Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 215-218 recibido 30-04-2013 revisado 16-06-2013 aceptado 21-06-2013

OpeNER: Open Polarity Enhanced Named Entity Recognition

OpeNER: Reconocimiento de entidades nombradas con polaridad

Rodrigo Agerri
IXA NLP Group
UPV/EHU
rodrigo.agerri@ehu.es

Montse Cuadros Seán Gaines
HSLT, IP department
Vicomtech-IK4
{mcuadros,sgaines}@vicomtech.org

German Rigau
IXA NLP Group
UPV/EHU
german.rigau@ehu.es

Resumen: Actualmente existe una gran cantidad de empresas ofreciendo servicios para el análisis de contenido y minería de de datos de las redes sociales con el objetivo de realizar análisis de opiniones y gestión de la reputación. Un alto porcentaje de pequeñas y medianas empresas (pymes) ofrecen soluciontes específicas a un sector o dominio industrial. Sin embargo, la adquisición de la necesaria tecnología básica para ofrecer tales servicios es demasiado compleja y constituye un sobrecoste demasiado alto para sus limitados recursos. El objetivo del proyecto europeo OpeNER es la reutilización y desarrollo de componentes y recursos para el procesamiento lingüístico que proporcione la tecnología necesaria para su uso industrial y/o académico.
Palabras clave: Reconocimiento y Desambiguación de Entidades Nombradas, Coreferencia, Análisis de Sentimiento

Abstract: Currently there are a many companies offering Content Analytics and Social Internet Mining services for the purposes of Opinion Mining and Reputation Management. A high percentage of Small and Medium Enterprises (SMEs) are active offering niche solutions to specific segments of the market and/or domains. However, acquiring or developing the base qualifying technologies required to enter the market is an expensive undertaking that redirects the already limited resources of SMEs away from offering products and services that the market demands.The main goal of the OpeNER european project is the reuse and repurposing of existing language resources and data sets to provide a set of underlying technologies to the broader industrial and academic community.
Keywords: Named Entity Recognition and Disambiguation, Coreference, Opinion Mining

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[1]. Furthermore, another study by Complete Incorporated for the Tourist Domain showed that more than 80% of users preferred other users' opinions in order to make their buying decisions. In fact, it has been concluded that 97% of Internet users have read and been influenced by other users' opinions while planning a trip (Gretzel and Yoo, 2008). Obviously, 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 (see footnote 1). The role of user comments is of particular importance when there is little differentiation between the product and services on offer. Therefore, there is an objective necessity to manage and understand the knowledge conveyed by opinions. Opinion Mining consists of extracting and analysing, from unstructured text, opinions about products, people, events, institutions, etc. (Pang and Lee, 2008). In other words, the goal is to know "who" is speaking about "what", "when" and in "what sense" (Hu and Liu, 2004). More specifically, OpeNER will stress the importance of providing a good Name Entity Resolution system (Named Entity Recognition or NERC, Coreference and

[1] http://www.bazaarvoice.com/resources/stats

ISSN 1135-5948

Named Entity Disambiguation or NED) to feed the feature-based opinion mining systems with relevant information with respect to the entities about which the opinions are being expressed.

Currently there are a many companies offering Content Analytics and Social Internet Mining services for the purposes of Opinion Mining and Reputation Management. A high percentage of Small and Medium Enterprises (SMEs) are active offering niche solutions to specific segments of the market and/or domains. However, acquiring or developing the base qualifying technologies required to enter the market is an expensive undertaking that redirects the already limited resources of SMEs away from offering the products and services that the market demands.

The main goal of the OpeNER project[2] is the reuse and repurposing of existing language resources and data sets to provide a set of underlying technologies to the broader community. OpeNER will focus on the provision of a supplementary sentiment lexicon with culturally normalised and graduated values. NERC will also be addressed leveraging Linked Data with the aim of disambiguating the entity types recognised for the languages covered in the project: Spanish, English, French, German, Dutch and Italian. In the first year the project will be focused on a generic application domain, and later, adapted to the Tourism domain.

This will be achieved in conjunction with and End User Advisory Board (EUAB) composed of European Tourism Promotion Agencies, an online Tourism Portal and other interested parties. Furthermore, OpeNER will employ proven software from the Open Source community and develop an online development community thus ensuring long term viability beyond the project timeframe. In that way the benefits of the project will be adopted and extended to new domains and languages, OpeNER will strive to make the tools and techniques resulting from the project available under Open Source or Hybrid Licenses.

2 Objectives

OpeNER aims to provide enterprise 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. The key objectives of the project are the following: (i) Repurposing and/or developing of existing language resources and generation of a reference generic multilingual sentiment lexicon with cultural normalisation and scales; (ii) An extension lexicon for the Tourist sector in several languages (Spanish, Dutch, German, Italian, English and French); (iii) Named Entity Resolution (NERC, NED and Coreference) 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[3] resources such as DBpedia[4], etc; (iv) Development and open availability of validated reference Sentiment and Opinion Mining techniques and tools based on the results of the project; (v) Evaluation and Application of the project results in the cloud, principally in the tourism sector, with leading SMEs in the sector and with the support of several stakeholders as part of the End User Advisory Board (EUAB); (vi)Research and trialling of models that will ensure that the project results are self-sustainable and economically viable in the long term; (vii) Achievement of the projects objectives by repurposing and leveraging existing state of the art and established language resources.

3 Work Plan

In order to optimise the value of OpeNER technology, all the requirements along the value chain for the development and the exploitation of the project's objectives are directly represented in the project's Consortium, formed by 6 institutions from Italy, The Netherlands and Spain, with Vicomtech-IK4 as coordinator. The OpeNER Work Plan is structured in 8 Work Packages (WP), and can be divided in three blocks. Although we first describe every WP we will henceforth focus on the most relevant aspects to SEPLN, namely, those related to work packages 4-7:

1. **Management, Dissemination and Exploitation**: WP1 and WP8 led by Vicomtech-IK4. As this is an SME oriented project, the Dissemination and Exploitation of results will go beyond

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

[3] http://linkeddata.org

[4] http://dbpedia.org

the publication of scientific articles. It shall include industrial dissemination and exploitation also.

2. **System Design and Deployment**: WP2 and WP7 led by Synthema and Olery respectively. In order to ensure the future exploitation of the project by SMEs, the system design and deployment is crucial. Both Synthema and Olery have experience in software integration for industry related applications and products.

3. **NLP and Web techniques**: WP3 (Universidad del País Vasco/Euskal Herriko Unibertsitatea, UPV/EHU), WP4 (Consejo Nacional de Investigación de Italia, CNR), WP5 (Universidad Libre de Amsterdam, VUA) and WP6 (Vicomtech-IK4). Focused on Opinion Mining (WP5) and Named Entity Resolution (WP3) and any other basic NLP (WP6) and Web tools (WP4) required to perform those tasks.

OpeNER will provide language analysis tool chains for several languages to help researchers and companies make sense out of unstructured text via Natural Language Processing. It will consist of easy to install, improve and configure components to: (i) Detect the language of a text; (ii) Determine polarity of texts (sentiment analysis) and analysis of feature-based opinions; and (iii) Detect and classify the entities named in the texts and link them together via Named Entity Recognition, Coreference and Named Entity Disambiguation (e.g. President Obama or The Hilton Hotel). Besides the individual components, guidelines exists on how to add languages and how to adjust components for specific situations and topics. The following section will describe the English and Spanish OpeNER toolchains.

4 OpeNER NLP Pipelines

An OpeNER tool chain or pipeline consists of a broad mix of technologies glued together using Ruby. The prerequisites for running an OpeNER tool chain are the following: A GNU/Linux or Unix type operativing system (including MAC OS), Ruby 1.9.3+, Python 2.7+, Java 1.7+, Perl 5+. Every part of the OpeNER tool chain has individual dependencies, most of which are included in the components themselves. The OpeNER architecture consists of several building blocks called *components*, which can be used to build a tool chain called a *configuration*,

A *component* consists of a *kernel* which can be for example a POS tagger implemented in Java and a *glue* in Ruby to connect with other components. Figure 1 represents a possible flow of information between several components. Each component is configured to take the information it requires to perform a specific analysis from the previous module. KAF(Bosma et al., 2009)[5] is used as inter-component representation between the components. Each of the tool chains built are then deployed via Cloud Computing services such as Amazon Elastic Computing Cloud[6] (Amazon EC2).

As described in section 3, the NLP focus of OpeNER is on Named Entity Resolution (NERC, Coreference and NED) and Opinion Mining. The overall objective of Named Entity Resolution is to be able to recognise, classify and link every mention of a specific named entity in a text. A named entity can be mentioned using a great variety of surface forms (Barack Obama, President Obama, Mr. Obama, B. Obama, etc.) and the same surface form can refer to a variety of named entities: for example, the form 'San Juan' can be used to ambiguously refer to many toponyms, persons, a saint, etc. (e.g. see http://en.wikipedia.org/wiki/San_Juan). Furthermore, it is possible to refer to a named entity by means of anaphoric pronouns and coreferent expressions such as 'he', 'her', 'their', 'I', 'the 35 year old', etc. Therefore, in order to provide an adequate comprehensive account of named-entities in text it is necessary to recognise the mention of a named-entity, to classify it as a type (e.g. person, location, etc.), to link it or disambiguate it to a specific entity, and to resolve every form of mentioning or co-referring to the same entity in a text.

The *Opinion Mining* approach in OpeNER consists of three levels: (i) generation of polarity lexicons from WordNets for each language (ii) development of polarity systems at document and sentence level based on those lexicons and (iii) feature-based opinion mining based on supervised classification and feature extraction. For

[5] http://www.kyoto-project.eu

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

hotel reviews we will be looking at *features* such as room service, cleanliness, etc.

As we are working with 6 languages, it would be convenient, where possible, to use *one solution*, *one tool* and *one methodology* to provide most of the NLP annotation, including not only NERC, Coreference and NED, but also language identification, tokenisation, POS tagging, lemmatisation, and parsing. Otherwise, maintaining so many different tools for every annotation process would be far too cumbersome to provide easy-to-use integrated NLP pipelines in a virtual machine. Thus, every NLP component (except Opinion Mining) in the English and Spanish pipelines are being developed using the Apache OpenNLP API[7] for supervised Machine Learning based linguistic annotators: Sentence Segmentation, Tokenisation, Part of Speech tagging, NERC and Constituent Parsing. The Consortium is training new models for every component using the usual general domain datasets such as CoNLL and Evalita datasets for NERC, Penn Treebank WSJ for English POS and Constituent Parsing, Ancora[8] for Spanish POS and Constituent Parsing. Furthermore, lemmatisation is performed using word form and POS tags lookups in a dictionary for each language.

With respect to coreference, the Stanford multi-sieve pass system (Lee et al., 2013) is being re-implemented in such a manner that it facilitates its adaptation to other languages. The coreference module takes KAF containing POS and NERC annotated tokens and a constituent parse tree as input. The sieves are implemented in a way that the only requirements to adapt to another language is to change the POS and Parsing tagsets and a number of static dictionaries that contains information such as demonyms, gender, number, etc. Finally, the NED systems are being adopted from the English DBpedia Spotlight[9] (Mendes et al., 2011) which is based on DBpedia and Wikipedia for disambiguation of Named Entities.

5 Concluding Remarks

This paper presents OpeNER, a European project that will provide completely 'off-the-box' usable language analysis tool chains for six languages to make sense out of unstructured text via Natural Language Processing. These chains will easily be incorporated by SMEs in their workflow for applications such as Reputation Management and Information Access. Furthermore, on the second year of the project the toolchain will be adapted to process texts from the Tourist domain. To this purpose, the project will also investigate how the performance of the OpeNER toolchains can be improved by inter-relating with each other the various layers of linguistic annotation.

[7] http://opennlp.apache.org/

[8] http://clic.ub.edu/corpus/es/ancora

[9] https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 296451

References

Bosma, W., P. Vossen, A. Soroa, G. Rigau, M. Tesconi, A. Marchetti, M. Monachini, and C. Aliprandi. 2009. Kaf: a generic semantic annotation format. In *Proceedings of the Generative Lexicon (GL2009) Workshop on Semantic Annotation*, Pisa, Italy.

Gretzel, Ulrike and Kyung Hyan Yoo. 2008. Use and impact of online travel reviews. In *Information and communication technologies in tourism 2008*. Springer, page 3546.

Hu, M. and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, page 168177.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, pages 1–54, January.

Mendes, P. N., M. Jakob, A. Garca-Silva, and C. Bizer. 2011. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, page 18.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1135.