A Multi-Layered Framework for Building Multilingual Sentiment Lexicons

Mohammed Kaity and Vimala Balakrishnan
Department of Information System, Faculty of Computer Science & Information Technology, (University of Malaya), Kuala Lumpur, Malaysia
E-mail: moh.kaity@gmail.com; vimala.balakrishnan@um.edu.my

ABSTRACT
Sentiment analysis is a field of science that has contributed to extract and analyse public mood and views. However, it mainly focuses on building systems and resources in English, which caused delays in the exploit of more than 73% of data written in languages other than English. In this regard, this paper proposes a multi-layered framework for building multilingual sentiment lexicons. The layers divide based on the input resource to i) lexicon-based layer, ii) corpus-based layer, iii) human-based layer. Each layer sends its output as input to the next layer. Ultimately, a new lexicon of the target language is built and extended to be used in sentiment analysis tasks.

Keywords: Sentiment Analysis; Sentiment Lexicons; Multilingual; lexicon-based; corpus-based.

INTRODUCTION
In cyberspace, Internet users feel open and free to express their views and opinions about products, social issues, policies, and so on. Accordingly, electronic commerce such as Amazon, online reviewing sites, and social media such as Twitter and Facebook contain billions of comments and revisions that are written daily on the Internet. Thus, the Web has become a massive warehouse of people’s opinions and emotions (Liu, 2012). Sentiment analysis is the research area that deals with the analyzing and evaluating of those facts and opinions. It is considered as one of the most popular research areas that use NLP techniques, text analysis, and computational linguistics to identify the text polarity (i.e. positive, negative or neutral) based on the sentiment expressed in the text (Akhtar, Gupta, Ekbal, & Bhattacharyya, 2017). Because of the urgent need to know the trends of the public on a particular subject, sentiment analysis has become one of the most popular research areas over the past few years (Liu, 2012). Sentiment analysis systems deal with this huge amount of data and identify the users opinions and emotions (Lo, Cambria, Chiong, & Cornforth, 2016) (Liu, 2012). Although data are written in a number of languages, the biggest focus has been on the English language. In English, there are many resources of sentiment analysis whereas many of the other languages lack these sources. Therefore, sentiment analysis continues to be

1 http://www.internetworldstats.com/stats7.htm
A Multi-Layered Framework for Building Multilingual Sentiment Lexicons

one of the substantial research domains given its many challenges and language difficulties but also promise (Lo et al., 2016).

Sentiment classification approaches can be assembled into two main categories: lexicon-based and machine learning-based (Biltawi, Etaiwi, Tedmori, Hudaib, & Awajan, 2016). The classifiers that are based on sentiment lexicons are called lexicon-based or rule-based classifiers, while the machine learning-based classifiers depend on training datasets or annotated corpus. Lexicon-based approach is generally dependent on sentiment lexicons that build manually or semi-automatically (Lo et al., 2016). Creating these resources and making it available help researchers to construct training data for sentiment classification tasks. This task that aims at creating sentiment lexicons, dictionaries and corpora are called resources building. (Medhat, Hassan, & Korashy, 2014). Resources building could help to improve sentiment analysis tasks and increase classification accuracy, but actually, it is not a sentiment analysis task (Medhat et al., 2014). Thus, sentiment lexicons and annotated corpus are crucial to build any sentiment analysis system (Sun, Luo, & Chen, 2017).

A sentiment lexicon is basically a list of words and each word has semantic or polarity. They can be created either manually or semi-automatically. Generating them manually is very time consuming, but accurate. Differently, the semi-automatically methods usually combine manually and automated approaches to extract the sentiment lexicon words and terms (Biltawi et al., 2016). To that end, this paper proposes a multi-layered framework for building multilingual sentiment lexicons.

LITERATURE REVIEW

The existing sentiment lexicons did not cover most of the major languages. A few languages have some attempts to build their sentiment lexicons (Lo et al., 2016). Methods for building sentiment lexicons vary from being completely manual, semi-automatic, to limited automatic approaches (Nusko, Tahmasebi, & Mogren, 2016). Based on data sources, there are three main approaches as follow: lexicons-based approach, corpus-based approach, and human-based approach.

Due to the availability of many resources in the English language, a number of researchers have been adopted on these English resources (Dashtipour et al., 2016). One of the most important methods that benefited from the previous lexicons is the translation (Denecke, 2008). One of the earlier studies that rely on existing lexicons to construct a sentiment lexicon for non-English is the work that conducted by Yao, Wu, Liu, and Zheng (2006). Their proposed method was to build Chinese sentiment lexicon by using a bilingual dictionary to locate the sentiment orientation of the words. Mihalcea, Banea, and Wiebe (2007) benefited from the available tools and resources in English lexicon called Opinion-Finder (Wiebe & Riloff, 2005) to build a Romanian language sentiment lexicon. Moreover, Multilanguage translation approach was provided by Steinberger et al. (2012). They proposed an approach to build a sentiment lexicon called (triangulation) referring to the intersection of translation between three languages. Abdouaui, Azé, Bringay, and Poncelet (2016) presented an idea based on machine translation to build and elaborate a new French sentiment and emotion lexicon called (FEEL). There are other methods provided by some researchers. These methods may be similar or differ in some respects, but all of them based on translation; whether automatic or manual translation; to create the sentiment lexicons, such as (Hammer,
Bai, Yazidi, & Engelstad, (2014) and (Al-Twäresh, Al-Khalifa, & Al-Salman, 2016). When using machine translation, several errors may arise due to cultural differences about the sentiment orientations of words. On the other hand, some researchers used relationship-based approach to extract new words in order to build sentiment lexicon for non-English languages. This approach starts with a small group of core words (seeds) that expand by using the semantic relations between words (i.e. synonyms and antonyms) in an existing dictionary. Random walk (Rosell & Kann, 2010) and bootstrapping algorithms (Banea, Wiebe, & Mihalcea, 2008) have been used in this approach. Mahyoub, Siddiqui, and Dahab (2014) presented an algorithm that assigns sentiment scores to the entries found in the Arabic WordNet to create an Arabic sentiment lexicon. Their idea started from using small seed list of polar words (i.e. positive or negative words). After preparing the seed list, they applied a semi-supervised learning algorithm to increase the number of entries on the Arabic WordNet by exploiting the synset relations. Similarly, Nusko et al. (2016) presented their method to build a sentiment lexicon for the Swedish language. They started with a small group of core words (seeds) that were expanded by using the semantic relations between words in a lexical resource of modern Swedish called SALDO. In this approach, the obtained sentiment lexicons do not contain many words or shortcuts that are used on social networking sites.

Apart from the use of dictionaries and lexicons, some researchers have used the corpus to extract polar words from the text in order to construct sentiment lexicons. Remus, Quasthoff, and Heyer (2010) used the co-occurrence analysis of corpus that was collected from a business partner according to their paper. AraSenti-PMI is an Arabic sentiment lexicon that was built by using the pointwise mutual information (PMI) measure in a dataset of tweets by Al-Twäresh et al. (2016). The use of corpus faces a number of problems such as the lack of data pre-processing tools in many languages. Moreover, using the corpus to build sentiment lexicons requires a large annotated corpus volume.

Another idea based on encouraging people to build sentiment lexicon in their languages. This is done by creating crowdsourcing platforms to motivate people to answer questions or a puzzle in order to benefit from these answers in the construction of the sentiment lexicons (Hong, Kwak, Baek, & Moon, 2013). The systems often developed by using crowdsourcing platforms such as Amazon Mechanical Turk² (Mohammad & Turney, 2013) or by building games with a purpose (Hong et al., 2013). However, production of these lexicons is time-consuming and requires a large number of people and is costly.

RESEARCH DESIGN

Problem statement
Sentiment lexicons are valuable resources for opinion mining tasks for any language. However, research in building sentiment lexicons has mainly focused on the English language (Nusko et al., 2016). Due to the lack of those sentiment resources, there have been few studies in sentiment analysis in non-English languages (Dashtipour et al., 2016). Moreover, the current methods, such as machine translating, transfer learning, and bootstrapping based methods, have their respective drawbacks (Sun et al., 2017).

²https://www.mturk.com/
A Multi-Layered Framework for Building Multilingual Sentiment Lexicons

Objectives
This study aims to propose an integrated framework that can combine the three resources of building sentiment lexicons (lexicons, corpus, and human) to build large-scale lexicons for non-English languages. Additionally, to develop an effective method based on corpus to discover new polar words affecting sentiment analysis in non-English languages.

Methodology
A multi-layered framework integrating lexicons, corpus and the human effort is designed as shown in Figure 1 below.

The first layer is lexicon-based. Machine translation tools will be used to translate a source language lexicon (SentiWordNet3) in order to create an initial lexicon for the target language. The initial lexicon will be cleaned from any repeating entries. The initial lexicon will be the input of the second layer. Then, a large corpus will be crawled from social media (i.e. Twitter and Facebook) and it will be added to the second layer. This stage will be done repeatedly to expand the initial lexicon from the corpus by searching for new polar words or confirming the existing words. The output of this layer will be

---

3 http://sentiwordnet.isti.cnr.it/
CONCLUSION AND FUTURE WORK

The purpose of this work is to make use of all available resources for the construction of non-English sentiment lexicons. At the beginning, translating a sentiment lexicon from English into the target language. This will represent 50% of the final lexicon size. Then a large corpus will be collected to expand the lexicon and will represent 40% of the volume of the lexicon. Finally, the obtained lexicon will be reviewed by human experts to add or remove some polar words. Manual work will represent 10% of the effort to build the lexicon. In our future work, we aspire to develop a method based on our proposed framework and apply the method to a number of languages to build and expand non-English sentiment lexicons. Furthermore, we will evaluate our work by comparing it with the current methods.

ACKNOWLEDGMENTS

The first author is supported by Sheikh Thani bin Abdullah Al Thani Scholarship - Malaysia and Alandalus University for Science and Technology - Yemen.

REFERENCES


A Multi-Layered Framework for Building Multilingual Sentiment Lexicons


Mohammed Kaity and Vimala Balakrishnan
