Multilingual and Multi-Domain Opinion Mining and Sentiment Analysis

Proposers: Rodrigo Agerri and German Rigau
Contact: rodrigo.agerri@ehu.eus http://www.rodrigoagerri.net/

Description

Opinion Mining and Sentiment Analysis (OMSA) are crucial for determining opinion trends and attitudes about commercial products, companies reputation management, brand monitoring, or to track attitudes by mining social media, etc. Furthermore, given the explosion of information produced and shared via the Internet, especially in social media, it is simply not possible to keep up with the constant flow of new information by manual methods (Fernandez de Landa et al. 2019).

Early approaches to OMSA were based on document classification, where the task was to determine the polarity (positive, negative, neutral) of a given document or review (Pang 2008, Liu 2012). A well known benchmark for polarity classification at document level is that of (Pang et al. 2002). Later on, a finer-grained OMSA was deemed necessary (Pontiki et al. 2016, Agerri and Rigau 2018). This was motivated by the fact that in a given review more than one opinion about a variety of aspects or attributes of a given product is usually conveyed. Thus, Aspect Based Sentiment Analysis (ABSA) was defined as a task which consisted of identifying several components of a given opinion: the opinion holder, the target, the opinion expression (the textual expression conveying polarity) and the aspects or features. Aspects are mostly domain-dependent. In restaurant reviews, relevant aspects would include "food quality, price, service, restaurant ambience", etc. Similarly, if the reviews were about consumer electronics such as laptops, then aspects would include "size", "battery life", "hard drive capacity", etc.

In the review shown below there are three different opinions about two different aspects (categories) of the restaurant, namely, the first two opinions are about the quality of the food and the third one about the general ambience of the place. Furthermore, there are just two opinion targets because the target of the third opinion, the restaurant itself, remains implicit. Finally, each aspect is assigned a polarity; in this case all three opinion aspects are negative.

```
<sentence id="1016296:4">
    <text>Chow fun was dry; pork shu mai was more than usually greasy and had to
        share a table with loud and rude family</text>
    <Opinions>
        <Opinion target="Chow fun" category="FOOD#QUALITY" polarity="negative"
            from="0" to="8" pfrom=13 pto=16/>
        <Opinion target="pork shu mai" category="FOOD#QUALITY" polarity="negative"
             from="18" to="30" pfrom=53 pto=59/>
        <Opinion target="NULL" category="AMBIENCE#GENERAL" polarity="negative"
            from="0" to="0" pfrom=90 pto=103/>
    </Opinions>
</sentence>
```

The identification of each aspect for three opinions is modeled as different NLP tasks. Each of these tasks are highly dependent on the domain (restaurant reviews, sports, politics) and text genre (social media, newspaper

articles, etc.). This means that usually there is not training data available for many domains and language. This issue is particularly true of low resource languages such as Basque, but also for other languages such as Spanish.

Objectives

The candidate may choose between the following objectives:

1. New characterization of the Aspect Based Opinion Mining task.
2. Addressing implicit opinion targets.
3. Experiment with new transfer learning approaches for Opinion Mining in low resource languages (Basque).
4. Experiment with new dynamic, contextual word embeddings (Flair, Elmo, BERT, etc.) for multilingual opinion mining.
5. Semi-automatic generation of training data for domain and low-resourced languages.

The master thesis in Basque, English or Spanish.

Tasks and Plan

- December-January: Start of the project, defining the objectives and tasks.
- February: Start experiments. Optionally, it is recommended for the candidates to attend the "Seminar on language technologies. Deep Learning (LAP 18). https://ixa.si.ehu.es/master/programa_html
- March-May: Experiments and final development.
- June: Writing up.

References

Agerri R., Rigau G. (2018). Language independent sequence labelling for Opinion Target Extraction. Artificial Intelligence, 268 (2018) 85-95.

Akbik, A.; Blythe, D.; Vollgraf, R. (2018). Contextual string embeddings for sequence labeling.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL 2019.

Fernandez de Landa, J., Agerri, R., & Alegria, I. (2019). Large Scale Linguistic Processing of Tweets to Understand Social Interactions among Speakers of Less Resourced Languages: The Basque Case. MDPI: *Information*, *10* (6), 212. https://doi.org/10.3390/info10060212

Pang, B., Lee, L., 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2, 1–135.

Pang, B., Lee, L., Vaithyanathan, S., 2002. Thumbs up? sentiment classification using machine learning techniques, in: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. pp. 79–86.

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