SentiTagger - Automatically Tagging Text in OpinionMining-ML

Livio Robaldo, Luigi Di Caro, Alessio Antonini

Department of Computer Science, University of Turin
{robaldo,dicaro,antonini}@di.unito.it

Abstract. This paper presents `SentiTagger`, a research project proposal aiming at designing and implementing a computational system that automatically tag free text in `OpinionMining-ML` [1]. The latter is an XML-based formalism that has been proposed as a standard in the field of Sentiment Analysis.

Keywords: Sentiment Analysis, Opinion Mining

1 The Opinion Mining and the Limits of Current Systems

Opinion Mining, or Sentiment Analysis, can be generally defined as the extraction of users' opinions from texual data. The most relevant motivations behind the recent attraction on this task has to do with its interesting range of applications. For example, a product seller may be interested in knowing the customers' opinions about its products.

In computer, the discovery of *sentiments* and *opinions* that are contained in texts is involved on the use of Natural Language Processing (NLP) techniques (cf. [2]). At the current state of the art, NLP partially provides methods and approaches that can fit with these *emotion-based* kinds of information. Several electronic dictionaries for Sentiment Analysis like Senti-Wordnet [3] have been proposed so far. Nevertheless, the aggregation of simple associations *<word-sentiment>* does not take into account the high complexity of whole sentences, where the use of deep syntactic parsing becomes crucial in that sense.

In addition, in our opinion, the concepts of *sentiment* and *opinion* only cover one part of a bigger set of interesting information that can be relevant. The speaker/writer could point out details without ascribing any sentiment to them. For instance, he could point out that a certain restaurant made the take-away service available, without commenting anything about its efficiency, quality, and so on. Such objective information, that are clearly precious from the perspective of an Information Retrieval system, are usually denoted as "neutral" [4]. Finally, it seems that other kinds of information should be integrated in such models. For example, texts can contain suggestions, comparisons, questions, and so forth.

Therefore, from a computer scientist's perspective, Sentiment Analysis should be seen as an information extraction subtask, where the concept of *emotion* becomes less important than the concept of *facet that caused the emotion*. Furthermore, facets can have relations connecting them and so they can be organized

into an ontology (cf. [5], [6], and [7]). Then, sentiments, opinions, observations, suggestions and comparisons can refer to different concepts in the ontology, at different level of specificity.

In other words, it would be rather useful to have at disposal a formalism that allows to tag all relevant information and to organize them by decoupling relevant textual expressions from the facets those expressions refer to, and relate the former to the latter possibly collocating them within an ontology.
In the industry, there are some attempts to define such a formalism. But, to our knowledge, so far no one has ever tried to systematize and generalize the solutions found in order to share them with the scientific community, by making such solutions contextually independent, easy to extend, easy to integrate within heterogeneous computational systems, etc.

In the light of this, [1] proposed `OpinionMining-ML`, an XML-based formalism that can put some basis for the creation of a standard in the field of Sentiment Analysis. `OpinionMining-ML` will be presented in the next section. We propose here a research project aiming at designing and implementing a computational system able to automatically tag text in `OpinionMining-ML`.

2 `OpinionMining-ML` and `SentiTagger`

`OpinionMining-ML` is a *facet-oriented* annotation formalism. Facets are contextually relevant concepts about which the customers/owners of the restaurant could be interested in knowing what the commentators say. For instance, typical facets of the domain of restaurants are the *cuisine* (more or less tasty), the service (more or less polite), the price (more or less expensive) but also the ease of parking outside the restaurant, the availability of a take-away service, etc.
Obviously, the set and the granularity of the available facets varies depending on the domain and the customers' needs. For this reason, `OpinionMining-ML` organizes them into an ontology. Ontologies are still scarsely considered in Sentiment Analysis, while in `OpinionMining-ML` they have a crucial role, as they facilitate the management, organization, and retrieval of the annotated comments.
Once the ontology of facets is built, every portion of text that conveys an appraisal, observation, suggestion, comparison, etc. about a facet is annotated. Of course, in order to automatically identify the correct bounds of a portion of text referring to a facet, the use of a parser is crucial.
Two examples of comments taken from the corpus developed in [1] are:

1. Ottima pizza senza glutine! ;)
   [Excellent pizza without gluten!]
2. In qualche modo ricorda lo Shambala ma qui, secondo me, si mangia meglio.
   [In some sense it reminds the Shambala but here, in my view, you can eat better]

Let us assume, for simplicity, that (1)-(2) are about the same restaurant called "RestaurantX". The first module of `SentiTagger` has to identify the facets these comments are about. They are the "pizza", the "gluten-free food", the "cuisine"

of RestaurantX. RestaurantX itself is a facet, and also the Shambala restaurant and its cuisine, to which RestaurantX is compared.
The following ontology in `OpinionMining-ML` is then built:

```
<ONTOFACETS>
   <FACET id="1">RestaurantX</FACET>
   <FACET id="2">pizza served-at RestaurantX</FACET>
   <FACET id="3">gluten-free food served at RestaurantX</FACET>
   <FACET id="4">cuisine of RestaurantX</FACET>
   <FACET id="5">Restaurant Shambala</FACET>
   <FACET id="6">cuisine of Restaurant Shambala</FACET>
</ONTOFACETS>
```

Every facet has a unique id within the ontology, used for external references. The text within the tag `<FACET>` is a mere description that does not have any ontological value. Facets are concepts that need to be related to each other via additional relations. For instance, we state that the facet with id="4" is a feature of the facet with id="1" by adding the following assertion:

```
<FEATURE-OF id="1"><FACETREFERENCE>4</FACETREFERENCE></FEATURE-OF>
```

We do not report here the set of all additional relations that may be asserted on the facets above. See [1] for further details.
Once the ontology is built, it is possible to tag the text by attributing different portions of text to different facets. However, not all relevant portions of text convey positive or negative opinions about facets (called "appraisals" in `OpinionMining-ML`). Only the first comment in (1)-(2) contains an appraisal about the pizza served in RestaurantX. On the other hand, "senza glutine" is an observation of the kind of pizza served in RestaurantX. Although the latter is not an appraisal, it is considered relevant as well from the point of view of an Information Retrieval system, in that a celiac person could look in the web for restaurants compatible with his/her disease. Finally, the comment (2) contains two comparisons: one between RestaurantX and restaurant Shambala and the other between the cuisines of the two restaurants.
`OpinionMining-ML` provides tags for annotating the different linguistic expressions. A simplified version of the annotation of the two comments (1)-(2) is:

```
<COMMENT>
   <APPRAISAL polarity="positive">
      <FACETREFERENCE>2</FACETREFERENCE>
      Ottima pizza
   </APPRAISAL>
   <OBSERVATION>
      <FACETREFERENCE>3</FACETREFERENCE>
      senza glutine! ;-)
   </OBSERVATION>
</COMMENT>
```

```
<COMMENT>
   <COMPARISON>
      <FACETREFERENCE>1</FACETREFERENCE>
      <FACETREFERENCE>5</FACETREFERENCE>
      In qualche modo ricorda lo Shambala
   </COMPARISON>
   ma
   <COMPARISON>
      <FACETREFERENCE>4</FACETREFERENCE>
      <FACETREFERENCE>6</FACETREFERENCE>
      qui, secondo me, si mangia meglio.
   </COMPARISON>
</COMMENT>
```

In its general version, `OpinionMining-ML` allows to split the text into fragments, and then recollect and attribute them to the facets. The splitting of the text is based on its syntactic structure. This allows to deal with a broader range of expressions, involving coordinations or other complex linguistic phenomena.

For this reason, for automatically building documents in `OpinionMining-ML`, `SentiTagger` will exploit the Tule Parser, a rule-based dependency parser developed at the University of Turin [8]. It is currently one of the most effective dependency parsers for Italian.
Having at disposal the parsed trees of the text, and an ontology built offline depending on the domain (e.g., an ontology for the domain of restaurants, for processing comments from `http://www.2spaghi.it`), `SentiTagger` will be able to identify the facets the comments are about, and ascribing the proper textual expressions to them.

References

1. Robaldo, L., Di Caro, L.: OpinionMining-ML. Computer Standards & Interfaces **35** (2013)
2. Wiebe, J., Wilson, T., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation **1** (2005)
3. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: 7th conference on International Language Resources and Evaluation. Volume 25. (2010)
4. Go, A., Huang, L., Bhayani, R.: Twitter sentiment analysis. Entropy (2009)
5. Zhou, L., Chaovalit, P.: Ontology-supported polarity mining. JASIST **59** (2008)
6. Zhao, L., Li, C.: Ontology based opinion mining for movie reviews. In: KSEM. (2009) 204–214
7. Peñalver-Martínez, I., Valencia-García, R., Sánchez, F.G.: Ontology-guided approach to feature-based opinion mining. In: NLDB. (2011) 193–200
8. Lesmo, L.: The Rule-Based Parser of the NLP Group of the University of Torino. Intelligenza Artificiale **2** (2007) 46–47