Publishable summary

About OpeNER

The increasing importance of customer reviews and ratings on the Internet in the evaluation of products and services by potential customers, the OpeNER project aims to provide enterprises and society with base technologies for Cross-lingual Named Entity Recognition and Classification and Sentiment Analysis through the reuse of existing resources and the open development of complementary technologies. Focusing on the tourism sector, the fundamental aim of OpeNER is to allow users of the technologies to apply and contribute to a set of technologies, with a minimum total cost, allowing them to concentrate their efforts on other innovative areas of their business that meet immediate market need.

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

The key objectives of the project can be further listed as follows:

- Repurposing of existing language resources and generation of a reference generic multilingual sentiment lexicon with cultural normalisation and scales, and an extension lexicon for the tourism sector in different languages (Spanish, Dutch, German, Italian, English and French).
- Named Entity Recognition and Classification in the same set of target languages as the Sentiment Lexicon which is extensible to other languages by leveraging multilingual resources, such as Wikipedia and Linked Data.
- Development and open availability of validated reference Sentiment and Opinion Mining techniques and tools based on the results of the project.
- Validation of the project results will be done in the tourism sector, with leading SMEs in the field and the support of several stakeholders as part of the End User Advisory Board (a global tourism portal operator, several tourism destinations, etc.).
- Research and trialling of models that will ensure that the project results are self-sustainable and economically viable in the long term.

The goals of the project have been achieved by repurposing and leveraging existing state of the art and established language resources. Chief amongst these have been the semi-automatic generation of a generic multilingual sentiment lexicon via WordNet, EuroWordNet, MCR, ItalianWordNet and Wolf. A further extension to the lexicon has been built for the tourism domain and has allowed validation and testing.

Multilingual semantic resources that are tied to Linked Data have been leveraged for Named Entity Recognition and Classification. Wikipedia has been the principal source for NERC due to its suitability for the recognition of Toponyms, Proper Names, Company names, Eponyms, etc. The core NERC tools have been extended to the Tourism domain and fine-tuned for the recognition of NEs relevant to that domain and sentiment analysis of the properties related to those named entity types (e.g. cleanliness, price, service, location, etc. of an accommodation).

The project has used open standards based technologies and suggested a migration path from proprietary technologies that have a beneficial impact on further exploitation of the technology resulting from the project. An approach based on open standards will help guarantee the long term development and sustainability of the system after the project conclusion.

Identification of user requirements and architecture definition

During the first reporting period of the project the user requirements were identified, and those were reviewed and updated over the second reporting period. End users' needs, collected through questionnaires, teleconferences and the EUAB workshops, were translated into the initial general requirements and then into stratified topic requirements.

The OpeNER architecture design was identified and the infrastructure (Github, Drone.io) and methodology to be used throughout the project was circulated and agreed by all partners.

Development of basic components

Basic pre-processing tools such as language identifiers, tokenisers and part-of-speech taggers were early developed and integrated in the architecture to ensure the basis of all the core components to work properly. All these components were developed for all languages in the project (English, French, Spanish, Dutch, German and Italian). Additionally, the input and output formats were defined as raw text as first input in the chain, and KAF format as an input and output for all the following components in the chain. KAF (Knowledge Annotation Format), an xml-based format from the Kyoto project, was specified according to the OpeNER basis, and different libraries were built by the different partners to ensure correct codification. Additionally, a DTD was distributed through the consortium to verify the correctness of the KAF schema in all the different components.

Development of core components

During the first year of the project, the main core components were developed for all the languages in the project. Regarding the second year, domain adaptation tools for the main core technologies of the OpeNER project have been developed. Additionally, a set of lexicons, annotated datasets, and guidelines have been created to allow the adaptation of technology for a certain domain. The main core components are:

- NER: Named Entity Resolution tools for all languages in OpeNER. This involves Named Entity Recognition and Classification (NERC), Coreference resolution and Named Entity Disambiguation (NED) tools for general and specific domains. Furthermore, Domain adaptation tools have also been developed.
- Datasets: A deep analysis of currently available social and semantic datasets that are relevant to the OpeNER project was performed. The goal of this analysis was to identify those tools that fit the tasks planned within OpeNER and to gauge differences between target languages. Moreover, the definition and the computation of specific indicators and metrics of popularity have also been studied, together with the reputation and their trends related to social network analysis as a complementary approach to sentiment analysis, the definition of a Domain Named Entity repository (Tourpedia) and further use of DBpedia Spotlight for general Named Entity Repository, and the development of a component (Entity Resource Linker) to query Linked Data catalogues to obtain all the knowledge available for a certain Named Entity.
- Sentiment: Development of multilingual opinion mining modules and the resources required in order to accomplish this duty. First, existing state-of-the-art resources and modules have been collected. Subsequently, existing techniques to create sentiment lexicons from existing

wordnets for English, Dutch, German, French, Italian and Spanish have been applied. A Sentiment Lexicon Markup format based on the current Wordnet-LMF proposal has been developed. Text mining tools to obtain sentiment values for words from text corpora have been created. Additionally, opinion mining processors for domain adaptation based on annotated datasets created within the project have also been developed. Finally, a set of tools aimed at allowing the adaption of a system to any domain have also been released.

- The software has been being wrapped up in an easy to use toolkit that can be used to create a domain extension of the generic sentiment lexicon.
- Datasets and Integration: A wide set of datasets have been crawled from Social Media as a result of a study carried out within the OpeNER project. These datasets proceed from Facebook, Google Places and Foursquare. Additionally, the integration of all the components has been performed in a suitable framework to evaluate the performance of the entire system.

Integration and testing

Regarding the integration process developed in OpeNER, a central point of integration, testing and deployment in an appropriate environment for early and frequent development, as well as a continuous testing and integration regime has been defined. This central point has been essential for the beta and final GM deployment of OpeNER as a set of libraries, documentation, APIs, Sample Applications and, most importantly, a user development community with supporting collaborative tools such as source control, Wiki and Bug tracking. For this purpose, existing platforms have been used (Github, Drone.io, Amazon AWS). Besides, a central entrance point has been provided to the end users to access the OpeNER technology.

As a result, OpeNER has provided a setup of the development environments for all developers individually, as well as the support *collaborative* development platform and tools. Consequently, OpeNER has succeed in the integration of all components into Cycle 1 and Cycle 2 prototypes. Additionally, the testing and demonstration processes have been accomplished by testing the developed prototypes with potential end-users, as well as the developer community in general. Of particular importance was the performance of the first Hackathon where OpeNER cycle 1 prototype was tested. Moreover, another Hackathon with a Tutorial session took place in LREC2014 for testing cycle 2, and finally the last Hackathon has been proposed for the testing of all the OpeNER tools.

Fostering the European tourism sector

Amongst the multiple challenges that the European tourism sector faces, such as the growing impact of innovation and communication technologies and the increasing global competition, improving the competitiveness of tourism in the EU plays a crucial role in strengthening the sector with a view to dynamic and sustainable growth. Within this context OpeNER develops an open platform for new ICT-based products and services to improve the competitiveness of European SMEs. OpeNER acts as an "ICT and tourism" platform for stakeholders to facilitate the adaption of the tourism sector and its business to market developments in new information technologies and improve their competitiveness by making the maximum use of possible synergies between the two sectors.

Taking into account the importance of the service dimension of the tourism sector, OpeNER results will provide managers with the information required for understanding customer needs and personalising its products on the basis of advanced profiling techniques. Thus, OpeNER fosters the advent of products and services that will boost the competitiveness of the European tourism sector as a whole. No doubt that provision of reliable and responsive services such as those planned in OpeNER surely enhance the competitive advantage of the European destinations and accommodation industry.

Besides, the open source approach will positively impact in the tourism industry, but specially the European industry and European citizens, fostering creative and thoughtful European citizens with ideas to start businesses.

Apart from that, OpeNER clearly raises awareness to the importance of knowledge and innovation in tourism through the creation of the End User Advisory Board, strengthening the use of new technologies by the coordinated action of public and private tourism actors, and encouraging the exchange of best practices within the actors in the tourism value chain.

End User Approach

From the very beginning OpeNER has involved end users in the project activities. Having into account the two type of end users identified: intermediaries (service integrators) and final end users (tourism service providers like hotels, tourism information providers of different public administrations); the impact on the first type has been tackled through the inclusion of the Dutch SME, OLERY and the Italian SYNTHEMA, service integrators SMEs within the OpeNER consortium. The interaction with the second type of end users is ensured through the End User Advisory Board (EUAB).

One of the key success factors within the project has been the organisation of End Users Workshop I in month 3 (devoted to brainstorm requirements for the OpeNER prototype), and End Users Workshop II at the beginning of month 13 (devoted to the evaluation of the first version of the prototype and their positive results). The increasing number of interested potential end users participating in the workshops is a clear sign of the growing and positive impact that the project has reached among those.

OpeNER will raise awareness to the importance of knowledge and innovation in tourism through the creation and work of the EUAB, strengthening the use of new technologies by the coordinated action of public and private tourism actors, and encouraging the exchange of best practices within the actors in the tourism value chain.

Needless to say that OpeNER will positively impact the scientific, technical and market positioning of the project partners.

Project portal: http://www.opener-project.org