OPEN POLARITY ENHANCED NAMED ENTITY RECOGNITION

Instrument: Collaborative Project
Thematic Priority: FP7- ICT -2011.4.1
Grant Agreement: 296451

ANNUAL PUBLIC REPORT 2012

<table>
<thead>
<tr>
<th>Deliverable Number</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Annual Public Report</td>
</tr>
<tr>
<td>Version</td>
<td>1.0</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>PU</td>
</tr>
<tr>
<td>Date</td>
<td>15 November 2012</td>
</tr>
<tr>
<td>Status</td>
<td>Final version</td>
</tr>
</tbody>
</table>

EC Distribution

Project Partners: VICOMTECH-IK4, UPV/EHU, VUA, SYNTHEMA, OLERY, CNR

Contributors: VICOMTECH-IK4 (Seán Gaines)Every effort has been made to ensure that all statements and information contained herein are accurate; however the Partners accept no liability for any error or omission in the same.

© Copyright in this document remains vested in the Project Partners
Table of content

1. INTRODUCTION .................................................................................................................................................. 1

2. SUMMARY OF ACTIVITIES ................................................................................................................................. 1
   2.1. STATE OF THE ART ANALYSIS ....................................................................................................................... 2
   2.2. SYSTEM DESIGN ............................................................................................................................................... 4
   2.3. FUNCTIONAL REQUIREMENTS ....................................................................................................................... 4
   2.4. USER SCENARIOS ............................................................................................................................................ 4
   2.4. END USER ADVISORY BOARD (EUAB) .......................................................................................................... 5
   2.5. EUAB WORKSHOP ........................................................................................................................................ 6
   2.6. WORKSHOP WITH LOCAL END USERS ...................................................................................................... 6

3. FUTURE WORK OR EXPLOITATION PROSPECTS .............................................................................................. 7

4. PROJECT PROGRESS .............................................................................................................................................. 7
1. Introduction

Customer reviews and ratings on the Internet are increasing importance in the evaluation of products and services by potential customers. In certain sectors, it is even becoming a fundamental variable in the purchase decision. A recent Forrester study showed more than 30% of Internet users have evaluated products online, and that 70% of those studied end user generated reviews.

This trend will continue with the growth of Social Media and access to Information and Communication Technologies (ICT). Consumers tend to trust the opinion of other consumers, especially those with prior experience of a product or service, rather than company marketing. The role of user comments is of particular importance when there is little differentiation between the product offers.

Sentiment Analysis and Opinion Mining are established, although nascent, fields of research, development and innovation. The goal is always broadly the same; to know “Who” is speaking about “What”, “When” and in “What sense”.

These factors have led to a burgeoning industry with a plethora of companies offering Sentiment Analysis services in Social Media. While most offer a generic service, in typically just one language, several companies have specialised offering services specific to tourism due to its bounded domain, demonstrable value, and the high level of adoption of Internet technologies by both suppliers and consumers. It is also an application domain with limited scope and variation, and a high dependency on multilingual sentiment analysis and detection and classification of a wide range of common Named Entities.

Named Entity Recognition and Classification (NERC) are important in determining roles. Once multilingualism and cultural skew are introduced, the complexity of the challenge increases manifold. OpeNER will create base technologies for Cross-lingual NERC and Sentiment Analysis that will enable industry users to both 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.

The OpeNER project will provide a rich Named Entity Data Source in a simple, structured and standardised format. The Named Entity Detection will be capable of marking Named Entities in the same format irrespective of the text under analysis or the language of the text. The project will also provide linking modules that are capable of matching locally detected Named Entities with generic data.

2. Summary of activities

The project is currently in Month 5. The OpeNER kick off meeting took place in July 2012 at the European Commission, in Luxembourg.

In month 3, September 2012, a second Steering Committee, two technical project meetings and the first End User Advisory Board (EUAB) Workshop took place.

For the first four months of the project, the technical work of the consortium has focused on system definition in WP2, and the survey studies in WP3, WP4, and WP6. Of particular importance at this first stage of the project, has been the work related to WP2 DESIGN – System Design and User Requirements Modelling.

Regarding the dissemination activities included in WP8, the main activities have been the setup of the End User Advisory Board (EUAB), and the first EUAB workshop and the establishment of the OpeNER online presence.
2.1. State of the Art Analysis

In month 2, August 2012, the state of the Innovation analysis milestone has been reached, with surveys of existing tools and datasets:

An in-depth survey of the current state of the art, data sources, tools and technology related to Named Entity Recognition and Classification (NERC), coreference resolution, and Named Entity Disambiguation (NED) for English, French, Dutch, German, Spanish and Italian.

The goal is to feed into the functional requirements and system design and establish the needs for linked data and processing techniques required for NERC, coreference resolution and NED. The study confirms:

- The needs for common repository of entities which will be developed in WP4 (prototype 1 in deliverable D4.31), not only for general Named Entity Resolution but also for its extension to the touristic domain.
- The existence of sufficient tools and data for NERC in the general domain.
- The need for a decision to be made as to whether different systems are going to be used for each language or if a single approach will be adopted, given that there is not a system or approach which currently provides coreference resolution for all six languages in OpeNER.
- The need to create a repository of entities for the tourist domain (in WP4) for the development of Named Entity Resolution systems and for OpeNER technology in general.

An analysis of currently available social and semantic datasets that are relevant to the OpeNER project, that:

- Recognizes the evidence that multilingual named entity recognition and classification is an open task and a lot of research is going on to produce datasets and tools suitable for a variety of languages.
- The importance of the multilingualism issue, when considering Web extracted text and the OpeNER project deals with Web extracted texts in 6 different languages (English, Dutch, German, French, Spanish and Italian). English is obviously the language with the largest number of dedicated resources, but finding resources for the others is not an issue.
- The vast majority of social datasets are potentially available for all the languages of the OpenNER project.
- Wikipedia is one of the most exploited multilingual resources for named entities recognition and classification. A wide variety of research works focusing on multilingualism uses Wikipedia as a source for concepts and relations: it is available in 250 languages, it is updated constantly and its content is more structured than a simple text. The most exploited features are: the links between two articles in the same language; links between an article and its category in a Category page; links to Redirect and Disambiguation pages; interwiki links, between an article and its equivalent in another language.
- Looking at available services for semantic datasets we spotted the lack of them for Dutch:
A survey of available resources and technologies for sentiment analysis in the 6 OpenNER languages

- Overviews of the findings are shown in the following table. There are many resources and tools for English, but hardly any for other languages. The tools can be easily ported to other languages, provided that we deliver the training data or the lexicons. Developing the training data and lexicons appears to be the first priority for the project. For Italian, there are for example no freely available sentiment lexicons. In addition, there is a need for evaluation data that is compatible across the different languages so that we can inform the users about the quality.

## LEXICONS

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>NL</th>
<th>DE</th>
<th>FR</th>
<th>ES</th>
<th>IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic</td>
<td>9</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>HOTEL</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

• 5 commercial, 6 research-only, 8 open-source, 3 unknown

## CORPORA

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>NL</th>
<th>DE</th>
<th>FR</th>
<th>ES</th>
<th>IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>HOTEL</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

• 15 open-source, 3 unknown

## TOOLS

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>NL</th>
<th>DE</th>
<th>FR</th>
<th>ES</th>
<th>IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>HOTEL</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
<td>N.A.</td>
</tr>
</tbody>
</table>

• 2 commercial, 5 research-only, 3 open-source, 2 unknown

### Table 1. Services classified by language.

<table>
<thead>
<tr>
<th>Language</th>
<th>Available services for semantic datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>AlchemyAPI, DBpedia Spotlight, OpenCalais, Zemanta, Wikimeta, TAGME</td>
</tr>
<tr>
<td>Dutch</td>
<td>-</td>
</tr>
<tr>
<td>German</td>
<td>AlchemyAPI</td>
</tr>
<tr>
<td>French</td>
<td>AlchemyAPI, OpenCalais, Wikimeta</td>
</tr>
<tr>
<td>Spanish</td>
<td>AlchemyAPI, DBpedia Spotlight, OpenCalais, Wikimeta</td>
</tr>
<tr>
<td>Italian</td>
<td>AlchemyAPI, TAGME</td>
</tr>
</tbody>
</table>
2.2. System Design

A proposal of the OpeNER system designed architecture has been defined.

The objective for Cycle 1 is to implement the base core technologies required for the project. The modules will be implemented only for the project targeted languages (English, German, French, Italian, Spanish and Dutch) and only for a generic domain.

2.3. Functional Requirements

The functional and non functional requirements for the first Cycle of OpeNER project were defined. The surveys have been crucial for the definition of the requirements necessary to the first prototypes to be delivered in month 10 of the project, deliverable D3.21. Both the functional and non-functional requirements can be changed during the life cycle of the project.

In the first End User Workshop of the project, which was held in San Sebastian on the 13th and 14th of September 2012 (deliverable 8.4.1), the main recommendations concerned from the Board were:

- The extendibility to further languages, in particular Portuguese (both from Portugal and Brazil), Rumanian and Latin-American Spanish, as well as other languages such as Arabic, Russian (this was considered very important by Casa400), Hindi and Mandarin.

- The possibility for OpeNER to design and evaluate marketing campaigns on defined countries, regions, cities, with focus on specific languages. Especially, it would be interesting to target certain markets with specific requirements such as elderly Chinese and baths. In this case, the challenge will be to translate campaigns and messages to concepts, sets of entities, or grades of sentiment.

- The possibility to fine-tune the system on cultural differences among tourists. This has a clear influence on the sentiment analysis of several reviews, as these differences are reflected in expectations and preferences (this aspect will be dealt with in Cycle 2)

- The attention to sources other than reviews, such as conversations that normally take place in their micro-sites. In this case, the system should provide information about the evolution and trends of several entities among time to see, for example, the variations on the themes of the conversations (multi-document co-reference).

2.4. User Scenarios

Taking into account main recommendations of the end users, some user scenarios were set:

- The user wants to search for reviews or comments expressing consumer’s attitudes about a particular Entity (brand, product, service, etc). The search is focused on a particular Named Entity and all its properties.

- The user wants to search for reviews or comments expressing opinions and sentiments about a certain area. Thus the system should concentrate for services based in that area.
• The user wants to search for specific properties concerning services and analyze consumers attitudes and polarities about them.

• The user wants to mine for sentiment and opinions on their service by a specific class of customers (e.g. elderly Dutch travellers) in order to target their campaigns on them.

Such scenarios determine the functionalities required by the OpeNER services:

• Take raw input data potentially containing relevant information in:
  o several languages (Spanish, Dutch, German, Italian, English and French)
  o several sources (reviews in forums, conversations in social media, ..)

• Extract the text, normalize and perform basic linguistic analysis using NLP tools

• Analyze the text in order to extract all the relevant information that is:
  o recognize and classify named entities (is it a city, a hotel, an attraction, a business..)
  o recognize coreferences

• Recognize properties (such as for instance - for an accommodation - cleanliness, price, service, location, etc. of an accommodation)

• Mine for sentiments and opinions referring to the named entities in the document or to properties thereof and extract them in a structured way

• Display the results in a form that allows querying for Entities, properties, typology of consumer, area monitor results of searches for particular targets over time.

2.4. End User Advisory Board (EUAB)

The End Users Advisory Board (EUAB) of the OpeNER project was set in month 3, September 2012, with the principal aim of assisting in the definition of requirements, testing and assessing impact of the project. The EUAB will be in charge of the major decisions regarding to the tools development for Opinion Mining and Sentiment Analysis within the OpeNER project.

The first EAUB consisted of:

• Tourism Organisations (Destination Management Organisations)
  o National Tourism Organizations (Segittur)
  o Regional Tourism Organizations (Fondazione Sistema Toscana)
  o Local Tourism Organizations (Amsterdam Tourism and Convention Board)

• Online Tourism Portals
  o TripAdvisor
  o Trivago

• Additional experts
  o Yahoo Research
2.5. EUAB Workshop

The first EUAB workshop was held in San Sebastian on the 13th and 14th of September 2012. In addition to the Consortium members, the following experts were invited:

- Mr. Hans Vught, Casa400, Manager director of Hotel Casa400 as well as member of the iAmsterdam (Tourism Board)
- Mr. Carlos Romero, Spanish State Agency for Innovation in Tourism
- Dr. Jordi Atserias, Yahoo Research
- Dr. Alexandra Balahur

Principal conclusion of the EUAB

During debate with the EUAB and a review of End Users defined some of the requirements they would like to have in the final version of the OpeNER prototype. The list has been filtered and organized following some of the key aspects of OpeNER.

- Target Languages: There is consensus regarding further European languages such as Portuguese (both Iberian and Brazilian), Rumanian and Latin-American Spanish, as well as other languages such as Arabic, Russian (this was considered very important by Casa400), Hindi and Mandarin.
- Named Entities: In the view the End-User Advisory Board has been the definition of the entities that are most related to a tourism destination. It should be mentioned that the identification of the NE will be related to the strategies of the destination. There are several things that were mentioned such as events, places, attractions, point in time, poor weather events, activities, disasters.
- Data Mining on Entities: SEGITTUR discussed the potential use of OpeNER to design and evaluate marketing campaigns on defined countries, regions, cities or even languages. Targeting certain markets with specific requirements such as elderly Chinese and Baths in Hotel rooms rather than showers. In this case, the challenge will be to translate campaigns and messages to concepts, sets of entities, or grades of sentiment.
- Another important aspect that OpeNER will work on is the cultural differences among tourists. This is a clear influence on the sentiment analysis of several reviews, as these differences are reflected in expectations and preferences. There are many aspects such as locations (near monuments) or cleanliness that are very subjective and depend on the cultural background of the tourist.
- In the view of the EUAB, the project should not only think about reviews but also about conversations that normally take place in their micro-sites. In this case, the system should provide information about the evolution and trends of several entities among time to see, for example, the variations on the themes of the conversations (multi-document co-reference).

2.6. Workshop with Local End Users

As a complementary activity, several potential Local End Users were invited to join the consortium meeting during the 14th of the September 2012. The meeting was held in the San Sebastian Technology Park. During the event a presentation of the project to the local End Users and a free debate about requirements, needs, advantages and disadvantages of the proposed
solutions of OpeNER was made. This audience included: Mondragon University LKS Consulting (Mondragon Group) BitText Technologies, Alianzo Networks, Overalia, Mapalia, S21Sec Labs, Basquetour, The Movie and Representatives of the Basque Government Department of Tourism.

This event was covered by major press outlets and was subject of a one full page report in the El Pais, Global Innovator's Supplement. El Pais is an international paper of record.

3. Future Work or Exploitation Prospects

Principal RTD activities will focus on:

- Further definition of the global system, defining more the architecture and components design for Cycle 1.
- Defining the most appropriate development lifecycle to successfully build the OpeNER system.
- Defining the performance and functional criteria that the Cycle 1 prototype needs to meet.

The next Steering Committee and project technical face to face meetings will be held in January 2013. The second, expanded, EUAB workshop will be held in month 12, June 2013, with principal aim of evaluating the first version of the OpeNER prototype.

4. Project Progress

Deliverables submitted

D1.21 Project operating and, quality and risk procedures
D2.11 System requirements for Cycle 1 (SYNTHEMA)
D3.11 Survey of existing tools and data sets (EHU/UPV)
D4.11 Survey of existing tools and data sets (CNR)
D5.11 Survey of tools and data sets (EHU/UPV)
D6.11 State of the innovation report (VICOM)
D8.11 EUAB Setup and First OpeNER Workshop and Resort (VICOM)
D8.31 Project Dissemination and Communication Plan (VICOM)

Milestones reached

MS1 State of the art analysis