitable than *txiki* and for the synset ili-30-01467534-a, the word *xehe* "minuscule" is more appropriate. So, in these cases, the variants for the concepts should be added from another resource.

4. Conclusion and Future Work

In this work we have explored a method to enrich the BWN using *SentiTegi*, the sentiment lexicon in Basque. *SentiTegi* contains words with semantic information (sentiment valences, in this case) which are useful in the enrichment of the BWN with the help of a dictionary. In fact, in addition to the match of the definition, if the evaluation (positive or negative) of the meaning of synsets matches with the evaluation of the word (positive or negative sentiment valence), the word of the sentiment lexicon is valid for the BWN. This proves that *SentiTegi* and our methodology are good starting points for the enrichment of the BWN. However, we have found some cases where the direct addition of the word to BWN is doubtful. This leads us to think that criteria still need to be analysed and revised.

In future work, we would like to apply the Appraisal Theory (Martin and White, 2003) to this process of enrichment of the BWN. The Appraisal Theory is useful to categorize the type of subjectivity of words with sentiment valence. Indeed, not all the words with sentiment valence express the same sentiment. Some of them express opinions (for example, "hate") and others express sentiments (for instance, "happy"). The annotation of sentiment words with this theory would help to identify better the synsets that would match with them.

5. Acknowledgements

This work has been partially funded by the the project DeepReading (RTI2018-096846-B-C21) supported by the Ministry of Science, Innovation and Universities of the Spanish Government, Ixa Group-consolidated group type A by the Basque Government (IT1343-19) and BigKnowledge – *Ayudas Fundación BBVA a Equipos de Investigación Científica 2018*.

6. Bibliographical References

Agirre, E., Aldezabal, I., Etxeberria, J., Izagirre, E., Mendizabal, K., Pociello, E., and Quintian, M. (2006). A methodology for the joint development of the basque wordnet and semcor. In *LREC*, pages 23–28.

Aldezabal, I., Artola, X., de Ilarraza, A. D., Gonzalez-Dios, I., Labaka, G., Rigau, G., and Urizar, R. (2018). Basque e-lexicographic resources: linguistic basis, development, and future perspectives. In *Workshop on eLexicography: Between Digital Humanities and Artificial Intelligence*.

Alkorta, J., Gojenola, K., Iruskieta, M., and Taboada, M. (2017). Using lexical level information in discourse structures for basque sentiment analysis. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 39–47.

Alkorta, J., Gojenola, K., and Iruskieta, M. (2018). Sentitegi: Semi-manually created semantic oriented basque lexicon for sentiment analysis. *Computación y Sistemas*, 22(4).

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Lrec*, volume 10, pages 2200–2204.

Brooke, J., Tofiloski, M., and Taboada, M. (2009). Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of the international conference RANLP-2009*, pages 50–54.

Elhuyar Hizkuntza Zerbitzuak. (2013). *Elhuyar hiztegia: euskara-gaztelania, castellano-vasco*. Elhuyar.

Euskaltzaindia. (2016). *Euskaltzaindiaren Hiztegia*. Euskaltzaindia.

Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Gatti, L., Guerini, M., and Turchi, M. (2015). SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis. *IEEE Transactions on Affective Computing*, 7(4):409–421.

Martin, J. R. and White, P. R. (2003). *The language of evaluation*, volume 2. Springer.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Otegi, A., Imaz, O., Díaz de Ilarraza Sánchez, A., Iruskieta Quintian, M., and Uria Garin, L. (2017). Analhitza: a tool to extract linguistic information from large corpora in humanities research.

Pociello, E., Agirre, E., and Aldezabal, I. (2011). Methodology and Construction of the Basque WordNet. *Language resources and evaluation*, 45(2):121–142.

Sarasola, I. (2005). *Zehazki: gaztelania-euskara hiztegia*. Alberdania.

Siddharthan, A., Cherbuin, N., Eslinger, P. J., Kozlowska, K., Murphy, N. A., and Lowe, L. (2018). WordNet-feelings: A Linguistic Categorisation of Human Feelings. *arXiv preprint arXiv:1811.02435*.

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