Expert Systems with Applications 41 (2014) 5984–5994

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Building layered, multilingual sentiment lexicons at synset and lemma levels

Fermín L. Cruz *, José A. Troyano, Beatriz Pontes, F. Javier Ortega

Department of Languages and Computer Systems, University of Seville, Av. Reina Mercedes s/n, 41012 Sevilla, Spain

ARTICLE INFO

Keywords:
Sentiment analysis
Multilingual sentiment lexicons
Spanish resources for sentiment analysis

ABSTRACT

Many tasks related to sentiment analysis rely on sentiment lexicons, lexical resources containing information about the emotional implications of words (e.g., sentiment orientation of words, positive or negative). In this work, we present an automatic method for building lemma-level sentiment lexicons, which has been applied to obtain lexicons for English, Spanish and other three official languages in Spain. Our lexicons are multi-layered, allowing applications to trade off between the amount of available words and the accuracy of the estimations. Our evaluations show high accuracy values in all cases. As a previous step to the lemma-level lexicons, we have built a synset-level lexicon for English similar to SentiWordNet 3.0, one of the most used sentiment lexicons nowadays. We have made several improvements in the original SentiWordNet 3.0 building method, reflecting significantly better estimations of positivity and negativity, according to our evaluations. The resource containing all the lexicons, ML-SentiCon, is publicly available.

1. Introduction

Sentiment analysis is a modern subdiscipline of Natural Language Processing which deals with subjectivity, affects and opinions in texts (a good survey on this subject can be found in Pang & Lee (2008) and Liu & Zhang (2012)). It is a very active research area, since opinions expressed on the Internet by users constitute a very valuable information for governments, companies and consumers, and its large volume and the high rate of appearance require automated analysis methods. Detection of subjectivity, text classification based on the overall sentiment expressed (positive vs. negative), or extraction of individual opinions and their participants, are three of many tasks addressed. Some of these tasks rely on *sentiment lexicons* as a component of the solutions.

A sentiment lexicon is a lexical resource containing information about the emotional implications of words. Commonly, this information refers to the prior polarity (positive vs. negative) of words, i.e. the positive or negative nature of words, regardless of context. For example, the word "good" has a positive prior polarity, although it may be used in a negative sentence ("His second album is not so good"). In this paper we present new sentiment lexicons for English, Spanish and other three official languages in Spain. The lexicons are multi-layered, allowing applications to trade off between the amount of available words and the accuracy of the estimations of their prior polarities. As a previous step, we have reproduced the method proposed by Baccianella, Esuli, and Sebastiani (2010) to build SentiWordNet 3.0, one of the most used sentiment lexicons nowadays. We have introduced several improvements to the original method, affecting positively the accuracy of the resource obtained, according to our evaluations.

We believe that the resource containing all the lexicons, ML-SentiCon, can be useful in many sentiment applications for both English and Spanish. The automatic method proposed here could also be reproduced for new languages, whenever WordNet versions for those languages are available. This is advantageous in that it allows to quickly obtain sentiment lexicons for new languages that lack such resources. However, it should also be noted that any lexicon constructed by automatic or semi-automatic methods must be used with caution, as they will inevitably contain errors (words incorrectly labelled as positive or negative). In this sense, it is a good practice to have the lexicons reviewed by native speakers. In the case of ML-SentiCon, layers 1–4 have been completely reviewed. Although the remaining layers have not been reviewed, evaluations based on statistically representative random sample indicate a tolerable error rate up to layer 7 (see Section 4.3 for details).

The structure of the paper is as follows. In Section 2, we review some related works on sentiment lexicons, including a description

* Corresponding author. Address: Escuela Técnica Superior de Ingeniería Informática, Av. Reina Mercedes s/n, 41012 Sevilla, Spain. Tel.: +34 954 55 62 33.
E-mail address: fcruz@us.es (F.L. Cruz).

http://dx.doi.org/10.1016/j.eswa.2014.04.005
0957-4174/

of the method used to build SentiWordNet 3.0. Some references to works on non-English sentiment lexicons are also included. In Section 3, we describe our SentiWordNet-based method, and compare the lexicon obtained with the original SentiWordNet 3.0. Section 4 explains the steps followed to obtain the layered, multilingual sentiment lexicons, and shows some results concerning the evaluation of the resource. Finally, in Section 5 we point out some conclusions and final remarks.

2. Related works

There exist many works that deal with the creation of sentiment lexicons with different approaches. General Inquirer (Stone, Dunphy, & Smith, 1966) can be considered, among other things, the first sentiment lexicon. It is a hand-made lexicon constituted by lemmas. Lemmas are semantic units that can appear in multiple lexicalized forms, e.g. the verb "approve" is a lemma that can be found in texts with different inflections, like "approved" or "approving". General Inquirer includes a great amount of information (syntactic, semantic and pragmatic) related to each lemma. Among all this information, there are 4206 lemmas which are tagged as positive or negative. In spite of its age, General Inquirer is still widely used in many works on Sentiment Analysis.

MPQA Subjectivity Lexicon (Wilson, Wiebe, & Hoffmann, 2005) is an example of a piece of work based on General Inquirer. In particular, it is a lexicon which comprises, in addition to the positive and negative words from General Inquirer, a set of automatically compiled subjective words (Riloff & Wiebe, 2003) and also other terms obtained from a dictionary and a thesaurus. It totals 8221 words, whose polarities (positive, negative or neutral) were manually annotated. The resulting list contains 7631 positive and negative elements, and it is very heterogeneous as it is comprised of both lemmas and inflections. As in General Inquirer, this list does not include multi-words, i.e. terms constituted by more than one word.

According to the number of cites, the two most used lexicons nowadays are Bing Liu's Opinion Lexicon (Hu & Liu, 2004; Liu, Hu, & Cheng, 2005) and SentiWordNet (Baccianella et al., 2010; Esuli & Sebastiani, 2006). They are two very different approaches and, to some extent, opposed. Bing Liu's lexicon is formed by 6800 inflections, including mispellings and slangs (informal expressions frequently used on the Internet). On the other hand, SentiWordNet is built based on WordNet (Fellbaum, 1998), a lexical resource where the basic units, the so-called synsets, comprise a set of words which share the same meaning. Bing Liu's lexicon is built using an automatic method, but the lists of positive and negative words have been manually updated until the current version available on the web, which dates from 2011. On the contrary, SentiWordNet assigns real values, between 0 and 1, representing positive or negative polarities to each of the +100 K synsets of WordNet. These values have been automatically computed based on two sets of positive and negative seeds, respectively.

It is worthy to note the difference between word-level and lemma-level lexicons, like General Inquirer, MPQA Subjectivity Lexicon or Bing Liu's Opinion Lexicon, and the synset-level lexicons like SentiWordNet. The first ones are formed by terms with semantic ambiguity due to the polysemy of many words. On the contrary, the synset-level lexicons do not have this problem because their basic units univocally represent one meaning. Nevertheless, the use of this kind of lexicons makes it necessary to pre-process the texts with a Word Sense Disambiguation tool, which has a relatively low accuracy nowadays. Most of the works using SentiWordNet compute the polarity at the level of words or lemmas by aggregating the polarity values from all the respective synsets (Agrawal et al., 2009; Denecke, 2008; Desmet & Hoste, 2013; Kang, Yoo, & Han, 2012; Martín-Valdivia, Martínez-Cámara, Perea-Ortega, & Alfonso Ureña-López, 2012; Saggion & Funk, 2010; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). In our work, we face the building of lexicons of both types: at a synset level, adding some improvements to the method implemented for SentiWordNet 3.0; and also at a lemma level, using the values computed in the synset-level lexicon.

2.1. SentiWordNet 3.0

The current version of SentiWordNet (Baccianella et al., 2010) assigns positivity and negativity values between 0 and 1 to each synset in WordNet 3.0. It uses an automatic method divided in two steps (Fig. 1).

In the first step, the polarities of synsets are estimated individually: various ternary classifiers are trained, which are able to classify each synset as positive, negative or neutral, depending on the words contained in the definition of the synset (WordNet provides a definition, the so-called "gloss", for each synset). Starting from some positive and negative seeds, and after applying an expansion method, different training sets are obtained. Then, standard techniques for text classification are applied (tf-idf vectorial representation of the glosses, plus SVM and Rocchio algorithms). Finally, the resulting classifiers are applied to each synset in WordNet and their positivity and negativity scores are computed from the outputs of each classifier.

In the second step, these scores are globally refined. A graph of synsets is built, where nodes n_i correspond to each synset s_i, and an edge from n_i to n_j is created if, and only if, the synset s_i appears in the definition of the synset s_j. Note that WordNet glosses are non-disambiguated texts, so it is necessary the use of Princenton WordNet Gloss Corpus,[1] a resource where WordNet glosses are partially disambiguated. This graph is defined as a part of the inverse flow model (Esuli & Sebastiani, 2007). The intuition behind is the assumption that those synsets whose definition contains positive synsets are likely to be positive, and analogously for the negative ones. In the inverse flow model, a variation of the random-walk algorithm PageRank (Page, Brin, Motwani, & Winograd, 1998) is applied to the graph. This algorithm propagates the positivity scores computed in the previous step through the edges of the graph, in order to obtain the positivity values for each synset. The same process is applied to the negativity scores in a second computation of the algorithm.

2.2. Non-English sentiment lexicons

Although there are not many sentiment lexicons for other languages than English, the number is growing slowly. There exist works focused on the creation of sentiment lexicons for very diverse languages, such as Hindu and French (Rao & Ravichandran, 2009), Arabian (Abdul-Mageed, Diab, & Korayem, 2011), German (Clematide & Klenner, 2010), Japanese (Kaji & Kitsuregawa, 2007), Chinese (Lu, Song, Zhang, & Tsou, 2010; Xu, Meng, & Wang, 2010), Romanian (Banea, Mihalcea, & Wiebe, 2008) and Spanish. Two Spanish lexicons are automatically built in Brooke, Tofiloski, and Taboada (2009) from an English sentiment lexicon by using two resources: a bilingual dictionary[2] and Google Translator.[3] The authors do not show any evaluation of the resulting lexicons, but they provide the results obtained by a sentiment classification tool based on them. A similar technique is used in Molina-González, Martínez-Cámara, Martín-Valdivia, and Perea-Ortega (2013), where an automatic translation process (from English

[1] http://wordnet.princeton.edu/glosstag.shtml.
[2] http://www.spanishdict.com.
[3] http://translate.google.com.