SentiTur: Building Linguistic Resources for Aspect-Based Sentiment Analysis in the Tourism Sector

Soluna Salles and Aroa Orrequia-Barea

The use of linguistic resources beyond the scope of language studies, i.e. commercial purposes, has become commonplace since the availability of massive amounts of data and the development of tools to process them. An interesting focus on these materials is provided by Sentiment Analysis (SA) tools and methodologies, which attempt to identify the polarity or semantic orientation of a text, i.e., its positive, negative, or neutral value. Two main approaches have been made in this sense, one based on complex machine-learning algorithms and the other relying principally on lexical knowledge *(*Taboada *et al*., 2011). Lingmotif is an example of lexicon-based SA tool offering polarity classification and other related metrics, together with an analysis of the target segments evaluated (Moreno-Ortiz, 2017). Sentiment has been shown to be domain-specific to a large extent (Choi & Cardie, 2008) and it is therefore necessary to study and describe how sentiment is expressed not only in general language, but also in specialized domains. The availability of annotated, domain-specific corpora could greatly enhance the capacity of SA tools.

Furthermore, the demand for a more fine-grained approach requires the identification of specific domain terminology, allowing the recognition of target terms associated with the polarity (Liu, 2012). Most available SA corpora are annotated at the document level, which allows systems to be trained to return the overall orientation of the text. However, more detail is necessary: what aspects exactly are being praised or criticized? This type SA is known as Aspect-Based Sentiment Analysis (ABSA), and attempts to extract more fined-grained knowledge. ABSA has attracted the attention of recent SemEval shared-tasks (Pontiki *et al.*, 2015).

We propose the creation of the SentiTur corpus, a bilingual (Spanish-English), aspect-annotated corpus of user reviews covering three sectors of the tourism industry: accommodation, catering, and car rental. Reviews obtained from online platforms (Tripavisor and Booking, among others) are being annotated according to an annotation schema by means of the collaborative annotation tool Brat (Stenetorp *et al*., 2012). Five annotators work initially in a preliminary dataset, so as to validate the schema and the methodology proposed. Inter annotator agreement (IAA) is computed measuring the pairwise agreement among annotators by the Kappa coefficient (K) (Siegel, 1988). These results are used to revise and adjust the annotation schema before it is employed to annotated the general corpus. Finally, the corpus is offered in standard XML format suitable as input of Sentiment Analysis tools.

Key words: SentiTur, Aspect-based Sentiment Analysis, Sentiment Analysis, tourism.

References

Choi, Y., & Cardie, C. (2008). Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis, (October), 793–801.

Liu, B. (2012). *Sentiment Aanalysis and Opinion Mining: Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Moreno-Ortiz, A. (2017). Lingmotif: Sentiment Analysis for the Digital Humanities. In *Proceedings of the EACL 2017 Software Demonstrations, Valencia, Spain, April 3-7 2017* (pp. 73–76).

Pontiki, M., Galanis, D., & Papageorgiou, H. (2015). SemEval-2015 Task 12: Aspect-Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, June 4-5, 2015*.

Siegel, S. (1988). *Nonparametric Statistics for the Behavioral Science*. McGraw-Hill.

Stenetorp, P., Pyysalo, S., Topi, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23-27 2012*. (pp. 102–107).

Taboada, M., Brooke, J., & Voll, K. (2011). Lexicon-Based Methods for Sentiment Analysis, (September 2010).