Automatic Expansion of a Social Judgment Lexicon for Sentiment Analysis

Mário J. Silva, Paula Carvalho, Carlos Costa and Luís Sarmento

DI–FCUL–TR–2010–08

DOI:10455/6694

(http://hdl.handle.net/10455/6694)

December 2010

Automatic Expansion of a Social Judgment Lexicon for Sentiment Analysis

Mário J. Silva[1], Paula Carvalho[1], Carlos Costa[1] and Luís Sarmento[2]

[1] University of Lisbon, Faculdade de Ciências, LASIGE,
Campo Grande, Lisboa, Portugal
{pcc, mjs,cc}@fc.ul.pt
[2] Faculdade de Engenharia da Universidade do Porto,
Rua Dr. Roberto Frias, Porto, Portugal
las@fe.up.pt

Abstract. We present a new method for automatically enlarging a sentiment lexicon for mining social judgments from text, i.e., extracting opinions about human subjects. We use a two-step approach: first, we find which adjectives can be used as human modifiers, and then we assign their polarity attribute. To identify the human modifiers, we developed a set of hand-crafted lexico-syntactic rules representing elementary copular and adnominal constructions where such predicates can be found. These are applied to a large n-gram collection, gathering evidence about human/not-human adjective behavior. The adjective and rule frequencies are then used as input features to a statistical classifier. In the second and final stage, we assign polarities to human adjectives, by exploring the graph of synonyms. We calculate the shortest distances of each adjective with unknown polarity to the human adjectives of each sentiment class with prior polarity in the initial lexicon. The computed distances for each adjective are then used as the input features of a statistical polarity classifier. The application of the proposed method to a collection of manually annotated adjectives showed its effectiveness for developing a fine-grained sentiment lexicon for social judgment, which can be seen as a human domain-specific alternative to general-purpose lexicons. We focus on the polarity classification of Portuguese human adjective predicates, but the studied method could be applied to other languages and semantic predicates. The lexicon of human adjectives with polarities produced with this method is now available as an open resource.

Keywords: Sentiment Analysis, Opinion Mining, Sentiment Lexicon, Human Predicates

1 Introduction

Synonyms have identical semantic orientation, which has motivated different authors to explore synonymy in language resources for automatically enlarging their sentiment lexicons. In general, lexicon-based approaches start from a confined list of manually annotated polar words, which are then used as seeds for finding new polar items. Classical methods typically explore *WordNet* [1] to determine the polarity of new sentiment words with which seed words are semantically related. However, many languages, including Portuguese, do not have comprehensive and publicly available *WordNets*, leaving that option out.

Typically, the polarity of known polar words is propagated to the elements belonging to the same *synsets*. However, such procedure is fallible because identical lexical forms (or homographs) can make part of different *synsets*, which in turn could present a diversity of meanings and polarities. Additionally, the majority of publicly available lexical resources describing synonymy only provide the potential synonyms for a given word, without defining the context(s) where they have an identical interpretation. For example, the adjective *fresh* can be used as modifier of a human noun, and it can be replaced by adjectives such as *impertinent* or *impudent*, which have a negative semantic orientation. On the other hand, it can modify non-human nouns, exhibiting in that case an opposite polarity. For example, when combined with an abstract noun such as *portrayal*, *fresh* is interpretable as *new* or *novel*, conveying a positive semantic orientation.

The above example illustrates that the polarity of predicates can change according to the syntactic-semantic nature of the nouns they relate to. This means that polarity assignments will only be successful if such combination constraints are taken into account.

We present a methodology for automatically enlarging a sentiment lexicon for mining social judgments from text, i.e., extracting opinions and sentiments about human subjects. In this paper, we address Portuguese human adjective predicates (i.e. adjectives modifying human nouns), but the methodology proposed is language independent and can be applied to a wider range of syntactic-semantic predicates.

We start from (i) a lexicon of adjectives, which is partially classified by a linguist with semantic category (human, not-human) and polarity (positive, negative, neutral), and (ii) a set of publicly-available synonym dictionaries.

We use a two-step approach to accomplish such goal: first, we find which adjectives can be used as human modifiers, and, next, assign them a polarity attribute.

The identification of human adjectives is performed through the combination of a linguistic-based strategy, for extracting human adjective candidates from corpora, and a machine learning strategy, for automatically selecting the best human candidates, and classifying them accordingly. In particular, we design a small library of hand-crafted lexico-syntactic rules representing elementary copular and adnominal constructions where human adjectives can be found. These rules are applied to a large *n-gram* collection, in order to gather evidence about human/not-human adjective behavior. The adjective and rule frequencies are then used as features to train a human classifier.

In the second and final step, we create an expanded polarity lexicon of human adjectives, derived from exploring a graph of adjective synonyms built as the union of multiple open thesauri. We calculate the distance in the graph of each adjective whose polarity is unknown to the human adjectives already classified in initial lexicon. The computed distances are then used to generate input features for an automatic polarity classifier.

The remainder of the paper is organized as follows: Section 2 presents related works; Section 3 describes the input linguistic resources used in our experiments; Sections 4 and 5 detail the methods used for identifying human adjectives and classifying their polarities, respectively; evaluation results are presented and discussed in Section 6; Section 7 presents our conclusions.

2 Related Work

Initial approaches to the construction of sentiment lexica aimed at identifying the subjective lexical units in general language and determining their semantic orientation (or polarity). The pioneer work of Hatzivassiloglou and McKeown tackles the problem of determining the semantic orientation of adjectives by exploring the co-occurrence of positive and negative adjectives with other expressions in corpora [2]. Such co-occurrences are namely sought in the scope of copulative constructions (whose constituents tend to be coherent in terms of polarity) and adversative constructions (whose constituents tend to exhibit different polarity). After combining these constraints across many adjectives, the authors use a clustering algorithm that separates the adjectives into different groups, which are then labeled as positive or negative. Turney and Littman propose a bootstrapping method for inferring the semantic orientation of new polar words by computing the *pointwise mutual information* (PMI) between each target word and an initial set of previously classified positive and negative paradigm words (or seeds) [3]. On the other hand, Riloff et al. use two bootstrapping algorithms that exploit extraction patterns to learn sets of subjective nouns [4]. The authors show that algorithms typically used for automatically generating extraction patterns to identify words belonging to a semantic class, can also be effective in identifying subjective words. This approach is based on the hypothesis that words of the same semantic class or category tend to occur in similar contexts.

Significant research on sentiment lexicon construction has also explored WordNet (e.g. [5], [6], [7] and [8]) and other lexical resources for languages where WordNet is not available [9], to acquire new polar words. For example, Kamps et al. try to determine sentiments of adjectives in WordNet by measuring the relative distance of each adjective to the reference positive and negative words “good” and “bad”, respectively [5]. Kim and Hovy built a sentiment classifier that also uses WordNet, but performs a tri-polarity classification (positive, negative and neutral), based on a manually annotated dataset composed of verbs and adjectives [6].

Rao and Ravichandran treat polarity detection as a semi-supervised label propagation problem in a graph [9]. Takamura et al. exploit the gloss information associated to words in dictionaries for determining their semantic orientation [10]. They construct a lexical network by linking two words whenever one of them appears in the gloss of the other word. Semantic orientations are regarded as spins of electrons, and the mean field approximation is used to compute the approximate probability function of the system.

Despite the availability of a number of sentiment lexicons, especially for English, it has being argued that their use is frequently unsatisfactory, because they do not reflect domain-specific lexical usage [11]. Hence, different approaches have been proposed to create domain-dependent polarity lexicons, instead of general-purpose lexicons. For instance, Fahrni and Klenner propose a two-stage method for determining the sentiment analysis of adjectives in a given domain [12]. First, they use the Wikipedia for automatically detecting the candidate targets associated to adjectives in a given domain, followed by a bootstrapping approach to determine the target-specific adjective polarity. Choi and Cardie propose a method based on integer linear programming that can adapt an existing lexicon into a new specific one [11]. They consider the relations among words and opinion expressions to derive the most likely polarity of each lexical item for a given domain.

Kanayama and Nasukawa use manually-crafted syntactic patterns to identify polar atoms, which correspond to minimal human-understandable syntactic structures that specify polarity of clauses in a given domain [13]. They use clause level context coherency to find candidate words from sentences that appear consecutively with sentences containing seed sentiment words, based on the assumption that same polarities tend to appear successively in contexts.

In this paper, we propose a methodology for developing a fine-grained sentiment lexicon, which can be seen as a human-specific alternative to general-purpose lexicons. Polarities in this lexicon are assigned based on the syntactic-semantic category of targets of sentiment. The methodology only handles human adjectives presently, but it could be adapted to other syntactic and semantic categories following identical principles. As Riloff et al. and Kanayama and Nasukawa, we make use of manually-crafted lexico-syntactic patterns, but ours do not require any POS tagging or parsing. Similar to Kim and Hovy, we use a manually annotated adjectives lexicon for assigning positive, negative and neutral polarity to adjectives linked to these ones by a synonymy relationship. As suggested by Rao and Ravichandran, we use information aggregated from several publicly-available thesauri. This helps us dealing with the lack of a comprehensive WordNet-like resource for Portuguese, the target language of our work.

3 Input Linguistic Resources

The identification of human adjective candidates relies on a small set of lexical resources: a lexicon of adjectives (Section 3.1) and dictionaries of person names, professions and organizational titles (Section 3.2). These are used for extracting candidate adjectives from a large *n-gram* collection (Section 3.4). The lexicon of adjectives is also used, together with a dictionary of synonyms (Section 3.3), in the polarity assignment stage.

3.1 Lexicon of Adjectives

The tasks of recognition, classification and evaluation of polar human adjectives in our environment use a lexicon of 24,792 adjectival lemmas, which are partially annotated with their semantic category (human, not-human) and polarity (positive, negative, neutral).

In detail, 4,546 entries include information about their possible semantic category: 4,034 adjectives were manually assigned to the human attribute and the remaining

511 lemmas to the non-human attribute. Human adjectives are characterized as co-occurring with a human subject (e.g. *the prime-minister is popular*). On the contrary, this type of subject is interdicted with non-human adjectives (e.g. ?*the prime-minister is sporadic*).

Human adjectives were manually labeled with their prior polarities, which may be positive (1), negative (-1) or neutral (0). In terms of polarity distribution, 56% of the entries were labeled as negative (2,242 lemmas), 19% as positive (785 lemmas) and the remaining 25% as neutral (1,009 lemmas).

Polar adjectives were mostly collected from a lexico-syntactic database of human intransitive adjectives available in contemporary European Portuguese [14]. Such resource systematically describes the syntactic and semantic properties of 4,250 lemmas and morpho-syntactically associated nouns (e.g. *stupid/stupidity*).

3.2 Dictionaries of Names and Professions

In addition to the above lexicon, the patterns that we have identified make also use of a set of dictionaries that we have created from public information resources:

Names and surnames. These were collected from the public lists of placed secondary teacher names in the 2009 recruitment, available from the Portuguese Ministry of Education website. We obtained a list of 562 proper names and 1,388 surnames. First names correspond to the first element in name combinations. After removing all possible prepositions and conjunctions, we extracted all the tokens from name combinations after the first two (in Portuguese countries, many people have two given names), and classified them as surnames.

Professions and Professional Titles. We created a dictionary composed by 383 lemmas (~1,200 inflected forms) denoting a professional title (*engineer*) or organizational position (*prime-minister*). Dictionary entries have been semi-automatically compiled from news corpora, by exploring syntactic structures where such type of nouns typically occurs (e.g. in apposition to a human named entity).

3.3 Dictionary of Synonyms

To expand the original polarity lexicon of human adjectives, we explored synonymy among adjectives in different publicly-available thesauri for Portuguese. We specifically used PAPEL 2.0 [15], TeP [16] and DicSin.

PAPEL is composed by 96,538 lemmas, which are organized in 141,752 binary semantic relations. Those relations were semi-automatically extracted from one of the most popular Portuguese dictionaries (Dicionário *PRO de Língua Portuguesa* from Porto Editora). It contains 79,161 pairs of synonyms, 19,073 corresponding to adjectives.

TeP is a manually developed electronic thesaurus for Brazilian Portuguese. It follows the basic organizing principles of WordNet. The latest version, Tep 2.0, is composed by 43,119 lemmas organized in 19,888 *synsets*. After converting *synsets* into binary relations and removing duplicates, we obtained 77,988 pairs of synonyms. From those, 19,066 are classified as adjectives.

DicSin is a lexical database developed under the Brazilian project *BrOffice.org*, which aims to maintain and update the *OpenOffice* database of synonyms and antonyms, through a community collaborative effort. The version used in our experiments was downloaded in May 2010 and contains 22,387 lemmas and 54,813 pairs of synonyms, from which 16,194 involve adjectives.

The above mentioned resources comprise 87,327 different lemmas; from those, 35,925 correspond to adjectives, distributed in 136,913 pairs of synonyms, 36,326 involving adjectives, as summarized in Table 1.

Table 1. Synonymy distribution in the lexical resources.

Resources	Synonyms	Adjectival Synonyms
Papel	79,161	19,073
Tep	77,988	19,066
DicSin	54,813	16,194

3.4 N-gram Corpus

In our experiments, we explore WPT05, a collection of over 10 million documents from the Portuguese web [17]. We used the *n-grams* (and their frequencies) generated from the documents in the collection with language automatically identified as Portuguese (~ 7 million documents, 26 Gb of text).

We filtered the tokens with length > 32. In this corpus, we did not exclude n-grams from the set based on their frequency. Given that we will be looking for the occurrence of multiple patterns with a given lemma in our classification process, low frequency n-grams can combine to produce high frequency patterns.

This corpus contains a large and representative sample of the Portuguese documents available on the Web, including a comprehensive range of types and genres of texts. We have totals of 8 million unigrams, 501 million trigrams, 984 million tetragrams and 1,321 million pentagrams in this resource.

4 Identification of Human Adjectives

To identify human adjective candidates, we create a library of hand-crafted lexico-syntactic patterns representing elementary copular and adnominal constructions, where such predicates can be found. These are then applied to a large n-gram collection, to gather evidence about adjective behavior. The adjective and pattern frequencies in the n-gram corpus are then used as input features to a binary classifier, which is trained with the manually labeled adjectives.

4.1 Lexico-Syntactic Patterns

We identified 29 distinct lexico-syntactic patterns, which were matched against the WPT05 trigrams, quadrigrams and pentagrams (see Table 2). These patterns apply only to masculine and feminine singular forms, and some of them differ very slightly from each other, depending, for example, on the nature of the subject involved.

In the patterns, we match the subject position against dictionaries of Portuguese person names, human generic nouns (HREF), such as *pessoa* (*person*) and *indivíduo* (*individual*), profession and position names (ERGO), such as *primeiro-ministro* (*prime-minister*) and *professor* (*professor*), to the third-person singular pronouns *ele* (*he*), *ela* (*she*), *você* (*you*), which were coded as IREF, and to the second-person singular pronoun *tu* (*you*). Regarding names, we seek patterns having *first names* (N) and combinations of "*first-name surname"* (N S) and "*surname surname"* (S S).

Within these constructions, adjectives can relate to their subject through the elementary copulative verbs (COP) *ser* and/or *estar* (both translatable by *to be*) and other variants (namely, *andar*, *continuar*, *ficar*, *permancer*, *encontrar-se*, *mostrar-se*, *reveler-se*, *tornar-se* and *viver*).

Adjectives can also occur in a characterizing indefinite construction, i.e. preceded by an indefinite article, or in a predicative construction supported by the verb *ser* (*to be*). Furthermore, adjectives can be found in adnominal position, post-modifying or pre-modifying the noun they are correlated. We also account for the possibility of adjectives filling the head of a cross-construction, linking to the noun they modify by the preposition *de* (*of*).

Pre-modification of adjectives is also considered. We use of a small list of the adverbs and intensifiers (MODIF) that usually co-occur with human adjectives (e.g. *muito* (*very*), *particularmente* (*particularly*), *verdadeiramente* (*truly*)).

We confine the tense in predicative constructions to simple present, past and future (third-person singular). The constructions invoking the second-person singular (*tu, you*) only consider simple present verb forms, the most representative when using direct address pronouns.

Results. Together, the patterns used in our experiments matched 191,782 different sequences in the WPT05 n-grams, containing 8,579 different adjectival lemmas (Table 2).

Table 2. Lexico-syntactic patterns: patterns starting by 3, 4 and 5 apply, respectively, to 3-grams, 4-grams and 5-grams.

ID	Pattern	Matches	ID	Pattern	Matches
301	N COP ADJ	14,474	401	N S COP ADJ	7,240
302	IHREF COP ADJ	27,655	402	S S COP ADJ	4,466
303	HREF COP ADJ	9,157	403	N COP MODIF ADJ	2,755
304	ERGO COP ADJ	23,337	404	N é um|uma ADJ	2,218
305	tu és|estás ADJ	1,412	405	o|a ADJ do|da N	2,659
306	um|uma HREF ADJ	11,489	406	o|a ADJ do|da ERGO	5,846
307	um|uma ERGO ADJ	31,933	407	tu és|estás MODIF ADJ	616
308	um|uma ADJ HREF	1,206	408	HREF COP MODIF ADJ	2,238
309	um|uma ADJ ERGO	13,497	409	ERGO COP MODIF ADJ	4,170
	Sub-total	**134,160**	410	IREF COP MODIF ADJ	11,994

501	N S COP MODIF ADJ	805	411	IREF é um|uma ADJ	3,957
502	S S COP MODIF ADJ	465	412	HREF é um|uma ADJ	1,102
503	N S é um|uma ADJ	1,088	413	ERGO é um|uma ADJ	2,146
504	S S é um|uma ADJ	667	414	um|uma HREF MODIF ADJ	4,822
505	o|a ADJ do|da S S	698	415	um|uma ERGO MODIF ADJ	4,030
	Sub-total	**3,723**		**Sub-total**	**60,259**
	Total				**191,782**

The most productive patterns apply to trigrams (68%). On the contrary, the patterns applying to pentagrams only get 2% of the matched patterns. The most representative syntactic structure in trigrams includes an adjective in post-adnominal position that modifies an ergonym (Pattern 307). In quadrigrams, the most productive pattern accounts for a copular construction, where the adjective, preceded by an intensifier or quantifier, modifies a person pronoun (Pattern 410). In general, the most productive subject in our patterns corresponds to an ergonym or a personal pronoun.

An inspection of the obtained results showed that some constructions were unsurprisingly mismatched due to the productive homography in language, particularly in what concerns adjectives and nouns. This fact is particularly expressive in patterns 405, 406 and 505, which represent a cross-construction, where an adjective fills the head of a typical noun phrase in Portuguese.

4.2 Automatic Classification

To refine the results provided by the lexico-patterns and filter out potential erroneous cases, we first explored, for each candidate human adjective, (i) the number of matches in the corpus, and (ii) the number and type of instantiated patterns. For example, the adjective *ventilado* (*ventilated*), illustrated below, only matches Pattern 302, just once, while *impotente* (*impotent*) has a total of 97 matches, instantiating 9 different patterns that apply to trigrams and quadrigrams.

ventilado; { '302':1}
impotente; {301':4, '302':25, '303':11, '304': 24, '306':14, '307':11, '401':3, '408':4', 410':1}

According to these data, it is reasonable to infer that the adjective *impotent* is more prone to be considered as a valid human adjective than *ventilado*. On the other hand, it can also be assumed that the adjectives not matching any pattern in the entire collection are either rare in language or they have not a human behavior. In order to better assess these cases, we also associate to each potential adjective its frequency in the whole corpus.

We train a statistical classifier to automatically distinguish high-human evidence (HHE) adjectives from low-human evidence (LHE) adjectives. The automatic classification is performed based on the following identified features: (i) frequency of each pattern, (ii) total number of instantiated patterns, (iii) frequency of matches, and (iv) lemma frequency in the n-gram collection. These attributes are associated to each recognized adjective from our original lexicon, including those whose semantic category is already known.

In our experiments, the training set was composed by 4,042 entries, distributed in the following proportion: 2,580 HHE adjectives, and 1,462 LHE adjectives. The HHE adjectives correspond to the lemmas in the corpus labeled as human in the original lexicon. On the other hand, LHE adjectives include both the adjectives recognized in the corpus that are assigned to the not-human attribute in the original lexicon, and the adjectives that do not occur in the n-gram collection, regardless of their prior semantic classification (human/not-human) in the original lexicon.

5 Polarity Assignment

Once the human lemmas are identified, using the approach above described for human adjectives, we focus on assigning polarities to the lemmas with high human evidence.

The procedure starts by deriving a synonym graph, called a *syngraph*, where the nodes are the previously identified human lemmas and the edges represent synonymy relationships between lemmas. Each node is named as the concatenation of a lemma, its grammatical category and semantic class. This combination, which we designate henceforth as a qualified lemma, *qualiflemma*, was previously applied to the normalization of entries in dictionaries. For instance, the lemma *impostor* (in the initial lexicon) was converted into the qualiflemma *impostor:adj:hum*, and the synonymy relationship *brilhante SYN_OF célebre* (in PAPEL) became *brilhante:adj:hum SYN_OF célebre:adj:hum*. Having a network of qualiflemmas enables the assignment of different polarities to the same lemma used in different contexts (human/not-human, adjective/noun), and prevents propagation of synonymy relations between lemmas of distinct categories. Additionally, our polarity assignment software is able to handle all the lemmas of different categories that we may have at the same time.

To automatically assign polarities to the unlabelled qualiflemmas, we train a statistical classifier, which explores a feature vector extracted from the *syngraph*. The feature vector is derived from the distances of each qualiflemma to the nearest nodes with assigned polarities. In our experiments, we used 80% of the polar qualiflemmas for generating the syngraph, and saved the remaining 20% for learning a model to assign polarities to lemmas based on the feature vectors.

5.1 Feature Extraction

In the syngraph we have nodes with polarities, {-1, 0, 1, null}, where *null* designates unassigned polarity. We wish to learn a model that predicts the polarity of a node with null polarity given the polarity information of its neighborhood. A qualiflemma in the syngraph with unassigned polarity may possibly have adjacent nodes exhibiting the four distinct polarities, making the decision complex. A situation where all the adjacent nodes have null polarity is quite common. However, we can attempt to observe across the adjacent nodes and assign polarities based on the polarities of the cloud of the connected qualiflemmas in the synonym graph. We capture that information by computing the shortest-paths and distances to the nearest nodes with assigned polarity.

The distances are computed using Dijkstra's shortest-path algorithm on a modified syngraph, to which we added three start nodes, labeled “1”, “-1” and “0”, each representing a polarity value. These are directly connected to all the nodes representing qualiflemmas with the same assigned polarity. The distances from each qualiflemma q, to each of these three start nodes correspond to the $dpos_q$, $dzer_q$ and $dneg_q$ features used in the subsequent statistical classification.

Besides these features, we also calculate three polarity weights ($wpos_q$, $wzer_q$, $wneg_q$) as the sum of the inverses of the distances of each node to the corresponding start node. For $wpos_q$, we have:

$$wpos_q = \sum_i \frac{1}{1 + dpos_i} \tag{1}$$

where the i represent the nodes adjacent to q.

5.2 Polarity Assignment

With the above features calculated for the 20% of the *qualiflemmas* and their polarities (which have been blanked for generating the *syngraph*) we learn a classification model for assigning polarities.

We start by partitioning the initial set or *qualiflemmas* into five folds, in order to create five different models. We use the 20% of the *qualiflemmas* in a fold for learning the model and the remaining for generating a *syngraph*. As the *syngraphs* for each fold are different, we also observe differences in the resulting models; the class probabilities for a given lemma obtained with distinct models having no polarity assigned nodes close-by could be almost identical, whereas such models would become quite different from another model trained with a *syngraph* having neighboring nodes with prior polarities. To compensate for these differences, we combine the results of the 5 models in the polarity assignment phase, by assigning the final polarity with a simple boosting technique (the final polarity is assigned by majority vote among the models, with ties broken by picking the polarity class with highest average probability).

6 Evaluation

We use in all experiments the C4.5 decision tree predictive model implementation of the Weka toolkit with default parameters and 5-fold cross validation [17]. To reduce the impact of unbalanced data in automatic classification, we use the SMOTE (Synthetic Minority Oversampling Technique) filter implemented in Weka, which creates new minority class examples by interpolating between existing minority instances.

In the polarity assignment experiments, we used 36,326 pairs of synonyms, which were obtained from the publicly-available resources described in Section 3.4. The derived *syngraph* contains 5,063 nodes, of which 1,989 have a prior polarity (500 positive, 380 neutral, 1,109 negative). An inspection of this graph of human adjectives shows that deciding on the polarities based on the synonyms is not trivial. The graph is highly connected: its order is 7.15 and the counts of nodes directly connected to a node of positive/neural/negative polarity are 4,340, 4,460 and 3,731, respectively. In Table 3, we give the average and standard deviation of the weights in the used *syngraph* of human adjectives.

Table 3. Distance weights statistics in the *syngraph*.

Weight	Avg.	Stdev.
Wpos	3.01	3.59
Wzer	3.29	3.68
Wneg	3.58	4.16

6.1 Human Classifier

The generated models are able to correctly predict the semantic category of adjectives in 94% of the cases, with a recall also of 94% (Table 4). The observed recall is identical in both cases. These evaluation results reinforce both the adequacy of methodology proposed, and the pertinence of the previously identified features for classifying adjectives as having LHE (Low Human Evidence) or HHE (High Human Evidence).

Table 4. Metrics on automatic semantic classification.

SC	Precision	Recall	F-measure
HHE	0.94	0.95	0.94
LHE	0.94	0.94	0.94
Average	0.94	0.94	0.94

6.2 Lexical Expansion

Table 5 provides statistics on the lexicon of human adjectives before and after the application the expansion method introduced in this paper.

Table 5. Lexicon of adjectives, before and after lexical expansion (LE)

Adjectives	Pos	Neg	Neut	All
Before LE	785	2,242	1,009	4036
After LE	1,359	3,826	1,702	6887
Gain	574	1,584	693	2,851
Gain (%)	73%	71%	69%	71%

We started with a large set of human adjective lemmas with initial polarities and restricted the expansion to those lemmas with high evidence of usage in social judgments in our corpus. However, we were still able to uniformly expand the lexicon by 7% for all polarity classes with good accuracy.

6.2 Polarity Classifier

After lexical expansion, we randomly selected 10% (285) of the automatically annotated adjectives and had them manually annotated by the same annotators and using the same criteria used for manually classifying the adjectives used for training the model. The contingency matrix summarizing the results is shown in Table 6.

Table 6. Contingency Table for the Automatic Polarity Classification.

	Positive	Neutral	Negative	Total	Recall
Positive	45	18	5	68	66%
Neutral	13	39	18	70	56%
Negative	9	30	108	147	73%
Total	67	87	131	285	
Precision	67%	45%	82%		

The learned model has an accuracy of 67%. The Chi-square test of independence for this contingency table indicates a significance level of 0.1% (p-value < 0,001). We can also observe that the model is more precise in classifying negative adjectives (precision of 82%) than positive adjectives (67%). The most problematic cases involve neutral polarity, which is, in average, correctly assigned only in 45% of the cases. Similarly, the highest recall is obtained for negative polarities (73%), and the lowest for neutral polarities (56%).

If we adopted a binary (positive/negative) classification, instead of having a tri-polarity classification, the performance of correct assignments would be significantly higher, because adjectives identified as neutral were accounted as positive/negative. However, as noted by Kim and Hovy, this tri-classification is crucial to prevent the classifier from assigning either positive or negative sentiment to weak opinion-bearing words [6].

Social judgments in the Web, especially in user-generated content, are often negative [18]. For opinion mining, the correct identification of a positive opinion is more critical than the detection of several negative opinions. In the above presented evaluation, our main concern was assessing the proposed methodology. We have observed that alternative learning algorithms to the decision trees used in the experiments, such as support vector machines, could produce better polarity classifiers under the proposed methodology. Hence, an interesting development to this work would involve studying how to obtain a model offering improved overall accuracy and optimized for better predicting the polarity of positive lemmas.

6 Conclusion

In this paper, we described a method for automatically expanding a sentiment lexicon for mining social judgments from text human subjects. The experiments show that the automatic identification and classification of human adjectives in a lexicon can be carried out successfully by obtaining evidence about their use in large corpora. This can be performed using specific sets of handcrafted lexico-syntactic patterns describing the contexts where such categories are expected to occur. The frequency of matches recognized by those patterns, together with the adjective frequency in corpora, proved to be important and a distinctive feature for automatically distinguishing high-human evidence adjectives from low-human evidence adjectives.

We believe that the method proposed above could also be successfully applied to other semantic categories and different languages, provided that their human evidence patterns are created. However, other analyses need to be completed before reaching that conclusion, even for adjective lemmas, such as measuring how accuracy is affected by the size of the initial lexicon or by the percentage of lemmas in that lexicon reserved for the synonyms graph construction.

The lexicon can be downloaded from the SentiLex-PT01 page at http://xldb.fc.ul.pt/wiki/SentiLex-PT01_in_English.

7 Acknowledgements

This work was supported in part by FCT (Portuguese research funding agency) through its LASIGE Multi-annual support, grant UTA-Est/MAI/0006/2009 (project REACTION), the post-doc scholarship SFRH/BPD/45416/2008 and the doctoral scholarship SFRH/BD/65972/2009.

References

1. C. Fellbaum, editor. 1998. WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press.
2. V. Hatzivassiloglou and K. McKeown. 1997. «Predicting the Semantic Orientation of Adjectives». In Proceedings of ACL. Madrid, Spain.
3. P. Turney, and M. Littman. 2003. «Measuring praise and criticism: Inference of semantic orientation from association». ACM Transactions on Information Systems, 21(4), 315-346.
4. E. Riloff and J. Wiebe. 2003. «Learning extraction patterns for subjective expressions». In Proceedings of the EMNLP.
5. J. Kamps, M. Marx, R. J. Mokken, and M. de Rijke. 2004. «Using WordNet to measure semantic orientation of adjectives». In Proceedings of LREC.
6. S-M. Kim and E. Hovy. 2006. «Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text». In Proceedings of ACL/COLING Workshop on Sentiment and Subjectivity in Text, Sydney, Australia.
7. A. Esuli and F. Sebastiani. 2006. «Determining term subjectivity and term orientation for opinion mining». In Proceedings EACL, Trento, Italy.

8. N. Godbole, M. Srinivasaiah and S. Skiena. 2007. «Large-scale sentiment analysis for news and blogs». In Proceedings of the International Conference on WebLogs and Social Media (ICWSM 07), Colorado, USA.
9. D. Rao and Deepak Ravichandran. 2009. «Semi-Supervised Polarity Lexicon Induction». In Proceedings of EACL.
10. H. Takamura, T. Inui, and M. Okumura. 2005. «Extracting semantic orientation of words using spin model». In Proceedings of ACL.
11. Y. Choi and C. Cardie. 2009. «Adapting a Polarity Lexicon Using Integer Linear Programming for Domain-Specific Sentiment Classification». In Proceedings of EMNLP.
12. A. Fahrni and M. Klenner. 2008. «Old wine or warm beer: Target-specific sentiment analysis of adjectives». In Symposion on Affective Language in Human and Machine, AISB Convention.
13. H. Kanayama Tetsuya Nasukawa. 2006. «Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis». In Proceedings of ACL.
14. P. Carvalho. 2007. Análise e Representação de Construções Adjectivais para Processamento Automático de Texto. Adjectivos Intransitivos Humanos. PhD Thesis, University of Lisbon.
15. H. Gonçalo Oliveira, D. Santos and P. Gomes. 2010. «Extracção de relações semânticas entre palavras a partir de um dicionário: o PAPEL e sua avaliação». In Linguamática 2.1.
16. E. Maziero, T. Pardo, A. Felippo e B. Dias-da-Silva. 2008. «A base de dados lexical e a interface web do tep 2.0 - thesaurus eletrônico para o português do Brasil». In Proceedings of TIL.
17. David Batista, Mário J. Silva. 2010. A statistical study of the WPT05 crawl of the Portuguese Web. FALA 2010: VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop. Vigo, Spain.
18. I. H. Witten and E. Frank. 2005. Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, CA.
19. L. Sarmento, P. Carvalho, M. J. Silva and E. de Oliveira. 2009. «Automatic creation of a reference corpus for political opinion mining in user-generated content». In TSA'09 - 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement, Hong Kong.