Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 28 (2014) 56–75

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis☆

Alexandra Balahur[a,*], Marco Turchi[b,1]

[a] European Commission Joint Research Centre, IPSC, GlobeSec, OPTIMA, Via E. Fermi 2749, Ispra, Italy
[b] Fondazione Bruno Kessler-IRST, Via Sommarive, 18, Povo, Trento, Italy

Received 27 August 2012; received in revised form 25 February 2013; accepted 27 March 2013
Available online 18 April 2013

Abstract

Sentiment analysis is the natural language processing task dealing with sentiment detection and classification from texts. In recent years, due to the growth in the quantity and fast spreading of user-generated contents online and the impact such information has on events, people and companies worldwide, this task has been approached in an important body of research in the field. Despite different methods having been proposed for distinct types of text, the research community has concentrated less on developing methods for languages other than English. In the above-mentioned context, the present work studies the possibility to employ machine translation systems and supervised methods to build models able to detect and classify sentiment in languages for which less/no resources are available for this task when compared to English, stressing upon the impact of translation quality on the sentiment classification performance. Our extensive evaluation scenarios show that machine translation systems are approaching a good level of maturity and that they can, in combination to appropriate machine learning algorithms and carefully chosen features, be used to build sentiment analysis systems that can obtain comparable performances to the one obtained for English.

Keywords: Multilingual sentiment analysis; Opinion mining; Machine translation; Supervised learning

1. Introduction

Together with the increase in the access to technology and the Internet, the recent years have shown a steady growth of the volume of user-generated contents on the Web. The diversity of topics covered by this data (also containing expressions of subjectivity) in the new textual types such as blogs, fora, microblogs, has been proven to be of tremendous value to a whole range of applications, in Economics, Social Science, Political Science, Marketing, to mention just a few. Notwithstanding these proven advantages, the high quantity of user-generated contents makes this information hard to access and employ without the use of automatic mechanisms. This issue motivated the rapid and steady growth in interest from the natural language processing (NLP) community to develop computational methods to analyze subjectivity and sentiment in text. Additionally, apart from the research on sentiment analysis in the context of user-generated contents, studies have also focused on developing methods for sentiment analysis in newspaper articles. This

☆ This paper has been recommended for acceptance by Prof. R.K. Moore.

* Corresponding author. Tel.: +39 0332 785808.
E-mail addresses: alexandra.balahur@jrc.ec.europa.eu (A. Balahur), turchi@fbk.eu (M. Turchi).

[1] Work developed while working at the European Commission Joint Research Centre, Ispra, Italy.

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.03.004

task is especially relevant to the online reputation management of public figures and organization and to monitoring the reaction to the events described in mainstream media. As such, different methods have been proposed to deal with these phenomena for the distinct types of text and domains, reaching satisfactory levels of performance for English. Nevertheless, for certain applications, such as news monitoring, the information in languages other than English is also highly relevant and cannot be disregarded. Additionally, systems dealing with sentiment analysis in the context of monitoring must be reliable and perform at similar levels as the ones implemented for English.

Although the most direct solution to these issues of multilingual sentiment analysis would be the use of machine translation systems, researchers in sentiment analysis have been reluctant to using such technologies due to the low performance they used to have. However, in the past years, the performance of machine translation systems has steadily improved. Public or open access solutions (e.g. Google Translate,[2] Bing Translator[3]) offer more and more accurate translations for frequently used languages.

Bearing these thoughts in mind, in this article we study the manner in which sentiment analysis can be done for languages other than English, using machine translation. In particular, we will study this issue in three languages – French, German and Spanish – using three different machine translation systems – Google Translate, Bing Translator and Moses (Koehn et al., 2007) and different machine learning models.

We employ these systems to obtain training and test data for these three languages and subsequently extract different features that we employ to build different machine learning models using Support Vector Machines Sequential Minimal Optimization – SVM SMO – (Platt, 1999). We additionally employ meta-classifiers to test the possibility to minimize the impact of noise (incorrect translations) in the obtained data. To have a more precise measure of the impact of quality translation on this task, we create Gold Standard sets for each of the three languages, by translating the data with the Yahoo translation system[4] and subsequently manually correcting the output.

Our experiments show that machine translation systems are reaching a reasonable level of maturity so as to be employed for multilingual sentiment analysis and that for some languages (for which the translation quality is high enough) the performance that can be attained is similar to that of systems implemented for English, in terms of weighted *F*-measure.

2. Related work

The work presented herein is related to two different directions of research in NLP: multilingual sentiment analysis and the use of machine translation for multi- and cross-lingual tasks in NLP. The contributions in these two research directions that are relevant to the present research are presented in the following subsections.

2.1. Multilingual sentiment analysis

Most of the research in subjectivity and sentiment analysis was done for English. However, there were some authors who developed methods for the mapping of subjectivity lexicons to other languages. To this aim, Kim and Hovy (2006) use a machine translation system and subsequently employ a subjectivity analysis system that was developed for English to create subjectivity analysis resources in other languages. Ahmad et al. (2007) use the topical distributions in different languages to detect important sentiment phrases in a multilingual setting, starting from the idea that words with a lower frequency are more representative of the topic and searching for sentiment-related terms around those. Inui and Yamamoto (2011) employ machine translation and, subsequently, sentence filtering to eliminate the noise obtained in the translation process, based on the idea that sentences that are translations of each other should contain sentiment-bearing words that have the same polarity. Mihalcea et al. (2007) propose a method to learn multilingual subjective language via cross-language projections. They use the Opinion Finder lexicon (Wilson et al., 2005) and use two bilingual English-Romanian dictionaries to translate the words in the lexicon. Another approach was proposed by Banea et al. (2008b). To this aim, the authors perform three different experiments – translating the annotations of the MPQA corpus, using the automatically translated entries in the Opinion Finder lexicon and the third, validating

[2] http://translate.google.it/.
[3] http://www.microsofttranslator.com/.
[4] http://www.babelfish.com/.