Natural Language Engineering **1** (1): 1–26. Printed in the United Kingdom

A syntactic approach for opinion mining on Spanish reviews

DAVID VILARES, MIGUEL A. ALONSO and

CARLOS GÓMEZ-RODRÍGUEZ

Departamento de Computación, Universidade da Coruña
Campus de Elviña, 15071 A Coruña, Spain
emails: {david.vilares, miguel.alonso, carlos.gomez}@udc.es

(*Received 12 June 2013*)

Abstract

We describe an opinion mining system which classifies the polarity of Spanish texts. We propose an NLP approach that undertakes pre-processing, tokenisation and POS tagging of texts to then obtain the syntactic structure of sentences by means of a dependency parser. This structure is then used to address three of the most significant linguistic constructions for the purpose in question: intensification, subordinate adversative clauses and negation. We also propose a semi-automatic domain adaptation method to improve the accuracy of our system in specific application domains, by enriching semantic dictionaries using machine learning methods in order to adapt the semantic orientation of their words to a particular field. Experimental results are promising in both general and specific domains.

1 Introduction

Asking for or giving an opinion to someone we know is something that most people do at some point in their lives. With the explosion of Web 2.0 and the rise of blogs, forums and social networks millions of users express their views about various topics on these sites. They discuss current issues and praise, compare or complain about products, services and even people. These opinions are especially useful for the business sector, because they make it possible to find out how people perceive a product. For example, it enables business organizations to undertake major market research by polling blogs, forums and social networks in order to obtain a general idea of public sentiment for or against a specific product. Opinion extraction is also useful because it allows us to know which aspects of a product are working and which have to be improved (*e.g. 'The mobile XXX is too big'*). The economic benefits that can be derived from this knowledge are obvious, so the market has begun to demand solutions to analyse this enormous flow of opinions. However, manual monitoring is not an option, given the complexity of the analysis and the exponential growth of the number of reviews that appear on the web. In this respect, sentiment analysis,

This version of the article has been accepted for publication, after peer review, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1017/S1351324913000181

also known as opinion mining (OM)[1], is a growing field of research focussed on automatic processing of the subjective information present in texts, where one of the main tasks is *polarity* classification, *i.e.*, to determine whether the opinion expressed is positive, negative, neutral or mixed.

In this regard, we present a sentiment analyser prototype for reviews in Spanish language which uses dependency parsing to resolve the polarity of a given text. The practical application of our proposal is supported by the good results obtained after testing in different corpora.

The remainder of this article is organized as follows. We begin by presenting related research in Section 2, focusing on the polarity classification task. In Section 3 we provide a broad introduction to our proposal, detailing the treatment of the linguistic aspects considered. Section 4 describes a semi-automatic domain adaptation method to improve performance in a specific field, taking movies as an example. Experimental results are shown in Section 5. Finally, we present conclusions and future work in Section 6.

2 Related work

The development of effective OM systems requires overcoming a number of challenges (Pang and Lee, 2008). Firstly, it is necessary to identify subjective content within a text. This task is not trivial in specific review sites like Epinions[2] or Ciao[3], but it is an even more difficult problem in Twitter[4] or blogs, where both subjective and objective content is published, and sentiment can often be expressed in a very subtle manner, making it difficult to identify by terms considered in isolation (Pang and Lee, 2008). Proposed approaches range from the use of simple hints such as emoticons (Pak and Paroubek, 2010) to the use of complex combinations of lexico-syntactic features that capture basic characteristics of conversation structure across modalities (Murray and Carenini, 2011).

Secondly, OM systems must classify the overall sentiment of a given text, known as its polarity. Traditionally, this has been considered as a binary classification task (positive vs. negative). However, it is also possible to formulate a more fine-grained categorisation (Sidorov et al., 2012), such as differentiating between positive, negative and neutral opinions.

The polarity classification task has traditionally been tackled from two different perspectives: supervised machine learning (ML) approaches and non-supervised semantic-based methods. ML solutions involve building classifiers from a collection of annotated texts (Pang et al., 2002), where each text is usually represented as a bag-of-words. It is also common to include some linguistic-related processing for

[1] Following Pang and Lee (2008), section 1.5, we use these terms interchangeably, although *opinion mining* was initially associated with web search and information retrieval and *sentiment analysis* referred to the automatic analysis of subjective texts.

[2] `http://www.epinions.com`

[3] `http://www.ciao.com`

[4] `http://www.twitter.com`

preparing features (Montejo-Ráez et al., 2012; Bakliwal et al., 2012), such as lemmatisation, stemming or stop word removal. Classifiers of this kind perform well in the domain where they have been trained, but their accuracy drops markedly in other areas, because they are highly domain dependent (Aue and Gamon, 2005). Semantic-based methods (Turney, 2002) involve the use of dictionaries where different kinds of words are tagged with their semantic orientation (SO). To classify polarity, these methods obtain the words present in a text and aggregate their SO in a given way (Taboada et al., 2011). In contrast with ML approaches, semantic-based methods are more domain independent, although their performance can still vary from one domain to another.

There also exist hybrid approaches such as those by Greene and Resnik (2009), Nakagawa et al. (2010), Zhang et al. (2009), Joshi and Penstein-Rosé (2009) and Wu et al. (2009), that use NLP to approximate semantic properties automatically and use them as features for a supervised ML system. Contrary to these methods, we present a novel unsupervised approach based on dependency parsing. Our aim is to use dependency structures to combine the SO of lexical items in a principled way, since effective sentiment analysis requires not only considering words individually, but also taking into account linguistic constructions that can change the overall sentiment.

One of such constructions are *valence shifters* (Kennedy and Inkpen, 2006) such as negation and intensification. The usual way of dealing with them is by means of heuristics. With respect to negation, Pang et al. (2002) assume that the scope of negation includes the words between the negator and the first punctuation mark after the negation term. Yang (2008) proposes flipping the SO of words in the vicinity of a negation, which includes a great number of words to the right of a negator. Fernández Anta et al. (2012) uses a conservative technique for processing negations and intensifications on Twitter, before training a classifier to resolve the polarity: if the pre-processing algorithm finds a valence shifter, it changes the SO of the three terms following the shifter. The English SO-CAL system (Taboada et al., 2011) uses POS tagging information to identify the scope of negation. Moreover, it works with intensifiers and also identifies and discards the SO present in subjunctive and conditional forms, the main types of *irrealis* mood. Irrealis is usually used for expressing perceptions or non-factual opinions, thereby distorting sentiment calculation on texts.

In contrast with these purely lexicon-based heuristics, Jia et al. (2009) propose a procedure that uses a parse tree and a collection of rules for identifying the scope of negation terms. Their experimental results show an improvement over other methods in accuracy of sentiment analysis.

In addition to these linguistic aspects, sentiment analysis must take into account the poor text quality of web reviews. Moreover, users often employ ungrammatical phenomena: emoticons, replication of characters[5] (Bakliwal et al., 2012) or overuse

[5] *e.g.* 'it's beautifuuuul'.

of upper case[6] (Saralegi Urizar and San Vicente Roncal, 2012). These are phenomena that frequently appear in order to give greater emphasis to phrases, thus complicating the analysis of these texts.

Finally, we would like to emphasise that although research on OM has been very active in the last decade, most of the work has focused on texts written in the English language. For Spanish, few OM systems have been proposed, the most relevant being The Spanish SO-CALculator (Brooke et al., 2009), a semantic-based model that uses a collection of semantic dictionaries to calculate the sentiment present in common nouns, adjectives, adverbs and verbs. Lexical and grammatical differences between Spanish and English can impact the performance of OM systems designed for these languages. In this context, Boiy and Moens (2009) have studied the impact of language particularities of English, Dutch and French in the performance of OM techniques. They show that it is more difficult to extract the correct sentiment from Dutch and French sentences and that performance improves when using language-specific features. One of the reasons for this is the higher degree of morphological inflection present in these languages, reflected in proportions of respectively 15% and 14% unique words in the corpus compared to 10% for English. This means that more training examples are required to accurately identify all sentiment patterns. Other language-specific problems regard the use of compounds in Dutch and the spelling in French, with accents omitted in an inconsistent way, and the divergence of French texts from formal language. As Spanish and French belong to the same family of Indo-European Western Romance languages, we conjecture than most remarks made by Boiy and Moens (2009) for French can be applied to Spanish.

3 A syntactic approach to semantic orientation calculation

Most OM systems are typically lexicon-based or ML-based solutions that do not take into account the relations between words because they cannot interpret the syntactic structure of texts. In order to try to overcome these limitations, it is common to implement heuristics to simulate a comprehension of negation, intensification and other linguistic constructions, but these often fail, given the complexity of human language. As an alternative, in this article we propose an unsupervised dependency parsing based method for determining the semantic orientation of texts written in Spanish. To this end, we employ Natural Language Processing (NLP) techniques in order to obtain the syntactic structure of sentences.

As a first step, we use a pre-processor for the treatment of some special cases, such as:

- *Unification of compound expressions.* There are many compound expressions in Spanish like *'sin embargo'* ('however') or *'en absoluto'* ('not at all'), that must usually be interpreted as single units of meaning. To find them, we use a dictionary of compound expressions, extracted from the Ancora corpus (Taulé

[6] *e.g.* 'I LOVE THAT FILM'.

et al., 2008).[7] If the pre-processing algorithm identifies a group of these words, it unifies them into a single token (*e.g. 'en absoluto'* becomes *'en_absoluto'*).

- *Normalization of punctuation marks.* People do not usually respect punctuation rules in forums or social networks. This is a handicap for the rest of processing, especially tokenisation. To resolve this, pre-processing homogenises all punctuation mark representation by adding blanks when required (*e.g.* '*I like it,but it is too expensive*' becomes '*I like it, but it is too expensive*').

As a second step, we tokenise sentences and words, dealing with abbreviations, words that start sentences and punctuation marks that indicate their end to then apply POS tagging. The next step consists of running the Brill tagger (Brill, 1992) to then apply an affix-based tagger to try to annotate unknown tokens. Both taggers were trained using 90% of Ancora POS tags as the training set and the remaining 10% as the development set. An additional challenge for Spanish word-category disambiguation is that the use of accents is commonly ignored by people when writing in a web environment (*e.g. 'rapido'* instead of *'rápido'* ('fast')). To improve practical performance of our tagger, we have expanded the training set: we cloned each training sentence to obtain its equivalent without any acute accent.

We evaluated both the regular tagger and the tagger trained with the cloned set. We obtained an accuracy of 95.86% and 95.71% on the test set (which has not been cloned), respectively, but we have observed that the regular tagger performs poorly on web texts. We hypothesise this is due to the fact the Ancora corpus is correctly written, which is not the case of the majority of the web reviews. However, we have observed that our cloned tagger was able to tag these type of reviews correctly. Table 1 shows how both taggers annotate the sentence of the SFU Spanish Reviews Corpus:[8] *'No he tenido tiempo de escribir sobre el y ya esta estropeado'*, which is not correctly written. The correct Spanish sentence would be *'No he tenido tiempo de escribir sobre él y ya está estropeado'*, which translates to *'I had no time to write about it and it is already broken'*. The issue is that Spanish language uses these diacritical accents to distinguish the meaning of *'el'* ('the', determiner) from *'él'* ('it', pronoun), and the meaning of *'esta'* ('this', determiner) from *'está'* ('is', verb). As we can see, the regular tagger fails on these words, but ours is able to tag them satisfactorily. Although removing accents increases ambiguity, the POS tagger trained on the cloned corpus seems to be able to solve it based on neighbour words.

Once these steps have been performed, we use dependency parsing (Kübler et al., 2009; Gómez-Rodríguez et al., 2011) for analysing the syntactic structure of each given sentence. In particular, we have used MaltParser (Nivre et al., 2007) and the Ancora corpus to train a dependency parser based on the *Nivre arc-eager* algorithm (Nivre, 2008). As a result, we obtain a *dependency tree* for each sentence, consisting of a set of *head/dependent* binary relations, called *dependencies*, between words. Each dependency has a label with a given *dependency type*, which denotes

[7] The Ancora corpus is described in some detail in Section 5.1

[8] A description of the SFU Spanish Review Corpus can be found in Section 5.1.

Table 1. *POS-tagging example both with our tagger (O) and with a regular tagger (R) trained with Ancora: adverb (*r*), verb (*v*), noun (*n*), preposition (*s*), pronoun (*p*), conjunction (*c*), determiner (d)*

	No	he	tenido	tiempo	de	escribir	sobre	el	y	ya	esta	estropeado
	not	have	had	time	of	to write	about	it	and	already	is	broken
O	r	v	v	n	s	v	s	p	c	s	v	v
R	r	v	v	n	s	v	s	**d**	c	s	**d**	v

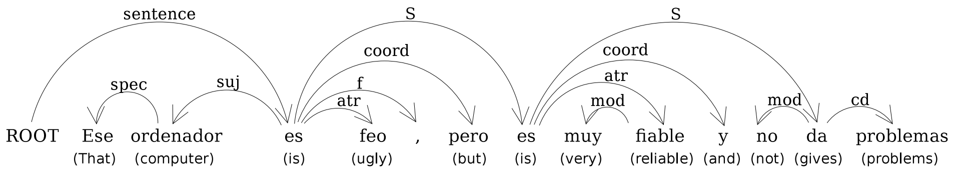

Fig. 1. Dependency parsing for a Spanish sentence

the existing syntactic relation between head and dependent. To simplify computational implementation, an artificial ROOT node is added as the first word of each sentence. We trained and evaluated our parser model using the same splitting used by the tagger. We achieved a LAS[9] of 81.79% and an UAS[10] of 86.76%, which is a competitive accuracy for Spanish. The best-performing system among the 19 participants in the CoNLL-X shared task (Buchholz and Marsi, 2006) reported a LAS of 82.25% and an UAS of 86.05% (note that, since that task used different training and test corpora, this should be taken as a rough indicator of performance and not as a direct comparison between parsers). This means that we have a solid base from which to reliably detect relevant syntactic phenomena like intensification, subordinate adversative clauses and negation; and misdetections are likely to be infrequent enough to not have a large impact in our system's performance. A more precise estimation of this impact could be obtained by task-oriented evaluation, but this would require a costly manual annotation process (Volokh and Neumann, 2012).

As an example, Figure 1 shows a dependency tree for the Spanish sentence *'Ese ordenador es feo, pero es muy fiable y no da problemas'*, which translates to *'This computer is ugly, but it is very reliable and does not give any problem'*, using Ancora dependency labels. We provide an individual literal translation below each word. For example, the adjective *'feo'* ('ugly') is a dependent of type *atr* (attribute) of the verb *'es'* ('is'). We will use this sentence as a running example throughout the paper.

Finally, we rely on SODictionariesV1.11Spa (Brooke et al., 2009) to carry out sentiment analysis. This is a collection of dictionaries for subjective common nouns, adjectives, adverbs and verbs where each word is annotated with its SO, between -5 (the most negative) and +5 (the most positive). Actually, the entries are lemmas

[9] Labelled Attachment Score: Percentage of words that have their head and their dependency type correctly assigned.

[10] Unlabelled Attachment Score: Percentage of words that have their head correctly assigned.

Table 2. *Piece of SODictionariesV1.11Spa*

Dictionary	Word	Semantic orientation
Adjectives	feo (ugly)	-3
Adjectives	fiable (reliable)	2
Nouns	problema (problem)	-2
Nouns	eficacia (effectiveness)	3
Intensifiers	muy (very)	0.25

and the SO corresponds to a generic assignment, without considering a specific domain. It also provides a dictionary of intensifiers, where the label assigned to each intensifying expression represents the value (positive or negative) of modification. SODictionariesV1.11Spa was created by merging the translation of The English SO-CALculator dictionaries (Taboada et al., 2011) and the manual list of subjective words extracted from the SFU Spanish Reviews Corpus. Table 2 shows some entries in these dictionaries.

Generic SO assignments perform well in a large number of domains, but they can be inadequate for a specific field. For example, *'war'* is apparently a clearly negative word, but its polarity can change or be non-existent in some contexts, like in the sentence *'Saving Private Ryan is a film about war'*. In Section 4 we describe in detail a procedure to adapt (and expand) those dictionaries to a particular domain.

3.1 Baseline

Our starting point is equivalent to a purely lexical approach, as we calculate the SO of a sentence by taking the common nouns, adjectives, adverbs and verbs stored in SODictionariesV1.11Spa into account. The SO of each word spreads recursively to the upper levels of the dependency tree until ROOT is reached. Each head node aggregates the sentiment of its children. Syntactic constructions such as negation, subordinate adversative clauses or intensification are not considered at this time.

Figure 2 shows a sentiment analysis on the running example applying this initial approach. Boldface is used to indicate words with an associated sentiment and dashed lines show how their SO is propagated to the top of the sentence. The sentence in the example is generally perceived as slightly positive, but this initial proposal classifies it as negative, because there are syntactic constructions that have been not considered in the base system, such as the negation *'no'* ('not'), the intensification *'muy'* ('very') or the adversative subordinate conjunction *'pero'* ('but').[11] In the following sections we describe how we deal with them and how we include these valence shifters on our approach.

[11] Throughout the article we will use italics to represent all the linguistic aspects that can shift the sentiment of a sentence

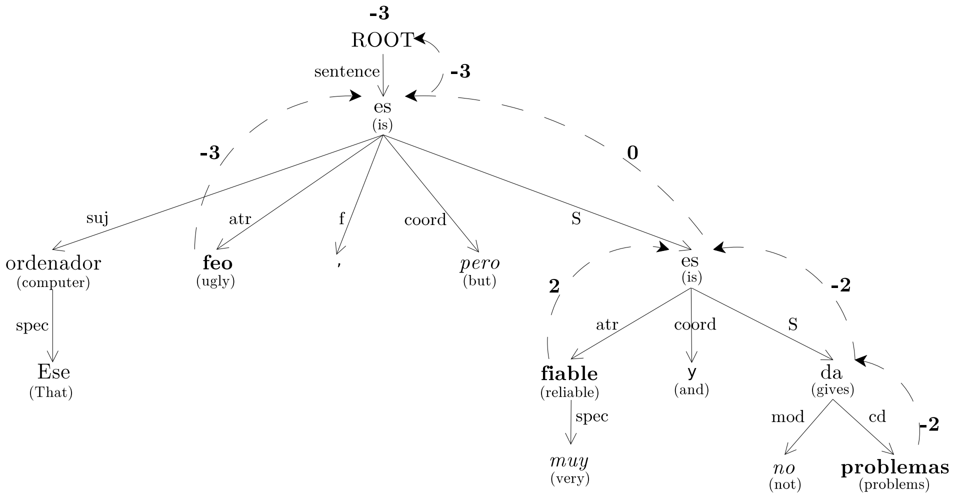

Fig. 2. Semantic orientation analysis over a dependency tree

3.2 Treatment of intensification

An intensifier is a word or an expression which plays the role of a valence shifter in a sentence. There are two types according to their category: *amplifiers* and *downtoners*. The former maximize semantic orientation of one or more tokens, such as *'muy'* ('very'); whereas the latter decrement it, *e.g.* , *'en absoluto'* ('not at all') or *'poco'* ('little').

In some respects, our treatment of intensification is similar to that of Taboada et al. (2011), in the sense that amplifiers and downtoners are modeled as SO modifiers. Each intensifier has an associated percentage, positive if it is an amplifier and negative if it is a downtoner. However, we use syntactic dependencies to identify the scope of an intensifier; whenever an adverb is a dependent of a specifier (*spec, espec*) or an adjunct (*cc,sadv*) type, we take that word as a valence shifter and its head as the exact scope to be shifted. Figure 3 illustrates the effect on the sentiment calculation in the running example once the treatment of intensification is incorporated. We take *'fiable'* ('reliable') as an intensified word, because its dependent node is an adverb and it is labelled with the dependency type *spec*. To calculate the sentiment of this piece of the sentence, we retrieve the original SO of *'fiable'*, which is 2, and we increase it by 25%, the percentage associated to the amplifier *'muy'* ('very'): $2 * (1 + 0.25) = 2.5$. Also, it is possible to nest the effect of two or more intensifiers to shift the SO of a term. Nested intensifiers are labelled with the *spec* dependency type and their head node is always another intensifier. In this case, we calculate the final valence shift by aggregating the percentages associated to different intensifiers, subsequently applying the resulting percentage to a token. For example, in *'en absoluto muy fiable'* ('not very reliable at all'), where *'en absoluto'* ('not at all') has an associated percentage of -1, we would calculate the semantic orientation of that expression as $2 * (1 + (0.25) + (-1)) = 0.5$.

Finally, there are other ways of emphasising an idea. Exclamation marks make

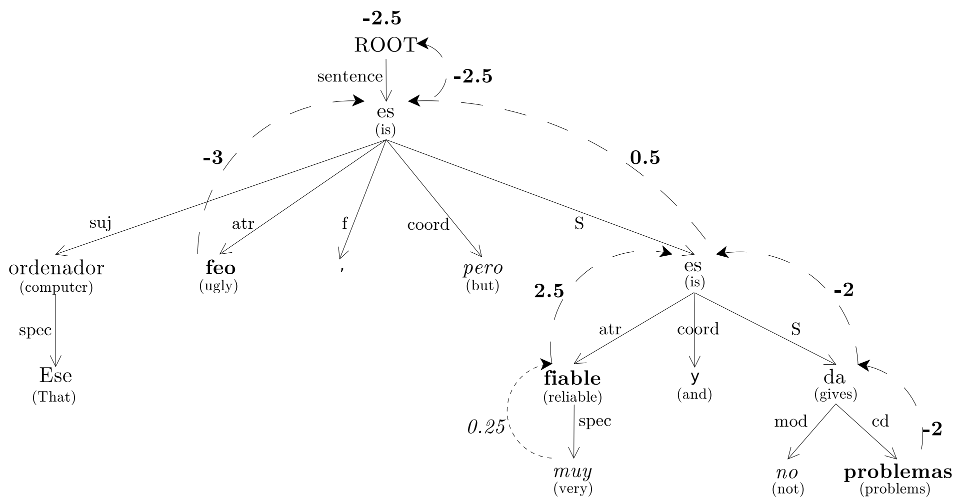

Fig. 3. Semantic orientation analysis with treatment of intensification

it possible to indicate a stronger conviction or a salient word in a sentence. For treating this phenomenon, we included *'!'* in the dictionary of intensifiers with a percentage value of 0.5 and we added the *f* dependency type (used for punctuation marks) to the algorithm for detecting intensified expressions.[12]

3.3 Treatment of subordinate adversative clauses

A subordinate adversative clause expresses an event or fact that is the opposite to that of the main clause. In an OM context, we hypothesise that these type of constructions are a way of restricting, excluding or amplifying the sentiment reflected by both the main and subordinate clauses. We consider subordinate adversative clauses as a special case of intensification, but involving clauses, not individual terms. For example, the sentence *'The actor acted badly but the movie was great'* is perceived as slightly positive because the conjunction *'but'* implicitly gives more importance to the subordinate adversative clause *'the movie was great'*, while the main clause is partially ignored.

In this respect, we distinguish two different types of adversative conjunctions, as is pointed out in Campos (1993), Chapter 3. The first type, *restrictives*, increase the sentiment of the subordinate clause and decrease the SO of the main clause. The second type, *exclusives*, ignore totally the sentiment reflected in the main clause. Unfortunately, the Ancora corpus uses different dependency trees and dependency types for representing different adversative clauses. In this work, we only treat sentences that are uniformly structured: we take *'pero'* ('but') and *'mientras'* ('while') as restrictive conjunctions and *'sino'* ('but rather') and *'sino que'* ('but on the

[12] Unlike English, Spanish uses *'¡'* to begin exclamatory sentences, but it is customary to omit it in a web environment, and for this reason it has not been considered here.

Table 3. *Weights of restrictive and exclusive conjunctions*

Type of conjunction	Weight for main clause	Weight for subordinate clause
Restrictive	0.75	1.4
Exclusive	0	1

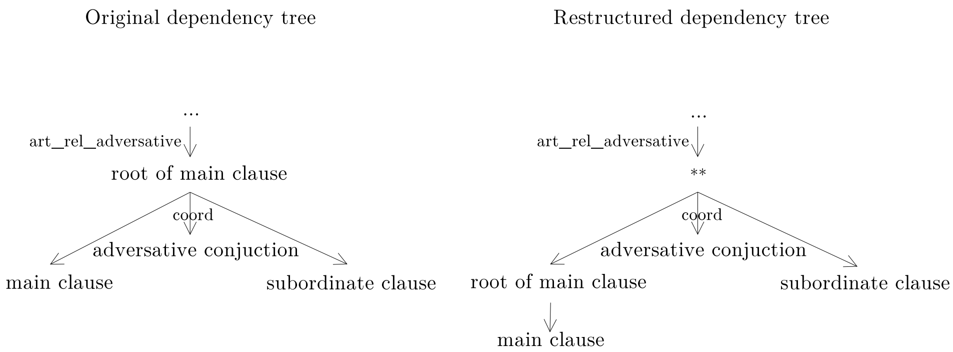

Fig. 4. Reorganization of subordinate adversative clauses

other hand') as exclusives. Table 3 illustrates how we weight both types.[13] In order to homogenise in the future all syntactic representations of the subordinate adversative clauses, we carried out a reorganisation of dependency trees, as shown in Figure 4. Moreover, it simplifies our SO calculation algorithm to weight both the main and subordinate clauses. For this purpose, we include an artificial node, called SAC, at the top of subordinate adversative clauses; and a new dependency type, *art_rel_adversative*, to identify syntactically the beginning of this type of clause.

Figure 5 shows the reorganisation of our running example and how we calculate the sentiment of a sentence once the treatment of adversative subordinate clauses is incorporated. Thus, our sentiment analyser would identify an artificial node, would decrease the SO accumulated in the main clause by 25% (multiplying by 0.75) and amplify by 40% (multiplying by 1.40) the sentiment of the subordinate sentence: $0.75 * (-3) + 1.40 * (0.50) = -1.55$.

[13] The weights have been empirically established over the SFU Spanish Review Corpus. We tested values between 0 and 2 both for main and subordinate clauses using steps of 0.15 and 0.2, respectively.

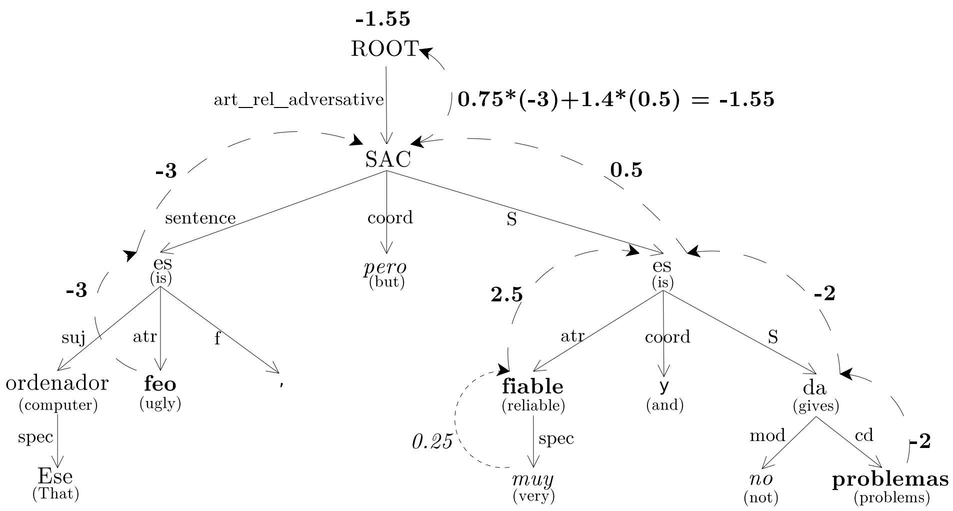

Fig. 5. Semantic orientation analysis with treatment of adversative clauses

3.4 Treatment of negation

The most common and simple way to negate a sequence of tokens in Spanish is the adverb *'no'* ('no'/'not'), but other terms such as *'sin'* ('without') or *'nunca'* ('never') are frequently employed. However, some types of Spanish sentences usually require the use of double negatives to make a negative sentence. In this respect, words like *'nada'* ('nothing'), *'ninguno'* ('none') or *'nadie'* ('nobody') are commonly preceded by *'no'*. Moreover, the difference between a negation term and a downtoner is diffuse. Tokens like *'apenas'* ('barely') or *'casi'* ('almost') could easily be classified in either of these two categories. We have chosen to consider these type of expressions as intensifiers and therefore we only consider explicitly as negators the adverbs *'no'*, *'nunca'* and *'sin'*, which cover a great number of negative sentences. Our treatment of a negation consists of two basic steps: 1) *identify the scope of a negation term* and 2) *modify the semantic orientation of affected tokens.*

3.4.1 Scope identification

The procedure for identifying the scope of a negation depends on the adverb used in the phrase.

The syntactic structure used in Ancora for representing an adverb *'sin'* assures us that its child node should be the scope of negation, without needing to analyse the dependency type. But we cannot assume the same for the negators *'no'* and *'nunca'*. Normally they are represented as leaf nodes and the candidate scope of negation always involves a head node or a collection of sibling nodes, so we require a more complex algorithm for their treatment. We use a procedure based on Jia et al. (2009), which uses a parse tree and a collection of special rules to identify the scope of each negation. Firstly, the candidate scope for a negator is identified. Then, the exact scope is determined by searching *delimiters* by means of a syntactic heuristic procedure. A delimiter is a token that has the capability to eliminate some words from the candidate scope of a negation term. We have adapted this procedure to profit from the additional information provided by the syntactic structure of the sentence. We use dependency types to directly extract the exact scope without identifying delimiter words. When a token has a negator *'no'* ('not') or *'nunca'* ('never') as a child node and it is a dependency of type *'neg'* or *'mod'*; we try the collection of syntactic heuristic rules shown at Figure 6, in the following order:[14]

1. *Subjective parent rule*: Whenever a parent node of a negation term has sentiment, only that node is negated. Figure 6.a shows how we take the scope when this rule matches. For example, in the sentence *'he does not praise my work'*, the negation *'not'* depends on *'praise'*, which is included as a subjective word in the SO dictionaries, so we consider this term as the scope of the negation.
2. *Subject complement/Direct object rule*: Whenever a branch at the same level as a negation node is labelled with a dependency of type subject complement (*atr*) (*e.g. 'the meal is not good'*) or a direct object (*cd*) (*e.g. 'the meal does not look good'*), our sentiment analyser negates that branch, as we show in Figure 6.b.
3. *Adjunct rule*: Whenever a negation term has an adjunct branch (*cc*) at the same level, the sentiment of that branch is shifted. If there is more than one adjunct, only the first one is negated, as shown in Figure 6.c. For example, in the sentence *'he does not work efficiently on Fridays'*, our method takes the mood adjunct (*'efficiently'*) as the scope of the negation, because it is the nearest to the negation.
4. *Default rule*: Figure 6.d shows how when none of the previous rules matches, we consider as scope the sibling branches of a negator.

We now explain in more detail the treatment of negation in the running example. In Figure 7 we can see that the word *'no'* has as its head the verb *'da'* ('gives'). Our method first tries to apply the *subjective parent rule*, but in this case, this is not a subjective node, so that rule is ignored. Then, our procedure continues with the *direct object rule*, which matches, because there is a direct object dependent (identified by *cd*) at the same level as the negation, so this rule is applied and takes *'problemas'* ('problems') as the scope of negation.

3.4.2 Polarity flip

There are several ways of taking into account the effect of negation. Machine learning methods tend to unify the negator and the negated word into a single feature (in this way *'not good'* becomes *'not_good'*) (Sidorov et al., 2012). Another possibility is to change a given number of words following the negator (Yang, 2008; Fernández Anta et al., 2012), a method that can even be generalised to deal with the intensifiers in the same way.

On the other hand, the simplest way to negate a word in semantic approaches is

[14] Only the first matching rule is applied.

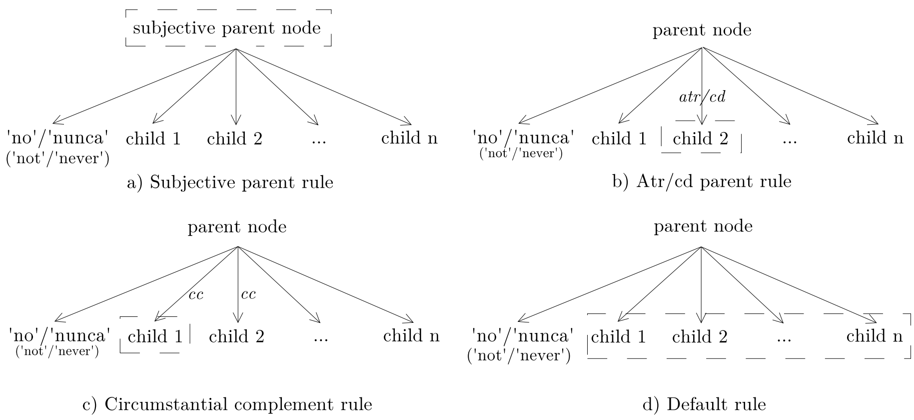

Fig. 6. Heuristic rules to identify the scope of negation

to invert the SO (*e.g.* if SO(good) = 2 then SO(not good) = -2). The main drawback of this method is that it is not coherent with human intuition. For example, if the SO of *'fascinating'* is 5 the sentiment of *'not fascinating'* would be -5, when it could even be considered a slightly positive expression.

Our polarity flip algorithm follows a shift negation method where the SO value is shifted toward the opposite polarity by a fixed amount: following Taboada et al. (2011), we have chosen a flip value of 4 for the adverbs *'no'* ('not') and *'nunca'* ('never'). Figure 7 shows how the SO of the scope of negation, which is *'problemas'* ('problems') as we saw in the previous section, is modified by this amount. The word *'problemas'* has a SO of -2, and the phrase *'no da problemas'* has a SO of $-2 + 4 = 2$

For the adverb *'sin'* ('without'), based on our experimental setup, we have chosen a value of 3.5. We hypothesise this kind of negation as being less potent, given that its scope is fairly local. Experimental results showed an improvement in accuracy when carrying out this strategy.

3.5 Other features

Along with the syntactic-based issues, there are other factors that can influence the overall sentiment, such as the discourse structure (Pang and Lee, 2008). The order in which authors express their opinions can change the sentiment polarity. It is customary that the final sentences of a text play the role of a summary or conclusion, giving implicitly more emphasis to this part of the document. To simulate this phenomenon, our proposal increases the sentiment of the last three sentences of a given review. We chose a value of 0.75 based on experimental evidence. Also we note that by increasing the SO of the nouns, adjective, verbs and adverbs of SODictionariesV1.11Spa by 20% our approach improved the performance on our development corpus. Thus, the SO considers values between -6 and 6. This modification is applied both to the hand-created dictionaries and to the automatically

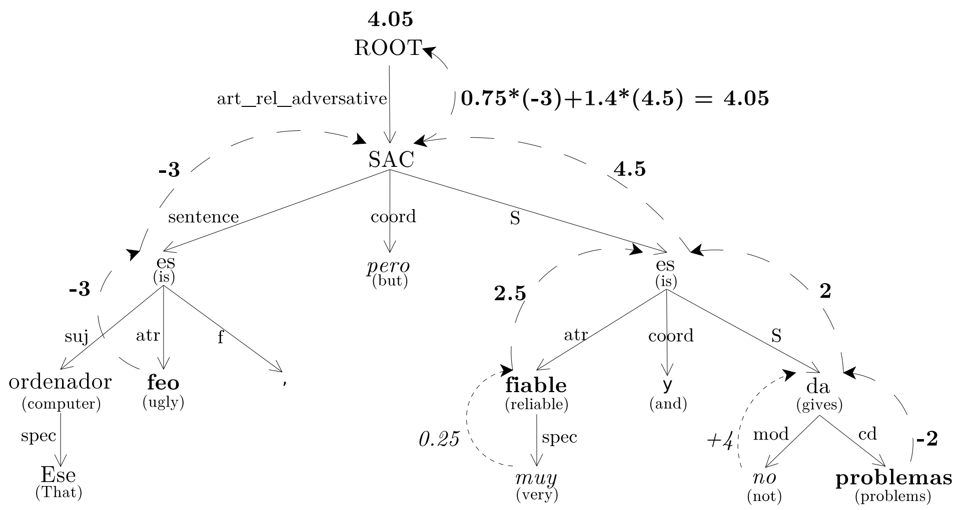

Fig. 7. Semantic orientation analysis with treatment of negation

enriched dictionaries explained in Section 4.2. All the strategies that improved the performance of our proposal were included in our final version.

The motivation of all these optional features was experimental, taking the SFU Spanish Review Corpus as the development set, but they also work satisfactorily on other long text corpora, as we show in Section 5.

4 Semi-automatic domain adaptation

As explained previously, the generic SO of dictionaries can be inadequate in a particular domain. Entertainment contexts are some of the typical fields where this phenomenon occurs more frequently. In this section, we provide a semi-automatic method to adapt and enrich semantic dictionaries to a specific area and we use CorpusCine, a corpus of Spanish movie reviews, as an example. In Section 5.1 we detail the content of this corpus. In Section 5.4 we illustrate how our adaptation method improves the performance for this domain.

4.1 Title preprocessing

Movie titles often seek to arouse an emotion or an interest in the possible audience, and so they usually contain subjective words. This is a disadvantage when automatically classifying the polarity of a text, because it can distort their categorisation. For example, in a review of the movie *'Monsters, Inc'*, the title will probably appear a number of times in the text. Each occurrence would be classified as unfavourable because *'monster'* is listed in our dictionary as a word with negative polarity.

To resolve this problem we pre-process movie titles. For each movie review, we unify its title into a single unit of meaning (*e.g. 'Monsters, Inc'* becomes *'Monsters,_Inc'*). Moreover, a large proportion of quoted items refer to other movie titles, so we unify such expressions too. We are aware that some instances will not match movie titles, but rather represent an irony or words used in a different sense than usual. However, we do not consider this to be a drawback, because it is a way of discarding phenomena that would probably require special treatment. The analysis of this type of quoted excerpts is left as future work.

4.2 Enrichment of semantic dictionaries

Our aim is to learn the polarity of subjective words in a given domain. This implies discovering words which are not present in the generic dictionary and also to adapt the polarity of words already present in the dictionary to their use in the specific field.

For the first task, we learn which tokens are good polarity classifiers in the area in question by extracting the most representative words in that domain relying on WEKA, an ML environment (Hall et al., 2009). This software has an explorer with a panel for attribute selection. It provides different search methods and different evaluation metrics for identifying the most relevant features in a particular classification task. In this respect, we have taken *InfoGainAttributeEval* as attribute evaluator and *Ranker* as the search method. *InfoGainAttributeEval* evaluates the value of an attribute by measuring the information gain with respect to the class (in our case, the classes are positive sentiment and negative sentiment) and *Ranker* ranks attributes by their individual evaluations. Once we have classified the attributes, we need to give an SO to each selected word. We hypothesise that if an attribute appears more frequently in positive than in negative texts, that feature must be positive, and vice versa. If a word is positive we calculate its SO with the equation (1), and if it is negative we employ equation (2).

(1) $$OS_{word_i} = \frac{\log_2(\frac{x_i+1}{y_i+1})}{\log_2 z} \times 5$$

(2) $$OS_{word_i} = \frac{\log_2(\frac{y_i+1}{x_i+1})}{\log_2 w} \times -(5+\alpha)$$

The variable x_i represents the number of positive texts where $word_i$ appears, and y_i the number of negative texts. The variable z represents the maximum value x_i/y_i and w the maximum coefficient y_i/x_i, for all i. The resulting values are normalised between 5 and -5, in order to make them comparable with the values in SODictionariesV1.11Spa. The words with an SO close to 0 will represent neutral terms. Finally, the parameter α is employed to give more relevance to negative words. We need to create a 'pessimistic' dictionary to improve performance and counteract the 'optimistic'tendency of CorpusCine, a characteristic widely explained in other studies, as we will show in Section 5.2. A possible option to do this is to increase the semantic orientation of negative words. Another equivalent option consists of including more negative than positive words. Both perspectives will be analysed in Section 5.4.

After creating the domain dictionary, we must merge it with the generic dictionary. We hypothesise that if the SO is less than 0.5 in absolute value, the word is not a clear subjective word, so we discard it. For the rest of the words in the domain dictionary, we check the generic semantic orientation of the word in SODictionariesV1.11Spa. If it does not have a generic SO specified in that dictionary or it has a different sign than the domain specific SO obtained, the latter prevails. If both the generic and the adapted SO have the same sign, then the generic SO prevails. This means that our method will only change the SO of words that are clearly used with a non-standard polarity in the target domain, but it will not try to adjust the exact SO value for words where the obtained sign matches the one in the dictionary. As an example, Table 4 shows the top five representative informative attributes in the movie domain while Table 5 shows some words of the movie domain that have changed their semantic orientation with respect to the general dictionary.

Table 4. *Best features in a movie domain*

Ranking	Word	Generic SO	Movie domain SO
1	perfecto (perfect)	4	1.808
2	obra (work)	5	1.139
3	maestro (masterly)	0	1.760
4	imprescindible (indispensable)	4	3.259
5	peor (worse)	-2	-1.712

Table 5. *Semantic orientations adapted to a movie domain*

Word	Generic SO	Movie domain SO
violencia (violence)	-5	1.511
guerra (war)	-2	1.310
zombi (zombie)	-1	0.730
kryptonita (kryptonite)	0	-1.981
bestseller	4	-1.250

5 Evaluation

In this section we describe the corpora employed to develop and evaluate our approach and we show and compare the performance obtained for our system and other approaches on SFU Spanish Reviews and CorpusCine.

5.1 Corpora

To train both the POS tagger and dependency parser we have employed Ancora (Taulé et al., 2008). Ancora are two multilingual corpora created from newspaper articles of 500,000 words each: a Catalan corpus and a Spanish corpus. Among other things, it is tagged with lemma, part of speech and dependency information in CoNLL-X format (Buchholz and Marsi, 2006).

In order to test the performance of our opinion mining system, we used three annotated corpora:

- The SFU Spanish Review Corpus (Brooke et al., 2009) is a collection of 400 Spanish reviews on cars, hotels, washing machines, books, cell phones, music, computers, and movies from the ciao.es web site. Each category has a total of 25 favourable and 25 unfavourable reviews. As usually happens in reviewing web sites, texts have unstressed words, unrecognised abbreviations and ungrammatical sentences. This allows us to evaluate our proposal in a real and complex environment. Moreover, The Spanish SO-CAL was developed on this corpus, so their lexical-based approach and our dependency parsing-based method can be compared.
- CorpusCine Reviews (Cruz Mata, 2011) is a collection of 3,878 movie reviews written in Spanish from the muchocine.net web page. Each document is rated between one and five stars, where one is the most negative rating and five the most positive. There are 351 one-star, 923 two-star, 1,253 three-star, 890 four-star and 461 five-star reviews. We classify one or two-star documents as negative. Three-star reviews are discarded because we consider them as neutral or mixed reviews. This is a widely accepted strategy that has been employed in other studies (Cruz Mata, 2011) and corpora, like the SFU Spanish Review Corpus[15]. Documents ranked with four or five stars are taken as positive reviews.
- HOpinion[16] is a collection of 17,934 hotel reviews extracted from www.tripadvisor.es, rated between one and five stars. There are 841 one-star, 1269 two-star, 3468 three-star, 6244 four-star and 6112 five-star reviews. We followed the same strategy that in CorpusCine to evaluate it, discarding three-star texts.

5.2 Results on SFU Spanish Reviews Corpus

Table 6 shows the performance of our system with a number of different options on the SFU Spanish Review Corpus. All features contribute to performance. One of the most important improvements in accuracy comes from the treatment of negation. As we can see, before incorporating this feature our approach favours positive classifications. This likely happens as the result of a human tendency to positive language (Kennedy and Inkpen, 2006). People usually negate positive sentences to

[15] This issue is detailed on the readme file of www.sfu.ca/∼mtaboada/download/downloadCorpusSpa.html
[16] http://clic.ub.edu/corpus/hopinion

Table 6. *Performance with different configurations of our proposal over SFU Spanish Reviews Corpus*

Category	Neg. accuracy	Pos. accuracy	Tot. accuracy
Initial proposal	0.310	**0.925**	0.618
+intensification	0.450	0.870	0.660
+adversative clauses	0.455	0.885	0.670
+negation	**0.745**	0.765	0.755
Final proposal	0.740	0.830	**0.785**

express an unfavourable opinion. For example, it is common to use expressions like *'not good'* instead of *'bad'* or *'I don't like it'* instead of *'I dislike it'*. Even after processing negation terms, a lexicon-based system such as the English SO-CAL increases the final SO of any negative expression by 50% to overcome that positive bias, improving its performance by around 6% with this strategy. However, in our current implementation that feature gave no benefit. This suggests to us that our negation algorithm performs well, at least in a general context.

Table 7 shows the performance of our final approach on each subcorpus of the SFU Spanish Review Corpus. As we can see, there are significant differences in performance depending on the category. For domains where quality criteria are reasonably objective, such as hotels, computers or washing machines, our proposal performs well (over 80% accuracy), because the generic SO is usually adequate. But the same is not true for entertainment domains such as movies, books and music, where performance falls below the average. We believe this is mainly due to the problem of generic semantic orientations, as we have discussed throughout the paper, which primarily affects these type of domains. Moreover, movies or books are contexts where personal tastes are particularly important. For example, the fragment *'is a low-budget movie'* is in principle a negative sentence, but it could be positive for a person who loves B movies. This makes it difficult to assign a semantic orientation according to the sentiment of users, even for a particular domain.

Table 8 compares the performance of various methods on the SFU Spanish Review Corpus. Our syntactic proposal improves the accuracy of The Spanish SO-CAL by about 6%, even though the SO-CAL is a system with more functionality (*e.g.* treatment of irrealis). This suggests that parsing is useful in order to resolve the polarity of a given text. In particular, we believe that an effective treatment of negation requires a more complex algorithm than a purely lexicon-based technique.

We also compare our proposal with an ML method. More specifically, we have trained a Support Vector Machine (SVM) as a classifier. We have relied on WEKA to build it, using libSVM (Chang and Lin, 2011). Specifically, we chose an SVM of type C-SVC, a radial basis function as the kernel type and a value of 1 for the cost parameter. Testing was done with 10-fold cross-validation. Data was pre-processed in order to change the words to their lowercase form, and we have employed the

Table 7. *Performance of final version over SFU Spanish Reviews Corpus*

Category	Neg. accuracy	Pos. accuracy	Tot. accuracy
Hotels	**0.92**	0.88	**0.90**
Computers	0.80	**0.92**	0.86
Washing machines	0.88	0.76	0.82
Cell phones	0.72	0.88	0.80
Cars	0.68	0.80	0.74
Music	0.64	0.88	0.76
Books	0.64	0.84	0.74
Movies	0.64	0.68	0.66

Table 8. *Performance on SFU Spanish Reviews for various methods*

Method	Neg. accuracy	Pos. accuracy	Tot. accuracy
Our proposal	0.7400	**0.8300**	**0.7850**
SVM + our SO as feature	**0.7490**	0.7700	0.7594
The Spanish SO-CAL			0.7425
SVM	0.7230	0.7270	0.7250

output word counts as the weighting factor. Over the SFU Spanish Review Corpus, our syntax-driven analyser provides better accuracy than the SVM, reinforcing the idea that ML approach is not the best technique to build a general domain polarity classifier, at least when performing a binary classification[17]. Finally, we tested a hybrid approach, labelled on Table 8 as 'SVM + our SO as feature': we analysed each text with our proposal and we included the SO obtained as a feature for the SVM. However, the resulting accuracy was worse than the one obtained with our system alone.

5.3 Results on HOpinion

Table 9 shows the performance on HOpinion. Results are similar to those obtained on the hotel category of the SFU Spanish Review Corpus, achieving an accuracy of 0.8938.

We also built an SVM classifier specific to HOpinion, applying lemmatisation to

[17] We tested various configurations with different weightings factors and different types of preprocessing, but we only show the configuration who achieved the best performance. Results are similar that the presented on the same corpus by Brooke et al. (2009).

Table 9. *Performance on HOpinion for various methods*

Method	Neg. accuracy	Pos. accuracy	Tot. accuracy
SVM	0.5800	**0.9930**	**0.9328**
Our proposal	**0.7294**	0.9218	0.8938
SVM_{sfu}	0.6770	0.7940	0.7766

Table 10. *Performance on CorpusCine for various methods*

Method	Neg. accuracy	Pos. accuracy	Tot. accuracy
SVM	**0.8440**	**0.8220**	**0.8328**
Our proposal with domain adaptation	0.7997	0.8024	0.8011
(Cruz Mata, 2011)[a]	0.8250	0.7250	0.7750
Our proposal	0.4804	0.7935	0.6415
SVM_{sfu}	0.6250	0.6130	0.6179

[a] Only a portion of CorpusCine has been evaluated

the texts, using tf-idf[18] as weighting factor and selecting the default configuration of WEKA for the SVM (type C-SVC, a radial basis function as the kernel type and 1 as the cost parameter). We used 10-fold cross-validation to evaluate it, achieving an accuracy of 0.9328. This supervised classifier did not satisfactorily learn negative reviews due to the low number of unfavourable opinions in the corpus. Finally, we evaluated an SVM trained with the SFU Spanish Review corpus (SVM_{sfu}) on HOpinion. In this case, we did not apply lemmatisation, as we did in the classifier trained on the SFU Spanish Reviews, and we changed each word to its lowercase form and used their total output count as the weighting factor.

5.4 Results on CorpusCine

Table 10 shows the performance on CorpusCine obtained by the different approaches explained. Moreover, we included the results obtained by a supervised approach presented in (Cruz Mata, 2011). This specific domain method uses five morphosyntactic patterns to extract sentiment bigrams using multiple seed words (Turney, 2002) to then calculate their SO. It provides a supervised technique which uses an optimal threshold for categorising favourable and unfavourable texts.

We also built an SVM classifier specific to CorpusCine. We pre-processed the data with the title pre-processing explained in Section 4.1 and we applied lemmatisation. Also, we used tf-idf as the weighting factor. We selected the default configuration

[18] We tested other weighting factors such as the binary or the total occurrence of each term, but we achieved the best performance using tf-idf.

Table 11. *Detailed performance on CorpusCine for our generic and adapted proposal*

Polarity	Number of stars	Our proposal	Our proposal with domain adaptation
Negative	1	0.6923	0.9003
	2	0.3948	0.7614
Positive	4	0.7933	0.7674
	5	0.7939	0.8698

of WEKA for the SVM (type C-SVC, a radial basis function as the kernel type and 1 as the cost parameter). We used 10-fold cross-validation to evaluate it. Moreover, as we did with HOpinion, we evaluated an SVM trained with the SFU Spanish Review corpus (SVM_{sfu}) on CorpusCine. The performance drops below our generic approach, which reinforces the idea that ML methods are highly domain dependent. In contrast, our generic proposal shows a performance similar to that obtained on the 'movie' category of the SFU Spanish Review Corpus, a result that confirms the domain independence of the proposal.

Finally, to test our proposal with dictionaries adapted to the movies domain we have also used 10-fold cross-validation. For each fold we extracted around 22.000 attributes (there are many more positive than negative attributes) from WEKA and for each one we built a dictionary using the training set, and we tested it against the development set.

As we can see, our adapted approach improves the performance obtained with our generic approach by about sixteen basis points. Moreover, we neutralise the positive bias that our generic system presented on CorpusCine. Table 11 compares, in greater detail, the performance of our proposal on CorpusCine, before and after adapting it to the movie domain.

We have observed that unfavourable reviews had a high presence of condescending and ironic expressions, complicating the semantic analysis of those texts. To overcome this, we chose to build a dictionary where negative words had more relevance. Figure 8 shows how different weightings for the negative words (the parameter α explained in equation 2), and the different number of positive and negative entries in our specific semantic movie dictionary, affect the performance. We identify each graphic with a notation P-N, where P means that for that case of study we have only considered the first P percent of the positive attributes extracted from WEKA and the first N percent of the negative ones. For example, *75-25* would represent a case where we only employed 75% of the best positive classifiers and only the first 25% of the negative ones. Note that for each weight, the number of negative words is different, because with a higher negative weighting there are more negative words with an SO greater than 0.5 in absolute value, the threshold value established in Section 4.2. Below we provide a brief explanation for each graphic included in Figure 8:

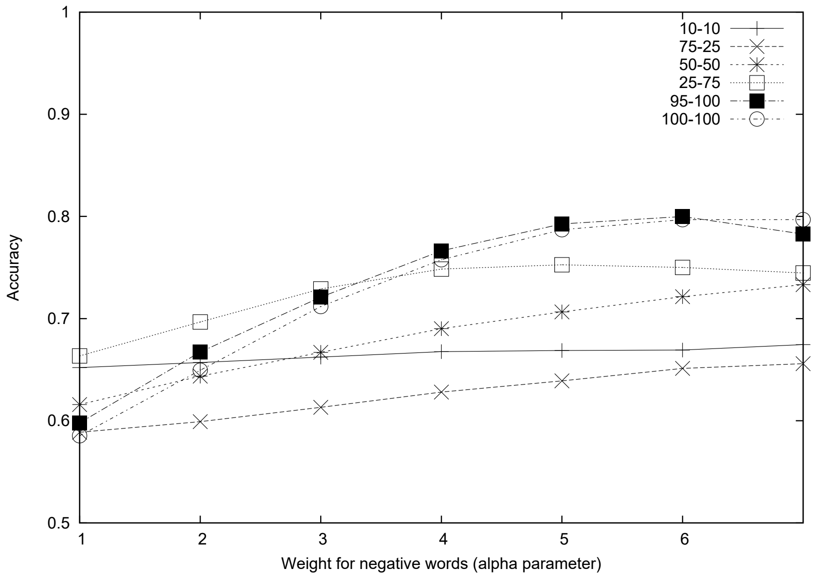

Fig. 8. Accuracy on CorpusCine increasing negative word weighting

- *10-10*: The improvement in performance is minimum. Most of the words included in the dictionary are already present in the SODictionariesV1.11Spa and have the same polarity, so our system takes few words from the specific domain dictionary.
- *75-25*: This configuration does not work well, due to the optimistic trend to favour positive classifications that our initial proposal presents in this particular corpus.
- *50-50*: The behaviour is similar to that explained in the previous point (75-25). Although we employ 50% both for positive and negative words, the dictionary extracted from WEKA has many more positive attributes, so this is also a configuration that favours positive classification. However, we can see how by increasing negative weightings we obtain a good final performance.
- *25-75*: With this setup we obtain a good baseline, but performance decreases when we employ high negative weightings, because our system becomes too favourable to negative classifications.
- *95-100*: This was the best setup. We achieved an accuracy of 0.8011 with a negative weighting factor of 5.5.
- *100-100*: As in the 50-50 configuration, with large negative weightings we can obtain a high performance and counteract the optimistic tendency.

5.5 Discussion of results

Results on HOpinion and CorpusCine are similar to those obtained on the corresponding domain in the SFU Spanish Reviews, which indicates that our approach can be generalisable to different data sets. We showed that in a general domain our syntactic approach works better than ML approaches, but the same is not always true for a specific field, where the semantic dictionaries are affected by a low recall, which has been also pointed out by other authors (Zhang et al., 2011). However, the enrichment of semantic dictionaries applied over CorpusCine suggests that semantic-based approaches can be as effective on specific fields as ML methods are, provided they have access to appropriate semantic dictionaries.

6 Conclusions and future work

In this article, we have described a syntactic-based method for sentiment analysis of Spanish reviews. We use dependency-based methods to treat some relevant linguistic aspects in OM, such as intensification, subordinate adversative clauses and negation. Two sets of experiments were performed to compare our method to other existing techniques. Experimental results on a general domain corpus show that our syntactic proposal improves over ML and lexicon-based approaches. Moreover, we performed an evaluation over a specific domain corpus (movies), where ML techniques obtain a much better baseline accuracy than semantic approaches, due to the invalidity of the generic semantic orientations. We have proposed a semi-automatic method to enrich and adapt the semantic dictionaries to a particular field, and we have applied it to our model. Experiments show a good performance, obtaining an accuracy close to that of ML classifiers and improving over other existing domain specific systems.

With respect to future work, our system could be improved by taking more semantic information into account. In particular, we will try to identify irrealis clauses in order to distinguish the consumers' experience from their desires and conjectures. The treatment of negation could be improved by taking into account variations and special cases of negative constructions in Spanish, such as the nesting presence of the adverb *'ni'* ('nor') (*e.g. 'El restaurante no es ni bueno ni barato'* ('The restaurant is neither good nor cheap')). Also, we plan to explore microblogging social networks, such as Twitter. Tweets need special pre-processing for an effective semantic analysis (*e.g.* hashtags, retweets, favourites, high presence of ungrammatical constructions, *etc*) and irony or sarcasm have a particular relevance. In this respect, Reyes et al. (2012) and Reyes et al. (2013) could help enrich our proposal.

Acknowledgments

Research reported in this article has been partially funded by Ministerio de Economía y Competitividad and FEDER (Grant TIN2010-18552-C03-02) and by Xunta de Galicia (Grants CN2012/008, CN2012/319). We would like to thank Maite Taboada for giving us access to SODictionariesV1.11Spa

References

Aue, Anthony, and Gamon, Michael. 2005. Customizing Sentiment Classifiers to New Domains: A Case Study. In: *RANLP*.

Bakliwal, Akshat, Arora, Piyush, Madhappan, Senthil, Kapre, Nikhil, Singh, Mukesh, and Varma, Vasudeva. 2012. Mining sentiments from Tweets. Pages 11–18 of: *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. WASSA '12. Stroudsburg, PA, USA: Association for Computational Linguistics.

Boiy, Erik, and Moens, Marie-Francine. 2009. A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval*, **12**(5), 526–558.

Brill, Eric. 1992. A simple rule-based part of speech tagger. Pages 112–116 of: *Proceedings of the workshop on Speech and Natural Language*. HLT'91. Stroudsburg, PA, USA: Association for Computational Linguistics.

Brooke, Julian, Tofiloski, Milan, and Taboada, Maite. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish. Pages 50–54 of: *Proceedings of the International Conference RANLP-2009*. Borovets, Bulgaria: ACL.

Buchholz, Sabine, and Marsi, Erwin. 2006. CoNLL-X shared task on multilingual dependency parsing. Pages 149–164 of: *Proceedings of the Tenth Conference on Computational Natural Language Learning*. CoNLL-X'06. Stroudsburg, PA, USA: Association for Computational Linguistics.

Campos, Héctor. 1993. *De la oración simple a la oración compuesta: Curso Superior de Gramática Española*. Georgetown University Press.

Chang, Chih-Chung, and Lin, Chih-Jen. 2011. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**(3), 27:1–27:27.

Cruz Mata, Fermín L. 2011. *Extracción de opiniones sobre características: Un enfoque práctico adaptado al dominio*. Ph.D. thesis, Universidad de Sevilla.

Fernández Anta, Antonio, Morere, Philippe, Núñez Chiroque, Luis, and Santos, Agustín. 2012. Techniques for Sentiment Analysis and Topic Detection of Spanish Tweets: Preliminary Report. In: *TASS 2012 Working Notes*.

Gómez-Rodríguez, Carlos, Carroll, John, and Weir, David. 2011. Dependency parsing schemata and mildly non-projective dependency parsing. *Comput. Linguist.*, **37**(3), 541–586.

Greene, Stephen, and Resnik, Philip. 2009 (June). More than Words: Syntactic Packaging and Implicit Sentiment. Pages 503–511 of: *NAACL'09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, Boulder, Colorado.

Hall, Mark, Frank, Eibe, Holmes, Geoffrey, Pfahringer, Bernhard, Reutemann, Peter, and Witten, Ian H. 2009. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, **11**(1), 10–18.

Jia, Lifeng, Yu, Clement, and Meng, Weiyi. 2009. The effect of negation on sentiment analysis and retrieval effectiveness. Pages 1827–1830 of: *Proceedings of the 18th ACM conference on Information and knowledge management*. CIKM'09. New York, NY, USA: ACM.

Joshi, Mahesh, and Penstein-Rosé, Carolyn. 2009 (Aug.). Generalizing dependency features for opinion mining. Pages 313–316 of: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. ACL, Suntec, Singapore.

Kennedy, Alistair, and Inkpen, Diana. 2006. Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence*, **22(2)**, 110–125.

Kübler, Sandra, McDonald, Ryan, and Nivre, Joakim. 2009. *Dependency Parsing*. Morgan & ClayPool Publishers.

Montejo-Ráez, A., Martínez-Cámara, E., Martín-Valdivia, M. T., and Ureña López, L. A. 2012. Random walk weighting over sentiwordnet for sentiment polarity detection on Twitter. Pages 3–10 of: *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*. WASSA '12. Stroudsburg, PA, USA: Association for Computational Linguistics.

Murray, Gabriel, and Carenini, Giuseppe. 2011. Subjectivity detection in spoken and written conversations. *Natural Language Engineering*, **17**(3), 397–418.

Nakagawa, Tetsuji, Inui, Kentaro, and Kurohashi, Sadao. 2010 (June). Dependency Tree-Based Sentiment Classification using CRFs with Hidden Variables. Pages 786–794 of: *NAACL HLT'10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Proceedings of the Main Conference.* ACL, Los Angeles, CA.

Nivre, Joakim. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, **34**(4), 513–553.

Nivre, Joakim, Hall, Johan, Nilsson, Jens, Chanev, Atanas, Eryigit, Glsen, Kübler, Sandra, Marinov, Svetoslav, and Marsi, Erwin. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, **13**(2), 95–135.

Pak, A., and Paroubek, P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA).

Pang, Bo, and Lee, Lillian. 2008. *Opinion Mining and Sentiment Analysis.* Hanover, MA, USA: now Publishers Inc.

Pang, Bo, Lee, Lillian, and Vaithyanathan, Shivakumar. 2002. Thumbs up? Sentiment classification using machine learning techniques. Pages 79–86 of: *Proceedings of EMNLP.*

Reyes, Antonio, Rosso, Paolo, and Buscaldi, Davide. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, **74**(Apr.), 1–12.

Reyes, Antonio, Rosso, Paolo, and Veale, Tony. 2013. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, **47**(1), 239–268.

Saralegi Urizar, Xabier, and San Vicente Roncal, Iaki. 2012. Detecting Sentiments in Spanish Tweets. In: *TASS 2012 Working Notes.*

Sidorov, Grigori, Miranda-Jiménez, Sabino, Viveros-Jiménez, Francisco, Gelbukh, Alexander, Castro-Sánchez, Noé, Velásquez, Francisco, Díaz-Rangel, Ismael, Suárez-Guerra, Sergio, Treviño, Alejandro, and Gordon, Juan. 2012. Empirical Study of Opinion Mining in Spanish Tweets. In: *LNAI 7629-7630.*

Taboada, Maite, Brooke, Julian, Tofiloski, Milan, Voll, Kimberly, and Stede, Manfred. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, **37**(2), 267–307.

Taulé, Mariona, Mart, M. Antónia, and Recasens, Marta. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In: Calzolari, Nicoletta, Choukri, Khalid, Maegaard, Bente, Mariani, Joseph, Odjik, Jan, Piperidis, Stelios, and Tapias, Daniel (eds), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).*

Turney, Peter D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Pages 417–424 of: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* ACL '02. Stroudsburg, PA, USA: ACL.

Volokh, Alexander, and Neumann, Günter. 2012. Task-oriented dependency parsing evaluation methodology. Pages 132–137 of: *IRI.*

Wu, Yuanbin, Zhang, Qi, Huang, Xuanjing, and Wu, Lide. 2009 (Aug.). Phrase Dependency Parsing for Opinion Mining. Pages 1533–1541 of: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.* ACL, Singapore,.

Yang, Kiduk. 2008. WIDIT in TREC 2008 Blog Track: Leveraging Multiple Sources of Opinion Evidence. In: Voorhees, E. M., and Buckland, Lori P. (eds), *NIST Special Publication 500-277: The Seventeenth Text REtrieval Conference Proceedings (TREC 2008).*

Zhang, Changli, Zeng, Daniel, Li, Jiexum, Wang, Fei-Yue, and Zuo, Wanli. 2009. Sentiment Analysis of Chinese Documents: From Sentence to Document Level. *Journal of the American Society for Information Science and Technology*, **60**(12), 2474–2487.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., and Bing, L. 2011. Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis.