Linguistically motivated Language Resources for Sentiment Analysis

Voula Giouli **Aggeliki Fotopoulou**
Institute for Language and Speech Processing, Athena RIC
{voula;afotop}@ilsp.athena-innovation.gr

Abstract

Computational approaches to sentiment analysis focus on the identification, extraction, summarization and visualization of emotion and opinion expressed in texts. These tasks require large-scale language resources (LRs) developed either manually or semi-automatically. Building them from scratch, however, is a laborious and costly task, and re-using and repurposing already existing ones is a solution to this bottleneck. We hereby present work aimed at the extension and enrichment of existing general-purpose LRs, namely a set of computational lexica, and their integration in a new emotion lexicon that would be applicable for a number of Natural Language Processing applications beyond mere syntactic parsing.

1 Introduction

The abundance of user-generated content over the web has brought about the shift of interest to the opinion and emotion expressed by people or groups of people with respect to a specific target entity, product, subject matter, etc. The task of sentiment analysis involves determining the so-called *private states* (beliefs, feelings, and speculations) expressed in a particular text or text segment as opposed to factual information. More precisely, it is focused on the following: (a) *identification of sentiment expressions* in textual data and their *classification* as appropriate, and (b) *recognition* of *participants* in the private state, as for example, the entities identified as the *Source* and *Target* of the emotion. More recently, aspect-based sentiment analysis has also been in the focus of research (Wilson, 2008).

Traditionally, classification of sentiment expressions is usually attempted in terms of the general notion of *polarity* defined as *positive, negative and neutral*. Traditional approaches to text classification based on stochastic methods are quite effective when applied for sentiment analysis yielding quite satisfactory results. However, certain applications require for more fine-grained classifications of sentiment i.e. the identification of emotional states such as *anger, sadness, surprise, satisfaction*, etc. in place of mere recognition of the polarity. Such applications might be the identification of certain emotions expressed by customers (i.e., satisfaction, or dissatisfaction) with respect to some product or service, or the analysis of emotions and feelings described by users in blogs, wikis, fora and social media (Klenner at al., 2009). In this respect, stochastic approaches fail to recognize multiple or even conflicting emotions expressed in a document or text segment. In these cases, linguistic (syntactic and semantic knowledge) is necessary in order to assess the overall polarity of a clause and or the feeling expressed in it.

The paper is organised as follows: In section 2 we present the aims and scope of the specific work; section 3 gives an overview of related work on affective LRs, whereas section 4 gives an account of the LRs developed within the framework of Lexicon – Grammar. Our efforts towards enriching the existing resources with semantic information and re-purposing them are presented in sections 5 and 6 respectively, while section 7 outlines our conclusions and prospects for future research.

2 Aims and scope

We present work aimed at extending, enriching and re-purposing existing LRs, the ultimate goal being

Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing, pages 39–45,
Coling 2014, Dublin, Ireland, August 24 2014.

their integration in a tool for sentiment analysis. In specific, a suite of computational lexica developed within the framework of Lexicon – Grammar (LG) and treating *verbal and nominal predicates* denoting emotion were used. These resources were initially constructed manually as a means to describe general language, and they bear rich linguistic information that would be otherwise difficult to encode in an automatic way, namely (a) subcategorisation information, (b) semantic and distributional properties, and (c) syntactic transformations of the predicates. Within the current work, semantic information that is meaningful for sentiment analysis was also added to lexicon entries. The final resource was then used to bootstrap a *grammar of emotions*. This grammar is a rule-based approach to sentiment analysis aimed at capturing and modeling linguistic knowledge that is necessary for the task at hand.

The work presented here was based on a previous study (Giouli et al., 2013), making further extensive use of the Hellenic National Corpus (HNC), a large reference corpus for the Greek language (Hatzigeorgiou et al, 2000). Additionally, a suite of specialized corpora that were developed to guide sentiment studies in multimodal (Mouka et al., 2012) and in textual (Giouli and Fotopoulou, 2013) data was used. Thus, the resulting *Greek Sentiment Corpus*, that amounts to c. ~250K tokens, comprises audiovisual material (movies dialogues), and texts selected manually from various sources over the web. More particularly, the online edition of two newspapers along with a news portal were searched on a daily basis for the identification and selection of commentaries dealing with a set of predefined topics; Greek blogs and fora were also used as sources for text collection. The aforementioned corpus was annotated at the sentence and phrase level for opinion and emotion, and was subsequently used to populate the sentiment lexicon under construction. Moreover, initial steps were made towards creating a rule-based system for the identification of sentiment expressions in texts and computing the overall phrase polarity in context on the basis of corpus evidence.

3 Related work

A number of large-scale lexica appropriate for sentiment analysis have been developed either manually or semi-automatically. These range from mere word lists to more elaborate resources. General Inquirer (Stone et al. 1966), the Subjectivity lexicon integrated in OpinionFinder (Wiebe et al., 2005), and SentiWordNet (Esuli and Sebastiani 2006) are examples of such affective lexica. On the other hand, WordNet-Affect (Strapparava and Valitutti 2004), an extension of WordNet Domains, is linguistically oriented as it comprises a subset of *synsets* that are suitable to represent affective concepts in correlation with *affective words*. A set of A-labels is used to mark concepts representing emotions or emotional states, moods, eliciting emotions situations, and emotional responses. Finally, EmotiNet (Balahur et al, 2011) is a knowledge base (KB) for representing and storing affective reaction to real-life contexts and action chains described in text.

From a purely linguistic perspective – yet with a view to Natural Language Processing - substantial work has been devoted to the semantic classification of verbal predicates denoting emotion in (Mathieu, 1999). In this work, verbs denoting emotional states and evaluative stances should also be classified according to the so-called *semantic field'.* Verbs were, thus, categorized into homogenous semantic classes which share common syntactic properties; this classification is claimed to facilitate semantic interpretation.

Statistical approaches to sentiment analysis feature a "bag-of-word" representation (Hu and Liu, 2004). Rule-based systems, on the other hand, exploit linguistic knowledge in the form of syntactic/lexical patterns for computing polarity in context. In most cases, negative particles and modality are reported as the most obvious shifters that affect sentiment polarity (Polanyi and Zaenen 2006, Jia *et al*. 2009, Wiegand *et al*. 2010, Benamara et al., 2012). Finally, compositionality features have been explored for the computation of multiple or conflicted sentiments on the basis of deep linguistic analysis (Moilanen and Pulman, 2007), (Neviarouskaya et al., 2009), (Klenner et al., 2009).

4 Lexicon – Grammar tables

4.1 Lexicon – Grammar framework

The Lexical Resources hereby exploited were initially constructed in accordance with the Lexicon-Grammar (LG) methodological framework (Gross 1975), (Gross 1981). Being a model of syntax limited to the *elementary sentences* of the form *Subject – Verb – Object*, the theory argues that the unit of

meaning is located at the sentence rather than the word level. To this end, linguistic analysis consists in converting each elementary sentence to its predicate-argument structure. Additionally, main complements (subject, object) are separated from other complements (adjuncts) on the basis of formal criteria; adverbial complements (i.e., prepositional phrases) are considered as crucial arguments only in the case that they characterize certain verb frames:

(1) John removed the cups *from the table*.

To cater for a more fine-grained classification, and the creation of homogenous word classes, this formal syntactic definition is further coupled with *distributional properties associated with words*, i.e., types of prepositions, features attached to nouns in subject and complement positions, etc. A set of transformation rules, construed as equivalence relations between sentences, further generate equivalent structures. It becomes evident, therefore, that the resulting resources are rich in linguistic information (syntactic structure, distributional properties and permitted transformational rules), which is encoded formally in the so-called LG tables.

4.2 The Lexicon – Grammar of verb and noun predicates denoting emotion

Within the LG framework, 130 noun predicates denoting emotions (*Nsent*) in Modern Greek were selected and classified into 3 classes, according to their syntactic and distributional properties (Fotopoulou & al., 2008). The 1st class comprises nouns of interpersonal relations with an obligatory prepositional complement and a conversed construction, as for example *θαυμασμός (= admiration)*. The 2nd class are indicative of an external cause including a non obligatory prepositional complement, as for example *φόβος (= fear)*. The 3rd class without complements have a static character, as for example *ευτυχία (= happiness)*. Identification of the specific light verbs (or support verbs, *Vsup*) they select for was also performed. Furthermore, their distributional properties and their co-occurrence with specific verbs expressing diverse modalities (aspect, intensity, control, manifestation or verbal expression) have also been encoded in a formal way. These properties reveal the restrictions nouns impose on the lexical choice of verbs.

Furthermore, 339 Greek verbal predicates denoting emotion (*Vsent*) have been selected from various sources (i.e. existing reference lexicographic works and corpora) and were subsequently classified in five LG tables. Classification was performed on the basis of the following axes: (i) syntactic information (i.e, subcategorisation information); (ii) selectional restrictions (+Hum/ -Hum) imposed over their Subject and Object complements; and (iii) transformation rules. More precisely, as far as syntactic structure is concerned, the predicates under consideration were identified to appear in both transitive and intransitive constructions being represented as *N0 V N1* and *N0 V* respectively. Certain verbs also allow for a prepositional phrase complement represented as *N0 V Prep N1*[1] configurations. A close inspection over the data revealed the relationship between the N0 or N1 complements that denote the *Experiencer* of the emotion (i.e., the entity feeling the emotion). In two of the resulting classes the *Experiencer* is projected as the structural Subject of the verb, whereas the *Theme* or *Stimulus* is projected as their structural object. Similarly, the remaining 3 classes realize the *Theme/Stimulus* as the subject and the *Experiencer* as their object, their distinguishing property being their participation in unaccusative and middle constructions, the latter being linked to the implicit presence of an Agent (middle) and the absence of an Agent (unaccusative). These properties have been checked for the whole range of lexical data based on both linguistic introspection and corpus evidence.

A number of Harrisian constructions and transformations (Harris, 1951; 1964; 1968) have been extensively utilized within the LG formalism to define syntactically related and semantically equivalent structures. Apart from passivisation and middle alternation constructions - also relevant to emotion predicates - the restructuring transformation has been accounted for (Guillet and Leclère, 1981):

(2) Ο Γιάννης θαυμάζει *τη Μαρία για το θάρρος της*.
The John admires the Maria for the courage-her.
John admires Maria for her courage.

[1] Adopting the LG notation, N0 denotes a Noun in *Subject* position of a given verb V, whereas, N1 denotes its *Object*.

(3) Ο Γιάννης θαυμάζει <u>*το θάρρος της Μαρίας*</u>.
The John admires the courage the Maria-of
John admires Maria's courage.

Moreover, each verbal predicate was also coupled with morphologically-related *adjectives* and *nouns*, and the alignment of semantically equivalent nominal, verbal and adjectival structures was performed thereof. A number of semantically equivalent paraphrases of the verbs with the morphologically related nouns and adjectives were also encoded in the tables.

Finally, following the same methodology, a set of 2,500 verbal multi-word expressions denoting emotions were identified from corpora and classified in 13 categories according to their syntactic structure. The final resource comprises a total of ~3000 entries, organized in 21 LG tables with lemmas inter-connected via the tables relative to verbs.

5 Semantic classification of emotion predicates

Semantic classification of the verbal predicates has also been performed on the basis of their underlying semantics. In this way, the syntactic and distributional properties encoded in the LG tables have been coupled with semantic information that defines an affective taxonomy. These properties were added as columns in the tables that describe the verb predicates. Our goal was to group together predicates that are synonyms or near synonyms and to create an affective taxonomy hierarchical organized. To this end, certain abstractions and generalizations were performed where necessary for defining classes of emotion types.

Initially, 59 classes of emotion-related-senses were identified. At the next stage, a number of iterations followed aimed at grouping together senses that are semantically related. This procedure resulted in the identification of a set of senses that may be used as taxonomy of emotions. Following practices adopted in similar endeavours (i.e., Mathieu, 1999), each class was further assigned a tag that uniquely identifies the respective class. The following classes (19 classes) were identified: *anger, fear, sadness, disgust, surprise, anticipation, acceptance, joy, love, hate, disappointment, indifference, shame, envy, jealousy, relaxedness, respect, resentment,* and *remorse*.

Next, each entry was further specified as regards the specific relation that holds between the entry and the emotion type it belongs to. A set of properties were then defined for which each entry was then examined, namely: *FeelEmotion, EmotionManifestation, Behaviour,* and *EntailsEmotion*.

At a more abstract level, entries were further assigned a value for the semantic property *polarity*. Following previous works (Mathieu and Fellbaum, 2010**)**, the encoding caters for the *apriori polarity* of the emotion denoted which subsumes one of the following values: (a) *positive, i.e.* predicates which express a pleasant feeling; (b) *negative,* i.e., predicates which express an unpleasant feeling; (c) *neutral,* and (d) *ambiguous,* i.e., predicates expressing a feeling the polarity of which is *context-dependent* (e.g., surprise).

Moreover, to better account for the semantic distinction between near synonyms that occur within a class such as *φοβάμαι (= I am scared), πανικοβάλλομαι (=panic), etc.*, entries are further coupled with the feature *intensity* with possible values: *low, medium, high, uncertain*. Intensity was attributed to the lexical items on the basis of linguistic introspection and the definitions of lexical entries.

6 Transforming Lexicon-Grammar tables to a grammar of emotions

Being initially developed to serve as a means of linguistic description, this framework has, never-theless, been proved to be applicable for the construction of robust computational lexica. And although it has been claimed (Mathieu, 2008) that the information is not *directly* exploitable for NLP applications due to the fact that certain pieces of information are not formally encoded or are *implicit*, a number of works (Hathout and Namer 1998, Danlos and Sagot 2009) have successfully managed to reformat LG tables in efficient large-scale NLP lexica.

To this end, we have tried to exploit information available in the tables and make the mappings that are necessary for the task of sentiment recognition. On the one hand, subcategorisation information with respect to selectional restrictions imposed over the Subject and Object of the verbal predicates was exploited. Once a verbal predicate has been identified, the constituent either in Subject or Object

position that is also assigned a (+Hum) property corresponds unambiguously to the *Experiencer* of the emotion depending on the class it belongs to (i.e., SubjectExperiencer or Object Experiencer). Similarly, the NP in *Object* position of verbs that pertain to the 2nd class *αγαπώ (=love)* corresponds to the *Target* of the emotion. All other constituents correspond to the *Trigger* or *Cause.*

On these grounds, initial steps towards building a rule-based component that identifies emotion verbal and nominal predicates in texts along with the participating entities, namely the *Experiencer* and *Target* of the emotion expressed have been performed. To this end, a library of *local grammars* (Constant, 2003) for emotion predicates has been constructed modeling structures in the annotated corpus. Local grammars (also referred to in the literature as *graphs*) are algebraic grammars formulated as combinations of sequences of grammatical symbols in the form of regular expressions that describe natural language. In this sense, they are a powerful tool to represent the majority of linguistic phenomena in an intuitive manner. Moreover, they are compiled into finite state transducers that transform input text by inserting or removing special markers. Rules are sequentially applied to the text using longest match. We made use of the UNITEX platform (Paumier, 2013) for creating the graphs and then compiling them into finite state transducers. UNITEX consists of three modules, namely, corpus handling, lexicon development and grammar development that are integrated into a single intuitive graphical user interface. Based on the Lexicon-Grammar tables developed for the verbal predicates (c.f. section 2 above), we initially created five parameterized graphs manually; these graphs depict the syntactic and semantic properties of the predicates. At the next stage, a set of graphs was constructed automatically using UNITEX, each one representing the syntactic and semantic properties of a given predicate.

It should be noted, however, that LG tables provide descriptions at an abstract level. To remedy this shortcoming, a number of graphs and sub-graphs describing a wide range of syntactic phenomena (noun phrase, coordination, modifiers, negation, and valency shifters) were constructed manually. The set of graphs comprises a grammar applied to the text as a cascade for the identification of the *emotive* predicate, being either verbal or nominal, its *polarity* and the participants of the emotion event that can be identified from the underlying structure – namely the *Experiencer* and the *Theme* and the *Cause*.

7 Conclusions and future work

We have described work aimed at enriching, re-purposing and re-using already available LRs for a new task, namely identification of emotion expressions in texts. The existing lexica carry rich linguistic information which has been mapped onto categories that are meaningful for the task. Our efforts have been oriented towards developing a rule-based system that efficiently will eventually recognise emotion expressions in texts and the participants in the emotion event.

Future work has been planned already, consisting of the exploitation of other properties that are encoded in the LG tables, as for example the restructuring property as a facet of the aspect-based sentiment analysis and the conversion of the enriched LG tables to a standardised lexical format. Finally, the validation of the final resource is due against the manually annotated corpus.

Acknowledgments

The research within the project *LangTERRA: Enhancing the Research Potential of ILSP/"Athena" R.C. in Language Technology in the European Research ERA* leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013).

References

Alexandra Balahur and Jesús M. Hermida and Andrés Montoyo and Rafael Muñoz. 2011. EmotiNet: A Knowledge Base for Emotion Detection in Text Built on the Appraisal Theories. In R. Muñoz et al. (Eds.): *Natural Language Processing and Information Systems, Lecture Notes in Computer Science*, Volume 6716, Springer-Verlag Berlin Heidelberg 2011, pp 27-39.

Farah Benamara, Baptiste Chardon, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2012. How do Negation and Modality Impact on Opinions? In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics, ExProM '12*, Jeju, Republic of Korea, 2012, pp 10–18.

Matthieu Constant. 2003. *Grammaires locales pour l'analyse automatique de textes : méthodes de construction et outils de gestion*. Thèse de doctorat, Université de Marne-la-Vallée.

Laurence Danlos and Benoît Sagot. 2009. Constructions pronominales dans Dicovalence et le lexique-grammaire: Integration dans le Lefff. Actes du 27e Colloque international sur le lexique et la grammaire.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, in *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italy, pp 417-422.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Aggeliki Fotopoulou, Marianna Mini, Mavina Pantazara and Argiro Moustaki. 2008. La combinatoire lexicale des noms de sentiments en grec modern. In *Iva Novacova & Agnes Tutin (eds), Le lexique des émotions. ELLUG*, Grenoble.

Voula Giouli and Aggeliki Fotopoulou. 2012. Emotion verbs in Greek. From Lexicon-Grammar tables to multi-purpose syntactic and semantic lexica. In *Proceedings of the XV Euralex International Congress (EURALEX 2012)*. Oslo, Norway.

Voula Giouli and Aggeliki Fotopoulou. 2013. Developing Language Resources for Sentiment Analysis in Greek. In *Proceedings of the Workshop "The semantic domain of emotions: cross-domain and cross-lingual considerations. From words to phrases/text and beyond*". Workshop organized within the framework of the *International Conference in Greek Linguistics. ICGL*, Rhodes.

Voula Giouli, Aggeliki Fotopoulou, Effie Mouka, and Ioannis E. Saridakis. 2013. Annotating Sentiment Expressions for Lexical Resourcres. In Blumenthal, Peter, Novakova, Iva, Siepmann, Dirk (eds.), *Les émotions dans le discours. Emotions in discourse*. Frankfurt, Main et al.: Peter Lang.

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

Maurice Gross. 1981. Les bases empiriques de la notion de prédicat sémantique. *Langages* 15, 7-52.

Allain Guillet and Christian Leclère. 1981. La restructuration du sujet. *Langages* 65. Paris, France.

Zelling S. Harris. 1951. *Methods in Structural Linguistics*. The University of Chicago Press, Chicago.

Zelling S. Harris. 1964. The Elementary Transformations. In *T.D.A.P. University of Pennsylvania* 54, Pennsylvania.

Zelling S. Harris. 1968. *Mathematical Structures of Language*. Wiley, New York.

Nabil Hathout and Fiammetta Namer. 1998. Automatic Construction and Validation of French Large Lexical Resources: Reuse of Verb Theoretical Linguistic Descriptions. In *Proceedings of the Language Resources and Evaluation Conference, Grenada, Spain*.

Nick Hatzigeorgiu, Maria Gavrilidou, Stelios Piperidis, George Carayannis, Anna Papakostopoulou, Anna Spiliotopoulou, Anna Vacalopoulou, Penny Labropoulou, Elena Mantzari, Harris Papageorgiou, and Iason Demiros. 2000. Design and Implementation of the Online ILSP Greek Corpus. In *Proceedings of the 2nd Language Resources and Evaluation Conference (LREC, 2000)*, Athens, Greece.

Lifeng Jia, Clement Yu and Weiyi Meng. 2009. The effect of Negation on Sentiment Analysis and Retrieval Effectiveness. In *Proceedings of the 18th ACM conference on Information and knowledge management*, Hong Kong, pp. 1827-1830.

Manfred Klenner, Stefanos Petrakis and Angela Fahrni. 2009. Robust Compositional Polarity Classification. In *Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria

Yvette Yannick Mathieu. 1999. Un classement sémantique des verbes psychologiques. *Cahiers du C.I.EL*: pp.115-134

Yvette Yannick Mathieu. 2008. Navigation dans un texte à la recherche des sentiments. *Linguisticae Investigationes*. 31:2, pp. 313-322.

Yvette Yannick Mathieu and Christiane Fellbaum, 2010. Verbs of Emotion in French and English. *Emotion*, vol. 70, 2010.

Karo Moilanen and Stephen Pulman. 2007. Sentiment Composition. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria, 2007, pp 378–382.

Effie Mouka, Voula Giouli, Aggeliki Fotopoulou, and Ioannis E. Saridakis. 2012. Opinion and emotion in movies: a modular perspective to annotation. In *Proceedings of the 4th International Workshop on Corpora for Research on Emotion, Sentiment & Social Signals (ES³ 2012)*. Istanbul, Turkey.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2009. Compositionality Principle in Recognition of Fine-Grained Emotions from Text. In *Proceedings of the International Conference on Weblogs and Social Media*, AAAI, San Jose, USA, May 2009, pp. 278–281.

Sébastien Paumier. 2003. *UNITEX User Manual*.

Livia Polanyi and Annie Zaenen. 2006. Contextual Valence Shifters. In Shanahan, J., Qu, Y., and Wiebe, J. *Computing Attitude and Affect in Text: Theory and Applications*. Berlin: Springer, pp. 1-10.

Philip J. Stone and Earl B. Hunt. 1963. A computer approach to content analysis: studies using the General Inquirer system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference. Detroit, Michigan, pp. 241-256.*

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. In *Proceedings of Language Resources and Evaluation Conference* (LREC 2004), pp. 1083-1086.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *In Proceedings of HLT-EMNLP-2005.*

Teresa Wilson. 2008. Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States. University of Pittsburgh. Available at: http://mpqa.cs.pitt.edu/data/TAWilsonDissertationCh7Attitudes.pdf. [Accessed November 2011]

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow and Andres Montoyo. 2010. A survey on the Role of Negation in Sentiment Analysis. In: *Proceedings of NeSp-NLP '10, pp. 60-68*.