MultiBooked: A Corpus of Basque and Catalan Hotel Reviews Annotated for Aspect-level Sentiment Classification.

Jeremy Barnes, Patrik Lambert, Toni Badia
Pompeu Fabra University, Barcelona, Spain
{firstname.lastname}@upf.edu

Abstract

While sentiment analysis has become an established field in the NLP community, research into languages other than English has been hindered by the lack of resources. Although much research in multi-lingual and cross-lingual sentiment analysis has focused on unsupervised or semi-supervised approaches, these still require a large number of resources and do not reach the performance of supervised approaches. With this in mind, we introduce two datasets for supervised aspect-level sentiment analysis in Basque and Catalan, both of which are under-resourced languages. We provide high-quality annotations and benchmarks with the hope that they will be useful to the growing community of researchers working on these languages.

Keywords: basque, catalan, sentiment analysis, aspect-level, under-resourced, opinion mining, cross-lingual

1. Introduction

Sentiment analysis has become an established field with a number of subfields (aspect-level sentiment analysis, social media sentiment analysis, cross-lingual sentiment analysis), all of which require some kind of annotated resource, either to train a machine-learning based classifier or to test the performance of proposed approaches.

Although much research into multi-lingual and cross-lingual sentiment analysis has focused on unsupervised or semi-supervised approaches (A.R. et al., 2012; Perez-Rosas et al., 2012; Gao et al., 2015), these techniques still require certain resources (linked wordnets, seed lexicon) and do not generally reach the performance of supervised approaches.

In English the state-of-the-art for binary sentiment analysis often reaches nearly 90 percent accuracy (Tai et al., 2015; Kim, 2014; Irsoy and Cardie, 2014), but for other languages there is a marked drop in accuracy. This is mainly due to the lack of annotations and resources in these languages. This is especially true of corpora annotated at aspect-level. Unlike document- or tweet-level annotation, aspect-level annotation requires a large amount of effort from the annotators, which further reduces the likelihood of finding an aspect-level sentiment corpus in under-resourced languages. We are, however, aware of one corpus annotated for aspects in German (Klinger and Cimiano, 2014), although German is not a particularly low-resource language.

The movement towards multi-lingual datasets for sentiment analysis is important because many languages offer different challenges, such as complex morphology or highly productive word formation, which can not be overcome by focusing only on English data.

The novelty of this work lies in creating corpora which cover both Basque and Catalan languages and are annotated in such a way that they are compatible with similarly compiled corpora available in a number of languages (Agerri et al., 2013). This allows for further research into cross-lingual sentiment analysis, as well as introducing the first resource for aspect-level sentiment analysis in Catalan and Basque. The corpus is available at `http://hdl.handle.net/10230/33928` or `https://jbarnesspain.github.io/resources/`.

2. Related Work

In English there are many datasets available for document- and sentence-level sentiment analysis across different domains and at different levels of annotation (Pang et al., 2002; Hu and Liu, 2004; Blitzer et al., 2007; Socher et al., 2013; Nakov et al., 2013). These resources have been built up over a period of more than a decade and are currently necessary to achieve state-of-the-art performance.

Corpora annotated at fine-grained levels (opinion- or aspect-level) require more effort from annotators, but are able to capture information which is not present at document- or sentence-level, such as nested opinions or differing polarities of different aspects of a single entity. In English, the MPQA corpus (Wiebe et al., 2005) has been widely used in fine-grained opinion research. More recently, a number of SemEval tasks have concentrated on aspect-level sentiment analysis (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016).

The Iberian peninsula contains two official languages (Portuguese and Spanish), as well as three co-official languages (Basque, Catalan, and Galician) and several smaller languages (Aragonese, Gascon). The two official languages do have available resources for sentiment at tweet-level (Villena-Román et al., 2013; Arruda et al., 2015), as well as at aspect-level (Agerri et al., 2013; Villena-Román et al., 2015; Almeida et al., 2015). The co-official languages, however, have almost none. The authors are aware of a small discourse-related sentiment corpus available in Basque (Alkorta et al., 2015), as well as a stance corpus in Catalan (Bosco et al., 2016). These resources, however, are limited in size and scope.

3. Data Collection

In order to improve the lack of data in low-resource languages, we introduce two aspect-level sentiment datasets to the community, available for Catalan and Basque. To collect suitable corpora, we crawl hotel reviews from `www.booking.com`. Booking.com allows you to search for reviews in Catalan, but it does not include Basque. Therefore, for Basque we crawled reviews from a number of other websites that allow users to comment on their stay[1]

Many of the reviews that we found through crawling are either 1) in Spanish, 2) include a mix of Spanish and the target language, or 3) do not contain any sentiment phrases. Therefore, we use a simple language identification method[2] in order to remove any Spanish or mixed reviews and also remove any reviews that are shorter than 7 tokens. This finally gave us a total of 568 reviews in Catalan and 343 reviews in Basque, collected from November 2015 to January 2016.

We preprocess them through a very light normalization, after which we perform tokenization, pos-tagging and lemmatization using Ixa-pipes (Agerri et al., 2014).

Our final documents are in KAF/NAF format (Bosma et al., 2009; Fokkens et al., 2014). This is a stand-off xml format originally from the Kyoto project (Bosma et al., 2009) and allows us to enrich our documents with many layers of linguistic information, such as the pos tag of a word, its lemma, whether it is a polar word, and if so, if it has an opinion holder or target. The advantage of this format is that we do not have to change the original text in any way.

4. Annotation

For annotation, we adopt the approach taken in the OpeNER project (Agerri et al., 2013), where annotators are free to choose both the span and label for any part of the text.

4.1. Guidelines

In the OpeNER annotation scheme[3] (see Table 1 for a short summary), an annotator reads a review and must first decide if there is any positive or negative attitudes in the sentence. If there are, they then decide if the sentence is on topic. Since these reviews are about hotels, we constrain the opinion targets and opinion expressions to those that deal with aspects of the hotel. Annotators should annotate the span of text which refers to:

- **opinion holders**,
- **opinion targets**,
- and **opinion expressions**.

If any opinion expression is found, the annotators must then also determine the polarity of the expression, which can be STRONG NEGATIVE, NEGATIVE, POSITIVE, or STRONG POSITIVE. As the opinion holder and targets are often implicit, we only require that each review has at least one annotated opinion expression.

For the strong positive and strong negative labels, annotators must use clues such as adverbial modifiers ('very bad'), inherently strong adjectives ('horrible'), and any use of capitalization, repetition, or punctuation ('BAAAAD!!!!!') in order to decide between the default polarity and the strong version.

Is there a positive / negative attitude?	yes/no
Is the sentence on topic ?	yes/no
Is it to the point?	yes/no
IF YES TO ALL, ANNOTATE:	
What is the span of the expression?	choose span
Is the expression positive or negative?	choose
Is the expression strong?	choose
Is there an explicit target?	yes/no
If yes, what is the span?	choose span
Is there an explicit opinion holder	yes/no
If yes, what is the span?	choose span

Table 1: Simplified annotation guidelines.

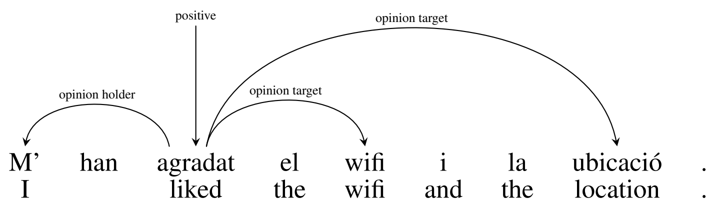

Figure 1: An opinion annotation following the annotation scheme detailed in Section 4.1..

4.2. Process

We used the KafAnnotator Tool (Agerri et al., 2013) to annotate each review. This tool allows the user to select a span of tokens and to annotate them as any of the four labels mentioned in Section 4.1..

The annotation of each corpus was performed in three phases: first, each annotator annotated a small number of reviews (20-50), after which they compared annotations and discussed any differences. Second, the annotators annotated half of the remaining reviews and met again to discuss any new differences. Finally, they annotated the remaining reviews. For cases of conflict after the final iteration, a third annotator decided between the two.

The final Catalan corpus contains 567 annotated reviews and the final Basque corpus 343.

4.3. Dataset Characteristics

The reviews are typical hotel reviews, which often mention various aspects of the hotel or experience and the polarity towards these aspects. An example is shown in Example Statistics for the two corpora are shown in Table 2.

4.4. Agreement Scores

Common metrics for determining inter-annotator agreement, e.g. Cohen's Kappa (Cohen, 1960) or Fleiss' Kappa (Fleiss, 1971), can not be applied when annotating sequences, as the annotators are free to choose which parts

[1]We took reviews from a total of 35 different websites, including www.airbnb.com, www.atrapalo.com, www.nekatur.net, www.rentalia.es, www.toprural.es, and www.tripadvisor.com.

[2]We use the count of stopwords to predict the probability that a review is written in Spanish, Catalan, or Basque.

[3]http://www.opener-project.eu/

	Catalan	Basque
Number of Reviews	567	343
Average length in tokens	45	46.9
Number of Targets	2762	1775
Number of Expressions	2346	2328
Number of Holders	236	296

Table 2: Corpus Statistics

	Catalan	Basque
Number of Reviews	567	343
Targets	.767	.739
Expressions	.716	.714
Holders	.121	.259
Polarity (MSE)	1.53	2.7

Table 3: Inter-annotator agreement scores. *AvgAgr* score is reported for targets, expressions and holders and averaged mean squared error is reported for polarity.

of a sequence to include. Therefore, we use the *agr* metric (Wiebe et al., 2005), which is defined as:

$$\mathrm{agr}(a||b) = \frac{|A \text{ matching } B|}{|A|} \quad (1)$$

where a and b are annotators and A and B are the set of annotations for each annotator. If we consider a to be the gold standard, agr corresponds to the recall of the system, and precision if b is the gold standard. For each pair of annotations, we report the average of the agr metric with both annotators as the temporary gold standard,

$$\mathrm{AvgAgr}(a,b) = \frac{1}{2}\big[\mathrm{agr}(a||b) + \mathrm{agr}(b||a)\big] \quad (2)$$

Perfect agreement, therefore, is 1.0 and no agreement whatsoever is 0.0. Similar annotation projects (Wiebe et al., 2005) report $AvgAgr$ scores that range between 0.6 and 0.8 in general.

For polarity, we assign integers to each label (Strong Negative: 0, Negative: 1, Positive: 2, Strong Positive: 3). For each sentence of length n, we take the mean squared error (MSE),

$$\text{Mean Squared Error} = \frac{1}{n}\sum_{i=1}^{n}(A-B)^2 \quad (3)$$

where A and B are the sets of annotations for the sentence in question. This approach punishes larger discrepancies in polarity more than small discrepancies, i.e. if annotator 1 decides an opinion expression is STRONG NEGATIVE and annotator two that the same expression is POSITIVE, this will be reflected in a larger MSE score than if annotator 2 had chosen NEGATIVE. Perfect agreement between annotators would lead to a MSE of 0.0, with the maximum depending on the length of the phrase. For a phrase of ten words, the worst MSE possible (assuming annotator 1 labeled all words STRONG POSITIVE and annotator 2 labeled them STRONG NEGATIVE) would be a 9.0. We take the mean of all the MSE scores in the corpus.

Inter-annotator agreement is reported in Table 3.

The inter-annotator agreement for target and expressions is high and in line with previous annotation efforts (Wiebe et al., 2005), given the fact that annotators could choose any span for these labels and were not limited to the number of annotations they could make. This reflects the clarity of the guidelines used to guide the annotation process.

The agreement score for opinion holders is somewhat lower and stems from the fact that there were relatively few instances of explicit opinion holders. Additionally, Catalan and Basque both have agreement features for verbs, which could be considered an implicit mention of the opinion holder. This is not always clear, however. Finally, the mean squared error of the polarity scores shows that annotators generally agree on where and which polarity score should be given. Again, the mean squared error in this annotation scheme requires both annotators to choose the same span and the same polarity to achieve perfect agreement.

5. Difficult Examples

During annotation, there were certain sentences which presented a great deal of problems for the annotators. Many of these are difficult because of 1) **nested opinions**, 2) **implicit opinions reported only through the presence or absence of certain aspects**, or 3) **the difficulty to identify the span of an expression**. Here, we give examples of each difficulty and detail how these were resolved during the annotation process.

(1) Hotela bikaina zen , nahiz eta bertako
Hotel.ABS.SG great.ABS.SG be , although there.from
langileak ez bereziki jatorrak izan.
workers.ABS.PL not particularly friendly.ABS.PL were
'The hotel was great, although the workers there were not particularly friendly.'

In the Basque sentence in Example 1, we can see that there are two distinct levels of aspects. First, the aspect 'hotel', which has a positive polarity and then the sub-aspect 'workers'. We avoid the problem of deciding which is the opinion target by treating these as two separate opinions, whose targets are 'hotel' and 'workers'.

(2) Igerilekua zegoen.
pool.ABS.SG was
'There was a pool.'

If there was an implicit opinion based on the presence or absence of a desirable aspect, such as the one seen in Example 2, we asked annotators to identify the phrase that indicates presence or absence, i.e. 'there was', as the opinion phrase.

(3) Langileek emandako arreta
workers.ERG.PL given.COMP attention.ABS.SG
bikaina zen .
excellent.ABS.SG was
'The attention that the staff gave was excellent.'

Finally, in order to improve overlap in span selection, we instructed annotators to choose the smallest span possible that retains the necessary information. Even after several iterations, however, there were still discrepancies with difficult examples, such as the one shown in Example 3, where the opinion target could be either 'attention', 'the attention', or 'the attention that the staff gave'.

6. Benchmarks

In order to provide a simple baseline, we frame the extraction of opinion holders, targets, and phrases as a sequence labeling task and map the NAF tags to BIO tags for the opinions in each review. These tags serve as the gold labels which will need to be predicted at test time. We also perform classification of the polarity of opinion expressions.

For the extraction of opinion holders, targets, and expressions we train a Conditional Random Field[4] (CRF) on standard features for supervised sequence labeling (word-, subword-, and part-of-speech information of the current word and previous words). For the classification of the polarity of opinion expressions, we use a Bag-of-Words approach to extract features and then train a linear SVM classifier[5]

For evaluation, we perform a 10-fold cross-validation with 80 percent of the data reserved for training during each fold. For extraction and classification, we report the weighted F_1 score. The results of the benchmark experiment (shown in Table 4) show that these simple baselines achieve results which are somewhat lower but still comparable to similar tasks in English (Irsoy and Cardie, 2014). The drop is not surprising given that we use a relatively simple baseline system and due to the fact that Catalan and Basque have richer morphological systems than English, which were not exploited.

	Catalan	Basque
Targets	.64	.57
Expressions	.52	.54
Holders	.56	.54
Polarity	.80	.84

Table 4: Weighted F_1 scores for extraction of opinion targets, expressions and holders, as well as the weighted F_1 for classification of polarity.

7. Conclusion

In this paper we have presented the MultiBooked corpus – a corpus of hotel reviews annotated for aspect-level sentiment analysis available in Basque and Catalan. The aim of this annotation project is to allow researchers to enable research on supervised aspect-level sentiment analysis in Basque and Catalan, as well as provide useful data for cross- and multi-lingual sentiment analysis. We also provide inter-annotator agreement scores and benchmarks, as well as making the corpus available to the community.

8. Bibliographical References

Agerri, R., Cuadros, M., Gaines, S., and Rigau, G. (2013). OpeNER: Open polarity enhanced named entity recognition. In *Sociedad Española para el Procesamiento del Lenguaje Natural*, volume 51, pages 215–218.

Alkorta, J., Gojenola, K., Iruskieta, M., and Perez, A. (2015). Using relational discourse structure information in basque sentiment analysis. In *Proceedings of the 5th Workshop on RST and Discourse Studies at SEPLN(2015)*, pages 1–10.

Almeida, M. S. C., Pinto, C., Figueira, H., Mendes, P., and Martins, A. F. T. (2015). Aligning opinions: Cross-lingual opinion mining with dependencies. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 408–418.

A.R., B., Joshi, A., and Bhattacharyya, P. (2012). Cross-lingual sentiment analysis for Indian languages using linked wordnets. In *Proceedings of COLING 2012: Posters*, pages 73–82.

Arruda, G. D. d., Roman, N. T., and Monteiro, A. M. (2015). An annotated corpus for sentiment analysis in political news. In *Proceedings of the 2015 Symposium in Information and Human Language Technology*, pages 101–110.

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447.

Bosco, C., Lai, M., Patti, V., Pardo, F. M. R., and Rosso, P. (2016). Tweeting in the debate about catalan elections. In *Proceedings of the 2016 LREC workshop on Emotion and Sentiment Analysis Workshop (ESA)*, pages 67–70.

Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., and Aliprandi, C. (2009). KAF: A generic semantic annotation format. In *Proceedings of the Generative Lexicon (GL2009) Workshop on Semantic Annotation*, Pisa, Italy.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurements*, 20:37–46.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Fokkens, A., Soroa, A., Beloki, Z., Ockeloen, N., Rigau, G., Robert van Hage, W., and Vossen, P. (2014). Naf and gaf: Linking linguistic annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*.

Gao, D., Wei, F., Li, W., Liu, X., and Zhou, M. (2015). Cross-lingual sentiment lexicon learning with bilingual word graph label propagation. *Computational Linguistics*, 41(1):21–40, March.

Hu, M. and Liu, B. (2004). Mining opinion features in customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*, pages 168–177.

Irsoy, O. and Cardie, C. (2014). Deep recursive neural networks for compositionality in language. In *Advances in*

[4]We use the implementation available in *sklearn_crfsuite*.

[5]We use the liblinear implementation from *sklearn*.

Neural Information Processing Systems, volume 3, pages 2096–2104.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Klinger, R. and Cimiano, P. (2014). The USAGE review corpus for fine grained multi lingual opinion analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2211–2218.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013). Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Perez-Rosas, V., Banea, C., and Mihalcea, R. (2012). Learning sentiment lexicons in spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3077–3081.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., and Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., and Eryiğit, G. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the EMNLP 2013*, pages 1631–1642.

Tai, K. S., Socher, R., and Manning, C. D. (2015). Improved semantic representations From tree-structured long short-term memory networks. In *Association for Computational Linguistics 2015 Conference*, pages 1556–1566.

Villena-Román, J., Lana-Serrano, S., Martínez-Cámera, E., and González-Cristóbal, J. C. (2013). Tass - workshop on sentiment analysis at sepln. *Procesamiento del Lenguaje Natural*, 50(0):37–44.

Villena-Román, J., Martínez-Cámera, E., García Morera, J., and Jiménez Zafra, S. M. (2015). Tass 2014 - the challenge of aspect-based sentiment analysis. *Procesamiento del Lenguaje Natural*, 54(0):61–68.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation (formerly Computers and the Humanities)*, 39(2/3):164–210.

9. Language Resource References

Rodrigo Agerri and Josu Bermudez and German Rigau. (2014). *IXA pipeline: Efficient and Ready to Use Multilingual NLP tools*. European Language Resources Association (ELRA), 1.0.