# Linked-Data based Domain-Specific Sentiment Lexicons

**Gabriela Vulcu, Raul Lario Monje, Mario Munoz, Paul Buitelaar, Carlos A. Iglesias**

Insight, Centre for Data Analytics, Galway, Ireland
Paradigma Tecnologico
Universidad Politecnica de Madrid, Spain
gabriela.vulcu@deri.org, rlario@paradigmatecnologico.com, mmunoz@paradigmatecnologico.com,
paul.buitelaar@deri.org, cif@dit.upm.es

**Abstract**

In this paper we present a dataset componsed of domain-specific sentiment lexicons in six languages for two domains. We used existing collections of reviews from Trip Advisor, Amazon, the Stanford Network Analysis Project and the OpinRank Review Dataset. We use an RDF model based on the lemon and Marl formats to represent the lexicons. We describe the methodology that we applied to generate the domain-specific lexicons and we provide access information to our datasets.

**Keywords:** domain specific lexicon, sentiment analysis

## 1. Introduction

Nowadays we are facing a high increase in the use of commercial websites, social networks and blogs which permit users to create a lot of content that can be reused for the sentiment analysis task. However there is no common way to representing this content that can be easily exploited by tools. There are many formats for representing the reviews content and different annotations.

The Eurosentiment project [1] aims to developing a large shared data pool that bundles together scattered resources meant to be used by sentiment analysis systems in an uniform way.

In this paper we present domain-specific lexicons organized around domain entities described with lexical information represented using the lemon[2](McCrae et al., 2012) format and sentiment words described in the context of these entities whose polarity scores are represented using the Marl[3] (Westerski et al., 2011) format. Our language resources dataset consists of fourteen lexicons covering six languages: Catalan, English, Spanish, French, Italian and Portuguese and two domains: Hotel and Electronics. Part of the lexicons are built directly from the availale review corpora using our language resource adaptation pipeline and part using an intermediary result of sentiment dictionaries built semi-automatically by our project partners paradigma Tecnologico.

In section 2. we list the datasources that we used to build the lexicons. In section3. we describe the methods, tools and algorithms used to build the lexicons. In section 4. we provide details about the RDF structure of our lexicons conversion.

## 2. Datasources

We used 10000 aspect-based annotated reviews from the Trip Advisor [4] reviews dataset and 600 reviews from the Electronics dataset from Amazon [5]. The TripAdviseor data contains rated reviews at aspect level. Listing 1 shows the TripAdvisor data format:

Listing 1: TripAdvisor data format.

```
<Author>everywhereman2
<Content>THIS is the place to stay at
    when visiting the historical area
    of Seattle. ...
<Date>Jan 6, 2009
<img src="http://cdn.tripadvisor.com/
    img2/new.gif" alt="New"/>
<No. Reader>-1
<No. Helpful>-1
<Overall>5
<Value>5
<Rooms>5
<Location>5
<Cleanliness>5
<Check in / front desk>5
<Service>5
<Business service>5
```

The Amazon electronics corpus consists of plain text reviews with custom ratings annotations. Listing 2 shows the Amazon electronics data format. The annotation [t] stands for the title of the review whereas the numbers in brackets stand for the rating of a certain aspect in the review.

Listing 2: Amazon electronics data format.

```
[t]the best 4mp compact digital
    available camera[+2]##this camera
    is perfect for an enthusiastic
    amateur photographer . picture[+3],
     macro[+3]##the pictures are razor-
    sharp , even in macro...
```

[1] http://eurosentiment.eu/
[2] http://lemon-model.net/lexica/pwn/
[3] http://www.gi2mo.org/marl/0.1/ns.html
[4] http://sifaka.cs.uiuc.edu/ wang296/Data/index.html
[5] http://www.cs.uic.edu/ liub/FBS/Reviews-9-products.rar

The industrial partners used the Stanford Network Analysis Project (SNAP)[6] and the OpinRank Review Dataset [7] (Ganesan and Zhai, 2011). The Stanford Network Analysis Project dataset consists of reviews from Amazon. The data spans a period of 18 years, including 35 million reviews up to March 2013. Reviews include product and user information, review-level ratings, and a plaintext review as shown bellow in Listing 3

Listing 3: SNAP data format.

```
product/productId: B00006HAXW
review/userId: A1RSDE90N6RSZF
review/profileName: Joseph M. Kotow
review/helpfulness: 9/9
review/score: 5.0
review/time: 1042502400
review/summary: Pittsburgh
review/text: I have all of the doo wop
    DVD's and this one is as good ...
```

The OpinRank dataset provides reviews using teh XML format and contains no ratings. The data format is described in Listing 4

Listing 4: OpinRank data format.

```
<DOC>
 <DATE>06/15/2009</DATE>
 <AUTHOR>The author</AUTHOR>
 <TEXT>The review goes here..</TEXT>
 <FAVORITE>User favorites things about
     this hotel.</FAVORITE>
</DOC>
```

We collected thousands of reviews (in English language, for both domains: Hotels and Electronics). It is important to remark that we do not publish these reviews; we publish the derive lexicons by provessing such reviews (i.e.: domain, context words, sentiment words). The language resources heterogeneity was one of the motivation of the EUROSENTIMENT project.

## 3. Method and Tools

One of the tasks of the EUROSENTIMENT[8] project is to develop a methodology that generates domain-specific sentiment lexicons from legacy language resources and enriching them with semantics and additional linguistic information from resources like DBpedia and BabelNet. The language resources adaptation pipeline consists of four main steps highlighted by dashed rectangles in Figure 1: (i) the Corpus Conversion step normalizes the different review corpora formats to a common schema based on Marl and NIF4; (ii) the Semantic Analysis step extracts the domain-specific entity classes and named entities and identifies links between these entities and concepts from the LLOD Cloud. It uses pattern-based term extraction algorithm with a generic domain model (Bordea, 2013) on each document, aggregates the lemmatized terms and computes their ranking in the corpus(Bordea et al., 2013) to extract entity classes that define the domain. We use AELA (Pereira et al., 2013) framework for Entity Linking that uses a DBpedia as reference for entity mentioning identification, extraction and disambiguation. For linking the entities to Wordnet we extend each candidate synset with their direct hyponym and hypernym synsets. Synset words are then checked for occurrence within all the extracted entity classes that define the language resource domain. (iii)The Sentiment Analysis step extracts contextual sentiments and identifies SentiWordNet synsets corresponding to these contextual sentiment words. We base our approach for sentiment word detection on earlier research on sentiment analysis for identifying adjectives or adjective phrases (Hu and Liu, 2004), adverbs (Benamara et al., 2007), two-word phrases (Turney and Littman, 2005) and verbs (Subrahmanian and Reforgiato, 2008). Particular attention is given to the sentiment phrases which can represent an opposite sentiment than what they represent if separated into individual words. For determining the SentiWordnet link to the sentiment words we identify the nearest SentiWordNet sense for a sentiment candidate using Concept-Based Disambiguation (Raviv and Markovitch, 2012) which utilizes the semantic similarity measure 'Explicit Semantic Analysis' (Gabrilovich and Markovitch, 2006) to represent senses in a high-dimensional space of natural concepts. Concepts are obtained from large knowledge resources such as Wikipedia, which also covers domain specific knowledge. We compare the semantic similarity scores obtained by computing semantic similarity of a bag of words containing domain name, entity and sentiment word with bags of words which contain members of the synset and the gloss for each synset of that SentiWordNet entry. We consider the synset with the highest similarity score above a threshold. (iv) the Lexicon Generator step uses the results of the previous steps, enhances them with multilingual and morphosyntactic (i.e. using the CELEX[9] dataset for inflexions) information and converts the results into a lexicon based on the lemon and Marl formats. Different language resources are processed with variations of the given adaptation pipeline. For example the domain-specific English review corpora are processed using the pipeline described in Figure 1 while the sentiment annotated dictionaries like the ones created by our industrial partner are converted to the lemon/Marl format using the Lexicon Generator step.

### 3.1. Paradigma Tecnologico sentiment dictionaries

Our project partner Paradigma Tecnologico used the SNAP and OpiRank review corpora to build the intermediary sentiment dictionaries linked to wordnet synset id following a semi-automatic approach that involved linguists. They used term frequency analysis on the reviews and we ranked the extracted terms based on their occurencies after filtering out hte stop words. These sorted lists were reviewed by linguists to filter only the domain-specific entities. The relevant entities are context entities (e.g. 'room', 'food' etc.) and sentiment words (e.g. 'clean', 'small' etc.).
Then they used a searching-chunking process to achieve the most relevant collocations of the corpora. This task consisted of identification of colocated context entities and

[6] http://snap.stanford.edu/data/web-Amazon-links.html
[7] http://archive.ics.uci.edu/ml/datasets/OpinRank+Review+Dataset
[8] http://eurosentiment.eu/
[9] http://celex.mpi.nl/

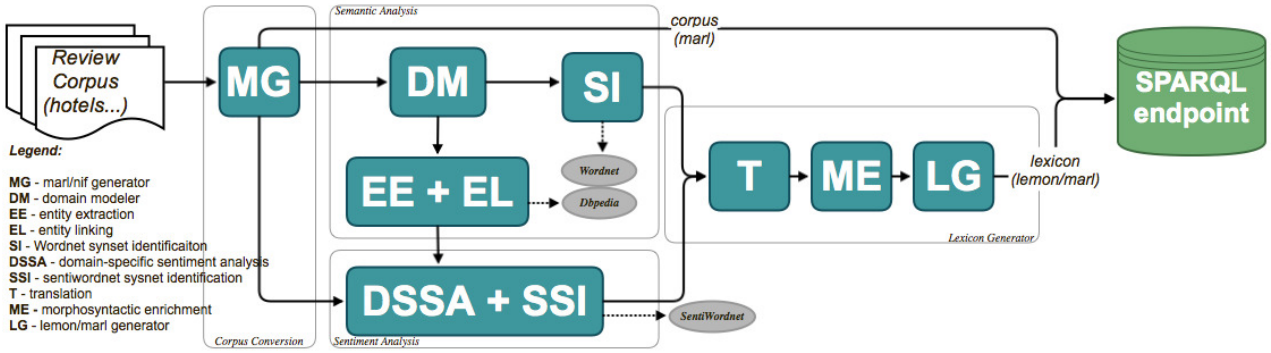


Figure 1: Methodology for Legacy Language Resources Adaptation for Sentiment Analysis.

sentiment words using a 3-word sliding window. The calculated collocations were reviewed again by linguists.
A simple web application helped the linguists to:

- Accept or reject the collocations. Do they make sense? Are they useful in this domain?
- When accepted, disambiguate the context entity and the sentiment word included in the collocation using Wordnet 3.0[10]. The linguists read the gloss and synonyms included in the corresponding synset and we chose the most agreed upon appropriate meaning (synset ID).
- Scoring the collocation from a sentiment perspective, in a [0..1] range [10]. Figure 2 shows a snapshot of the web application where linguists could provide their inputs for the sentiment scores.

A trade-off decision, between domain coverage and effort, was taken to include as many important domain entities as possible.
At the end of this process the resulted sentiment dictionaries are provided as CSV files, one file per language and domain, with the following fields:
$entity$, $entityWNid$, $entityPOS$, $sentiment$, $sentiWNid$, $sentiPOS$, $score$ where $entity$ is the context entity; $entityWNid$ is the wordnet synset id agreed for the $entity$; $entityPOS$ is the part-of-speach of the context entity; $sentiment$ is the sentiment word that occurs in the context of the $entity$; $sentiWNid$ is the Sentiwordnet id agreed for the sentiment word; $sentiPOS$ is the sentiment word's part-of speach adn finally $score$ is teh polarity score assigned to the sentimetn words bty the linguists.
As an example consider the following result from the 'Hotel' domain in English:
$04105893, n, room, 01676517, a, fantastic, 0.75$ . Here we see that the sentiment word fantastic is an adjective with the sysnet id 01676517 and has a polarity score of 0.75 in the context of the entity room which is a noun with the synset id 04105893.

[10]http://wordnet.princeton.edu/

Paradigma provided also sentiment dictionaries in other 5 languages for the two domains. The non-english dictionaries were built using MultiWordnet [11] translation based on the Wordnet synset ids from the English dictinoaries.

## 4. Lexicon

Both the results of our pipeline (up to the Lemon/Marl Generator component) and the sentiment dictionaries from Paradigma were converted to RDF using the RDF extension fo the GoogleRefine[12] tool to create the RDF lexicons. We used the folowing namespaces listed in Listing 5 : $lemon$ - the core lemon lexicon model, $marl$ - rdf properties for sentiment, $w$ - WordNet 3.0 synsets, $lexinfo$ - for part-of-speech properties, $ed$ - domain categories, $el$ - lexicon prefix, $ele$ - lexical entries prefix.
The URIs for the lexical entries are built from the $lee$ namespace and the name of the lexicalentry. For each lexical entry we add their written form and their language within a $lemon : CananicalForm$ object and their part-of-speach information using a $lexinfo$ object. For each different synset id of the same context entity we build a $lemon : sense$ For each sense we add the connections to other datasets using the lemon:reference property to refer to the Dbpedia and WordNet links. The sentiment words are represented similarly: for each sentiment word we create a lexical entry and for each of its distinct polarity values and synset pairs we create a different sense of the lexical entry. Diferently from the lexical entries generated for entity classes and named entities, the senses of the sentiment word lexical entries contain also the sentiment polarity values and polarity using Marl sentiment properties $marl : polarityValue$ and $marl : hasPolarity$ respectively.
Figure 3 shows an example of a generated lexicon for the domain 'hotel' in English. It shows 3 *lemon:LexicalEntries*: 'room ' (entity class), 'Paris' (named entity) and 'small' (sentiment word) which in the context of the lexical entry 'room' has negative polarity. Each of them consists of senses, which are linked to DBpedia and/or Wordnet concepts.

[11]http://multiwordnet.fbk.eu/english/home.php
[12]http://refine.deri.ie/

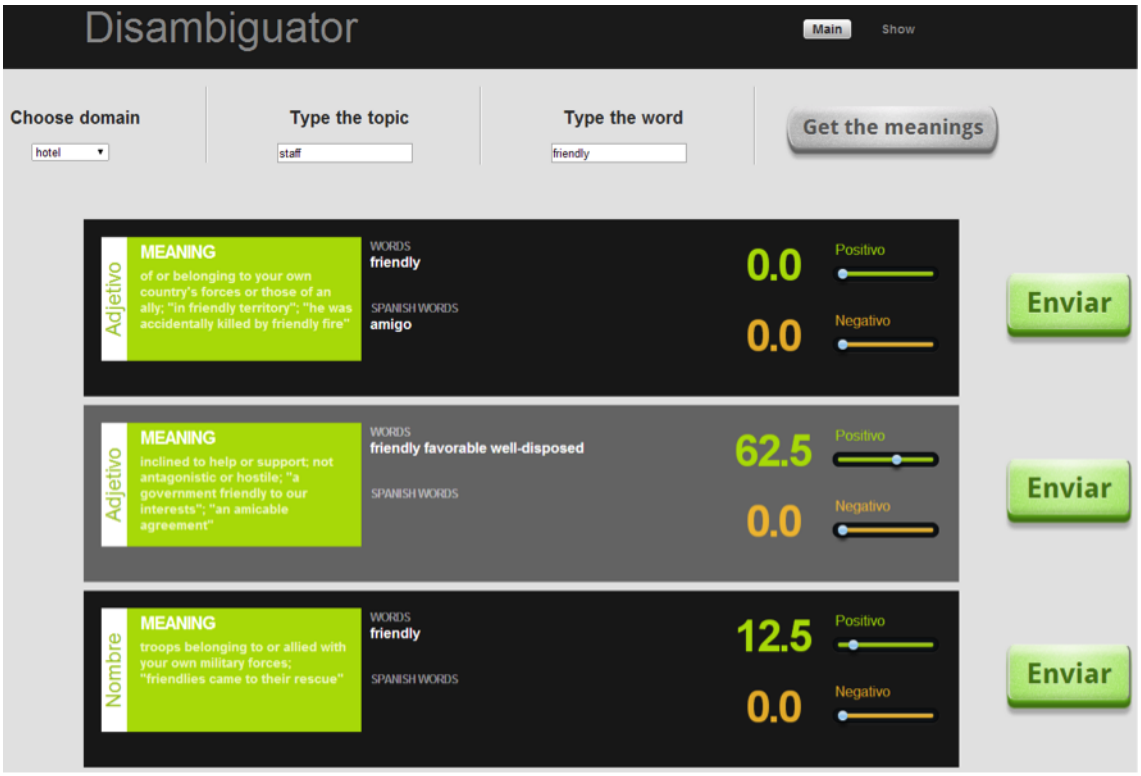


Figure 2: Snapshot of the Web applicatoin that allows linguists to specify the sentiment scores.

Listing 5: Namespaces used in the RDF lexicons..

```
lemon: http://www.monnet-project.eu/lemon
marl:  http://purl.org/marl/ns
wn:    http://semanticweb.cs.vu.nl/europeana/lod/purl/vocabularies/princeton/wn30
lexinfo: http://www.lexinfo.net/ontology/2.0/lexinfo
ed:    http://www.eurosentiment/domains
le:      http://wwww.eurosentiment.com/lexicon/<language>/
lee:   http://www.eurosentiment.com/lexicalentry/<language>/
```

We use named graphs to group the data from each lexicon. The URIs that we use for the named graphs are the lexicon URIs and they are built after the following pattern: $http://www.eurosentiment.com/ < domain > / < language > /lexicon/paradigma$ for the lexicons obtained fmor the sentiment dictionaries from Paradigma and $http://www.eurosentiment.com/ < domain > / < language > /lexicon/ < ta$ or $amz >$ for the lexicons obtained from the TripAdvisor adn Amazon corpora.

## 5. Availability

The domain-specific lexicons are loaded in a Virtuoso[13] SPARQL endoint which can be accessed from here: $http://140.203.155.231:8890/sparql$. We also installed the linked data frontend pubby[14] on top of this SPARQL endpoint to allow for easier browsing of the provided lexicons. For example one can start at the following link $http://140.203.155.231:8080/eurosentiment/$ to see the available lexicons. Then he/she can click on the uri of any of the lexicons to explore its lexical entries.

[13] http://virtuoso.openlinksw.com/

[14] http://wifo5-03.informatik.uni-mannheim.de/pubby/

## 6. Acknowledgements

This work has been funded by the European project EUROSENTIMENT under grant no. 296277.

## 7. References

Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and V. S. Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media*, ICWSM'07.

Georgeta Bordea, Paul Buitelaar, and Tamara Polajnar. 2013. Domain-independent term extraction through domain modelling. In *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence*, TIA'13, Paris, France.

Georgeta Bordea. 2013. *Domain Adaptive Extraction of Topical Hierarchies for Expertise Mining*. Ph.D. thesis, National University of Ireland, Galway.

Evgeniy Gabrilovich and Shaul Markovitch. 2006. Overcoming the brittleness bottleneck using wikipedia: En-

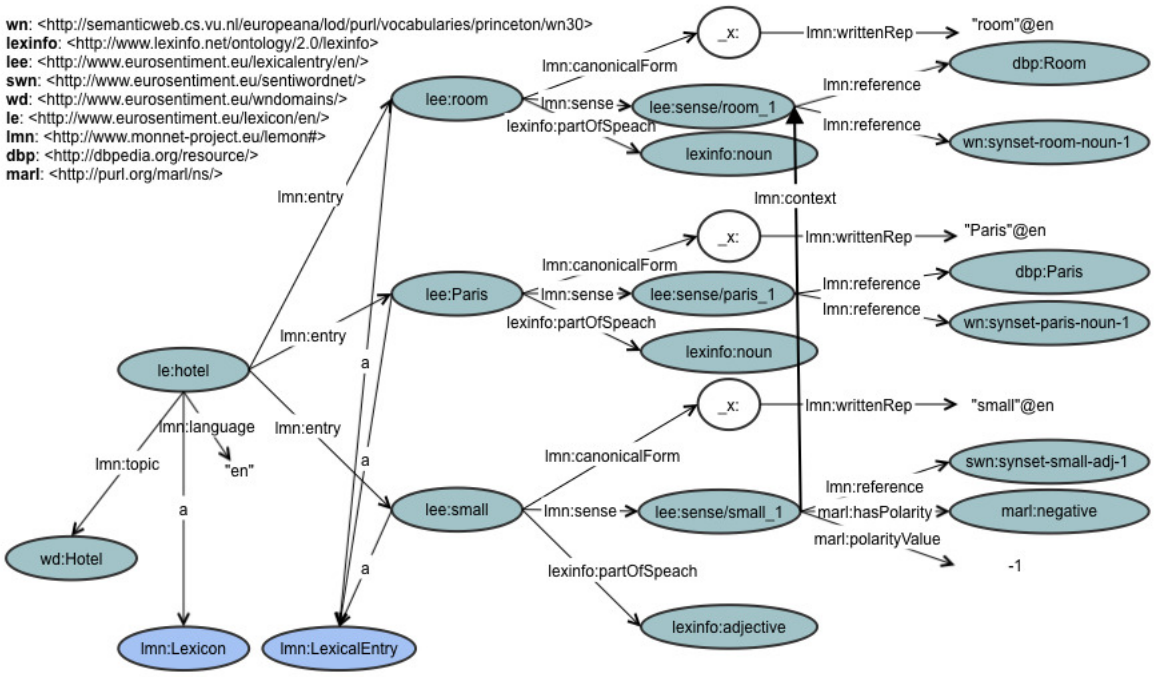


Figure 3: Example lexicon for the domain 'hotel' in English.

hancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence*, AAAI'06. AAAI Press.

Kavita Ganesan and ChengXiang Zhai. 2011. Opinion-based entity ranking. *Information Retrieval*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, New York, NY, USA. ACM.

John McCrae, Guadalupe Aguado de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asuncin Gmez-Prez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. 2012. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*.

Bianca Pereira, Nitish Aggarwal, and Paul Buitelaar. 2013. Aela: An adaptive entity linking approach. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, WWW'13, Republic and Canton of Geneva, Switzerland.

Ariel Raviv and Shaul Markovitch. 2012. Concept-based approach to word-sense disambiguation. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*.

V.S. Subrahmanian and Diego Reforgiato. 2008. Ava: Adjective-verb-adverb combinations for sentiment analysis. *Intelligent Systems*.

Peter D. Turney and Michael L. Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning*.

Adam Westerski, Carlos A. Iglesias, and Fernando Tapia. 2011. Linked Opinions: Describing Sentiments on the Structured Web of Data. In *Proceedings of the 4th International Workshop Social Data on the Web*.