Come Hack with OpeNER!

Workshop Programme

9:00 – 9:20 **Introduction by Workshop Chair**

9:20 – 10:30 **Tutorial: OpeNER technology**

10:30 – 11:00 **Coffee break**

11:00 –11:45 **Demo/Posters**

Carlo Aliprandi, Sara Pupi and Giulia di Pietro, *Ent-it-UP: a Sentiment Analysis system based on OpeNER cloud services*

Jordi Atserias, Marieke van Erp, Isa Maks, German Rigau and J. Fernando Sánchez-Rada, *EuroLoveMap: Confronting feelings from News*

Estela Saquete and Sonia Vázquez, *Improving reading comprehension for hearing impared students using Natural Language Processing*

Aitor García Pablos, Montse Cuadros, Seán Gaines and German Rigau, *OpeNER demo: Open Polarity Enhanced Named Entity Recognition*

Andoni Azpeitia, Alexandra Balahur, Montse Cuadros, Antske Fokkens and Ruben Izquierdo Bevia, *The Snowball effect: following opinions on controversial topics*

Stefano Cresci, Andrea D'Errico, Davide Gazzé, Angelica Lo Duca, Andrea Marchetti and Maurizio Tesconi, *Tour-pedia: a Web Application for Sentiment Visualization in Tourism Domain*

12:00 –13:00 **Lunch break**

12:00 – 16:00 **Hackathon**

16:00 – 16:30 **Coffee break**

16:30 – 17:30 **Results presentation**

Editors

Seán Gaines	Vicomtech-IK4
Montse Cuadros	Vicomtech-IK4

Workshop Organizers/Organizing Committee

Rodrigo Agerri	EHU/UPV
Montse Cuadros	Vicomtech-IK4
Francesca Frontini	CNR-ILC
Seán Gaines	Vicomtech-IK4
Ruben Izquierdo	VUA
Wilco van Duinkerken	Olery

Workshop Programme Committee

Carlo Aliprandi	Synthema
Andoni Azpeitia	Vicomtech-IK4
Aitor Garcia-Pablos	Vicomtech-IK4
Angelica Lo Duca	CNR-IIT
Isa Maks	VUA
Andrea Marchetti	CNR-IIT
Monica Monachini	CNR-ILC
German Rigau	EHU/UPV
Piek Vossen	VUA

Table of contents

Ent-it-UP: a Sentiment Analysis system based on OpeNER cloud services ,*Carlo Aliprandi, Sara Pupi and Giulia di Pietro* 1

EuroLoveMap: Confronting feelings from News, *Jordi Atserias, Marieke van Erp, Isa Maks, German Rigau and J. Fernando Sánchez-Rada* 5

Improving reading comprehension for hearing impared students using Natural Language Processing , *Estela Saquete and Sonia Vázquez* 8

OpeNER demo: Open Polarity Enhanced Named Entity Recognition, *Aitor García Pablos, Montse Cuadros, Seán Gaines and German Rigau* 12

The Snowball effect: following opinions on controversial topics, Andoni Azpeitia, Alexandra Balahur, Montse Cuadros, Antske Fokkens and Ruben Izquierdo Bevia, ... 15

Tour-pedia: a Web Application for Sentiment Visualization in Tourism Domain, *Stefano Cresci, Andrea D'Errico, Davide Gazzé, Angelica Lo Duca, Andrea Marchetti and Maurizio Tesconi* 18

Author Index

Atserias, Jordi 5
Aliprandi, Carlo 1
Azpeitia, Andoni 15
Balahur, Alexandra 15
Cresci, Stefano 18
Cuadros, Montse 12,15
Di Pietro, Giulia 1
D'Errico, Andrea 18
Fokkens, Antske 15
Gaines, Seán 12
García-Pablos, Aitor 12
Gazzé, Davide 18
Izquierdo-Beviá, Ruben 15
Lo Duca, Angelica 18
Maks, Isa 5
Marchetti, Andrea 18
Pupi, Sara 1
Rigau, German 5,12
Saquete, Estela 8
Sánchez-Rada, J.Fernando 5
Tesconi, Maurizio 18
Van Erp, Marieke 5
Vázquez, Sonia 8

Preface/Introduction

The OpeNER team is delighted to present a **Tutorial** and a **Hackathon** together in a one-day **workshop** on multilingual Sentiment Analysis and Named Entity Resolution using the OpeNER NLP pipelines as web services in the Cloud.

OpeNER hopes to repeat the success from the July 2013 Amsterdam Hackathon (http://www.opener-project.org/2013/07/18/opener-hackathon-in-amsterdam/) in which a broad spectrum of real end user SMEs, Micro-SMEs, Freelancers and even a few from technology giants, built creative applications using the OpeNER webservices. For examples of the applications built follow the URL provided above.

The proposed workshop will present briefly the project, and all the technology (http://opener-project.github.io/) multilingual NLP tools and resources created within the project. Additionally, it will be a slot for presentations of demos created before the Hackathon and presented in the call for papers.

The workshop will be complemented by a half day Hackathon. The Hackathon will encourage participants to form ad hoc multidisciplinary teams, brainstorm an idea, implement it and present a demo from which a winner will be picked by popular vote. Most of the "core developers" of the OpeNER pipeline technology will be available to help you out and get started.

All participants will be given access to the collateral needed such as NLP tools and resources in six languages beforehand from publicly deployed web services. As of writing the initial versions of the services are publically available at http://opener.olery.com. In order to present a demo or paper to the workshop the only thing that needs to be added is imagination.

Ent-it-UP
A Sentiment Analysis system based on OpeNER Cloud Services

Sara Pupi, Giulia Di Pietro, 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*).

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.

1.1 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* -,

[1] http://opener.olery.com/

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.

1.2 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
- 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

[2] http://voiceinteraction.pt

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).

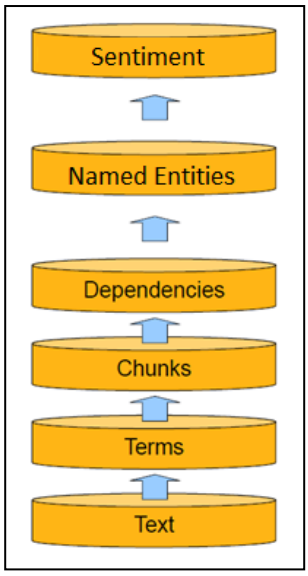

Figure 1. KAF Layers.

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.

1.3 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.

1.4 Data Visualization

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

Figure 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).

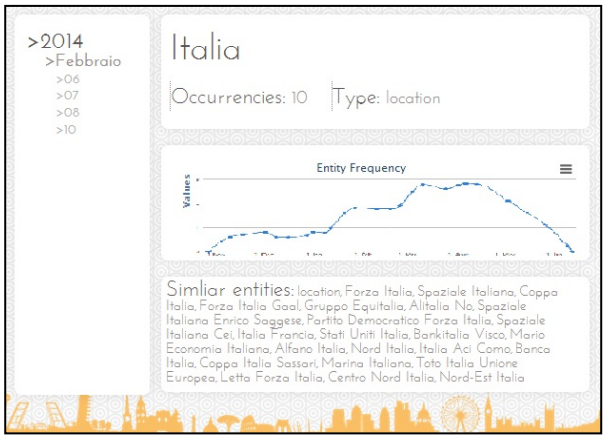

Figure 3. Ent-it-UP timeline

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).

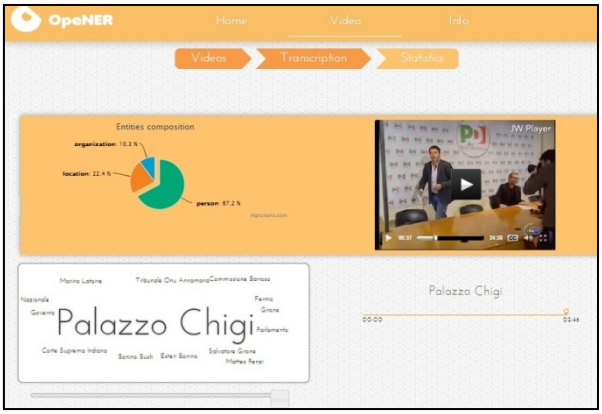

Figure 4. Ent-it-UP statistics

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).

Figure 5. Ent-it-UP transcription

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.

Acknowledgments

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

5. References

[1] Tesconi M., Francesco Ronzano, Salvatore Minutoli, Carlo Aliprandi and Andrea Marchetti: *"KAFnotator: a multilingual semantic text annotation tool":* 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] Wauter Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini and Carlo Aliprandi: *"KAF: a generic semantic annotation format"*, Proceedings of the 5th International Conference on Generative Approaches to the Lexicon GL 2009, Pisa, Italy, September 17-19, 2009.

[3] Carlo Aliprandi, Cristina Scudellari, Isabella Gallucci, Nicola Piccinini, Matteo Raffaelli, Arantza del Pozo, Aitor Álvarez, Haritz Arzelus, Renato Cassaca, Tiago Luis, Joao Neto, Carlos Mendes, Sérgio Paulo, Marcio Viveiros, *"Automatic Live Subtitling: state of the art, expectations and current trends"*, NAB Broadcasting Conference, Las Vegas, Nevada, United States, April 2014 (forthcoming).

EuroLoveMap: Confronting feelings from News

Jordi Atserias[1], Marieke van Erp[2], Isa Maks[2], German Rigau[3], J. Fernando Sánchez-Rada[4]

[1]Yahoo Labs Barcelona, [2]VU University Amsterdam, [3]The University of Basque Country,
[4]Universidad Politécnica de Madrid
jordi@yahoo-inc.com, {marieke.van.erp,e.maks}@vu.nl, german.rigau@ehu.es, jfernando@gsi.dit.upm.es

Abstract

Opinion mining is a natural language analysis task aimed at obtaining the overall sentiment regarding a particular topic. This paper presents a prototype that presents the overall sentiment of a topic based on the geographical distribution of the sources on this topic. The prototype was developed in a single day during the hackathon organised by the OpeNER project in Amsterdam last year. The OpeNER infrastructure was used to process a large set of news articles in four different languages. Using these tools, an overall sentiment analysis was obtained for a set of topics mentioned in the news articles and presented on an interactive worldmap.

Keywords: Opinion Mining, Visualisation, Hackathon

1. Introduction

Different topics are often presented in news from different perspectives. These perspectives may differ between countries and cultures, and are brought to the fore through different communication outlets. We aim to detect these opinions from news articles from different languages to compare the polarity profiles in different countries with respect to a particular topic. Within NLP research, there is a fair body of work on opinion and sentiment analysis (Pang and Lee, 2008; Liu, 2012). Several toolkits have been developed for the detection of polarity in text, but full multilingual opinion detection which includes the holder of the opinion and the target is still lagging. The OpeNER project plans to deliver an opinion detection tool that is trained on an annotated corpus of political news and aims at a sentence-based detection of opinion expressions with their holders and targets. For this demo, however, we use the rule-based opinion tagger that was available in June 2013.

This paper presents a prototype developed in a single day during the June 2013 hackathon organised by the OpeNER project (Agerri et al., 2013)[1] in Amsterdam.[2] OpeNER aims to detect and disambiguate entity mentions and perform sentiment analysis and opinion detection on the texts for six different languages (Maks et al., 2014). Team NAPOLEON used the OpeNER infrastructure[3] and web services[4] to obtain sentiment analyses for news articles in four different languages which were then aggregated into topics per country and presented visually on a map.

In the remainder of this contribution, we detail our system in Section 2., and present some examples in Section 3. We conclude with future work in Section 4.

2. Mining feelings from news using OpeNER

During the hackathon, we processed around 22,000 news articles in four different languages obtained from the RSS service of the European Media Monitor.[5] The content as well as some metadata of the newspaper articles was obtained before the hackathon. For this prototype, we decided to focus on English, Spanish, Italian and Dutch. For instance, the topic *gay marriage* was manually translated to the four languages and news articles relevant to this topic were collected and processed. An overall sentiment score was also obtained per language for each topic. Finally, the aggregated score for every topic-language pair was used for colouring a world map.

During the hackathon, we developed some software modules to process each news article through the OpeNER web services. In the remainder of this section, we detail the different steps in the workflow.

The OpeNER architecture consists of several Natural Language Processing (NLP) components. Each component is configured to take the information it requires to perform a specific analysis. KAF (Bosma et al., 2009) is used as linguistic representation. Each of the NLP processing pipelines is deployed as a Cloud Computing service using Amazon Elastic Computing Cloud[6] (Amazon EC2). Figure 1 presents an overview of the OpeNER components deployed as web services.

At the end of the different natural language processing pipelines, the extracted information is combined to obtain polarity clusters for the different topics selected.

Language Identifier: This component is responsible for detecting the language of an input news article and delivers it to the correct pipeline.

Tokenizer: This component is responsible for tokenising the text on two levels; 1) sentence level and 2) word level. This component is crucial for the rest of NLP components and is the first component in each language processing pipeline.

Part of Speech Tagger: This component is responsible for assigning to each token its morphological label, it also includes the lemmatisation of words. Combining the lemma and morphological label, later modules will consult a sentiment lexicon in order to assign polarity values to the words appearing in the news being processed.

Named Entity Recognition: This module provides Named Entity Recognition (NER) for the six languages covered by

[1] http://www.opener-project.org

[2] http://opener-fp7project.rhcloud.com/2013/07/18/opener-hackathon-in-amsterdam/

[3] http://opener-project.github.io/

[4] http://opener.olery.com/

[5] http://emm.newsbrief.eu/overview.html

[6] http://aws.amazon.com/ec2

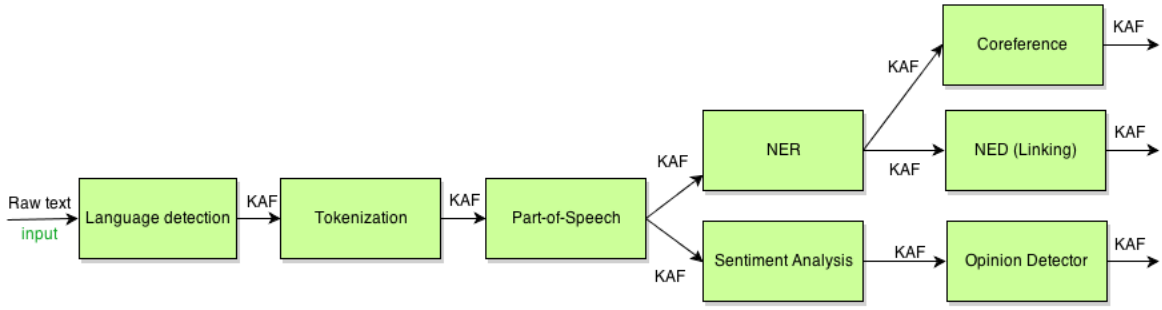

Figure 1: Overview of the components of the OpeNER pipeline

OpeNER and tries to recognize four types of named entities: persons, locations, organisations and names of miscellaneous entities that do not belong to the previous three groups.

Named Entity Linking: Once the named entities are recognised they can be identified or disambiguated with respect to an existing catalogue. This is required because the "surface form" of a Named Entity can actually refer to several different things in the world. Wikipedia has become the de facto standard as named entity catalogue. In OpeNER the NED component is based on the DBpedia Spotlight[7] which uses the DBpedia[8] as the resource for disambiguation entities.

Sentiment Analysis: The Opinion tagger we used is a rule and dictionary based tagger. It detects positive and negative polarity words (such as 'nice' and 'awful'), as well as intensifiers or weakeners (such as 'very' and 'hardly') and polarity shifters (such as 'not'). In addition, the module includes some simple rules that detect the holders and targets of the opinions related to the positive and negative polarity words.

Finally, the processed news in KAF format are stored and indexed using Solr[9] to easily query and retrieve the news articles about a selected topic. A web service was deployed to obtain json results grouping the scores detected by topic and language. The json results are then presented to the user in a world map.

3. Topics on EuroLoveMap

In order to test the prototype we manually selected a small number of topics in English, which were manually translated to Spanish, Italian and Dutch.[10] Table 1 presents the English topics and the corresponding translations in Spanish, Italian and Dutch used in the prototype[11].

Figure 2 presents a screenshot of the EuroLoveMap demo showing the extracted opinions on "gay marriage".

[7] http://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki

[8] http://dbpedia.org

[9] https://lucene.apache.org/solr/

[10] To scope the prototype, we decided to focus only on four out of the six project languages.

[11] The resulting demo can be found at http://eurolovemap.herokuapp.com/.

4. Future Work

As this is only a very first prototype built in a few hours during the previous OpeNER hackathon, there are several different avenues of research as well as engineering issues that spring from it.

To make the prototype more informative and useful for users interested in analysing trending opitions, possible extension to the prototype could be a trend line or the option to look at different snapshots of the EuroLoveMap. This could provide insights into how the opinions on the different topics evolve in different countries.

For selecting the news sources, we currently use language identification, but one preferably uses the publisher information as there may be news sources aimed at expats in languages different from the country's main language. This would not only be more precise, but also give us access to a host of background information about these sources that can be mined in order to obtain more fine-grained information. Different publishers can for example be classified as more left or right leaning. Having this information enables us to present a more fine-grained analysis of the different perspectives within a country. Information about the publisher or authors of the articles could be further mined to create authority and trust profiles using PROV-O(Moreau et al., 2012). Being able to bring up the actual text of the mined articles would make the EuroLoveMap a useful tool to for example communication scientists or anthropologists.

For this prototype, we manually selected the topics and translated them. Ideally, a system picks up on trending topics, for example by plugging into the European Media Monitor or Twitter trends and detecting which topics would be interesting to analyse. To translate these topics automatically one could imagine using DBpedia or a similar resource.

As processing the articles via the NLP pipelines is a time-consuming process, we are currently working with a static dump of processed articles. Research in for example the NewsReader[12] architecture is underway to optimise NLP pipelines further, but until then the most viable option for updating the demo would be with daily batches that are processed overnight.

[12] http://www.newsreader-project.eu

English	Spanish	Italian	Dutch
Berlusconi	Berlusconi	Berlusconi	Berlusconi
Boston	Boston	Boston	Boston
North Korea	Corea del Norte	Corea del Nord	Noord-Korea
Obama	Obama	Obama	Obama
Putin	Putin	Putin	Poetin
CIA	CIA	CIA	CIA
Snowden	Snowden	Snowden	Snowden
Spain	España	Spagna	Spanje
United States, US	Estados Unidos, E.E.U.U.	Stati Uniti	Verenigde Staten van Amerika, VS
Netherlands	Holanda	Olanda	Nederland, Holland
Italy	Italia	Italia	Italië
Germany	Alemania	Germania	Duitsland
Gay marriage, homosexual marriage	matrimonio homosexual, matrimonio gay	matrimonio gay	homohuwelijk

Table 1: Topics and translations

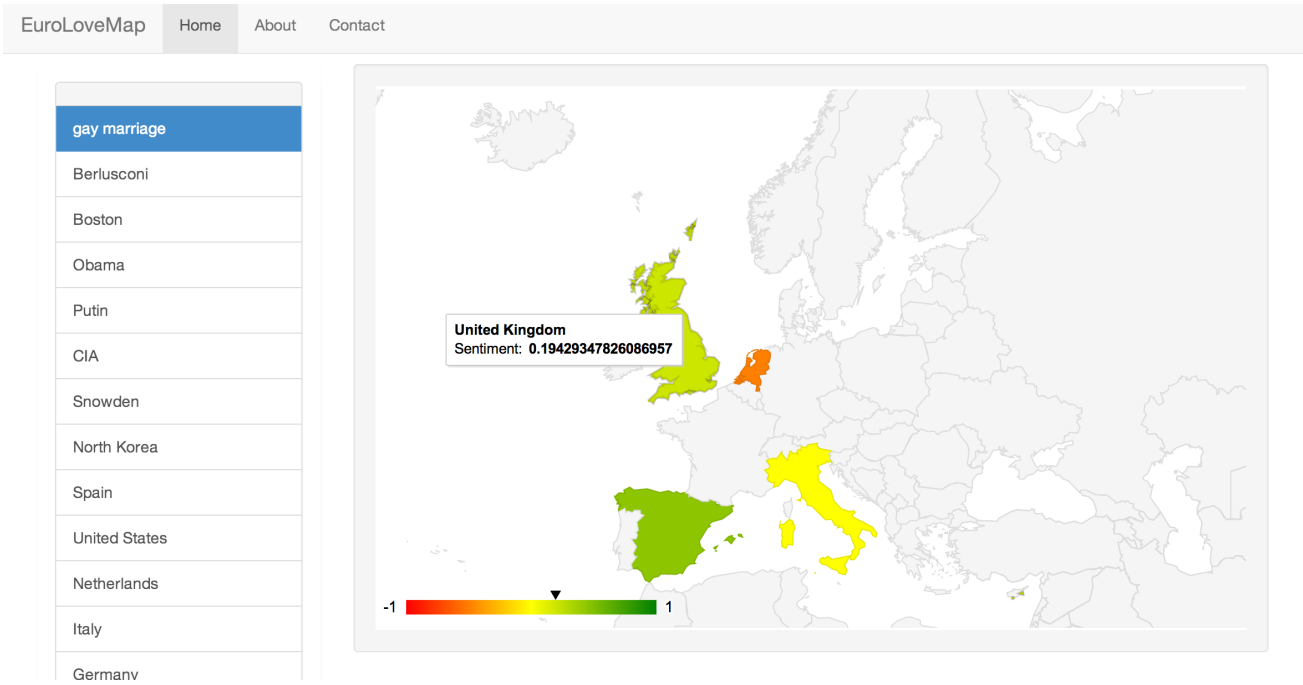

Figure 2: Screenshot of the EuroLoveMap demo showing the extracted opinions on "gay marriage"

Acknowledgements

This research is supported by the European Union's 7th Framework Programme via the OpeNER project (ICT 296541) and the NewsReader Project (ICT-316404).

5. References

Agerri, R., Cuadros, M., Gaines, S., and Rigau, G. (2013). Opener: open polarity enhanced named entity recognition. Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN'2013).

Bosma, Wauter, Vossen, Piek, Soroa, Aitor, Rigau, German, Tesconi, Maurizio, Marchetti, Andrea, Monachini, Monica, and Aliprandi, Carlo. (2009). Kaf: a generic semantic annotation format. In *Proceedings of the GL2009 Workshop on Semantic Annotation*.

Liu, Bing. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Maks, Isa, Izquierdo, Ruben, Frontini, Francesca, Azpeitia, Andoni, Agerri, Rodrigo, and Vossen, Piek. (2014). Generating polarity lexicons with wordnet propagation in 5 languages. In *In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, May.

Moreau, Luc, Missier, Paolo, Belhajjame, Khalid, B'Far, Reza, Cheney, James, Coppens, Sam, Cresswell, Stephen, Gil, Yolanda, Groth, Paul, Klyne, Graham, Lebo, Timothy, McCusker, Jim, Miles, Simon, Myers, James, Sahoo, Satya, and Tilmes, Curt. (2012). PROV-DM: The PROV Data Model. Technical report.

Pang, Bo and Lee, Lilian. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2).

Improving reading comprehension for hearing-impared students using Natural Language Processing

E. Saquete, S. Vázquez
University of Alicante
University of Alicante
stela@dlsi.ua.es, svazquez@dlsi.ua.es

Abstract

The main objective of this paper is developing a system, known as SimplexEduReading, capable of transforming educational texts in Spanish into very easy reading texts by using Natural Language Processing (NLP). This system is mainly oriented to support people with problems in reading comprehension, for instance, deaf people. The process of simplification and enrichment of texts consists of the automatically detection of linguistic features of input texts and: a) the reduction and removal of obstacles, but preserving in all cases the original meaning of the text, and b) the enrichment of texts using different open tools. To achieve our aim, at this point different NLP tools will be used: 1) OpeNER in order to detect name entities. Once detected, our system will provide extra information about them, for example, related images, information from Wikipedia or synonyms; 2) TERSEO in order to detect and resolve temporal expressions giving also a chronological timeline of the events; and 3) Wiktionary in order to clarify complex words. As further work Wordnet Domains will be used in order to detect the main domains of the text and therefore providing a context of the text to the reader.

Keywords: human language technologies, natural language processing, Name Entity Recognizer

1. Introduction

Reading comprehension, according to the PISA 2000 report, is defined as "the ability to understand, using and thinking about information from written texts, with the aim of achieving personal goals, developing the knowledge and the personal potential, and taking efficient part in the society.". For this reason, the reading comprehension has currently become one of the main important research issues in the psychology and education field. Previously research works have determined that the skills and conditions required to a properly reading comprehension are many and very complex. Moreover, the problem of reading comprehension becomes even higher for deaf people.

The main objective of this paper is developing a system (SimplexEduReading) capable of transforming educational texts in Spanish into very easy reading texts by using different Natural Language Processing (NLP) techniques. More specifically, in this paper we are integrating OpeNER [1] web services in order to determine the Name Entities of the text in Spanish.

2. Previous work

As published by the PISA report in 2006, the barely ability in reading comprehension is an increasing problem in this society, and it is higher for hearing impared people. This problem has been studied during years, not only from a lexical perspective, but also from a syntactic one (King and Quigley, 1985) (Berent, 1996) (LaSasso and Davey, 1987) (Paul and Gustafson, 1991). Previous studies have detected the following linguistic barriers that hearing impaired people find in reading comprehension:

1. Ambiguity problems: There are a lot of words that can have multiple meanings, and have a different sense depending on the context in which they appear. All of these polysemic words can lead to multiple problems for reading comprehension;

2. Limited vocabulary: This type of readers focuses fundamentally on common words, using very specific nouns and familiar verbs. Most of the times they have problems recognizing name entities and contextualizing them;

3. Complex sentences: Difficulties in the interpretation of complex syntactic structures, that are different from the basic syntactic structures like noun-verb-noun and subject-verb-object. Therefore, more complex structures like transitive active sentences, passive sentences or subordination increase the problem in reading comprehension;

4. Temporal reasoning: Difficulties in locating events in the temporal timeline: normally a text has different temporal points, and goes temporally backward and forward, implying temporal signals and temporal expressions interpretation for a full comprehension.

Previous works are focused in English problems with reading comprehension, and only some of them are focused in Spanish, not only at a lexical level (Mies, 1992), but also at a syntactic level (Stockseth, 2002). Besides, there are also some works in the educational field (Alegría and Leybaert, 1985) (Asensio and Carretero, 1989) (Mora, 1989). Within the Natural Language Processing research area, there are several works related to the extraction of sign languages from written and spoken texts with automatic and semi-automatic approaches (Parton, 2005) (Wu et al., 2004) (Duchnowsku et al., 2000). Furthermore, we want to stand out the MAS project (Manchón, 2001), which aim is to check the effects of using a multimedia tool to improve the reading comprehension using sign languages.

[1] http://opener.olery.com

3. Architecture of the system

The transformation perform by SimplexEduReading system implies:

- the automatic detection of specific linguistic features of the input texts that may interfere in the reading comprehension together with the automatic decrease and/or removal of these barriers, taking into account that the original meaning of the text has to be preserved
- the enrichment of these texts using resources like Simple Wikipedia[2], Wiktionary[3] or Google Images[4].

Natural Language Processing techniques will be applied to locate and eliminate these barriers, transforming them into much simpler elements or enriching the elements with additional simple information, thus facilitating the reading comprehension process. Such language barriers are derived from complex structures, ambiguity in terms, lack of context and problems with timeline, so the tool would generate supporting material by means of images, definitions of proper nouns extracted from online encyclopedias, timelines and resolution of temporal expressions to concrete dates and setting the context and topics of the original texts. The system integrates different NLP tools and open resources in order to transform the texts into simpler ones and enrich them with additional information.
We distinguish in the system three main parts:

- The lexical transformations, that include: 1) the recognition of Name Entities and the enrichment of them using Wikipedia and images, and 2) the detection of complex words whose definition will be provided.
- The syntactic transformations, that affected the transformation of complex syntactic structures to simpler ones.
- The semantic transformations, that include: 1) the temporal expressions recognition and resolution, and the graphical timeline representation, and 2) the detection of the context and main topics of the text.

Each implemented module will be shown in depth in next subsections. At this moment, the prototype system has been developed for Spanish. The interface is very simple (Figure 1). The user can upload a document and the system requires that the date of the input document is indicated in order to resolve the temporal expressions appearing in the text.
The modules that are part of the transformation of the text at this moment are:

1. Name Entity Recognition and Enrichment: The system is able to recognize this type of entities using OpeNER web services, providing the definition of the entity from an online encyclopaedia and a set of images related to the entity.

Figure 1: Tool Interface

2. Temporal Expressions Recognition and Resolution: The system, by means of TERSEO system, is able to recognize temporal entities and provide the exact date or period of dates that the expression is referring to.
3. Complex Words Recognition and Resolution: At this point, words are not marked as complex, but the user can get an online definition for all the words in the text if required.

These modules are explained in depth in following subsections.

3.1. Name Entity Recognition and Enrichment

In order to recognize and resolve Name Entities in the text, the web services provided by the OpeNER project[5] have been used. OpeNER is a project funded by the European Commission under the FP7 (7th Framework Program). Its acronym stands for Open Polarity Enhanced Name Entity Recognition. It is a two year duration project which officially started at July 2012, and finishes at July 2014. In OpeNER are collaborating partners from Italy, Holland and Spain.
OpeNER's main goal is to provide a set of ready to use tools to perform some natural language processing tasks, free and easy to adapt for SMEs to integrate them in their workflow. The OpeNER web services are available at http://opener.olery.com/. Specifically, in this system we have used three of them: a) the tokenizer[6]; b) the POS-tagger[7]; and c)the name entity recognizer[8].
Once the Name Entities are recognized, the system will mark them in blue colour font and it will allow the user to click them, if he or she wants to obtain additional information related to the Named Entity. For instance, Figure 2 shows an example of a text, where the Named Entities "Max Weber" and "Europa" are recognized.
As shown in Figure 2, by clicking the entity, the system opens a pop-up window, where the entity is explained. This

[2] http://simple.wikipedia.org
[3] http://es.wiktionary.org
[4] http://images.google.es
[5] http://www.opener-project.org/
[6] http://opener.olery.com/tokenizer
[7] http://opener.olery.com/pos-tagger
[8] http://opener.olery.com/ner

Figure 2: NE in the tool

information is automatically obtained from Wikipedia[9]. By clicking the button next to the entity, another pop-up window appears with a set of images related to the entity and automatically extracted from Google Images[10].

3.2. Temporal Expressions Recognition and Resolution

Temporal entities in the text are automatically recognized and resolved using a tool called TERSEO (Saquete et al., 2005). TERSEO system, whose architecture is shown in Figure 3 is a tool that performs the recognition and resolution of temporal expression in texts using a knowledge database that was manually created for Spanish and it was automatically extended to other languages like English and Italian

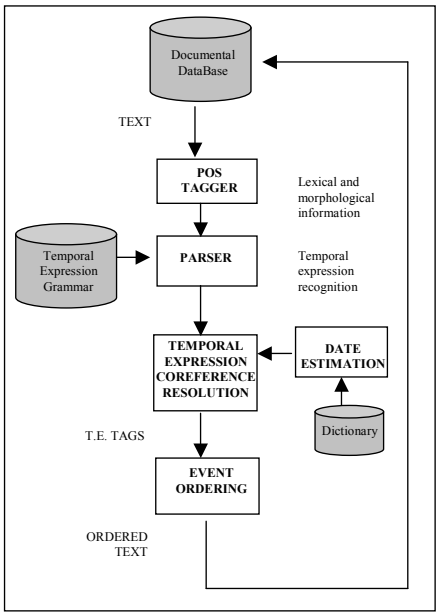

Figure 3: TERSEO system architecture

Given an input text, the system analyzes the text with a Postagger. A Postagger is a tagging program whose labels indicate a word's part of speech. With this information and a temporal expression grammar is able to recognize temporal expressions. After that, the expressions are resolved using the information stored in a resolution rule database. Finally, using the specific dates and periods obtained, the events are ordered in a timeline sequence. Our proposed system integrates TERSEO, indicating in a red colour link when a temporal expression appears in the text. If the user clicks on the expression, a pop-up window appears with the exact date or period of dates that the expression is referring to.

Figure 4: Temporal expressions in the tool

For instance, as shown in the example depicted in Figure 4, the temporal expression "1984" is found by calling to TERSEO system. By clicking in the temporal expression,a pop-up window is obtained with the exact date or period of dates the expression is referring to.

3.3. Complex Words Recognition and Resolution

At this moment, due to the fact that complex words are not being detected, all the words in the text could be clicked in order to obtain a definition from an online dictionary. In our system, we used Wiktionary.org[11]. This tool is a collaborative project to produce a free-content multilingual dictionary. It aims to describe all words of all languages using definitions and descriptions. The information is presented following the same procedure as for the other types of entities (pop-up window).
In the example shown in Figure 5 the word "Carisma" (Charisma) has been clicked and the dictionary was automatically invoked, presenting the different definitions of the word.

4. Conclusions

Therefore, the main objective of this paper is designing and developing a system, called SimplexEduReading, that is able to transform Spanish Educational texts into texts that are easier to understand for hearing impaired people and enrich them with extra information and context. Different

[9] http://es.wikipedia.org
[10] http://images.google.es
[11] http://es.wiktionary.org/

Figure 5: Common entities in the tool

Natural Language Processing tools and techniques are used to detect the potential linguistic barriers. As mentioned before, such language barriers are derived from complex structures, ambiguity in terms, lack of context and problems with timeline, so the system would simplify complex text and generate supporting material by means of images, definitions of proper nouns extracted from online encyclopaedias, timelines and resolution of temporal expressions to concrete dates, definitions of common names with online dictionaries, or the main topics of the original text.

The existing technological tools in this field are mainly oriented to help reading comprehension using the sign language. However, our proposal, aims at helping users by simplifying and enriching the text, preserving also its meaning. In this manner, not only the lexical comprehension is facilitated, but also the syntactic and semantic comprehension of the text. At this moment, the system is able to provide the following supporting material by a given input educational text:1) For temporal entities: The system, by means of TERSEO system, is able to recognize temporal entities and provides the exact date or period of dates that the expression refers to; 2) For Named Entities or Proper Nouns: The system, by means of OpeNER web services, is able to recognize this type of entities, providing the definition of the entity from an online encyclopaedia and a set of images related to the entity; and 3) For all the common entities in the text, due to the fact that some of them could be difficult to understand, the system allows the user to directly obtain a definition from an online dictionary.

As further work, the system is able to detect the context of the text by means of Wordnet and Wordnet Domains. Besides, complex syntactic structures would be divided into simpler ones. It is important to emphasize that this type of tool is very helpful and appropriate in the educational area, especially for deaf students or people who are learning new languages. Nowadays, teachers and professionals are helping these people by manually performing this simplification task in order to make easier the reading comprehension for these students.

5. Acknowledgements

This paper has been supported by the University of Alicante with the project GRE1121 and by the Spanish government, Ministerio de Economía y Competitividad con número de referencia TIN201231224.

6. References

Alegría, J. and Leybaert, J. (1985). Adquisición de la lectura en el niño sordo: un enfoque psicolingüístico. *Investigación y Logopedia*.

Asensio, M. and Carretero, M. (1989). La lectura en los niños sordos. *Cuadernos de pedagogía*, 174.

Berent, G. (1996). The acquisition of English Syntax by Deaf Learners. *Handbook of Second Language Acquisition*, pages 469–506.

Duchnowsku, P., Lum, D., Krause, J., Sexton, M., Bratakos, M., and Braida, L. (2000). Development of speechreading supplements based on automatic speech recognition. *IEEE Trans Biomed Eng.*, 47(4):487–496.

King, C. and Quigley, S. (1985). Reading and Deafness. *Collegue Hill Press*.

LaSasso, C. and Davey, B. (1987). The relationship between lexical knowledge and reading comprehension for prelingually, profoundly hearing impaired students. *The Volta Review*, 89:211–220.

Manchón, A. F. (2001). La comprensión lectora en personas sordas adultas y el acceso a la Universidad. *ISAAC 2001: Odisea de la Comunicación. Segundas Jornadas sobre comunicación Aumentativa y Alternativa*.

Mies, B. (1992). El léxico en la comprensión lectora: Estudio de un grupo de alumnos sordos adolescentes. *Rev. Logop., Fon., Audiol.*, XII.

Mora, J. A. F. (1989). La lectura en el corriculum escolar del niño sordo. *Rev. Logop Fon Audiol*, IX.

Parton, B. (2005). Sign language recognition and translation: a multidisciplined approach from the field of the artificial intelligence. *J. Deaf Stud Deaf Educ.*

Paul, P. and Gustafson, G. (1991). Hearing-impaired students' comprenhension of hig-frequency multi-meaning words. *Remedial and special education*, 12:52–62.

Saquete, E., Muñoz-Guillena, R., and Martínez-Barco, P. (2005). Event ordering using TERSEO system. *Data and knowledge Engineering Journal*, 58.

Stockseth, D. R. (2002). Comprensión de la sintaxis española por lectores sordos chilenos. *Revista Signos*, 35.

Wu, C., Chiu, Y., and Guo, C. (2004). Text generation from Taiwanese Sign Language using PST-based language model for augmentative communication. *IEEE Trans Neural Syst. Rehabil Eng.*, 12(4).

OpeNER demo: Open Polarity Enhanced Named Entity Recognition

Aitor García-Pablos, Montse Cuadros, Seán Gaines
Vicomtech-IK4 research centre
Mikeletegi 57, San Sebastian, Spain
{agarciap, mcuadros,sgaines}@vicomtech.org

German Rigau
IXA Group
Euskal Herriko Unibertsitatea,
San Sebastian, Spain
{german.rigau}@ehu.es

Abstract

OpeNER is a project funded by the European Commission under the 7th Framework Programme. Its acronym means Open Polarity Enhanced Named Entity Recognition. OpeNER main goal is to provide a set of open and ready to use tools to perform some NLP tasks in six languages: English, Spanish, Italian, Dutch, German and French. In order to display these OpeNER analysis output in a format suitable for a non-expert human reader we have developed a Web application to display this content in different ways. This Web application should serve as a demonstration of some of the OpeNER modules capabilities.

Keywords: DEMO, OpeNER, named entity recognition, named entity linking, sentiment analysis

1. Introduction

OpeNER[1] is a project funded by the European Commission under the FP7 (7th Framework Programme, project reference: 296451). Its acronym means Open Polarity Enhanced Named Entity Recognition. It is a two year duration project which officially started at July 2012, and finishes at July 2014. In OpeNER are collaborating six partners; two from Italy, two from Holland and two from Spain.

OpeNER main goal is to provide a set of open and ready to use tools to perform some NLP tasks in six languages (English, Spanish, Italian, Dutch, German and French), free and easy to adapt for SMEs to integrate them in their workflow. Some of these NLP task are Named Entity Recognition and Classification, Named Entity Linking and Sentiment Analysis.

The OpeNER project analysis tools produce their output in XML following the KAF[2] format, acronym of Knowledge Annotation Framework (Bosma, et al., 2009). The KAF documents generated by the OpeNER NLP modules are very rich and complex XML files. Obviously, this makes difficult for a human reader to interpret the results by reading the KAF document without any further post-processing.

In order to display these KAF documents in a format suitable for a non-expert human reader we have developed a Web application capable of displaying the content of a KAF document in different ways. This Web application should serve as a demonstration of some of the OpeNER modules capabilities and to explore the results stored in KAF format.

[1] http://www.opener-project.org/

[2] https://github.com/opener-project/kaf/wiki/KAF-structure-overview

2. Visualization of KAF documents

KAF documents generated by OpeNER analysis modules are XML files with a complex structure. KAF documents are structured in different layers, each one corresponding to a different NLP task (e.g. "Text" layer for token related information, "Terms" layer for "term" related information like Part-of-Speech, lemma and polarity, "Entities" layer for Named Entity Recognition, etc.). In this format they are difficult to read and interpret by a human reader.

```
   […]
<text>
    <wf wid="w1" sent="1" para="1" offset="0"
        length="3">The</wf>
    <wf wid="w2" sent="1" para="1" offset="4"
        length="4">city</wf>
  </text>
  <terms>
    <!--The-->
    <term tid="t1" type="close" lemma="the"
          pos="D" morphofeat="DT">
            <span>
              <target id="w1" />
            </span>
    </term>
    <!--city-->
    <term tid="t2" type="open" lemma="city"
          pos="N" morphofeat="NN">
            <span>
              <target id="w2" />
            </span>
    </term>
     </terms>
   […]
```

Figure 1. A snippet of a KAF document

Among other information contained in a KAF document, there are some aspects that should be displayed in a way suitable for a human. These aspects are:

Detected Named Entities: inside a KAF document the detected named entities are stored in their own layer, with a pointer to the span of terms that hold the mention in the text. At the same time, each of these terms point to their respective token ("wf" element in the "text" layer) in the KAF document. To visualize this information the text must be reconstructed from these tokens adding an appropriated markup to the tokens that are inside a named entity span. This markup should allow displaying the text in a Web browser with the words belonging to a named entity highlighted (for example, rendering them with a different color) or to enrich the words adding an hyperlink or a tooltip.

Named entity types: KAF also contains information about the type of detected entities (e.g. person, organization, location). This information can be used to display the different entity types highlighted in different ways (for example, a color code for each entity type).

```
<entity eid="e24" type="person">
      <references>
        <!--Chris Stevens-->
        <span>
          <target id="t322" />
          <target id="t323" />
        </span>
      </references>
 </entity>
```

Figure 2. An example of a "entity" element from the "entities" layer in a KAF document

Linked entities: if a named entity is disambiguated and linked to an external knowledge-base (e.g. to DBpedia) KAF holds the linked resource URI. This information can be used to add hyperlinks to the linked resource when a linked named entity mention occurs, or use the linked resource attributes to gather and display further information about the entities.

```
<entity eid="e2" type="person">
      <references>
        <!--Obama-->
        <span>
          <target id="t15" />
        </span>
      </references>
      <externalReferences>
        <externalRef resource="spotlight_v1"
reference="http://dbpedia.org/resource/Barack_Obama"
/>
      </externalReferences>
 </entity>
```

Figure 3. An example of linked entity, a "entity" element with an "external reference" pointing to DBpedia

Polarity words: KAF contains information about the detected polarity of the words in a text. This information is inside the “terms” layer, which contain information and "term" level, like the Part-of-Speech tag or the lemma of the token pointed by a particular term. and can be used to highlight positive and negative words in a text. Other kind of words relevant to the sentiment, like the polarity shifters (i.e. words that reverse the polarity of other words, like "not") or the polarity intensifiers (i.e. words that intensify the polarity of other words, like "very") may also be highlighted.

```
<!--thanking-->
   <term tid="t116" type="open" lemma="thank"
         pos="V" morphofeat="VBG">
     <sentiment
          resource="VUA_olery_lexicon_en_lmf"
          polarity="positive" />
     <span>
       <target id="w116" />
     </span>
   </term>
```

Figure 4. A snippet of KAF showing the polarity annotation inside a "term" element

3. A Web demo for OpeNER

The main approach to display all this information in a more human-readable way is to reconstruct the original text from the KAF document, token by token, highlighting relevant information via HTML markup. Each token is checked against different KAF layers to find out if it belongs to a relevant category (i.e. a named entity, a positive or negative word, etc.). The final result of this highlighted reconstruction is displayed to the user, which can then read the full text and asses if the highlighted words are correctly annotated or not.

The Web application has been organized in three different top level menus, which can be selected clicking in the corresponding button in the top bar:

NERC/Sentiment demo contains a set of news texts for each of the six languages handled by OpeNER. Each text has been preprocessed with the OpeNER NLP tools, and the resulting KAF has been stored in a database. These KAF documents are employed to show different aspects of the analysis. The analysis results for each text are shown in a tabbed pane in the right side of the screen every time a text is selected. Each tab of this pane shows a different view of the selected text. There is a tab for the original text with no markup, a NERC tab with the Named Entities highlighted, a Sentiment tab with the polarities of the words highlighted and a KAF tab with the generated KAF document. In addition to these tabs, there are two tabs exploiting the links to DBpedia attached to some of the detected Named Entities. The first one, "Images", shows the thumbnails of the linked entities. The second one, "Map", shows the detected locations emplacement using the geo-coordinates from their respective linked DBpedia resource and Google Maps. If there are no linked entities in the selected text, or if DBpedia contains no thumbnail or geo-coordinates for the linked resources, these tabs will be empty.

Figure 5. A screenshot from the tabbed pane after selecting a text, showing highlighted words

EMM comparison demo is a named entity recognition comparison between OpeNER and the JRC European Media Monitor News Explorer. The EMM News Explorer gathers and analyzes news from a many sources and in a lot of languages. The EMM News Explorer Web site contains information about the different entities detected in the texts of the news inside each cluster. Some of those entities are disambiguated and linked to their respective Wikipedia page. As OpeNER can do a very similar task, it is a good chance to try to compare OpeNER with another working system. This view shows two list of detected entities for a same set of news texts. One list contains the entities detected in the EMM and the other list contains the entities detected in the OpeNER analysis. The entities that appear in the both lists are highlighted.

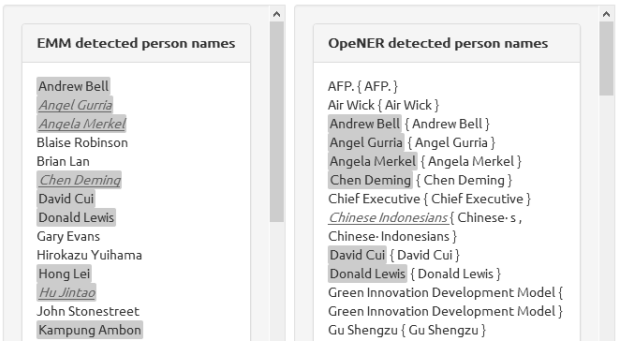

Figure 6. Screenshot showing the comparision between EMM and OpeNER detected entities

Custom text demo allows writing or copying a piece of text and launching the OpeNER NLP analysis on it. In this case the analysis is done in calling the OpeNER web services in real time, and the results is displayed immediately after the end of the analysis chain. The language can be selected manually or it can be detected by the OpeNER language identifier module. Some of the analysis steps can be selected or deselected using the corresponding check box. The result is displayed in the same kind of tabbed pane than in the "NERC/Sentiment demo" (see Figure 5), with the different analyzed information contained in KAF highlighted in the text.

The prototype for the demo[3] has been designed as a Web application using Java technologies like Apache Struts2[4] and Spring[5] Framework, and other Web programming related technologies like the Javascript JQuery[6] library. Being a Web application means that it can be accessed from a Web browser with no other requirement for an end user. All the OpeNER NLP modules are accessible as web services deployed by the OpeNER consortium[7]. These web services are accessed from the demo to perform the custom text analysis.

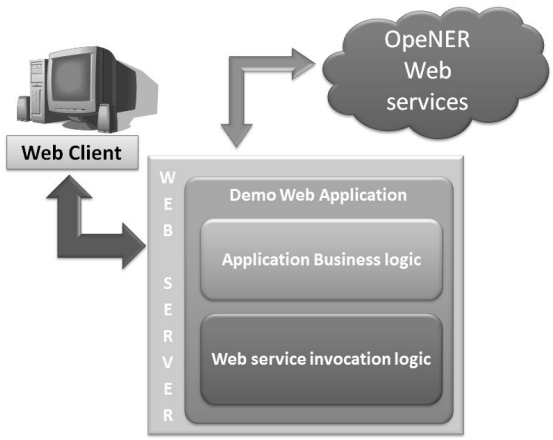

Figure 7. Basic architecture overview

4. Conclussions

In this paper we present a demo prototype that allows reading the result of the OpeNER analysis more easily for human readers. KAF represents a lot of information, but it becomes difficult to interpret by humans. Such a demo could be useful to explain how the different OpeNER modules work, for dissemination or teaching.

Acknowledgements

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

References

Rodrigo Agerri, Montse Cuadros, Seán Gaines and German Rigau (2013). OpeNER: Open Polarity Enhanced Named Entity Recognition, Sociedad Española para el Procesamiento del Lenguaje Natural,Volume 51

Bosma, W., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, Aliprandi, C. (2009). KAF: a generic semantic annotation format. In Proceedings of the GL2009 Workshop on Semantic Annotation.

[3] The prototype can be accessed from http://demo2-opener.rhcloud.com

[4] http://struts.apache.org/development/2.x/

[5] http://spring.io/

[6] http://jquery.com/

[7] The services can be accessed from http://opener.olery.com

The Snowball effect: following opinions on controversial topics

Andoni Azpeitia
Vicomtech-IK4
Donostia-San Sebastian, Spain
aazpeitia@vicomtech.org

Alexandra Balahur
European Commission Joint Research Centre
Ispra, Italy
alexandra.balahur@jrc.ec.europa.eu

Montse Cuadros
Vicomtech-IK4
Donostia-San Sebastian, Spain
mcuadros@vicomtech.org

Antske Fokkens
VU University
Amsterdam, Netherlands
antske.fokkens@vu.nl

Ruben Izquierdo
VU University
Amsterdam, Netherlands
ruben.izquierdo@vu.nl

Abstract

This paper describes a practical application of technology developed as part of the OpeNER project. This technology finds trending topics in different media sources and different languages. We use a rule-based opinion mining tool to analyze the global scandal on leaking data involving Edward Snowden. Results show a diversity of opinions depending on the language and the sources that were analyzed. Additionally, we found an interesting division between the opinions expressed in favor of Snowden's actions and in favor of the United States' reaction towards them.

1 Introduction

Since the emergence of Social Media sources and the global interest to know what people think about a specific topic, the field of opinion mining has become one of the most business-interesting areas in Natural Language Processing. This field studies the sentiment and opinions about objects, companies and events and is popular in fields such as Brand Monitoring and Social Opinion (Pang and Lee, 2008; Liu, 2012). For instance, a large number of companies build applications to help customers find out what the market thinks about them, what people think about their immediate competitors or simply to help them follow the news or social events that are of their concern. However, easy-to-use resources for cases where opinions come from multilingual sources are difficult to find. Most open-source and ready-to-use tools only support English. OpeNER(Agerri et al., 2013)[1] is a project funded by the European Commission under the FP7 (7th Framework Program). Its acronym stands for Open Polarity Enhanced Name Entity Recognition. It is a two year project which officially started in July 2012, and finishes in July 2014. OpeNER's main goal is to provide a set of ready-to-use tools to perform natural language processing tasks, free and easy to adapt and integrate in workflows of SMEs. More precisely, OpeNER aims to detect and disambiguate entity mentions and perform sentiment analysis and opinion detection on texts, for example, to be able to extract the sentiment and the opinion of customers about certain resources (e.g. hotels and accommodations) in Web reviews. OpeNER supports six different languages. We present a prototype that was developed during an OpeNER hackathon that uses OpeNER tools. The aim of this exercise was to explore what could be done in four hours (in terms of extracting multilingual opinions of a specific topic) of drafting and programming using these tools. In this paper, we describe a system that gathers and analyzes opinions on a trending topic (we took the "Snowden case" as an example, which was a very "hot" topic at the time) in different languages from different perspectives. This paper is organized as follows: Section 2 presents the methodology performed in the project, Section 3 shows the results of the opinions and Section 4 outlines the general conclusions obtained from the project.

2 Methodology

The project presented in this paper was divided in several blocks in order to make it feasible to implement in four hours. First, three main tasks were taken into account: acquiring datasets, data processing and visualization. These tasks formed the basic preparation blocks which were carried out in

[1] http://www.opener-project.eu

Language	Articles	Tokens
German	13	194
English	123	1903
Spanish	31	497
French	7	90
Italian	4	67
Dutch	7	122

Table 1: Number of articles and tokens per language

parallel by different members of the team cooperating and organizing themselves.

2.1 Data acquisition

Regarding the acquisition of datasets, we employed the RSS feeds provided by the Europe Media Monitor[2] site. The output RSS and Twitter were scraped in order to find all possible news and opinions in several languages related to the "snowball effect" created by Snowden's leaking of confidential news. News and Tweets containing the words "Snowden" in English, German, Dutch, Italian, Spanish and French were filtered. Table 2.1 shows the number of articles and total number of tokens scraped for each language. For some languages such as English, the amount of online data was bigger resulting in a higher number of articles related to Snowden.

2.2 Data processing

The datasets were obtained and stored in raw text format and subsequently processed using the available webservices from OpeNER.[3] OpeNER uses as standard input and output codification between the tools, an XML based format called KAF(Bosma et al., 2009).[4] The datasets were processed through a pipeline of tools to extract the opinions using the following tools in this particular order:

- Language Detector:[5] This component detects the language that predominates in the document.
- Tokenizer:[6] This component splits the words in the document in order to segment the content units from the punctuation marks.
- POS-tagger:[7] This component detects the morphological category of each word.
- Named Entity Recognizer:[8] This component detects the different entities in the document and categorizes them.
- Opinion mining: This component detects opinions expressed in the documents. There are two variations:
  - Ruled-based for Spanish, French, Italian and German.[9]
  - Machine Learning-based for English and Dutch.[10]

The opinion mining module returns single opinions found in the news, indicating what was the real opinion (the expression), what was this expression about (the opinion target), and who stated it (the opinion holder). The output of the analysis was transformed to JSON.[11] We aggregated single opinions for each of the languages in order to obtain an overall estimation of the amount of positive and negative opinions about Snowden that were found in each language. Furthermore, we collected opinions about the NSA and CIA. Because the NSA and CIA were the "opposing parties" in this conflict with Snowden, negative opinions about NSA and CIA contributed to positive opinions about Snowden and positive opinions on NSA and CIA increased the score of negative opinions about Snowden. Note that it is likely that most opinions about the CIA and NSA in our dataset can be seen in the context of the controversy about Snowden since we filtered our data set to exclusively include articles that explicitly mention Snowden.

2.3 Visualization

All agglomerated information obtained from the data processing was fed in a user-friendly inter-

[2] http://emm.newsbrief.eu/NewsBrief/clusteredition/en/latest.html
[3] http://opener.olery.org
[4] http://www.opener-project.org/kaf/
[5] http://http://opener.olery.com/language-identifier
[6] http://http://opener.olery.com/tokenizer
[7] http://http://opener.olery.com/pos-tagger
[8] http://http://opener.olery.com/ner
[9] http://http://opener.olery.com/property-tagger
[10] http://http://opener.olery.com/opinion-detector
[11] http://www.opener-project.org/json/

face which would highlight the main outcome of the analysis. In order to do this, the Django web toolkit[12] was used to generate bubbles in different colors. The bubbles show the strength of the opinions per language at first glance (big bubbles for more frequent opinions), and the polarity of these opinions (green for positive, grey for neutral and red for negative). Figure 1 shows the opinions on Snowden we extracted from text in different languages.

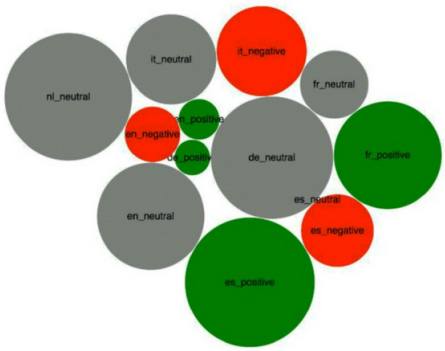

Figure 1: Overview of "Pro-Snowden" opinions in different languages and per polarity class - positive, negative, neutral

3 Results

The results in the graph show that the users posting opinions in different languages have different views on the Snowden case. English and Italian-speaking sources seem to be more negative about Snowden, while the Spanish and French-speaking sources are more positive. In other languages, we can see that the majority of the opinions are neutral. Being able to split the results into languages (and possibly, for the future, on text sources) may thus help to shed more light on the perception of specific populations of controversial topics. This can be useful to many real-world applications. It should be noted, however, that a four hour hackathon does not leave time for an elaborate evaluation. Future work will thus have to investigate the reliability of the opinions found by the OpeNER tools.

4 Conclusions

This paper presents a prototype system drafted in four hours that illustrates the possibility of analyzing mainstream and Social Media multilingual texts regarding a specific targeted topic and displaying the results of the analysis in a user-friendly way. The results in this case show that at first glance, there is a strong difference between the opinion on Snowden in terms of polarity between different languages. In this project, we have also demonstrated that OpeNER webservices were ready to use in an easy-to-plug-and-play way and performed the analysis fast enough to get meaningful results in a short period of time. Of course, the experiment was based in OpeNER's organized Hackathon in Amsterdam in limited time. Bearing this in mind, it is possible for the results to be relatively different in other conditions (e.g. by doing a more robust analysis of the datasets acquired and the results obtained) and a more elaborate evaluation is necessary to confirm the validity of our results.

[12] http://www.caktusgroup.com/services/custom-web-applications/

5 Acknowledgments

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

References

Rodrigo Agerri, Montse Cuadros, Seán Gaines, and German Rigau. 2013. OpeNER: Open Polarity Enhanced Named Entity Recognition. *Procesamiento del Lenguaje Natural*, 51(0).

Wauter Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. 2009. kaf: a generic semantic annotation format. In *GL*.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Rafael.

Bob Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Tour-pedia: a Web Application for Sentiment Visualization in Tourism Domain

**Stefano Cresci, Andrea D'Errico, Davide Gazzé,
Angelica Lo Duca, Andrea Marchetti, Maurizio Tesconi**

Institute of Informatics and Telematics,
National Research Council (CNR)
Via Moruzzi, 1 - Pisa, Italy
name.surname@iit.cnr.it

Abstract

Tour-pedia is a Web application which shows the sentiment of users about touristic locations in some of the most important cities and regions in Europe. It is implemented within the OpeNER project, which aims to provide a pipeline for processing natural language. More specifically, Tour-pedia, exploits the OpeNER pipeline to analyse users' reviews on places. All reviews are extracted from social media. Once analysed, each review is associated to a rate, which ranges from 0 to 100. The sentiment of each place is calculated as a function of all the reviews sentiment on that place. As a result, Tour-pedia shows all the places and their users sentiments on a map.

Keywords: tourism application, sentiment visualization, social media

1. Introduction

OpeNER (Open Polarity Enhanced Name Entity Recognition)[1] is a project funded under the 7th Framework Program of the European Commission. Its main objective is to implement a pipeline to process natural language. More specifically, OpeNER focuses on building a linguistic pipeline supporting six languages (English, Spanish, German, French, Italian, Dutch) that enables the identification and disambiguation of named entities and the analysis of sentiment in opinionated texts.

Within OpeNER, we have developed Tour-pedia, a Web application, which exploits the OpeNER pipeline in order to extract the sentiment of places related to tourism domain. In details, each place is associated to zero or more reviews extracted from social media (i.e. Facebook, Foursquare and Google Places). Each review is processed by the OpeNER pipeline and is associated to a rate, in order to extract its specific sentiment. The sentiment of a place is calculated as a function of all the sentiments of the reviews on that place.

As a result, Tour-pedia shows all the sentiments of all places on a map (as shown in Figure 1). This view allows a user to locate the best places with little effort. In practice, Tour-pedia guides the user in choosing the most suitable solution for his needs. In addition, it helps the user to overcome the most common problems of tourism web sites, such as the difficulty to retrieve detailed information about a specific entity in a single web site. In fact, users generally need to explore several pages on the web to extract the required information. Moreover, most of the tourism web sites analyze and show the reviews of a place from a unique source, and this cannot be representative of its overall overview on the Web. Finally, a user may have trouble in understanding, at a glance, the overall rating about a specific area or a certain entity.

Tour-pedia, instead, supplies a service where the user can find out every touristic information he needs in a unique web page. Infact, Tour-pedia provides the web page of a specific entity for the following social medias: Facebook, Foursquare and Google Places. It also includes different reviews from different sources and shows the reviews summary as an indicator of sentiment.

Currently Tour-pedia contains data about seven cities (Amsterdam, Barcelona, Berlin, Dubai, London, Paris and Rome) and one region (Tuscany). Tour-pedia contains more than 550.000 places divided into four categories: accommodations, restaurants, points of interest and attractions. An accommodation is a place where it is possible to sleep. A restaurant is a place where it is possible to eat and drink. A point of interest is a place where people can have public services, like in airport, railway station etc. Finally, an attraction is a place of entertainment, both for cultural purposes (museum, theater, cinema) and for sport or recreation (night club, swimming pool, golf, gym).

Tour-pedia is available at the following URL: http://www.tour-pedia.org/.

The remainder of the paper is organized as follows: in Section 2. we give an overview of some related work, while in Section 3. we describe the OpeNER project. In Section 4. we illustrate Tourpedia and its Graphical User Interface. Finally in Sections 5. and 6. we discuss some critical points and give our conclusions and future work.

2. Background

There are many initiatives having almost the same purpose of Tour-pedia such as (Dunlop et al., 2004) and (Kenteris et al., 2009).

Dunlop et al. designed and implemented a tourism information software, named Taeneb City Guide, for Personal Digital Assistant (PDA) and handheld computer. It is limited to the city of Taeneb. The main features of the application are the dynamic map interface, the dynamic information content and the community review system.

Kenteris et al. described the issues connected to a "Mobile tourism" application either in terms of networking capabilities of mobile either in terms of User Experience of the application (design, usability, portability). In addition, they implemented a prototye named myMytileneCity Guide.

[1] http://www.opener-project.org

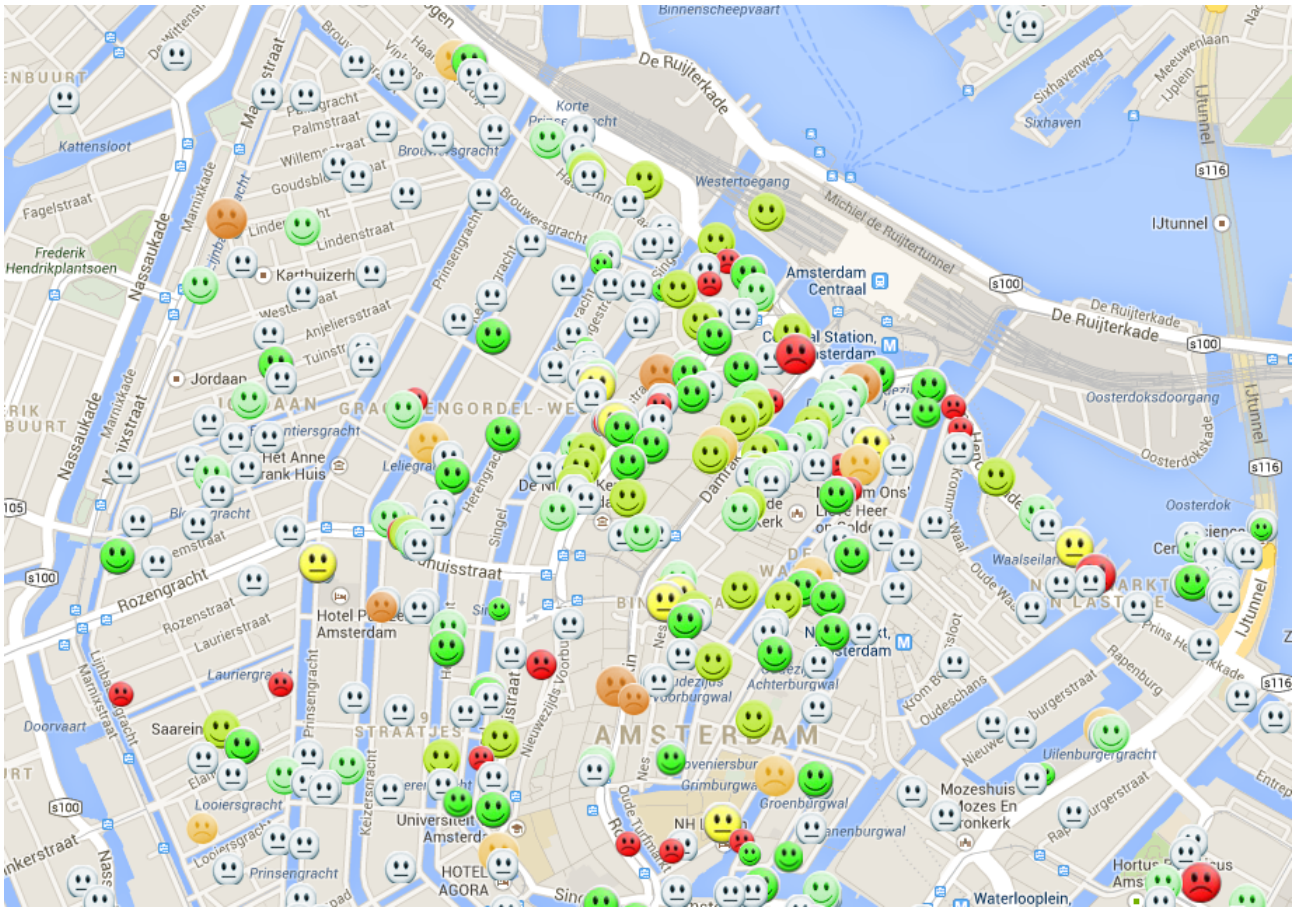

Figure 1: A snapshot of Tourpedia.

Name	Coverage	Categories	No. places	Social Media	Recommendation
Taeneb City Guide	Taeneb	tourism	n.a.	NO	YES
myMytileneCity Guide	Lancaster, UK	Acco, Rest, Attr	n.a.	NO	NO
The Hotel Map	all the World	Accommodation	$\geq$ 200.000	TA	NO
Google Hotel Finder	all the World	Accommodation	n.a.	B, E, etc	YES
Tour-pedia	Some parts of Europe	Acco, POI, Attr, Rest	$\geq$ 500.000	FS, FB, B, GP, I	YES

Table 1: Comparison among existing initiatives.

Other important works are: The Hotel Map and Google Hotel Finder. The Hotel Map[2] is a web application which shows hotels on a world map and allows the user to catch some information, like address, website and reviews from Travel Now[3] about the selected entity. However, the graphic seems very old and the website seems not so rich of informations.
Google Hotel Finder[4] is a Google service for room booking. After the specification of dates of vacation, the user can see the list of available hotels with their details of address, website, services, reviews and price.
The mentioned web sites offer a lot of information about accomodations but their attention is focused only on the booking of rooms or beds. It is difficult to find out data about user reviews or social media information related to a certain hotel; instead, this is the fundamental point of Tourpedia.

Table 1 shows a comparison among the four described initiatives plus Tourpedia. We mean B for Booking, E for Expedia, FS for Foursquare, FB for Facebook, GP for Google Places, TA for TripAdvisor and I for Instagram. The Hotel Map and Google Hotel Finder cover all over the world, Tour-pedia only a subset of Europe and the Taeneb City Guide and myMytilenCity Guide only a city. However, the number of places hosted by Tour-pedia is greater than those hosted by The Hotel Map. This means that Tour-pedia contains more places than The Hotel Map for the same location.

In addition, The Hotel Map and Google Hotel Finder are devoted only to accommodations, while Tour-pedia includes also POIs, attractions and restaurants.

[2]http://www.thehotelmap.net/
[3]http://travel.ian.com/
[4]https://www.google.com/hotels

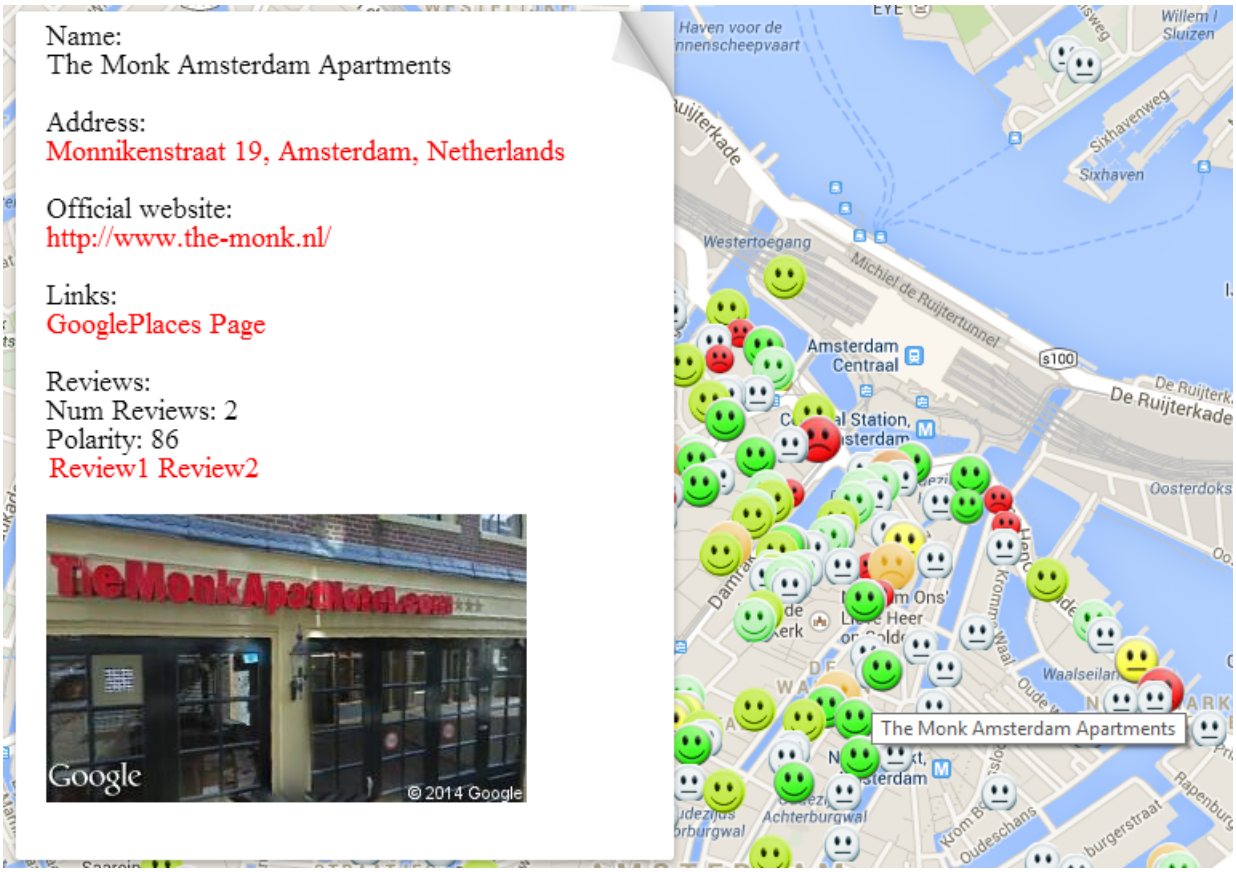

Figure 2: An example of infowindow.

Finally, Google Hotel Finder provides links to many external social media, that allow an user to book a room, whilethe others do not. However, Tour-pedia provides links to other kinds of social media, i.e. Facebook, Foursquare, Google Places and Booking.

3. OpeNER

The OpeNER project provides a set of ready-to-use modules for the processing of natural language.

OpeNER focuses on building a linguistic pipeline that supports six European languages: English, Spanish, German, French, Italian and Dutch, in order to enable the identification and disambiguation of named entities and the analysis of sentiment in opinionated texts. As a result, the project produces a framework for the extraction of the attitude of customers regarding given topics in online reviews. This means both the analysis of large quantities of data and the aggregation of such data for the benefit of professionals within the tourism industry.

The named entities that are found in reviews can belong both to the general domain (e.g. people, locations, dates) and to the tourism domain (e.g. accommodations, attractions, point of interests).

Tour-pedia exploits the OpeNER pipeline. First of all, a pre-analysis is done. A dedicated module elaborates the text of each review, by exploiting the following modules of the OpeNER pipeline: the language identifier, the tokenizer, the polarity tagger and the opinion detector. Firstly, the language identifier extracts the language of the review. Then, the tokenizer extracts tokens from the text of the review. After that, the pos-tagger extracts the parts of speech for each term in the review. The polarity tagger extracts the polarity of each term. The opinion-detector, eventually, extracts the opinion.

Once analysed, reviews about the same place are aggregated. Then, the sentiment about a place is calculated as a function of the extracted opinions on all the reviews about that place. Finally the sentiment of every place is shown on a map, as shown in Figure 1.

4. Tour-pedia

Recent research demonstrated that the APIs provided by Google Maps are very flexible (Pan et al., 2007). For this reason, Tour-pedia exploits them and emulates the navigation style of Google Maps[5], leveraging and enhancing its fundamental characteristics: a map which occupies the whole page; a simple menu placed over the map; a search bar embedded inside the map itself.

In order to draw users' attention on the map on the map (the focal point of the interface) all info-windows appear over the map, without subtracting too much space. Figure 2 shows an example of info-window for "The Monk Amsterdam Apartments".

Places appear on the map as smileys: the color and mood of the icons is determined by aggregating the sentiment ex-

[5]https://maps.google.com/

tracted from the reviews for that entity. A prevalent number of positive sentiment produces a green smiley, red is produced by a prevalent negative sentiment and different colors express intermediate ranges. White locations have no reviews available for evaluation. In addition, the size of the emoticon is proportional to the number of reviews for that entity, so big smileys mean many reviews and small smileys mean few reviews.

5. Discussion

As already said, Tour-pedia shows the result of sentiment on reviews extracted from Facebook, Foursquare and GooglePlaces. For this reason, sentiments resulting from the OpeNER analysis, reflect real users sentiments.
A mechanism to analyse reviews on the fly should be implemented, since new reviews continously add to existing ones. However, generally, the overall opinion about a place does not change frequently, unless the place itself changes something, i.e. adding new features or solving issues reported in past reviews. For this cause, it is quite reasonable that the reviews analysis could be done off-line.

6. Conclusions and Future Work

Tour-pedia is a practical example of how the OpeNER pipeline can be exploited. At the moment it shows the sentiment of places of seven cities and a region, but the final goal is to extend the project to the whole Europe.
In future, Tour-pedia could exploit also the module of named entity recognition developed within OpeNER, in order to show in the map the entities cited within the reviews.

Acknowledgements

This work has been carried out within OpeNER project, cofunded by the European Commission under the FP7 (7th Framework Programs Grant Agreement n. 296451).

7. References

MarkD. Dunlop, Piotr Ptasinski, Alison Morrison, Stephen McCallum, Chris Risbey, and Fraser Stewart. 2004. Design and development of Taeneb City Guide: From Paper Maps and Guidebooks to Electronic Guides. In AndrewJ. Frew, editor, *Information and Communication Technologies in Tourism 2004*, pages 58–64. Springer Vienna.

Michael Kenteris, Damianos Gavalas, and Daphne Economou. 2009. An innovative mobile electronic tourist guide application. *Personal and Ubiquitous Computing*, 13(2):103–118.

Bing Pan, JohnC. Crotts, and Brian Muller. 2007. Developing Web-Based Tourist Information Tools Using Google Map. In Marianna Sigala, Luisa Mich, and Jamie Murphy, editors, *Information and Communication Technologies in Tourism 2007*, pages 503–512. Springer Vienna.