Ent-it-UP

A Sentiment Analysis System Based on OpeNER Cloud Services

Sara Pupi, Giulia Di Pietro, and Carlo Aliprandi

Synthema Srl Via Malasoma 24 56121 Ospedaletto (Pisa) – Italy
{sara.pupi,giulia.dipietro,carlo.aliprandi}@synthema.it

Abstract. In this paper we present a web application that exploits OpeNER Cloud Services. Ent-it-UP monitors Social Media and traditional Mass Media contents, performing multilingual Named Entity Recognition and Sentiment Analysis. Since consumers tend to trust the opinion of other consumers, reviews and ratings on the internet are increasingly important. Given the huge amount of data flowing in the web, it has become necessary to adopt an automatic data analysis strategy, in order to understand what people think about a certain product, brand or topic. The goal of Ent-it-Up is to carry out statistics about retrieved entities and display results in a communicative, intuitive and user friendly interface. In this way the final user can easily have a hint about people opinions without wasting too much time in analyzing the huge amount of User-Generated Content.

Keywords: Reference Application, OpeNER, Named Entity Recognition and Classification, Sentiment Analysis, Social Media, User-Generated Content.

1 Introduction

Customer reviews and ratings on the internet are increasingly important in the evaluation of products and services by potential customers. In certain sectors, it is even becoming a fundamental variable in the purchase decision. Consumers tend to trust the opinion of other consumers, especially those with prior experience of a product or service, rather than trust company marketing opinions which are usually business oriented. Given the huge amount of data flowing in the web, it has become necessary to adopt an automatic data analysis strategy. It gives the possibility to understand what people think about a certain product, brand or topic without wasting too much time in exploring User-Generated Contents.

On the other hand, traditional Mass Media still play an important role in the way people get information. Opinion Mining in Media is a pretty new – but already consolidated - field of research. People operating in this sector aims to know *who* is speaking, about *what*, *when* and in *what sense*. **Named Entity Recognition and Classification** (NERC) are important in determining roles (*who, what* and *when*) while **Sentiment Analysis** (SA) is necessary to determine the attitude of a writer with respect to the overall contextual polarity of the text (*what sense*).

C. Stephanidis (Ed.): HCII 2014 Posters, Part II, CCIS 435, pp. 3–8, 2014.

OpeNER has created base technologies for Crosslingual NERC and Sentiment Analysis that are enabling industry users both to implement and contribute to a basic set of core technologies that all require and allow them to focus their efforts on providing tailored and innovative solutions at the rules and analysis levels. OpeNER aims to provide enterprise and society with online services for Crosslingual Named Entity Recognition and Classification and Sentiment Analysis.

In the paper we will present a new multimedia web application, **Ent-It-UP**, developed leveraging on OpeNER Cloud Services[1]. This application is a media monitoring solution for live analytics on User-Generated Contents (UGCs) and video contents.

2 Ent-it-UP Design

Ent-it-UP is an application accessible from the Web that provides users with a clear and effective visualization of the knowledge extracted from two different sources: User-Generated Contents and the transcriptions of videos. In the following sections we describe the necessary steps which will lead from the collected data to their communicative and intuitive visualization through the Ent-it-UP interface.

2.1 Data Harvesting

The first thing that has to be done is to collect the data and store them into a database. The data are taken from two different sources, in order to have the possibility to look at the same thing from two different point of view. In fact, the first source we take our data from are Social Media (such as blogs, forums, Online Travelling Agencies and so on) - which can be taken into account to know *what people think* -, and the second one are international news programs – which can be taken into consideration to know *what news say*. The first dataset needs to be pre-processed in order to delete noise and get clean text. On the other hand the news programs, needs to be processed by the SAVAS Speech Recognition Engine[2] in order to get transcriptions of the recorded videos. The system returns both an XML file and a plain text file. The XML contains information about words' timestamp and will be used to link transcribed text to the video itself. The raw text will be taken as input by OpeNER tools. The same happens to the UGC text previously cleaned.

All the data retrieved so far are stored on a MongoDB management system.

2.2 Data Annotation

The raw text files obtained are processed by the OpeNER Cloud Services which consist of a series of NLP tools, listed below.

- Language Identifier
- Tokenizer

[1] http://opener.olery.com/
[2] http://voiceinteraction.pt

- Tree Tagger
- Part-of-Speech Tagger
- Polarity Tagger
- Property Tagger
- Constituent Parser
- Kaf-Naf Parser
- Named Entity Recognition
- Scorer
- Named Entity Detection
- Opinion Detector

It is possible to use only some of the NLP tools or all of them. Of course, some basic analysis is required to provide implementation of Named Entity Recognition and Sentiment Analysis. This basic analysis can be performed by only two NLP tools, which are the Tokenizer (as far as the language of the text is known, otherwise the Language Identifier is required too) and the Part-of-Speech Tagger.

Thus, in order to implement Ent-it-UP functionalities, these are the four NLP tools that have been used:

- Tokenizer
- Part-of-Speech Tagger
- Named Entity Recognition
- Polarity Tagger

The result is a KAF (Knowledge Annotation Framework) [1][2] file which has an XML-like structure. It consists of several linguistic layers (Figure 1).

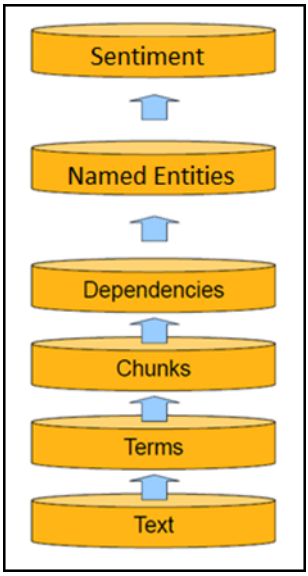

Fig. 1. KAF Layers

The annotated levels of the KAF that will be taken into account from the Ent-it-UP system are the *terms* level (from which it gets the word polarity) and the *named entities* level. These data are also added to the MongoDB database.

2.3 Data Processing

Once the raw texts have been transformed into KAF, they can be elaborated. Some PHP scripts perform queries to the MongoDB collections and return quantitative results such as entity frequency, entity occurrences and other metrics.

2.4 Data Visualization

The above mentioned results have now to be shown. Some of the functionalities offered by Ent-it-UP are the following.

Fig. 2. Ent-it-UP tagcloud

The user has the possibility to explore a general interactive tagcloud of the most frequent entities (Figure 2).

He can also explore an entity-focused report, which can be obtained by searching for a specific entity or choosing one of those shown in the tagcloud. The report includes the occurrences of the entity into the videos and its cross time frequency (Figure 3).

Fig. 3. Ent-it-UP timeline

If the user decide to focus his search on transcriptions of videos he can also explore a video-focused report choosing one specific video among those present in the collection. The report includes statistics about the entities composition (percentage of entities recognized in the video transcription that has been identified as `people`, percentage of entities identified as `organization`, and so on). This report also provides a tagcloud of those entities found in the video. The user can further choose to narrow down the tagcloud selecting the only category of entities he is interested in (Figure 4).

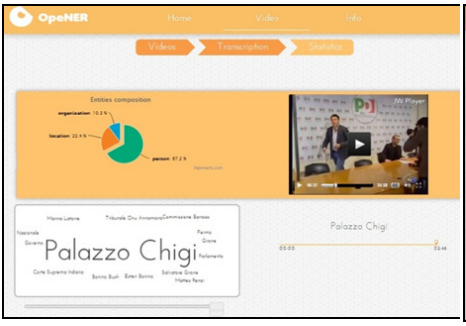

Fig. 4. Ent-it-UP statistics

Fig. 5. Ent-it-UP transcription

The user can also explore the transcription in which entities are marked with different colors according to their type (i.e. entities identified as `people` are colored in orange, entities identified as `location` are colored in green and so on). Terms with polarity (positive, negative or neutral) are also highlighted (respectively green, red and grey). The user can choose to highlight only entities (all types or just some), only sentiment, or both. Next to the transcription there is a player of the video. If the user wants to listen to the point in which a certain word is spoken he can just click that word in the text and the video will jump to that point (Figure 5).

3 Usage Case

In this section is presented a usage case in which both data sources are exploited. In the following case, in fact, UGC and video contents are both useful to the user, who can look at the same thing from two different point of view.
Suppose that the end user is interested in investigating what people think about a certain city. For example he wants to know how Paris is perceived by people. He could be interested in knowing what areas or features are the most mentioned and whether people love them or do not. On the other hand, he could be interested in having an overall insight of the city news events.

The user can access Ent-it-UP, search for the keyword *Paris* using one or the other dataset. In this way, he can get two different kind of information about Paris. In fact, choosing to use the UGC source he would probably get every-day-life information

about Paris (*what people think*). On the other hand, choosing the video source, the user would probably get information about the facts happening in Paris (*what news say*).

4 Conclusions

This paper has presented Ent-it-UP as reference application of the OpeNER project. We have presented how Ent-it-UP monitors Media contents, performing multilingual Named Entity Recognition and Sentiment Analysis on User-Generated Content and video transcriptions. After a short introduction we have described the Ent-it-UP design, identifying the main steps that leads from raw texts to some kind of knowledge. We have reported a usage case in which Ent-it-UP could be used to have an overall insight of a place. However it could be used to discover information also about a certain brand, person, organization and so on.

Ent-it-UP allows the user to focus on other activities rather than spend time analyzing the raw language resources.

Acknowledgment. This work is part of the OpeNER project which is funded by the European Commission 7th Framework Programme (FP7), grant agreement no 296451.

References

1. Tesconi, M., Ronzano, F., Minutoli, S., Aliprandi, C., Marchetti, A.: KAFnotator: a multilingual semantic text annotation tool. In: Proceedings of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, in Conjunction with the Second International Conference on Global Interoperability for Language Resources (ICGL 2010), Hong Kong, January 15-17 (2010)
2. Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., Aliprandi, C.: KAF: a generic semantic annotation format. In: Proceedings of the 5th International Conference on Generative Approaches to the Lexicon GL 2009, Pisa, Italy, September 17-19 (2009)
3. Aliprandi, C., Scudellari, C., Gallucci, I., Piccinini, N., Raffaelli, M.,, A., Álvarez, A., Arzelus, H., Cassaca, R., Luis, T., Neto, J., Mendes, C., Paulo, S., Viveiros, M.: Automatic Live Subtitling: state of the art, expectations and current trends. In: NAB Broadcasting Conference, Las Vegas, Nevada, United States (April 2014) (forthcoming)