Lingue e Linguaggi
Lingue Linguaggi 22 (2017), 137-168
ISSN 2239-0367, e-ISSN 2239-0359
DOI 10.1285/i22390359v22p137
http://siba-ese.unisalento.it, © 2017 Università del Salento

A HYBRID METHOD FOR THE EXTRACTION AND CLASSIFICATION OF PRODUCT FEATURES FROM USER GENERATED CONTENTS

ALESSANDRO MAISTO, SERENA PELOSI, MICHELE STINGO, RAFFAELE GUARASCI
UNIVERSITY OF SALERNO

Abstract – The research we present in this paper focuses on the automatic management of the knowledge about experience goods and services and their features, starting from real texts generated online by internet users. The details about an experiment conducted on a dataset of product reviews, on which we tested a set of rule-based and statistical solutions, will be described in the paper. The main goals are the review classification, the extraction of relevant product features and their systematization into product-driven ontologies. Feature extraction is performed through a rule-based strategy grounded on *SentIta*, an Italian collection of subjective lexical resources. Features and Reviews are classified thanks to a Distributional Semantic algorithm. In the end, we face the problem of the extracted knowledge organization by integrating the subjective information produced by the internet users within a product-driven ontology. The Natural Language Processing (NLP) tool exploited in the work is *LG-Starship*, a hybrid framework for Italian texts processing based on the Lexicon-Grammar theory.

Keywords: feature extraction; review classification; opinion mining; distributional semantics; feature ontology.

1. Introduction

Internet users and consumers can easily share their opinions with large and heterogeneous groups of people, replacing the power of traditional advertising channels. The information they share can modify the buyer expectations, especially with regard to *Experience Goods* (Nakayama *et al*. 2010); such as *movies* (Duan *et al*. 2008; Reinstein, Snyder 2005), *books* (Chevalier and Mayzlin, 2006), *videogames* (Bounie *et al*. 2005; Zhu, Zhang 2006), *hotels* (Nelson 1970; Ye *et al*. 2011) or *restaurants* (Zhang *et al*. 2010).

The rapid growth of the Internet drew the managers and business academics attention to the possible influences that this medium can exert on customers' information search behaviors and acquisition processes. In summary, the growth of the user generated contents and the eWOM (electronic Word of Mouth) can truly reduce the information search costs. On the other hand, the distance increased by e-commerce, the content explosion and the information overload typical of the Big Data age, can seriously hinder the achievement of a symmetrical distribution of the information, affecting not only the market of experience goods, but also that of search goods.

The largest amount of on-line data is semi structured or unstructured and, as a result, its monitoring requires sophisticated Natural Language Processing (NLP) tools, that must be able to pre-process textual data and automatically access their semantic content.

It is of crucial importance for both customers and companies to dispose of automatically extracted, analyzed and summarized data, which do not include only factual information, but also opinions regarding any kind of good they offer.

Companies could take advantage of concise and comprehensive customer opinion overviews that automatically summarize the strengths and the weaknesses of their products or services, with evident benefits in term of reputation management and customer relationship management. Customer information search costs could be decreased trough the same overviews, which offer the opportunity to evaluate and compare the positive and negative experiences of other consumers who have already tested the same products and services.

In this paper, focusing on the task of feature-based sentiment analysis, we discuss the possibility to associate the precision of rule-based linguistic methods and the effectiveness of statistical algorithms, in order to provide fine-grained visual summaries of opinionated user generated contents, easy to understand and consult for both marketers and consumers.

The work presented here is connected to three bigger projects: the construction of Lexicon-Grammar (LG) based sentiment lexical and grammatical resources for the Italian Language (see Section 4); the creation of a hybrid framework for the Italian NLP (see Section 5) and the formalization of the LG databases in machine-readable format (see Subsection 5.3) in order to develop an interactive NLP web application.

The result is an experiment conducted on user reviews (see Subsection 5.1) which has the main goals of the extraction of relevant product features, their classification and representation into semantic networks (see Section 5.2) and their systematization into product-driven ontologies (see Section 6).

Details about the Lexicon-Grammar theoretical framework and about the task of product feature extraction are respectively given in Section 2 and Section 3.

2. The Lexicon Grammar Theoretical Framework

With Lexicon-Grammar we mean the method and the practice of formal description of the natural language, introduced by Maurice Gross in the second half of the 1960s, who, during the verification of some procedures from the transformational-generative grammar (Chomsky 1965) laid the foundations for a brand new theoretical framework.

LG changed the way in which the relationship between lexicon and syntax was conceived before (Gross 1971, 1975). It has been underlined, for the first time, the necessity to provide linguistic descriptions grounded on the systematic testing of syntactic and semantic rules along the whole lexicon, and not only on a limited set of speculative examples.

In the LG methodology it is crucial the collection and the analysis of a large quantity of linguistic facts and their continuous comparison with the reality of the linguistic usages, by examples and counterexamples.

What emerges from the LG studies is that, associating more than five or six properties to a lexical entry, each one of such entries shows an individual behavior that distinguishes it from any other lexical item. However, it is always possible to organize a classification around at list one definitional property, that is simultaneously accepted by all the item belonging to a same LG class and, for this reason, is promoted as distinctive feature of the class.

The Lexicon-Grammar theory lays its foundations on the Operator argument grammar of Zellig S. Harris, the combinatorial system that supports the generation of utterances into the natural language. Saying that the operators store inside information regarding the sentence structures means to assume the nuclear sentence to be the minimum discourse unit endowed with meaning (Gross 1992b).

This premise is shared with the LG theory, together with the centrality of the distributional analysis, a method from the structural linguistics formulated for the first time by (Bloomfield 1933) and then perfected by (Harris 1970). The insight that some categories of words can somehow control the functioning of a number of actants through a dependency relationship called valency, instead, comes from (Tesnière 1959).

The Lexicon-Grammar framework offers the opportunity to create matches between sets or subsets of lexico-syntactic structures and their semantic interpretations. The base of such matches is the connection between the *arguments*, selected by a predicative item listed in *predicate tables*, and the *actants* involved by the same semantic predicate. In fact, as the Semantic Predicates Theory established, the whole set of syntactical structures of a given language (*Sy*) is connected with the entire collection of the semantic items of the same language (*Se*) by means of interpretation rules.

In general, the role played by the arguments of a given Predicate is not modified by the syntactic transformations in which their Predicate is involved. In order to semantically label the arguments in a correct way, they must be always carried to their original forms.

The LG framework uses a specific set of notion in order to describe sentences: *N*, that is followed by a number which specifies its nature (*N0* for the sentence formal subject, *N1* for the first complement and *N2* for the second complement), always indicates a nominal group; *V* represent the verbs; *Prep* stands for the prepositions and *Ch F* indicates the presence of completive or subjective clauses.

The choice of this paradigm is due to its compatibility with the purposes of the computational linguistics, that require a large amount of linguistic data in order to reach high performances in results. This data must be as much as possible, exhaustive, reproducible and well organized. Such richness in term of information opens the possibility to adapt the data to any kind of theoretical frameworks. Recent works based on the LG data are, for example, (Gardent 2005), (Tolone 2009) and (Sagot 2010).

3. The Task of Product Feature Extraction

Opinions are defined by (Liu 2010) as positive or negative views, attitudes, emotions or appraisals about a topic, expressed by an opinion holder in a given time. They are represented by a quintuple that involves an object of the opinion, its features, the positive or negative opinion semantic orientation, the opinion holder and the time in which the opinion is expressed.

The purpose of the sentiment analysis based on features is to provide companies with customer opinions overviews, which summarize the strengths and the weaknesses of the products and services they offer in an automatic way.

We can refer to both opinion objects and features with the term target (Liu, 2010), represented by the following function:

$$T=O(f)$$

Where the object can take anytime the shape of products, services, individuals, organizations, events, topics, etc., and the features are component or attributes of the object. Each object O is represented as a “special feature” and defined by a subset of features. It is formalized in the following way:

$$F = \{f1, f2, . . . , fn\}$$

Targets can be automatically discovered in texts through both synonym words and phrases *Wi* or indicators *Ii*:

$$Wi = \{wi1, wi2, . . . , wim\}$$

$$Ii = \{ii1, ii2, . . . , iiq\}$$

Discover the topic and the features of an opinionated document as well as its overall orientation is essential in order to discern the aspect of a product that must be improved, or whether the opinions extracted by the Sentiment Analysis applications are relevant to the product or not.

3.1. State of the Art on Feature-based Opinion Mining

Pioneer works on feature-based opinion summarization are (Hu, Liu 2004, 2006); (Carenini *et al*. 2005); (Riloff *et al*. 2006) and (Popescu, Etzioni 2007). Both (Popescu, Etzioni 2007) and (Hu, Liu 2004) firstly identified the product features on the base of their frequency and, then, calculated the Semantic Orientation of the opinions expressed on these features. In order to find the most important features commented in reviews (Hu, Liu 2004) used the association rule mining, thanks to which frequent itemsets can be extracted in free texts. Redundant and meaningless items are removed during a Feature Pruning phase.

(Hu, Liu 2006) presented the algorithm ClassPrefix-Span that aimed to find special kinds of patttern, the Class Sequential Rules (CSR), using fixed target and classes.

(Carenini *et al*. 2005) propose a method based on supervised and unsupervised approaches. Crude (learned) features are mapped into a User-Defined taxonomy of the entity's Features (UDF), which provided a conceptual organization for the information extracted. This method took advantages from a similarity matching, in which the UDF reduced the redundancies by grouping together identical features and then organized and presented information by using hierarchical relations).

(Riloff *et al*. 2006) used the subsumption hierarchy in order to identify complex features and, then, reduce the feature set by removing useless features, which have, for example, a more general counterpart in the subsumption hierarchy. The feature representations used for opinion analysis are n-grams (unigrams, bigrams) and lexicosyntactic extraction patterns.

(Popescu, Etzioni 2007) presented OPINE, an unsupervised feature and opinion extraction system, that used as corpus web pages in order to identify explicit and implicit features and relaxation-labelling methods to determin the Semantic Orientation of words. The system draws on WordNet's semantic relations and hierarchies for the individuation of the features (parts, properties and related concepts) and the creation of clusters of words.

(Ferreira *et al*. 2008) made a comparison between the likelihood ratio test approach (Yi *et al*. 2003) and the Association mining approach (Hu, Liu 2004).

Double Propagation (Qiu *et al*. 2009) focuses on the natural relation between opinion words and features. Because opinion words are often used to modify features, such relations can be identified thanks to the dependency grammar. Because these methods have good results only for medium-size corpora, they must be supported by other feature mining methods.

The strategy proposed by (Zhang *et al*. 2010) is based on “no patterns” and part-whole patterns (meronymy) which found noun phrases (“battery”, “a big screen”, “a cover”) and concept phrases (“the camera”, “mattress”, “the phone”) accompanied by verbs or prepositions. The verbs used are “has”, “have”, “include”, “contain”, “consist”, etc. “No” patterns are feature indicators as well. Examples of such patterns are “no noise” or “no indentation”.

(Somprasertsri, Lalitrojwong 2010) used a dependency based approach for the opinion summarization task. A central stage in their work is the extraction of relations between product features (“the topic of the sentiment”) and opinions (“the subjective expression of the product feature”) from online customer reviews. Adjectives and verbs have been used in this study as opinion words. The maximum entropy model has been used in order to predict the opinion-relevant product feature relation.

Because it is possible to refer to a particular feature using several synonyms, (Somprasertsri, Lalitrojwong 2010) used semantic information encoded into a product ontology, manually built by integrating manufacturer product descriptions and terminologies in customer reviews.

(Wei *et al*. 2010) proposed a semantic-based method that uses a list of positive and negative adjectives defined in the General Inquirer to recognize opinion words and, then, extracted the related product features in consumer reviews.

(Xia, Zong 2010) performed the feature extraction and selection tasks using word relation features, which seems to be effective features for sentiment classification because they encode relation information between words.

(Gutiérrez *et al*. 2011) exploited Relevant Semantic Trees (RST) for the word-sense disambiguation and measured the association between concepts, at the sentence-level, using the association ratio measure.

(Mejova, Srinivasan 2011) explored different feature definition and selection techniques (stemming, negation enriched features, term frequency versus binary weighting, n-grams and phrases) and approaches (frequency based vocabulary trimming, part-of-speech and lexicon selection and expected Mutual Information.

Concordance based Feature Extraction (CFE) is the technique used by (Khan *et al*. 2012). After a traditional pre-processing step, regular expressions are used to extract candidate features. Evaluative adjectives, collected on the base of a seed list from (Hu, Liu 2004), are helpful in the feature extraction task. In the end, a grouping phase found the appropriate features for the opinion's topic, grouping together all the related features and removing the useless ones. The algorithm used in this phase is based on the co-occurrence of features and uses the left and right feature's context.

According to (Khan *et al*. 2012), (Wei *et al*. 2010) and (Zhang, Liu 2011) selected candidate product features using noun phrases that appear in texts close to subjective adjectives. The centrepiece of the Khan's method is represented by hybrid patterns, Combined Pattern Based Noun Phrases (cBNP) that are grounded on the dependency relation between subjective adjectives (opinionated terms) and nouns (product features). Nouns and adjectives can be sometimes connected by linking verbs (e.g. "camera produces fantastically good pictures"). Preposition based noun phrases (e.g. "quality of photo", "range of lenses") often represents entity-to-entity or entity-to-feature relations. The last stage is the proper feature extraction phase, in which, using an *ad hoc* module, the noun phrases of the cBNP patterns have been designated as product features.

4. Anchoring the Feature Recognition on Evaluative Adjectives

In this paper we present the results of an experiment on feature based sentiment analysis, in which some of the more used statistical algorithms are applied to a corpus of opinionated reviews, that had already been preprocessed and syntactically parsed through a hybrid framework based on the Lexicon-Grammar theoretical assumptions: *LG-Starship* (Maisto 2017).

Before we start the description of our work, we must specify that, in the Semantic Predicates Theory and in the LG approach in general, the *predicativity* is not a property necessarily possessed by a particular class of morpho-syntagmatic structures, e.g. verbs, that carry information concerning person, tense, mood, aspect, but it is basically determined by the connection between elements (Giordano, Voghera, 2008; DeMauro, Thornton 1985). The concept of operator, in fact, does not depend on specific part of speech, therefore also nouns, adjectives and prepositions can possess the power to determine the nature and the number of the sentence arguments (D'Agostino 1992). Because only the verbs carry out morpho-grammatical information regarding the mood, tense, person and aspect, they must give this kind of support to non-verbal operators. The so called Support Verbs (Vsup) are different from auxiliaries (Aux), that instead support other verbs. Support verbs can be, case by case, substituted by stylistic, aspectual and causative equivalents.

In our experiment, we grounded the linguistic analysis on a subset of adjectives that from now on we will call *evaluative adjectives* (*AggVal*). In the next paragraph we will go in depth through the description of this kind of words, which have been selected from the Italian sentiment lexicon *SentIta* (Maisto, Pelosi 2014; Pelosi 2015).

Now we just anticipate the fact that, adjectives, as it is commonly recognized in literature (Hatzivassiloglou, McKeown 1997; Hu, Liu 2004; Taboada *et al*. 2006), seem to be the most reliable *semantic orientation* indicators among other Part-Of-Speech. This idea is confirmed by the composition of our corpus (see Section Corpus), if we consider that the 17% of the adjectives occurring in the corpus are polarized, compared to the 3% of the adverbs, the 2% of nouns and the 7% of verbs.

Moreover, considering all the opinion bearing words in the corpus, we notice that the adjectives' sentences cover 81% of the total number of occurrences (almost 5000 matches), while the adverbs, the nouns and the verbs reach, respectively, a percentage of 4%, 6%, and 2%. The remaining 7% is covered by the other sentiment expressions that, in any case, contribute to the achievement of satisfactory levels of Recall.

4.1. Adjectives expressing subjectivity in SentIta

SentIta is a sentiment lexical database that directly aims to apply the Lexicon-Grammar theory, starting from its basic hypothesis: the minimum semantic units are the elementary sentences, not the words (Gross 1975).

Therefore, in this work, the lemmas collected into the dictionaries and their Semantic Orientations are systematically recalled and computed into a specific sentence or phrase context. On the base of their combinatorial features and co-occurrences contexts, the SentIta lexical items can take the shapes of *operators*, the predicates, that can be verbs, nouns, adjectives, adverbs, multiword expressions, prepositions and conjunctions, or *arguments*, the predicate complements, that can be nominal and prepositional groups or entire clauses (Buvet *et al*. 2005; Elia 2014a).

Table 1 presents a summary, in term of percentage values, of the composition of the adjective dictionary in SentIta.

Adjectives	**Entries**
Positive Items in SentIta	1,358
Negative Items in SentIta	3,385
Intensifiers in SentIta	638
Neutral Adjectives in Sdic_it	28,664
Adjectives in Sdic_it	34,045

Table 1
Evaluative Adjectives of SentIta.

As exemplified above, the expressions in which we inserted the adjectives from SentIta are copulative constructions of the kind

N_0 *essere Agg Val*

where *Agg Val* represents an adjective that expresses an evaluation (Elia et al., 1981).
The verbs' equivalents included in this case are the following:

- aspectual equivalents: *stare* "to stay", *diventare* "to become", *rimanere, restare* "to remain";
- causative equivalents: *rendere* "to make";
- stylistic equivalents: *sembrare* "to seem", *apparire* "to appear", *risultare* "to result", *rivelarsi* "to reveal to be", *dimostrarsi, mostrarsi* "to show oneself to be".

Among the Italian LG structures that include adjectives we selected the following, in which polar and intensive adjectives occur with *essere* (Meunier 1984; Vietri 2004):

- Sentences with polar adjectives:
  - *N0 essere Agg Val*, *L'idea iniziale era accettabile*, "The initial idea was acceptable";
  - *V-inf essere Agg Val*, *Vedere quel film è stato demoralizzante,* "Watching this movie is demoralizing";
  - *N0 essere Agg Val di V-Inf*, *La polizia sembra incapace di fare indagini* "The police seems unable to do investigate"
  - *N0 essere Agg Val a N1*, *La giocabilità è inferiore alla serie precedente*, "The playability is worse than the preceding series";
  - *N0 essere Agg Val Per N1*, *Per me questo film è stato noioso,* "In my opinion this movie was boring"

- Sentences with adjectives as nouns intensifiers and downtoners:

- *N0 essere Agg Int di N1*, *Una trama piena di falsità,* “A plot filled with mendacity”

The support verb *avere* “to have” (and its equivalent *tenere*) has been observed into the structure *Nb Vsup Na V-a,* in which it is involved a special kind of nominal group subject that contains *noms appropriés* “appropriate nouns” *Napp* (Guillet, Leclère 1981; Harris 1970; Laporte 1997, 2012; Meydan, 1996, 1999).

Citing (Laporte 2012, p. 1), “A sequence is said to be appropriate to a given context if it has the highest plausibility of occurrence in that context, and can therefore be reduced to zero. In French, the notion of appropriateness is often connected with a metonymical restructuration of the subject.” and (Mathieu 1999b, p. 122), “*On considère comme substantif approprié tout substantive Na pour lequel, dans une position syntaxique donnée, Na de Nb = Nb*”.

We can clarify that “the notion of highest plausibility of occurrence of a term in a given context” (Laporte 2012) should not be interpreted in statistic terms or proved by searches in corpora, but just intuitively defined through the paraphrastic relation

Na di Nb = Nb

According to (Meydan 1996, p. 198), “the adjectival transformations with *Napp* (n.b. *(Na di Nb)Q essere V-a* =: *Il fisico di Lea è attraente* “The body of Lea is attractive”) can be put in relation through four types of transformations”, which correspond also to the structures included into our network of sentiment FSA. The obligatoriness of the modifiers and the appropriateness of the nouns are reflected in these transformations (Laporte 1997).

- Nominal constructions *Vsup Napp:*
  - *Nb Vsup Na V-a*, *Lea ha un fisico attraente* “Lea has an attractive body”

- Restructured sentences in which the GN subject is exploded into two independent constituents:
  - *Nb essere V-a Prep Na*, *Lea è attraente (per il suo + di) fisico* “Lea is attractive for her body”

- Metonymic sentences in which the *Napp* is erased:
  - *Nb essere V-a*, *Lea is attractive* “Lea is attractive”

- Constructions in which the Napp is adverbalized:
  - *Nb essere Na-mente V-a, Lea è fisicamente attraente* “Lea is physically attractive”

Moreover, into the Sentiment Analysis field, where the identification and the classification of the features of the opinion object even consist in a whole subfield of research, the *Napp* becomes a very advantageous linguistic device for the automatic feature analysis. See, for example, in which *Na (Napp)* is the feature and the *Nb* (human noun, *N-um*) is the object of the opinion.

Also on the base of their frequency in written and spoken corpora and in informal and formal speech, together with (Giordano, Voghera 2008), we consider verbless expressions syntactically and semantically autonomous sentences, which can be coordinated, juxtaposed and that can introduce subordinate clauses, just like verbal sentences. Among the verbless sentences available in the Italian language, we are interested here on those involving adjectives indicating appreciation (*Agg Val*), e.g. *Bella questa!* “Good one!” (DeMauro, Thornton 1985; Meillet 1906;).

In this Paragraph we also mention the use of the verbs of evaluation *Vval* (Elia *et al*. 1981), which represent a subclass of the LG class 43, grouped together through the acceptance of at least one of the properties *N1=:N1 Agg1 and N1=:Agg1 Ch F*. Examples are *giudicare* “to judge”, *trovare* “to consider”, *avvertire* “to notice”, *valutare* “to evaluate”, etc. Of course the *N1 Agg* here can include an *Napp,* that takes the shape of *(Na di Nb)1 Agg,* just as happens with the psychological predicates of Mathieu (1999b).

5. Experiment

5.1. The Dataset of Product Reviews

The first step of the experiment consisted in the Corpus Collection: the corpus dataset has been built using Italian opinionated texts in the form of users' reviews and comments found on e-commerce and opinion websites. It contains 600 text units (50 positive and 50 negative for each product class) and refers to three different domains, hotels, smartphones and videogames, for all of which different websites have been exploited. Each single review has been stored with a tag structured as follow:

C##P#

C indicates the category: H for hotels, V for videogames, C for smartphones; M for movies; B for books and C for Cars the category is followed by a numerical identity ranging from 00 to 50. The polarity of the opinion is expressed by a P for positive and N for negative followed by a number indicating the value of the opinion given by the user.

The composition of the reviews dataset is summarized in table 2.

Text features	**Cars**	**Smartphones**	**Books**	**Movies**	**Hotels**	**Games**	**Tot**
Neg docs	50	50	50	50	50	50	300
Pos docs	50	50	50	50	50	50	300
Text files	20	20	20	20	20	20	120
Word forms	17,163	19,226	8,903	37,213	12,553	5,597	101,655
Tokens	21,663	24,979	10,845	45,397	16,230	7,070	126,184

Table 2
Dataset of opinionated online customer reviews

5.2. Text Preprocessing

Part-of-speech (POS) tagging represents the essential baseline for any kind of further linguistic analyses. It is considered a “solved task”, with the state-of-the-art taggers achieving a precision of 97%-98% (Shen *et al*. 2007; Toutanova *et al*. 2003). Even though there are several resources available for the English language, the quantity of tools currently existing for the Italian language is very limited, especially if we consider the tools available for free.

We mention an Italian version of TreeTagger (Schmid 1995), an Italian model for OpenNLP (Morton *et al*. 2005), TagPro (Pianta, Zanoli 2007), CORISTagger (Favretti *et al*. 2002), Tanl POS tagger (Attardi *et al*. 2009), ensemble-based taggers (Dell’Orletta 2009) and Pattern tagger (Smedt, Daelemans 2012). Among the others, only TreeTagger, Pattern and OpenNLP are open source.

Due to this deficiency, we used in this work a brand new averaged perceptron POS Tagger, based on an algorithm widely used in many python libraries for the English language (NLTK,[1] Spacy[2]).

Furthermore, as regards lemmatization, we assumed instead that a morphologically rich language like Italian requires a lexicon-sensitive approach able to cope with the variety of wordforms and capable to provide high performances in term of precision. Therefore, we used a lemmatizer that takes advantage of the huge amount of linguistic data provided by the Italian Electronic Dictionaries developed by the researchers of the Department of Political, Social and Communication Sciences of the University of Salerno.

[1] http://www.nltk.org/
[2] http://spacy.io/

5.2.1. Pos Tagging and Lemmatization

In order to perform the Part-of-Speech Tag and the lemmatization of the corpus, we use the `Mr.Ling` Module of LG-Starship.

The PosTagger, called `Mr.Tag`, is based on the implementation of an Averaged Perceptron Tagger[3] for the English language, part of `TextBlob`[4] module and is written in Python 2.7. As the original model is optimized for English language, the original algorithm has been modified for the `Mr.Tag` Pos Tagger.

The set of features from the original model has been expanded, in order to make it more suitable for Italian language. The new features introduced this way are focused on the morpho-syntactic differences existing between the English and the Italian language. Then, a semi-supervised training phase has been performed on a 1 million-word tagged corpus extracted from the "*Paisa´ Corpus*" (Lyding *et al*. 2014). The tagset for the part of speech annotation is the one used for the "`DELA`" dictionaries (Elia 1995; Elia *et al*. 2010).

The Lemmatizer, called `Mr.Lemma`, is based on a set of dictionaries annotated in the DELAF, the Italian Electronic Dictionary of Flexed Forms. The DELAF, a dictionary including over 1 million of flexed Italian forms, has been divided into 6 sub-dictionaries in order to improve the performance of the algorithm. In detail, `Mr.Lemma` uses different dictionaries for each part of speech (Nouns, Verbs, Adjectives, Preposition and Determiners). In a first step, `Mr.Lemma` separates and lemmatizes Compound Prepositions, then it labels each word with the corresponding Tag obtained after the POS-Tagging phase and search in the correspondent dictionary the Lemma. Afterwards, a new iteration search the unrecognized words' lemmas in the other dictionaries.

Both the Postagger and the Lemmatizer have been tested on other section of the "*Paisá Corpus*" and reach respectively the 91.5% and the 92% of precision.

5.3 Enriched Lexical Resources

In order to make LGstarship capable to deal with the feature-based sentiment analysis we enriched its basic linguistic resources with some *ad hoc* dictionaries from the SentIta database, namely a list of evaluative adjectives, a small selection of support verb equivalents and a collection of evaluative verbs.

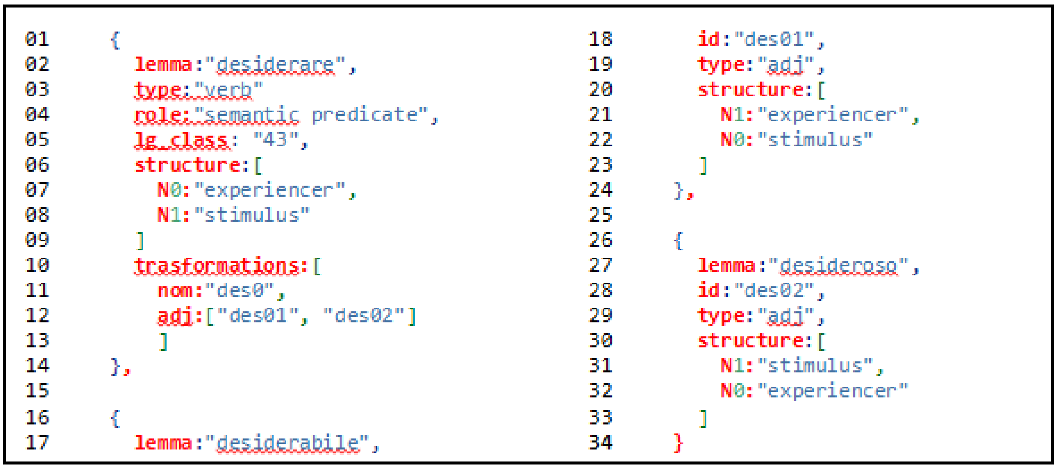
[3] http://www.nltk.org/_modules/nltk/tag/perceptron.html
[4] https://textblob.readthedocs.io/en/dev/

The idea is to work on the richness of the LG resources, making them compatible with popular programming languages, such as Java and Python in order to make the data stored into LG tables immediately usable in NLP automatic applications.

Here we present just an example from a wider framework that aims at the definitive conversion of the full LG databases into the Json (JavaScript Object Notation)[5] format.

We chose to use Json instead of the best known Xml for several reasons. First of all, given the large amount of data, the ultimate goal is to make these resources accessible and searchable using an interactive web application. The Json format is currently the best for use of data in web applications. Moreover, Json is smaller, faster and lightweight if compared to XML. It is, also, easier to read and write both for humans and machine.

Although it has never been used for linguistic purposes, we believe that Json format is particularly effective in the representation of linguistic structures, and compatible with some basic principles of the Lexicon-Grammar theory. Lexicon-Grammar argues that each lemma is tested on a variable number of properties, which can be accepted or not. There is a global reference, since each lemma has a specific behavior and which cannot be generalized. This perfectly matches with the Json philosophy, which by its nature is schemeless. Json schema is extremely elastic and allows to represent different structures and properties - such as linguistic ones - without the need to have a general reference schema.



```
{
  lemma:"desiderare",
  type:"verb"
  role:"semantic predicate",
  lg_class: "43",
  structure:[
    N0:"experiencer",
    N1:"stimulus"
  ]
  trasformations:[
    nom:"des0",
    adj:["des01", "des02"]
    ]
},

{
  lemma:"desiderabile",
  id:"des01",
  type:"adj",
  structure:[
    N1:"experiencer",
    N0:"stimulus"
  ]
},

{
  lemma:"desideroso",
  id:"des02",
  type:"adj",
  structure:[
    N1:"stimulus",
    N0:"experiencer"
  ]
}
```

Figure 1
Example of the *Json* Code

[5] http://www.json.org/.

According with the Semantic Predicate Theory (see Paragraph 2), we created a separated Json object for each semantic predicate. Every object has attributes expressing its semantic, transformational, distributional and structural properties and is linked to its possible nominalizations and adjectivalizations. Figure 1 shows a simplified version of the Json code for the semantic predicates formalization. In the example we formalized the psychological predicate *desiderare* "to desire", from the LG class 43 and its adjectivalizations *desiderabile* "desirable" and *desideroso* "desirous". Specifications about the semantic roles evoked by each object are relevant here because of the necessity to locate the product feature that in this case coincide with the *stimulus*.

5.4. Automatic Classification of Features and Reviews

5.4.1. Text Annotation and Syntactic Parsing

In order to perform the feature extraction on the reviews, we applied the Lexicon-Grammar based dependency parsers of the *LG-Starship Framework* enriched with the *SentIta* resources.

The parser, which is based on the lexical resource written in Json, uses predicates as anchors to determine the sentence structures by differentiating between arguments (essential complements) and unessential complements. In this work the evaluative adjectives have been used as clue for the extraction of the product features.

According to the sentence structures described in section 4.1, the parser tags evaluative adjectives as predicate and the proper argument as feature (information around the argument that plays the role of product feature are stored in the Json file, as stated in the previous paragraph). Obviously, such adjectives can appear also into nominal groups; in these cases they will be tagged as arguments and the features will be the modified nouns, e.g. *adatto a chi cerca una macchina grintosa* "suitable for those that search for a scrappy car".

Figure 2 shows examples of the parser's graphical representation. The sentence on the right, *la cucina dell'hotel era davvero fantastica*, "the food service of the hotel was truly amazing", presents a copulative structure with a modifier in the SV. The sentence on the left is a verbless sentence.

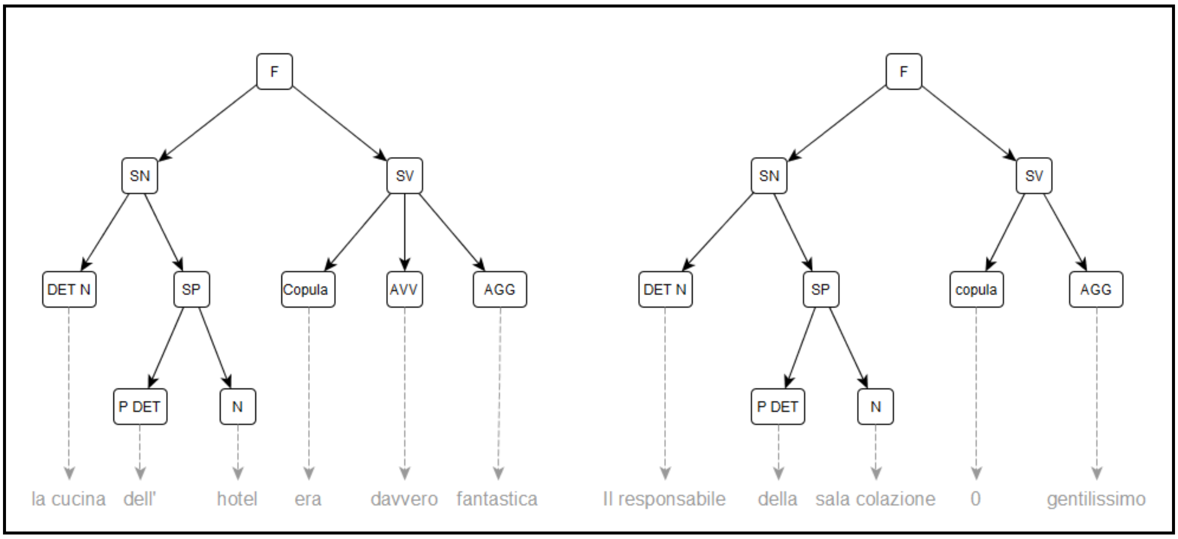


Figure 2
Graphical Representation of the Parser's output.

Figure 3 shows the semantic representation of the opinion quintuple described by (Liu 2010). The graph examines the structure of the opinion. The red shapes contain the information about the classified features and the polarity.

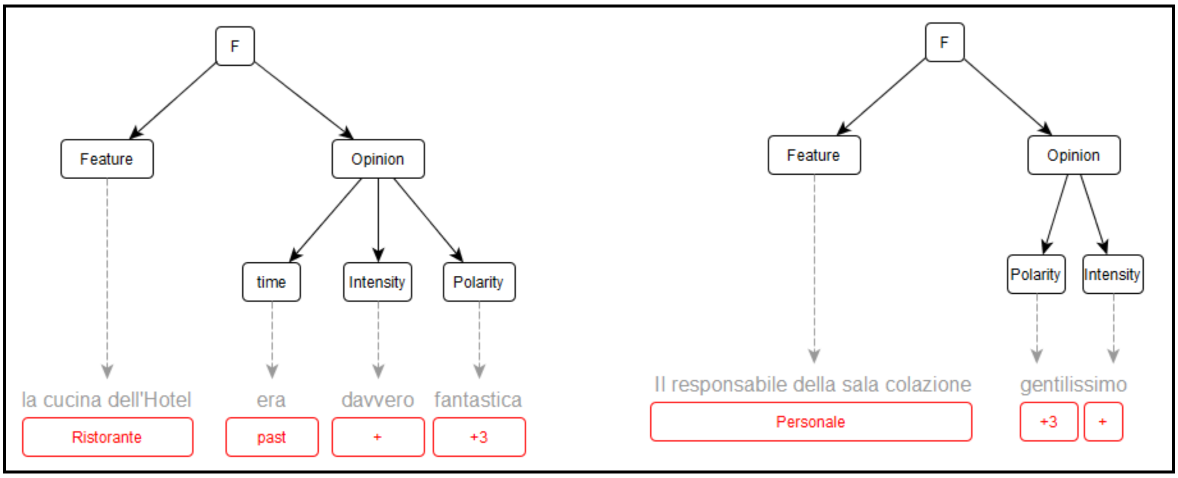


Figure 3
Graphical representation of the semantic tags given by the parser.

5.4.2 The Similarity Measure

The database of sentences that is the output of the parser has been used as input for the next step of the task which concerns the semantic expansion of the results. This stage carries out two operations: the review classification, for the definition of the opinion object, and the feature categorization for the description of the object's characteristics.

Similarity values are measured on the base of a large co-occurrence matrix that has been shaped from the analysis of a large corpus by the S-Space Package, a collection of algorithms for the creation of Semantic Spaces written in Java, developed by the Natural Language Processing Group at UCLA. The corpus on which the LG-Starship Semantic Module is based is a lemmatized version of about 45 million of words from the Paisà Italian Corpus.

The review classification is based on the expansion of the semantic network of each extracted sentence. The first step consists of the collection of features (almost always the subject *N0* of the sentence) and the opinion (expressed by the *Agg Val*) and the expansion of the sentence semantic networks, performed by extracting the 50 words that in the similarity matrix present the higher similarity values with opinions and features.

The algorithm creates a matrix of similarity values in which each row represents a sentence and each column a word. The generated semantic network can be visualized as a semantic graph, shown in figure 4. The graph, generated using Gephi (Bastian *et al*. 2009), emphasizes the nodes with the high weighted degree (calculated on connection rating), almost always adjectives.

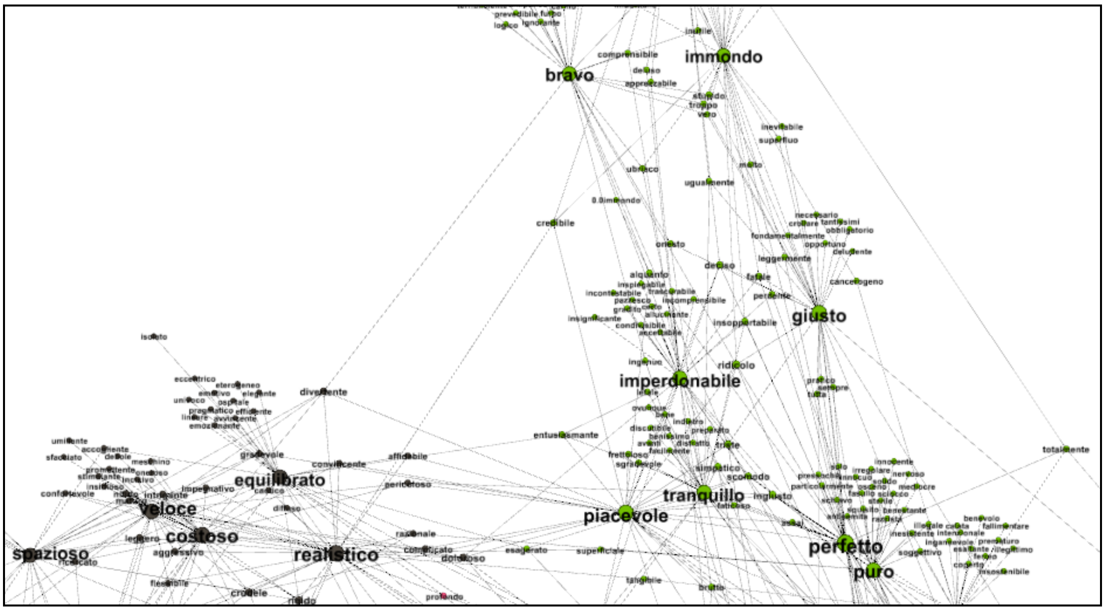


Figure 4
Extract of the word semantic graph of Hotel Reviews.

In a second step, the same semantic expansion algorithm has been tested on to a bigger corpus of 150 reviews of hotels, videogames and smartphones divided into 24 groups containing 5 reviews of a single topic and polarity. Each file has been numerated and named with the number, the initial of the topic (H for Hotels, F for Movies, V for Videogames, C for Smartphone, L for Books and A for Cars) and the polarity value (P for Positive and N for Negative).

Replacing feature nouns and evaluative adjectives used to extract similar words with the file name, we generate another matrix in which rows are files and columns are words.

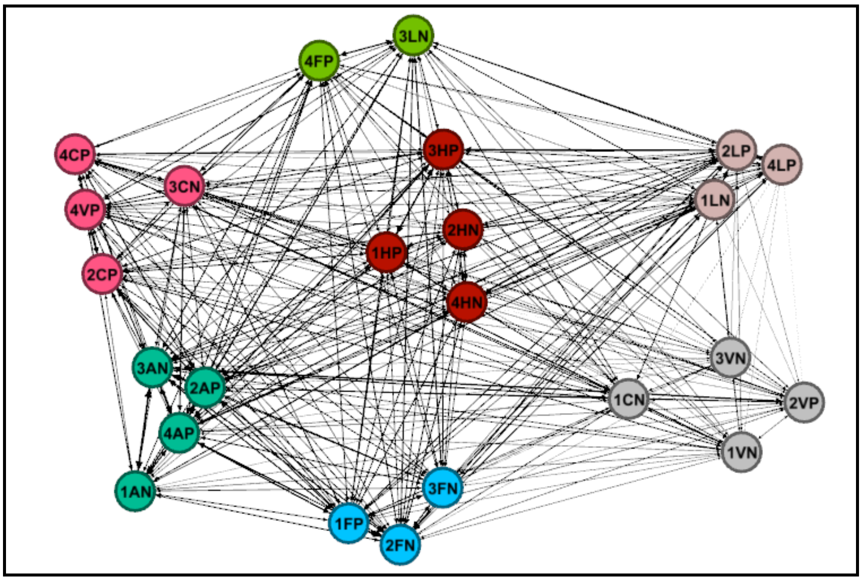


Figure 5
Text distance graph.

With the purpose of finding similarity between sentences, we applied a Cosine Similarity (Huang 2008) to the matrix's vectors and generate a graph in which each node corresponds to a sentence and each arc corresponds to the distance. Then, a Modularity Class algorithm (Newman 2006b) has been applied to the graph in order to highlight any group of sentences.

The Modularity Class algorithm partitions the graph on the base of similarity weights and finds internal communities. The result is shown in the figure 5.

As shown, the red community (hotels, A) and the emerald community (cars, A) include all the files with the expected topics. For what concern other groups it must underlined that smartphones and videogames communities present an error (4VP has been included in the Smartphones community and 1CN in the Videogames Community). Books (L) and movies (F) communities include three correct files and, the missing file of both topics form a different community represented in green colour.

As the well-classified categories are semantically distant from the rest of categories and the errors occur with books and films which could be included in a more general category of “stories” and smartphones and videogames which could be included in a “technology” category, the presence of this kind of errors has to be attributed to the semantic closeness between the topics.

We also used the Semantic Module of *LG-Starship* in order to calculate the mutual semantic similarity of each feature extracted from the corpus. Table 3 shows the similarity measure between different features. As it can be noticed, the features with higher similarity are the ones that possess the stronger semantic relation. We grounded the creation of a graph for the semantic representation of the features on this evidence: here each feature represents a node connected by a weighted arc to the most similar feature. This way, each node possesses several in-arcs but only one out-arc. In addition, the similarity has been calculated with a group of *Generic Features*, which have been inspired by the structured features contained in the reviews webpages. Examples of these generic features for the domain of the hotel are: *pulizia*, “cleaning”; *comodità*, “comfort”; *ambiente*, “location”; *stanze,* “room”; *personale*, “employees”.

Feature 1	Feature 2	Similarity
Colazione “breakfast”	Ristorante “restaurant”	0.907
Colazione “breakfast”	Arredamento “furnitures”	0.828
Colazione “breakfast”	Vista “view”	0.751

Table 3
Similarity values between pairs of features with different semantic relations.

Similarities between extracted features and generic features have been calculated as the average similarity between each extracted feature and a pair of words which represent the generic feature. The similarity between the feature *suite* and the generic feature *stanze* is 0.953842, which is the average value of the similarity *suite-camera* which is 0.958066 and the similarity *suite-stanza* which is 0.949617.

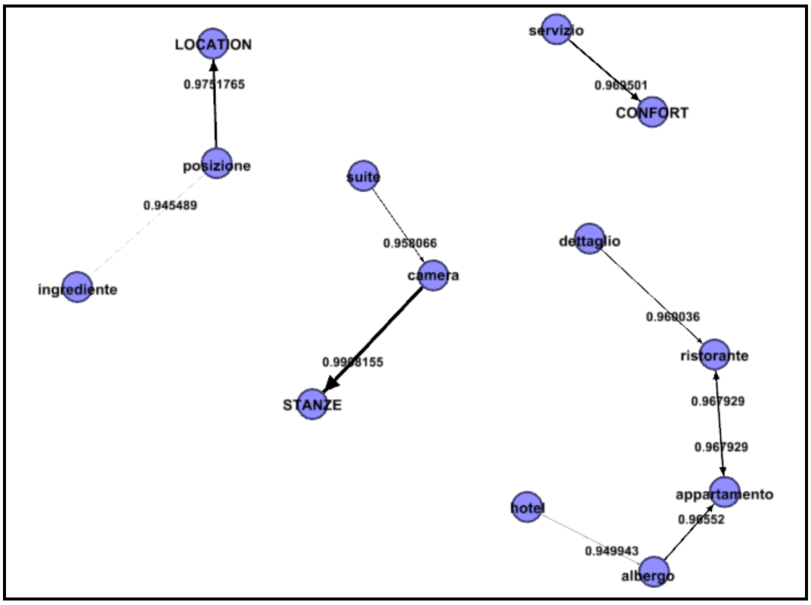


Figure 6
Feature Similarity graph of an hotel review.

By selecting connections with higher values, we generate the graph of Feature similarities shown in figure 6.

In order to classify each extracted feature, the algorithm proceeds in the following way: if the value of each in-arc is higher than the value of out-arcs, the feature is considered as a category and features pointing on it are considered as automatically belonging to this category. Contrariwise it is a sub-feature of another category. When two features point to each other, with a value higher than the value of their in-arcs, both features are considered categories.

In this way, if a feature points to a Generic Feature, the system assumes that it belongs to the respective category, as the feature *camera* and *suite* belong to the category *Stanze* in figure 6.

Anyway, the categories are not all established in advance, but some features with particularly high score, can became categories themselves. This happens with the word *appartamento*, which has been considered a new category in the example shown in figure 6.

6. Towards the Implementation of Opinion-oriented Product Ontologies

Once that meaningful information has been extracted through the method explained in the previous paragraphs, exploring how to store the thick data obtained from users' reviews represents a mandatory step toward the implementation of real time tools to be used both by e-businesses and customers. The results would be deeper marketing insights, for the former, and satisfactory web-browsing experiences, for the latter.

To this extent, creating opinion-oriented ontologies might be the best solution in terms of automatic treatment of fine-grained semantic knowledge.

As (Gruber 1993, p.1) stated, "an ontology is a specification of a representational vocabulary for a shared domain of discourse with definitions of classes, relations and functions". One of the main benefit of using an ontological approach is that the representation of a domain knowledge could be easily manipulated within specific entity relations and restrictions via object-oriented programming scripts.

(Daoud *et al*. 2009), building up on the seminal work of (Gauch *et al*. 2003), proposed an approach where graph-based models (issued from ontology) represent users' profiles. Subsequently, throughout the use of propagation scores and correlation measures, the authors analyzed new submitted queries, eventually bounding them to search results into user' active search sessions via final ranking.

Even though we can surely consider valid the methodology used by the mentioned scholars, trying to satisfy the queries of potential customers by predicting only their browsing paths – without exploiting the information contained within available users' review – does not seem to be the more suitable method, since this kind of approach does not consider at all the structural and abstract features related to goods and services and their evaluations. While trying to merge these pieces of information into a unique predictive model would issue a further level of complexity, an alternative solution in order to provide a better user experience (while improving at the same time the precision of internal search engines results) could be represented by an ontological approach where all the features of a product or service are bound to evaluative instances.

Furthermore, turning the perspective from customers' queries about goods to products and services themselves – and the opinions tied to them – it would better help business actors describing up to date marketing pictures about the reputation of the offered products.

An interesting approach of this kind is proposed by the work of (Wei, Gulla 2010) where the authors applied a feature-extraction algorithm in order to hierarchically build a sentiment ontology tree which describes the domain of digital cameras. In their approach every class of the ontology tree correspond to a feature and each classified feature holds two subclasses representing the negative and the positive polarity.

Despite (Wei, Gulla 2010) contemplated as fundamental the sentiment polarity, this approach is not exempt by downsides, amongst which the most significant is the tree dimension. As the scholars reported, increasing the range of the considered features would produce a decreased computational efficiency, for we are forced to ponder over alternative solutions that do not suffer from the same issue.

The first alternative has been suggested by (Sureka *et al*. 2010) and take advantage of *ConceptNet* in order to build and classify a domain-specific ontology to use for feature-extraction and sentiment classification tasks. *ConceptNet*, a semantic network based on the information collected manually into the OMCS database, is a directed graph in which nodes are concepts, and edges represent assertions of common sense about concepts as these of the following list: *CreatedBy*, *MadeOf*, *PartOf*, *DesireOf*, *DefinedAs*, etc. Since in this paper we are more focussed on enhancing marketing strategies and customers' satisfaction throughout the possible implementation of analytic tools, we do not aim at using ontologies as feature extraction steps as (Sureka *et al*. 2010) and many other scholars did. Nonetheless, we surely recognize that leveraging common-sense frameworks as *ConceptNet* et similia might represent a convenient choice because, as a consequence of the use of a structured knowledge base like the one exploited by (Sureka *et al*. 2010) we could automatically draw a knowledge domain picture, inferring both features and the functional relations without putting too much effort into the design phase.

6.1 A baseline for future implementation

The main purpose of the work drawn up in this Section is to outline hypotheses on how to take advantage of the feature extraction system exposed in 3.2, so as to increase the descriptive range and the inferential power of the ontology that should play the central role in a real-time analytic tool to develop further. Regarding the domain description, the preferred format would be the Web Ontology Language (OWL[6]).

For the purpose of the exposition we have created with Protegé, a basic ontology for the *Accommodation* domain from the hotels corpus of reviews.

The pros of using Protegé are several, but probably the most important is the possibility to create specific object properties, which are relationships statements occurring between two class members, and data object properties, which are additional information valid only for selected members of a class. In other words, not only using the object properties specification we could be able to expand the descriptional range of the ontology with more relations, compared to those available within semantic resources such as *ConceptNet*. and the like, but using the data object properties we could be even able to

[6] For a comprehensive explanation of the OWL see: https://www.w3.org/TR/2004/REC-owl-ref-20040210

represent particular features, that is to say features found to be true only for singular instances of a class member.

The Figure 7 displays an excerpt of the ontology expressing only the first two nodes of the hierarchy.

As it is possible to see, we have three classes describing the accommodation domain, plus one containing the evaluations:

- *Struttura ricettiva* “turist accomodation”: this class contains subclasses describing different kind of accommodation structures (Hotels, Bed&Breakfast, etc). The instances of this class hold the object-property has_service towards the subclasses of the superclass Features.
- *Features*: this class contains two subclasses describing generic or specific features offered by the accommodation structures. Among the generic features we have basic services offered by all the structures (Welcoming, Position, Cleanliness. etc), whereas among the specific features we have particular features offered by some of the structures (Spa, Restaurant, Conference Hall, etc.) The instances of the generic and specific features hold the object-property is_service_of towards the subclasses of the superclass Struttura ricettiva.
- *Personale* “staff”: this class contains subclasses each of which describes a different working position involved within accommodation structures (Receptionist, Chef, Cleaner, etc).
- *Valutazione* “evaluation”: The three superclasses *Struttura_ricettiva*, *Features* and *Personale*, via object-properties relations of the kind x_is_evaluated, point to this superclass which contains all the evaluative instances collected from the labelled texts. The superclass *Valutazione* – which is the core class of our opinion-oriented ontology – is organized into two subclasses, positive and negative. Both subclasses hold the data-property evaluation_score which ranges between the symbols {---, --, -, +, ++, +++}, meant as values of polar strength.

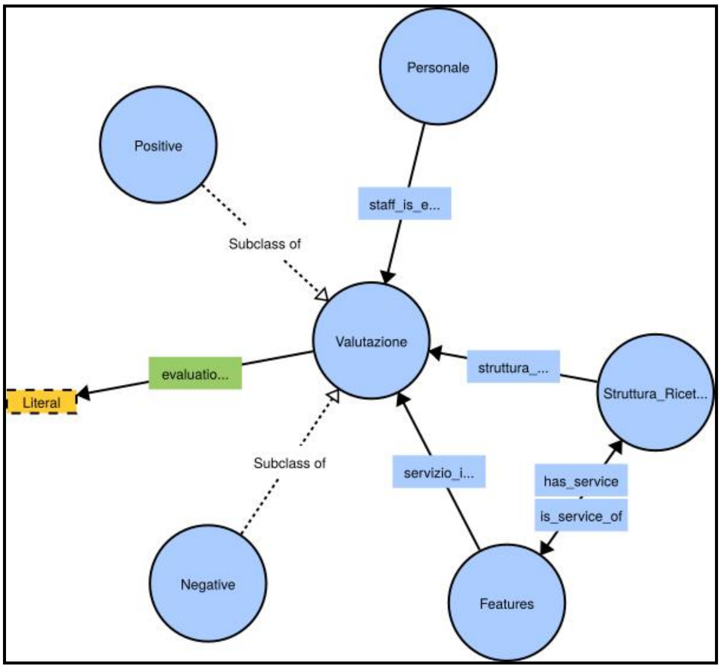


Figure 7
Excerpt of the basic ontology describing the accommodation domain. The resource should not be considered as definitive, for we will expand both the super/subclasses and the relationships occurring between them.

Once that a domain representation is available, the starting point of the algorithm governing a real time marketing tool should correspond to the automatically semantic annotated reviews.

Throughout object-oriented programming we could be able to automatically extract the semantic tags in order to populate classes of the pre-existing domain ontology. To this extent we will encode the semantic information into a format such as the Extensible Markup Language (XML). Let’s have now a look on a portion of a labelled review expressing opinions on some features of a Hotel:

> …<BENEFIT SCORE="3" TYPE="PULIZIA">La pulizia è eccellente</BENEFIT>. <BENEFIT SCORE="3" TYPE="LOCATION">La vista sul mare è splendida</BENEFIT> ma <DRAWBACK SCORE="2" TYPE="LOCATION">la notte purtroppo si sentiva qualche schiamazzo</DRAWBACK >….

As it has been already explained, all the evaluative text portions have a TYPE and a BENEFIT/DRAWBACK score. Every time the algorithm encounters a TYPE tag it has to collect the value as instance of one of the subclasses of the superclass *Features*. If the instance does not match any of the already existing features contained into the superclass, then the instance is automatically collected as new subclass of the specific class.

Once that the value of the TYPE tag has been correctly inserted within the ontology, the algorithm proceeds to extract the appraisals portions, which can be in turn words or simple clauses, collecting these items into the superclass Valutazione as positive if the word/statement has a BENEFIT SCORE, or NEGATIVE if it has a DRAWBACK SCORE.

After collecting the appraisal instance into the correct subclass, the algorithm processes the value of the score, assigning it to the already collected evaluation. For what concerns the class Struttura_ricettiva and Personale, both instances and relationships could be collected wether extracting them from metadata contained within the corpus of labelled reviews, or tracing them throughout exploitation of the power of rule-based inference engines such as *SweetRules*,[7] *JRuleEngine*,[8] *Drools*,[9] *Mandarax*,[10] *Apache Jena*[11] and similar. It should be noticed that for the sake of the exposition we have restricted the application hypothesis only to a single scenario, represented by the accommodation domain. Still we consider the gist of this practical proposition – which corresponds to a software governing an opinion-driven ontology system – as easily replicable in relation to other kind of semantic fields.

7. Conclusion

The present research handled the task of feature-based sentiment analysis with the purpose of automatically manage the knowledge about experience goods and services and their features, starting from real texts generated online by internet users.

The work is connected to three wider projects: the construction of Lexicon-Grammar (LG) based sentiment lexical and grammatical resources for the Italian Language; the creation of a hybrid framework for the Italian NLP and the formalization of the LG databases into a machine-readable format.

Here we presented an experiment conducted on a dataset of user generated contents in the form of product and services reviews.

We performed the extraction of relevant product features, their classification and their representation into semantic networks. We, furthermore, presented a baseline method for the feature systematization into product-driven ontologies.

Anchoring the statistical analysis of the corpus on the annotations produced by a fine-grained linguistic analysis we obtained satisfying results.

[7] http://sweetrules.semwebcentral.org.
[8] http://jruleengine.sourceforge.net/index.html.
[9] https://www.drools.org.
[10] http://mandarax.sourceforge.net.
[11] https://jena.apache.org.

The future lines of action of our research go in the direction of extension of the LG resources described in Json, for a wider and coverage of the LG analyses; the improvement of the syntactic parser, that aims at a better precision in the sentence annotation; and the definition of a method for the automatic population of the products ontologies on the base of the analyzed features.

The advantages of sophisticated NLP methods and software, and their ability to distinguish factual from opinionated language, are not limited to the ones discussed so far; but they are dispersed and specialized among different tasks and domains, such as: Ads placement, Question-answering: chance to distinguish between factual and speculative answers, Text summarization, Recommendation systems, Flame and cyberbullying detection, Literary reputation tracking, Political texts analysis, etc.

As concerns the limitations of this research, we mention, among others, the cases of irony, sarcasm and cultural stereotypes, which still remain open problems for the NLP in general and for the Sentiment Analysis in particular, since they can sometimes completely overturn the description of the sentences.

Bionotes: Alessandro Maisto received the PhD at the Department of Political, Social and Communication Science at the University of Salerno. He is member of Ass.I.Term. In 2014 he was Research Assistant at the University of Naples Federico II. In 2010 he received the master degree in Communication Theory at the University of Salerno and in 2013 he received the master degree in Artificial Intelligence at the Politecnical University of Madrid. Now he works as Research Fellow at the University of Salerno in a Computational Linguistics group. His research interests concerne Parsing Technologies and Distributional Semantics.

Serena Pelosi received the PhD at the Department of Political, Social and Communication Science at the University of Salerno. In 2014 she was Research Assistant at the University of Naples Federico II. In 2011 she received, summa cum laude, the master degree in Corporate and Public Communication at the University of Salerno. Now she works as Research Fellow at the University of Salerno in a Computational Linguistics group. Her research interests focus on Sentiment Analysis and Lexicon Development.

Michele Stingo is a PhD Student at the Department of Political, Social and Communication Science at the University of Salerno. In 2016 he received, summa cum laude, the master degree in Language Sciences at the Ca'Foscari University of Venice. His primary research interests focus on Pragmatic modelling and AI enhancement of industrial NLP solutions.

Raffaele Guarasci is a PhD Candidate at the Department of Political, Social and Communication Science at the University of Salerno. In 2015 he received, summa cum laude, the master degree in Digital Humanities at the University of Pisa. His research interests lie in Deception Detection and Distributional Semantics.

Authors' addresses: amaisto@unisa.it; spelosi@unisa.it; mstingo@unisa.it; rguarasci@unisa.it

References

Attardi G., Fuschetto A., Tamberi F., Simi M. and Vecchi, E. M. 2009, *Experiments in tagger combination: arbitrating, guessing, correcting, suggesting*, in “Poster and Workshop Proceedings of the 11th Conference of the IAAI”, page 10, Reggio Emilia, Italy.

Bacelar da Silva A.J. 2003, *The effect of instruction on pragmatic development: teaching polite refusals in English*, in “Second Language Studies” 22 [1], pp. 55-106.

Bastian M., Heymann S. and Jacomy M. 2009, *Gephi: an open source software for exploring and manipulating networks*, in “ICWSM” 8, pp. 361-362.

Biber D., Johansson S., Leech G., Conrad S. and Finegan E. 1999, *Longman Grammar of Spoken and Written English*, Longman, London.

Bloomfield, L. 1933, *Language*, University of Chicago Press, Chicago.

Bounie D., Bourreau M., Gensollen M. and Waelbroeck P. 2005, *The effect of online customer reviews on purchasing decisions: The case of video games*, in “Retrieved July” 8, p. 2009.

Buvet P.-A., Girardin C., Gross G. et Groud C. 2005, *Les prédicats d’affect*, in “LIDIL” 32, pp. 123-143.

Cameron D. 2005, *Language, Gender and Sexuality: Current Issues and New Directions*, in “Applied Linguistics” 26 [4], pp. 482-502.

Carenini G., Ng R. T. and Zwart E. (eds Clark P., Schreiber G.) 2005, Extracting knowledge from evaluative text, in “Proceedings of the 3rd international conference on Knowledge capture”, pp. 11-18. ACM, New York, NY, USA.

Carbonell J.G. 1979, *Subjective understanding: Computer models of belief systems*. Technical report, DTIC Document.

Carreras X. and Màrquez L. (eds Dagan I., Gildea D.) 2005, Introduction to the conll-2005 shared task: Semantic role labeling, in “Proceedings of the Ninth Conference on Computational Natural Language Learning”, pp. 152-164. Association for Computational Linguistics Stroudsburg, PA, USA.

Chen R. 2010, *Compliment and Compliment Response Research: a Cross-Cultural Survey*, in Trosborg A. (ed.), *Pragmatics Across Languages and Cultures*, Mouton de Gruyter, Berlin, pp. 79-102.

Chevalier J. A. and Mayzlin D. 2006, *The effect of word of mouth on sales: Online book reviews*, in “Journal of marketing research” 43, pp. 345-354.

Chomsky N. 1965, Aspects of the Theory of Syntax. The MIT press, Cambridge, Massachusetts.

Cogo A., Archibald A., Jenkins J. (eds.) 2011, *Latest trends in ELF research*, Cambridge Scholars Publishing, Cambridge.

Comrie B. 1976, *Aspect*, Cambridge University Press, Cambridge.

Daoud M., Tamine-Lechani L., Boughanem M. and Chebaro B. (eds. Shin S. Y., Ossowski S.) 2009, A session based personalized search using an ontological user profile, in “Proceedings of the 2009 ACM symposium on Applied Computing”, pp. 1732-1736. ACM, New York, NY, USA.

Dell’Orletta F. 2009, Ensemble system for part-of- speech tagging, “Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence”, 12th December 2009, Reggio Emilia, Italy. pp. 1-8.

D’Agostino E. 1992, *Analisi del discorso: metodi descrittivi dell’italiano d’uso*, Loffredo, Napoli.

De Longis R. 2001, *La Storia delle donne*, in Di Cori P., Barazzetti D. (a cura di), *Gli studi delle donne in Italia*, Carocci, Roma, pp. 299-320.

De Mauro T. e Thornton A. M. 1985, La predicazione: teoria e applicazione all'italiano, in

"Sintassi e morfologia della lingua italiana d'uso. Teorie e applicazioni descrittive, Atti del XVII Congresso internazionale di studi della SLI, Urbino, 11-13 settembre 1983", Bulzoni Editore, Roma. pp. 407-419.

Di Prospero B. (a cura di) 2004, *Il futuro prolungato*, Carocci, Roma.

Duan W., Gu B. and Whinston A. B. 2008, *The dynamics of online wordof-mouth and product sales – an empirical investigation of themovie industry*, in "Journal of retailing" 84, pp. 233-242.

D'Urso A. 2011, *Histoire des critiques du surréalisme et critique des Histoires du surréalisme. Pour une démystification de l'historiographie surréaliste*, in "Lingue e Linguaggi" 5, pp. 99-110.

Elia A., Martinelli M. e D'Agostino E. 1981, *Lessico e Strutture sintattiche. Introduzione alla sintassi del verbo italiano*, Liguori, Napoli.

Elia A. (eds. Salem A., Lebart L., Bolasco S.) 1995, Dizionari elettronici e applicazioni informatiche, in "In III Giornate internazionali di Analisi Statistica dei dati Testuali, JADT", CISU, Roma, pp. 55-6.

Elia A. 2014a, *Lessico e sintassi tra tempo e massa parlante*, in Marchese M.P., Nocentini A., *Il lessico nella teoria e nella storia linguistica*, Edizioni il Calamo, Roma, pp. 15-47.

Favretti R.R., Tamburini F. and De Santis C. 2002, *Coris/codis: A corpus of written italian based on a defined and a dynamic model*, in Wilson *et al.* (ed.), *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, Lincom-Europa, Munich, pp. 27-38

Ferreira L., Jakob N. and Gurevych I. 2008, *A comparative study of feature extraction algorithms in customer reviews*, in Das C.R. (ed.), *Proceedings of the 2008 IEEE International Conference on Semantic Computing*, IEEE Computer Society, Washington, pp. 144-151.

Gardent C., Guillaume B., Perrier G. and Falk I. 2005, *Maurice gross' grammar lexicon and natural language processing*, in Vetulani, Z. (ed.), *Proceedings of the 2nd Language and Technology Conference*, Springer-Verlag, Berlin, pp 120-123.

Gauch S., Chaffee J. and Pretschner A. 2003, *Ontology-based personalized search and browsing*, in "Web Intelligence and Agent Systems: An international Journal" 1 [3, 4], pp. 219-234.

Giordano R. e Voghera M. 2008, *Frasi senza verbo: il contributo della prosodia*, in Ferrari. A (a cura di), *Sintassi storica e sincronica dell'italiano. Subordinazione, coordinazione, giustapposizione.* Atti del X Congresso della Società Internazionale di Linguistica e Filologia Italiana, Cesati editore, Firenze, pp 1005-1024.

Gross M. 1971, *Transformational Analysis of French Verbal Constructions*, University of Pennsylvania.

Gross M. 1975, *Méthodes en syntaxe. Régime des constructions complétives*. Hermann, Paris.

Gross M. 1992b, *The argument structure of elementary sentences*, in "Language Research" 28, pp. 699-716.

Gruber T. R. 1993, *A translation approach to portable ontology specifications*, in "Knowledge acquisition" 5 [2], pp. 199-220.

Gutiérrez Y., Vázquez S. and Montoyo A. 2011, *Sentiment classification using semantic features extracted from WordNet-based resources*, in Balahur *et al.* (ed.), *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Association for Computational Linguistics, Stroudsburg, pp. 139-145.

Guillet A. et Leclère C. 1981, *Restructuration du groupe nominal*, in “Langages” 63, pp. 99-125.

Halliday M.A.K. and Hasan R. 1976, *Cohesion in English*, Longman, London.

Harris Z. S. 1970, *Discourse analysis*, in Harris Z.S. (ed.), *Papers in structural and transformational linguistics*, Reidel, pp. 313-347.

Hatzivassiloglou V. and McKeown K. R. 1997, *Predicting the semantic orientation of adjectives*, in Cohen, P.R. (ed.), *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the European chapter of the association for computational linguistics,* Association for Computational Linguistics, Stroudsburg, pp. 174-181.

Hollande F. 2012, *Changer de destin*, Robert Laffont, Paris.

Hu M. and Liu B. 2004, *Mining opinion features in customer reviews*, in “AAAI” 4, pp. 755-760.

Hu M. and Liu B. 2006, *Opinion feature extraction using class sequential rules*, in Nicolov *et al.* (ed.), *Computational Approaches to AnalyzingWeblogs*, AAAI Press, Menlo Park, pp. 61-66.

Huang A. 2008, *Similarity measures for text document clustering*, in Holland *et al.* (ed.), *Proceedings of the sixth New Zealand computer science research student conference,* NZCSRSC, Christchurch, pp. 49-56.

Khan K., Baharudin B. B. and City T. 2012, *Identifying product features from customer reviews using lexical concordance*, in “Research Journal of Applied Sciences Engineering and Technology” 4, pp. 833-839.

Laporte E. 1997, *L’analyse de phrases adjectivales par rétablissement de noms appropriés*, in “Langages” 31, pp. 79-104.

Laporte E. 1995. *Appropriate Nouns with Obligatory Modifiers* in “Language Research” 31 [2], pp.251-289.

Laver M., Benoit K. and Garry J. 2003. *Extracting policy positions from political texts using words as data*, in “American Political Science Review” 97, pp. 311-331.

Liu B. 2010, *Sentiment analysis and subjectivity*, in “Handbook of natural language processing” 2, pp. 627-666.

Lyding V., Stemle E., Borghetti C., Brunello M., Castagnoli S., Dell'Orletta F., Dittmann H., Lenci A. and Pirrelli V. 2014, *The paisa corpus of italian web texts*, in Bildhauer F. (ed.), *Proceedings of the 9th Web as Corpus Workshop*, Association for Computational Linguistics, Stroudsburg, pp. 36-43.

Maisto A. and Pelosi S. 2014, *A lexicon-based approach to sentiment analysis. The italian module for nooj*, in Monti *et al.* (ed.), *Formalising Natural Languages with Nooj: Selected papers from the NooJ 2014 International Conference*, Cambridge Scholar Publishing, Newcastle, pp. 37-49.

Maisto A. 2017, *A Hybrid Framework for Text Analysis*. Ph.D Thesis to be published. Department of Political, Social and Communication Sciences. University of Salerno, Italy.

Mathieu Y.Y. 1999, *Un classement sémantique des verbes psychologiques*, in Cortès C (ed.), *Cahiers du CIEL 1996-1997*, Université Paris Diderot, Paris, pp. 115-134.

Mejova Y. and Srinivasan P. 2011, *Exploring feature definition and selection for sentiment classifiers*, in Nicolov N. (ed.) *Proocedings of the fifth international conference on weblogs and social media*, AAAI Press, Menlo Park, pp. 546-549.

Meunier A. 1984, *La sémantique locative de certaines structures: N0 être adj*, in “Revue québécoise de linguistique” 13, pp. 95-121.

Meillet A. 1906, *La Phrase nominale en indo-européen*. Impr. nationale, Paris.

Meydan M. 1996, *Constructions adjectivales, substantifs appropriés et verbes supports*, in “Linx” 34, pp. 197–210. Centre de recherches linguistiques de Paris 10.

Meydan M. 1999, *La restructuration du gn sujet dans les phrases adjectivales à substantif approprié*, in “Langages”, pp 59-80.

Morton T., Kottmann J., Baldridge J. and Bierner G. 2005, *Opennlp: A java-based nlp toolkit*.

Moody L.A. 1999, *Religio-Political Insights of 19th Century Women Hymnists and Lyric Poets*. http://www.janushead.org/JHSumm99/moody.cfm (7.12.2010).

Mullen T. and Malouf R. 2006, *A preliminary investigation into sentiment analysis of informal political discourse*, in Nicolov N. (ed.), *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, AAAI Press, Menlo Park, pp. 159–162.

Nakayama M., Sutcliffe N. and Wan Y. 2010, *Has the web transformed experience goods into search goods?*, in “Electronic Markets” 20, pp. 251–262. Springer.

Newman M.E. 2006b, *Modularity and community structure in networks*, in “Proceedings of the national academy of sciences” 103(23), pp. 8577-8582.

Pelosi S. 2015, *Sentita and doxa: Italian databases and tools for sentiment analysis purposes*, in “Proceedings of the Second Italian Conference on Computational Linguistics” CLiC-it 2015, pp. 226–231. Accademia University Press.

Perelman C. et Olbrechts-Tyteca L. 1958, *Traité de l’argumentation. La nouvelle rhétorique*, P.U.F., Paris; trad. it. di Schick C., Mayer M. et Barassi E. 2001, *Trattato dell’argomentazione. La nuova retorica*, Einaudi, Torino.

Pianta E. and Zanoli R. 2007, *Tagpro: A system for italian pos tagging based on svm*, in “Intelligenza Artificiale” 4(2), pp. 8-9.

Piao S., Ananiadou S., Tsuruoka Y., Sasaki Y. and McNaught J. 2007, *Mining opinion polarity relations of citations*, in “InternationalWorkshop on Computational Semantics” (IWCS), pp. 366–371.

Popescu A.-M. and Etzioni O. 2007, *Extracting product features and opinions from reviews*, in “Natural language processing and text mining”, pp. 9–28. Springer.

Predelli S. 2010, *From the Expressive to the Derogatory: On the Semantic Role for Non-Truth-Conditional Meaning*, in Sawyer S. (ed.), *New Waves in Philosophy of Language*, Palgrave Macmillan, Houndmills/New York, pp. 164-185.

Qiu G., Liu B., Bu J. and Chen C. 2009, *Expanding domain sentiment lexicon through double propagation*, in “IJCAI” 9, pp. 1199–1204.

Reinkowski M. 2002, *Kulturerbe oder Erblast? Zum Status der Turzismen in den Staaten Südosteuropas, insbesondere des Bosnischen*, in “Mediterranean language review” 14 (2002), pp. 98-112.

Reinstein D. A. and Snyder C. M. 2005, *The influence of expert reviews on consumer demand for experience goods: A case study of movie critics*, in “The journal of industrial economics” 53, pp. 27–51.Wiley Online Library.

Reynolds K., Kontostathis A. and Edwards L. 2011, *Using machine learning to detect cyberbullying*, in “Machine Learning and Applications and Workshops, 2011 10th International Conference on” (ICMLA) 2, pp. 241–244. IEEE.

Riloff E., Patwardhan S. and Wiebe J. 2006, *Feature subsumption for opinion analysis*, in “Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing”, pp. 440–448. ACL.

Rosa J.G. 2001 (ed.), *No Urubuquaquá, no Pinhém*, Nova Fronteira, Rio de Janeiro.

Sagot B. 2010, *The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French*, in "7th international conference on Language Resources and Evaluation" (LREC 2010).

Schmid H. 1995, *Treetagger| a language independent part-of-speech tagger*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, pp. 43:28.

Schmitz B. 1975, *Sexism in French language textbook*, in Lafayette R. C. (ed.), *The Cultural Revolution in Foreign Language Teaching*, National Textbook Co., Skokie (IL), pp. 119-130.

Sebeok T. 1976, *Contributions to the Doctrine of Signs*, Indiana University Press, Bloomington; trad. it. di Pesaresi M. 1979, *Contributi alla dottrina dei segni*, Feltrinelli, Milano.

Seki Y., Eguchi K., Kando N. and Aono M. 2005, *Multi-document summarization with subjectivity analysis at duc 2005*, in "Proceedings of the Document Understanding Conference" (DUC).

Shen L., Satta G. and Joshi A. 2007, *Guided learning for bidirectional sequence classification*, in "ACL" 7, pp. 760-767. Citeseer.

Smedt T. D. and Daelemans W. 2012, *Pattern for python*, in "Journal of Machine Learning Research", 13(Jun), pp. 2063-2067.

Somprasertsri G. and Lalitrojwong P. 2010, *Mining feature-opinion in online customer reviews for opinion summarization*, in "J. UCS" 16, pp. 938–955.

Sureka A., Goyal V., Correa D. and Mondal A. 2010, *Generating domain-specific ontology from common-sense semantic network for target specific sentiment analysis*, in "Proceedings of the fifth international conference of the Global WordNet Association", pp. 1-8. Mumbai, India.

Taboada M., Anthony C. and Voll K. 2006, *Methods for creating semantic orientation dictionaries*, in "Proceedings of the 5th International Conference on Language Resources and Evaluation" LREC, Genova, Italy, pp. 427–432.

Terveen L., Hill W., Amento B., McDonald D. and Creter J. 1997, *Phoaks: A system for sharing recommendations*, in "Communications of the ACM" 40, pages 59–62. ACM.

Tesnière L. 1959, *Eléments de syntaxe structurale*. Klincksieck, Paris.

Thüne E.-M.e Leonardi S. 2009, *I colori sotto la mia lingua. Scritture transculturali in tedesco*, Aracne, Roma.

Tolone E. 2009, *Les tables du Lexique-Grammaire au format TAL*, in "MajecSTIC 2009" (pp. electronic-version).

Toutanova K., Klein D., Manning C. D. and Singer Y. 2003, *Feature-rich part-of speech tagging with a cyclic dependency network,* in "Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology" 1, pp. 173-180. ACL.

Vietri S. 2004, *Lessico-grammatica dell'italiano. Metodi, descrizioni e applicazioni*. UTET Università.

Wei C.-P., Chen Y.-M., Yang C.-S. and Yang C. C. 2010, *Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews*, in "Information Systems and E-Business Management" 8, pp. 149–167. Springer.

Wei W.and Gulla J. A. 2010, *Sentiment learning on product reviews via sentiment ontology tree*, in "Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics", pp. 404-413. ACL.

Xia R. and Zong C. 2010, *Exploring the use of word relation features for sentiment classification*, in “Proceedings of the 23rd International Conference on Computational Linguistics: Posters”, pp. 1336–1344. ACL.

Xiang G., Fan B., Wang L., Hong J. and Rose C. 2012, *Detecting offensive tweets via topical feature discovery over a large scale twitter corpus*, in“Proceedings of the 21st ACM international conference on Information and knowledgemanagement”, pp. 1980–1984. ACM.

Ye Q., Law R., Gu B. and Chen W. 2011, *The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings*, in “Computers in Human Behavior” 27, pages 634–639. Elsevier.

Yi J., Nasukawa, T., Bunescu R. and Niblack W. 2003, *Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques*, in “DataMining”, 2003. ICDM 2003. Third IEEE International Conference on, pp. 427–434.

Zhang L., Liu B., Lim S. H. and O’Brien-Strain E. 2010, *Extracting and ranking product features in opinion documents*, in “Proceedings of the 23rd international conference on computational linguistics: Posters”, pp. 1462–1470. Association for Computational Linguistics.

Zhang L. and Liu B. 2011, *Identifying noun product features that imply opinions*, in “Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers” 2, pp. 575–580. ACL.

Zhu F., Zhang X. 2006, *The influence of online consumer reviews on the demand for experience goods: The case of video games*, in “ICIS 2006 Proceedings”, page 25.