Generating Linked-Data based Domain-Specific Sentiment Lexicons from Legacy Language and Semantic Resources

Gabriela Vulcu, Paul Buitelaar, Sapna Negi, Bianca Pereira, Mihael Arcan, Barry Coughlan, J. Fernando Sanchez, Carlos A. Iglesias

Insight, Centre for Data Analytics, Galway, Ireland

gabriela.vulcu@insight-center.org, paul.buitelaar@insight-center.org, sapna.negi@insight-center.org, bianca.pereira@insight-center.org, mihael.arcan@insight-center.org, b.coughlan2@gmail.com,

Universidad Politecnica de Madrid, Spain

jfernando@gsi.dit.upm.es, cif@dit.upm.es

Abstract

We present a methodology for legacy language resource adaptation that generates domain-specific sentiment lexicons organized around domain entities described with lexical information and sentiment words described in the context of these entities. We explain the steps of the methodology and we give a working example of our initial results. The resulting lexicons are modelled as Linked Data resources by use of established formats for Linguistic Linked Data (lemon, NIF) and for linked sentiment expressions (Marl), thereby contributing and linking to existing Language Resources in the Linguistic Linked Open Data cloud.

Keywords: domain specific lexicon, entity extraction and linking, sentiment analysis

1. Introduction

In recent years, there has been a high increase in the use of commercial websites, social networks and blogs which permitted users to create a lot of content that can be reused for the sentiment analysis task. However the development of systems for sentiment analysis which exploit these valuable resources is hampered by difficulties to access the necessary language resources for several reasons: (i) language resource owners fear for losing competitiveness; (ii) lack of agreed language resource schemas for sentiment analysis and not normalised magnitudes for measuring sentiment strength; (iii) high costs for adapting existing language resources for sentiment analysis; (iv) reduced visibility, accessibility and interoperability of the language resources with other language or semantic resources like the Linguistic Linked Open Data cloud (i.e. LLOD). In this paper we are focusing on the second and the forth challenges by describing a methodology for the conversion, enhancement and integration of a wide range of legacy language and semantic resources into a common format based on the lemon[1](McCrae et al., 2012) and Marl [2] (Westerski et al., 2011) Linked Data formats.

1.1. Legacy Language Resources

We identified several categories of legacy language resources with respect to our methodology: domain-specific English review corpora, non-English review corpora, sentiment annotated dictionaries and Wordnets. The existing legacy language resources (gathered in the EUROSENTIMENT project [3]) are available in many formats and they contain several types of annotations that are relevant for the sentiment analysis task. The language resources formats range from plain text with or without custom made annotations, HTML, XML, EXCEL, TSV, CSV to RDF/XML. The language resources annotations are all or a subset of: *domain* - the broad context of the review corpus (i.e. 'hotel' is the domain for the TripAdvisor corpus); *language* - the language of the language resource; *context entities* - relevant entities in the corpus; *lemma* - lemma annotations of the relevant entities; *POS* - part-of-speach annotations of the relevant entities; *WordNet synset* - annotations with existing synsets from Wordnet of the relevant entities; *sentiment* - positive or negative sentiment annotation both at sentence level and or at entity level; *emotion* - more fine grained polarity values both expressed as numbers or as concepts from well defined ontologies; *inflections* - morphosyntactic annotations of the relevant entities.

1.2. Methodology for LR Adaptation and Sentiment Lexicon Generation

Our method 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 language resources to a common schema based on Marl and NIF[4]; (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; (iii) the Sentiment Analysis step extracts contextual sentiments and identifies SentiWordNet synsets corresponding to these contextual sentiment words; (iv) the Lexicon Generator step uses the results of the previous steps, enhances them with multilingual and morphosyntactic 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

[1] http://lemon-model.net/lexica/pwn/

[2] http://www.gi2mo.org/marl/0.1/ns.html

[3] http://eurosentiment.eu/

[4] http://persistence.uni-leipzig.org/nlp2rdf/

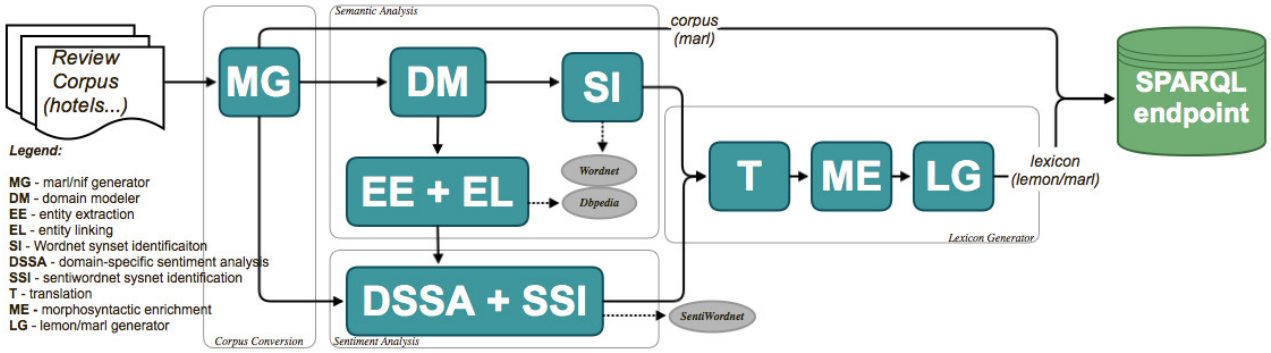

Figure 1: Methodology for Legacy Language Resources Adaptation for Sentiment Analysis.

processed using the pipeline described in Figure 1 while the sentiment annotated dictionaries are converted to the lemon/Marl format using the Lexicon Generator step. We detail these steps in the subsequent sections.

2. Corpus conversion

Due to the formats heterogeneity of the legacy language resources we need a common model that captures all the existing annotations in a structural way. The Corpus Conversion step adapts corpus resources to a common schema. We defined a schema based on the NIF and Marl formats that structures the annotations from the corpora reviews. For example each review in the corpus is an entry that can have overall sentiment annotations or annotations at the substring level. The Corpus Generator has been designed to be extensible and to separate the technical aspects from the content and formats being translated.

3. Semantic analysis

The Semantic Analysis step consists of: Domain Modeller (DM), Entity Extraction (EE), Entity Linking (EL) and Synset Identification (SI) components. The DM extracts a set of entity class using a 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). The EE and EL components are based on AELA (Pereira et al., 2013) framework for Entity Linking that uses a Linked Data dataset as reference for entity mentioning identification, extraction and disambiguation. By default, DBPedia and DBPedia Lexicalization (Mendes et al., 2011) are used as reference sources but domain-specific datasets could be used as well. The SI identifies and disambiguates WordNet synsets that match with the extracted entity classes. It extends each candidate synset with their direct hyponym and hypernym synsets. Then we compute the occurrence of a given entity class in each of these bag of words. We choose the synset with the highest occurrence score for an entity class.

4. Sentiment analysis

The Sentiment Analysis step consists of: Domain-Specific Sentiment Polarity Analysis (DSSA) and Sentiment Synset Identification (SSI) components. The DSSA component identifies a set of sentiment words and their polarities in the context of the entities identified in the Semantic Analysis step. The clause in which a entity mention occurs is considered the span for a sentiment word/phrase in the context of that entity. The DSSA is based 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 example, 'ridiculous bargain' represents a positive sentiment while 'ridiculous' could represent a negative sentiment. Sentiment words/phrases in individual reviews are assigned polarity scores based on the available user ratings. In case of language resources with no ratings we use a bootstrapping process based on Sentiwordnet that will rate the domain aspects in the review. We select the most frequent scores as the final sentiment score for a sentiment word/phrase candidate based on its occurrences in all the reviews. The SSI component identifies SentiWordNet synsets for the extracted contextual sentiment words. The sentiment phrases however, are not assigned any synset. Linking the sentiment words with those of SentiWordNet further enhances their semantic information. 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. The synset with the highest similarity score above a threshold is considered.

5. Lexicon generator

The Lexicon Generator step consists of: MorphoSyntactic Enrichment (ME), Machine Translation(T) and lemon/Marl Generator(LG) components. As WordNet does not provide

Sentiment	PolarityValue	Context
”good”@en	”1.0”	”alarm”@en
”damaged”@en	”-2.0”	”apple”@en
”amazed”@en	”2.0”	”flash”@en
”expensive”@en	”-1.0”	”flash”@en
”annoying”@en	”-1.5”	”player”@en

Table 1: Sentiment words the 'electronics' domain.

Sentiment	PolarityValue	Context
”warm”@en	”2.0”	”pastries”@en
”warm”@en	”2.0”	”comfort”@en
”warm”@en	”1.80”	”restaurant”@en
”warm”@en	”1.73”	”service”@en
”warm”@en	”0.98”	”hotel”@en

Table 2: Sentiment word 'warm' in the 'hotel' domain.

any morphosyntactic information (besides part of speech), such as inflection and morphological or syntactic decomposition, the ME provides a further process for the conversion and integration of lexical information for selected synsets from other legacy language resources like CELEX [5]. Next, the T component translates extracted entity classes and sentiment words in other languages using a domain-adaptive machine translation approach (Arcan et al., 2013). This way we can build sentiment lexicons in other languages. It uses the SMT toolkit Moses (Koehn et al., 2007). Word alignments are built with the GIZA++ toolkit (Och and Ney, 2003), where a 5-gram language model was built by SRILM with Kneser-Ney smoothing (Stolcke, 2002). We use two different parallel resources: the JRC-Acquis (Steinberger et al., 2006) available in almost every EU official language (except Irish) and the OpenSubtitles2013 (Tiedemann, 2012) which contains fan-subtitled text for the most popular language pairs. The LG component converts the results of the previous components (named entities and entity classes linked to LOD and sentiment words with polarity values) to a domain-specific sentiment lexicon represented as RDF in the lemon/Marl format. The lemon model was developed in the Monnet project to be a standard for sharing lexical information on the semantic web. The model draws heavily from earlier work, in particular from LexInfo (Cimiano et al., 2011), LIR (Montiel-Ponsoda et al., 2008) and LMF (Francopoulo et al., 2006). The Marl model is a standardised data schema designed to annotate and describe subjective opinions.

6. Working Example

Figure 2 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.

We applied our methodology on an annotated corpus of 10.000 reviews for the hotel domain and an annotated corpus of 600 reviews for the electronics domain. Table 1 shows an example of sentiment words from the 'electronics' domain, while Table 2 shows an example of different contexts of the sentiment word 'warm' with their corresponding polarities in the 'hotel' domain.

7. Future Work

We are currently working on evaluating the Semantic Analysis and Sentiment Analysis components by participating in the SemEval challenge [6] on aspect-based sentiment analysis. We also plan to investigate ways of linking the extracted named entities with other Linked Data datasets like Yago or Freebase. A next step for the use of our results is to aggregate sentiment lexicons obtained from Language Resources on the same domain.

8. Conclusions

In this paper we presented a methodology for creating domain-specific sentiment lexicons from legacy Language Resources, described the components of our methodology and provided example results.

9. Acknowledgements

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

10. References

Arcan, M., Thomas, S. M., Brandt, D. D., and Buitelaar, P. (2013). Translating the FINREP taxonomy using a domain-specific corpus. Poster presented at the Machine Translation Summit XIV, Nice, France.

Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D., and Subrahmanian, V. S. (2007). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media*, ICWSM'07.

Bordea, G., Buitelaar, P., and Polajnar, T. (2013). Domain-independent term extraction through domain modelling. In *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence*, TIA'13, Paris, France.

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

Cimiano, P., Buitelaar, P., McCrae, J., and Sintek, M. (2011). Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web*.

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., and Soria, C. (2006). Lexical markup framework (LMF) for NLP multilingual resources. In *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, Sydney, Australia. ACL.

Gabrilovich, E. and Markovitch, S. (2006). Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the 21st National Conference on Artificial Intelligence*, AAAI'06. AAAI Press.

[5] http://celex.mpi.nl/

[6] http://alt.qcri.org/semeval2014/

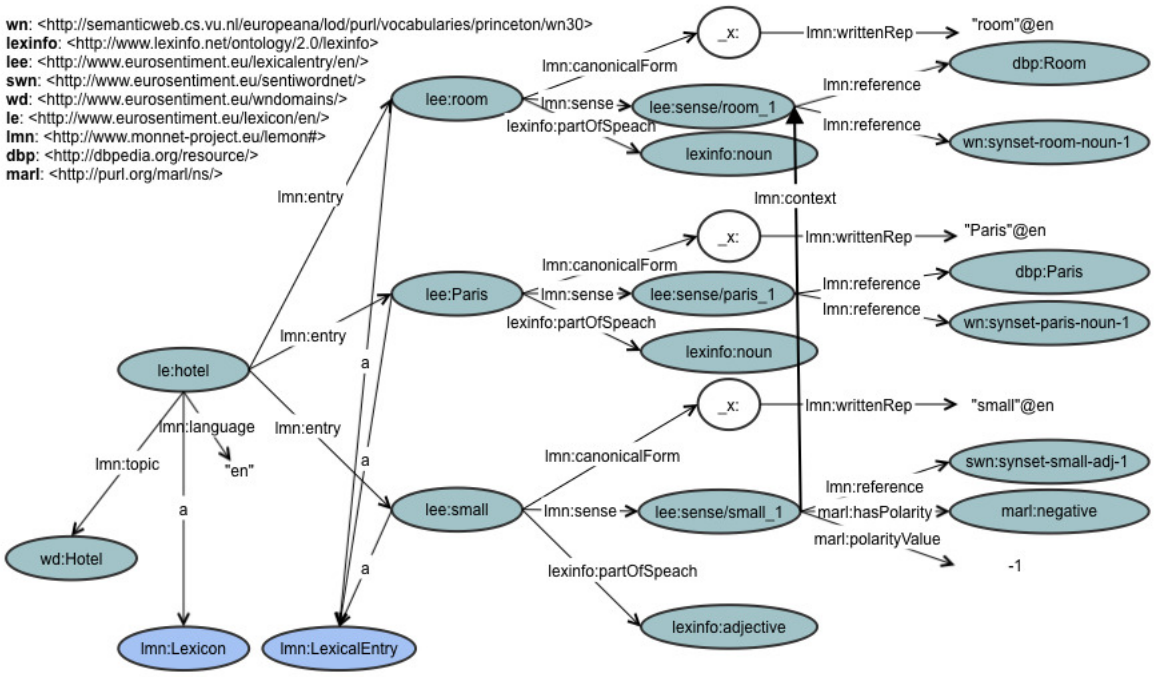

Figure 2: Example lexicon for the domain 'hotel' in English.

Hu, M. and Liu, B. (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.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, Stroudsburg, PA, USA. ACL.

McCrae, J., de Cea, G. A., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, New York, NY, USA. ACM.

Montiel-Ponsoda, E., de Cea, G. A., Gómez-Pérez, A., and Peters, W. (2008). Modelling multilinguality in ontologies. In *Poster at COLING'10*.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, March.

Pereira, B., Aggarwal, N., and Buitelaar, P. (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.

Raviv, A. and Markovitch, S. (2012). Concept-based approach to word-sense disambiguation. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., and Tufis, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC'06.

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing (ICSLP 2002*.

Subrahmanian, V. and Reforgiato, D. (2008). Ava: Adjective-verb-adverb combinations for sentiment analysis. *Intelligent Systems*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, LREC'12, Istanbul, Turkey. ELRA.

Turney, P. D. and Littman, M. L. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning*.

Westerski, A., Iglesias, C. A., and Tapia, F. (2011). Linked Opinions: Describing Sentiments on the Structured Web of Data. In *Proceedings of the 4th International Workshop Social Data on the Web*.