A Multi-lingual Annotated Dataset for Aspect-Oriented Opinion Mining

Salud María Jiménez Zafra[1], Giacomo Berardi[2], Andrea Esuli[2],
Diego Marcheggiani[2], María Teresa Martín-Valdivia[1], Alejandro Moreo Fernández[2]
[1]Departamento de Informática, Escuela Politécnica Superior de Jaén
Universidad de Jaén, E-23071 - Jaén, Spain
{sjzafra, maite}@ujaen.es
[2]Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
Consiglio Nazionale delle Ricerche, I-56124 - Pisa, Italy
{firstname.lastname}@isti.cnr.it

Abstract

We present the Trip-MAML dataset, a Multi-Lingual dataset of hotel reviews that have been manually annotated at the sentence-level with Multi-Aspect sentiment labels. This dataset has been built as an extension of an existent English-only dataset, adding documents written in Italian and Spanish. We detail the dataset construction process, covering the data gathering, selection, and annotation. We present inter-annotator agreement figures and baseline experimental results, comparing the three languages. Trip-MAML is a multi-lingual dataset for aspect-oriented opinion mining that enables researchers (i) to face the problem on languages other than English and (ii) to the experiment the application of cross-lingual learning methods to the task.

1 Introduction

Reviews of products and services that are spontaneously produced by customers represent a source of unquestionable value not only for marketing strategies of private companies and organizations, but also for other users since their purchasing decisions are likely influenced by other customers' opinions (Chevalier and Mayzlin, 2006).

Overall ratings (e.g., in terms of a five stars rating scale), and also aspect-specific ratings (e.g., the Cleanliness or Location of a hotel), are the typical additional information expressed by customers in their reviews. Those ratings help to derive a number of global scores to facilitate a first screening of the product or service at hand. Notwithstanding, users who pay more attention to a particular aspect (e.g., the Rooms of a hotel) remain constrained to manually inspect the entire text of reviews in order to find out the reasons other users argued in that respect. Methods for automatic analysis of the aspect-oriented sentiment expressed in reviews would enable highlighting aspect-relevant parts of the document, so as to allow users to perform a faster and focused inspection of them.

Previous work on opinion mining (Pang and Lee, 2008) has already faced the overall sentiment prediction (Pang et al., 2002), multiple aspect-oriented analysis (Hu and Liu, 2004), and fine-grained phrase-level analysis (Wilson et al., 2009). Most of the available opinion mining datasets contain only documents written in English, as this language is the most used on the Internet and the one for which more NLP tools and resources are available. (Hu and Liu, 2004) worked on the summarization of reviews by means of weakly supervised feature mining. (Täckström and McDonald, 2011) used a finer-grained dataset in which global polarity annotation is applied also to each sentence composing the document. Similarly did (Socher et al., 2013) with the Stanford Sentiment Treebank, which annotates each syntactically plausible phrase in thousands of sentences using annotators from Amazon's Mechanical Turk, annotating the polarity of phrases on a five-level scale. (Lazaridou et al., 2013) performed a single-label polarity annotation of elementary discourse units of TripAdvisor reviews, adopting ten aspect labels. (Marcheggiani et al., 2014) did a similar annotation work, using sentences as the annotation elements and adopting a multi-label polarity annotation, i.e., each sentence can be assigned to zero, one, or more than one aspect.

Cross-lingual sentiment classification (Wan, 2009; Prettenhofer and Stein, 2011) explores the scenario in which training data are available for

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2533–2538,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

a language that is different from the language of the test documents. Cross-lingual learning methods have important practical applications, since they allow to build classifiers for many languages reusing the training data produced for a single language (typically English), probably giving up a bit of accuracy, but compensating it with a large save in terms of human annotation costs.

Multi-lingual datasets are beneficial to the research community both as a benchmark to explore cross-lingual learning and also as resources on which to develop and test new NLP tools for languages other than English. Prettenhofer and Stein (2010) used a multi-lingual dataset focused on full-document classification at the global polarity level. Denecke (2008) used a dataset of 200 Amazon reviews in German to test cross-lingual document polarity classification using an English training set. Klinger and Cimiano (2014) produced a bi-lingual dataset (English and German), named USAGE, in which aspect expressions and subjective expressions are annotated in Amazon product reviews. In (Klinger and Cimiano, 2014) aspect expressions can be any piece of text that mentions a relevant property of the reviewed entity (e.g., *washer, hose, looks*) and are not categorical label, as in our dataset. The USAGE dataset is thus more oriented at information extraction rather than at text classification applications. Banea et al. (2010) used machine translation to create a multilingual version of the information-extraction oriented MPQA dataset (Wiebe et al., 2005) on six languages (English, Arabic, French, German, Romanian and Spanish).

In this paper we present Trip-MAML, which extends the Trip-MA[1] dataset of Marcheggiani et al. (2014) with Italian and Spanish annotated reviews. We describe Trip-MAML and report experiments aimed at defining a first baseline. Both the dataset and the software used in experiments are publicly available at `http://hlt.isti.cnr.it/trip-maml/`.

2 Annotation Process

We recall the annotation process adopted by Marcheggiani et al. (2014) for Trip-MA and the procedure we employed to extend it into Trip-MAML. We will use the national codes EN, ES, and IT, to denote the English, Spanish, and Italian parts of the Trip-MAML dataset, respectively. Note that EN coincides with Trip-MA.

2.1 English Reviews

The Trip-MA dataset was created by Marcheggiani et al. (2014) by annotating a set of 442 reviews, written in English, randomly sampled from the publicly available TripAdvisor dataset of Wang et al. (2010), composed by 235,793 reviews. Each review comes with an overall rating on a discrete ordinal scale from 1 to 5 “stars”. The dataset was annotated according to 9 recurrent aspects frequently involved in hotel reviews: Rooms, Cleanliness, Value, Service, Location, Check-in, Business, Food, and Building. The last two are not officially rated by TripAdvisor but were added because they are frequently commented in reviews. Two “catch-all” aspects, Other and NotRelated, were also added, for a total of 11 aspect. Aspect Other denotes opinions that are pertinent to the hotel being reviewed, but not relevant to any of the former nine aspects (e.g., generic evaluations like *Pulitzer exceeded our expectations*). Aspect NotRelated denotes opinions that are not related to the hotel (e.g., *Tour Eiffel is amazing*).

If a sentence is relevant to an aspect, the possible sentiment label values are three: Positive, Negative, and Neutral/Mixed[2]. Neutral/Mixed annotates subjective evaluations that are not clearly polarized (e.g., *The hotel was fine with some exceptions*).

2.1.1 Annotation protocol

Marcheggiani et al. (2014) relied on three human annotators to annotate each sentence of the 442 reviews with respect to polarities of opinions that are relevant to any of the 11 aspects. 73 reviews, out of 442, were independently annotated by all the annotators in order to measure the inter-annotator agreement, while the remaining 369 reviews were partitioned into 3 equally-sized sets, one for each annotator. Bias in the estimation of inter-annotator agreement was minimized by sorting the list of reviews of each annotator so that every eighth review was common to all annotators; this ensured that each annotator had the same amount of coding experience when labeling the same shared review.

[1] Marcheggiani et al. (2014) gave no name to their dataset, here we name it Trip-MA to identify its source and its multi-aspect nature.

[2] Marcheggiani et al. (2014) initially distinguished between implicit and explicit opinions but the human agreement was so low they removed this distinction from the schema.

	# Reviews	# Sentences	# Opinion-laden sentences
EN	442	5799	4810
ES	500	2620	2400
IT	500	2593	2392

Table 1: Number of reviews, sentences, and sentences with at least one opinion annotation.

2.2 Spanish and Italian Reviews

For the creation of ES and IT parts of the Trip-MAML dataset we followed the same annotation protocol of Marcheggiani et al. (2014), employing teams of three native speakers as annotators for each language. We crawled the Spanish and Italian reviews from TripAdvisor by accessing its websites with the '.es' and '.it' domains, which mostly contains reviews in the national language. From that domains we downloaded the reviews for the 10 most visited cities in Spain and Italy, respectively. We downloaded 10 reviews for every hotel of each city, obtaining a total of 17,020 reviews for Spanish and 33,325 for Italian. For each dataset, 500 reviews were selected by randomly sampling 50 reviews for each city. We thus obtained 139 unique reviews for each annotator, plus 83 reviews which all three annotators independently annotated.

We decided to annotate the aspects that were ratable on TripAdvisor at the time of our crawl (April 2015: Rooms, Cleanliness, Value, Service, Location, and Sleep Quality). Differently from the aspect schema in EN, we included the new aspect Sleep Quality, and we did not consider the missing aspects Check-in and Businnes, which are, in any case, the least frequent aspects in the Trip-MA dataset (see Table 2). We kept the additional aspects Food, Building, Other, and NotRelated, as they still appear frequently in the reviews. We adopted the same 3-values sentiment label schema of EN, i.e., Positive, Negative, or Neutral/Mixed.

Following the same procedure adopted by Marcheggiani et al. (2014), the Spanish and Italian annotator teams performed a preliminary annotation session on reviews not included in the final dataset. This preliminary activity was aimed at aligning the annotators' understanding about the labeling process for the different aspects, by sharing and solving any doubt that might arise during the annotation of some examples.

2.3 Statistics

Table 1 shows that English reviews have, on average, about double the number of sentences of Spanish and Italian reviews. This can be in part motivated by observing that the sentences in EN are, on average, 25% shorter than in ES and IT. Also, after a manual inspection of the data, we found that the EN part contains some reviews related to long vacations in resorts, thus describing in longer details the experience, while IT and ES reviews are mainly related to relatively short visits to classic hotels. However, the portion of opinionated sentences is similar across the three parts, indicating homogeneity in content, which is confirmed by the detailed aspect-level statistics reported in Table 2.

Both aspect and sentiment labels show imbalanced distributions that follow similar distributions across the three parts. The most frequent aspect in all collections is Other, followed by Rooms, Service, and Location. Building and Value are among the least frequent ones. The average value of the Pearson correlation between the lists of the shared aspects ranked by their relative frequency, measured pairwise among the three parts, is 0.795, which indicates a good uniformity of content among the parts. In all the three parts, Positive is the most frequent sentiment label, followed by Negative. Location is always the aspect with the highest frequency of positive labels.

3 Inter-annotator Agreement

We measured the inter-annotator agreement in two steps. The F_1 score measures the agreement on aspect identification, regardless of the sentiment label assigned. Then symmetric Macro-averaged Mean Absolute Error (sMAE^M) (Baccianella et al., 2009) measures the agreement on sentiment labels on the annotations for which the annotators agreed at the aspect level. Aspect NotRelated is not included in agreement evaluation, nor in the experiments of Section 4. sMAE^M is computed between each of the three possible pairs of annotators and then averaged to determine the agreement values reported in Table 3.

Agreement on aspect detection is higher for ES and IT than for EN. This difference is in part motivated by the fact that the two aspects that are missing in ES and IT have low agreement on EN, and the novel Sleep Quality aspect has instead a high agreement. However, also on the other aspects

		Other	Service	Rooms	Clean.	Food	Loc.	Check-in	Sleep-q.	Value	Build.	Busin.	NotRelated	*Total*
EN	Pos	893	513	484	180	287	435	93	-	188	185	23	63	*3344*
	Neg	353	248	287	66	127	51	56	-	87	62	3	40	*1377*
	Neu	167	40	111	5	82	38	12	-	35	22	4	350	*866*
	Total	*1413*	*801*	*882*	*251*	*496*	*524*	*161*	-	*310*	*269*	*30*	*453*	*5587*
ES	Pos	634	382	275	181	128	452	-	126	114	71	-	39	*2402*
	Neg	244	85	159	40	37	38	-	75	48	28	-	38	*792*
	Neu	46	19	62	6	32	22	-	6	18	7	-	4	*222*
	Total	*924*	*486*	*496*	*227*	*197*	*512*	-	*207*	*180*	*106*	-	*81*	*3416*
IT	Pos	582	415	267	259	207	389	-	103	135	77	-	50	*2484*
	Neg	189	74	110	65	56	22	-	50	27	43	-	15	*651*
	Neu	102	30	59	10	52	49	-	1	32	32	-	100	*467*
	Total	*873*	*519*	*436*	*334*	*315*	*460*	-	*154*	*194*	*152*	-	*165*	*3602*

Table 2: Number of opinion expressions at the sentence level of the datasets.

		Other	Service	Rooms	Clean.	Food	Loc.	Check-in	Sleep-q.	Value	Build.	Busin.	*Avg*
EN	F_1	.607	.719	.793	.733	.794	.795	.464	-	.575	.553	.631	*.675*
	sMAE^M	.308	.219	.191	.114	.234	.259	.003	-	.202	.150	.029	*.171*
ES	F_1	.789	.911	.854	.933	.882	.896	-	.895	.829	.538	-	*.836*
	sMAE^M	.174	.093	.133	.303	.120	.293	-	.000	.150	.184	-	*.161*
IT	F_1	.676	.812	.788	.913	.884	.856	-	.789	.858	.532	-	*.790*
	sMAE^M	.292	.166	.242	.114	.204	.204	-	.067	.292	.114	-	*.188*

Table 3: Inter-annotator agreement. F_1 on sentence-level aspect identification (higher is better). sMAE^M on sentence-level sentiment agreement (only on matching aspects, lower is better).

there is, in general, a higher or equal agreement in ES and IT with respect to EN, indicating that the formers two were annotated in a more consistent way. The agreement on assignment of sentiment label is rather similar across the whole dataset.

4 Experiments

The experiments we present here are aimed at defining a shared baseline for future experiments. For this reason we chose a relatively simple setup that uses a simple learning model and minimal linguistic resources. We used a sentence-level Linear Chain (LC) Conditional Random Field (Lafferty et al., 2001) as described by Marcheggiani et al. (2014). With respect to the features extracted from text, we used three simple features types: word unigrams, bigrams, and SentiWordNet-based features, which consist of a Positive and a Negative feature extracted every time the review contains a word that is marked as such in SentiWordNet (Baccianella et al., 2010). To use SentiWordNet on ES and IT, we used Multilingual Central Repository (Gonzalez-Agirre et al., 2012) and MultiWordNet (Pianta et al., 2002) to map sentiment labels to Spanish and to Italian, respectively.

Experiments were run separately on the EN, ES, and IT parts, leaving cross-lingual experiments to future work. On each part we built five 70%/30% train/test splits, randomly generated by sampling the reviews annotated by single reviewers (we left out reviews annotated by all the reviewers, as we consider that part of the dataset more useful as a validation set for the optimization of methods tested in future experiments). We then run the five experiments and averaged their results.

4.1 Evaluation Measures

As for the agreement evaluation (Section 3), we split the evaluation of experiments into two parts, aspect detection and sentiment labeling. For the sentiment labeling part we used simple Macro-averaged Mean Absolute Error (MAE^M, not the symmetric version) as the true dataset labels are the reference ones in this case, while in the annotator agreement case the two sets of labels have equal importance.

4.2 Results

Experiments on ES and IT obtain better F_1 values than on EN, indicating that the observed higher human agreement can be also explained by a lower hardness of the task when working with Spanish

		Other	Service	Rooms	Clean.	Food	Loc.	Check-in	Sleep-q.	Value	Build.	Busin.	*Avg*
EN	F_1	.482	.595	.626	.729	.541	.616	.230	-	.331	.281	.222	*.465*
	MAE^M	.549	.822	.641	.968	.585	.959	.264	-	.598	.471	.000	*.586*
ES	F_1	.520	.668	.766	.782	.567	.730	-	.416	.593	.215	-	*.584*
	MAE^M	.839	.737	.515	.377	.516	1.002	-	.395	.564	.000	-	*.549*
IT	F_1	.576	.747	.646	.770	.697	.757	-	.254	.630	.087	-	*.574*
	MAE^M	.707	.781	.809	.887	.829	.746	-	.053	.403	.000	-	*.579*

Table 4: Linear Chain CRFs experiments. F_1 on sentence-level aspect identification (higher is better). MAE^M on sentence-level sentiment assignment (only on correctly identified aspects, lower is better).

and Italian.

MAE^M values are all similar across languages, again confirming what has been observed on agreement. However, MAE^M values on experiments are sensibly worse than those measured on agreement, possibly due to the fact that we used very basic features, with limited use of sentiment-related information.

5 Conclusion

We have presented Trip-MAML a multi-lingual extension of Trip-MA, originally presented in (Marcheggiani et al., 2014). The extension process involved crawling and selecting the reviews for the two new languages, Spanish and Italian, and their annotation by a total of six native language speakers. We measured dataset statistics and inter-annotator agreement, which show that the new ES and IT parts we produced are consistent with the original EN part. We also presented experiments on the dataset, based on a linear chain CRFs model for the automatic detection of aspects and their sentiment labels, establishing a baseline for future research. Trip-MAML enables the exploration of cross-lingual approaches to the problem of multi-aspect sentiment classification.

Acknowledgments

This work has been partially supported by ATTOS project (TIN2012-38536-C03-0) from the Spanish Government.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2009. Evaluation measures for ordinal regression. In *Proocedings of the 9th Conference on Intelligent Systems Design and Applications (ISDA 2009)*, pages 283–287.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: Are more languages better? In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Judith A Chevalier and Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, 43(3):345–354.

Kerstin Denecke. 2008. Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, pages 507–512. IEEE.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual central repository version 3.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2525–2529.

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*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Roman Klinger and Philipp Cimiano. 2014. The usage review corpus for fine-grained, multi-lingual opinion analysis. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, May. ELRA.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*, pages 282–289.

Angeliki Lazaridou, Ivan Titov, and Caroline Sporleder. 2013. A bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1630–1639.

Diego Marcheggiani, Oscar Täckström, Andrea Esuli, and Fabrizio Sebastiani. 2014. Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In *Proceedings of the 36th European Conference on IR Research (ECIR 2014)*, pages 273–285.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 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.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the first international conference on global WordNet*, volume 152, pages 55–63.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1118–1127.

Peter Prettenhofer and Benno Stein. 2011. Cross-lingual adaptation using structural correspondence learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(1):13.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP 2013)*, volume 1631, page 1642.

Oscar Täckström and Ryan McDonald. 2011. Discovering fine-grained sentiment with latent variable structured prediction models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR 2011)*, pages 368–374.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 235–243.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010)*, pages 783–792.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433.