EMOCause: An Easy-adaptable Approach to Emotion Cause Contexts

Irene Russo **Tommaso Caselli** **Francesco Rubino**
ILC "A.Zampolli" – CNR
Via G. Moruzzi, 1- 56124 Pisa
{irene.russo}{tommaso.caselli}{francesco.rubino}@ilc.cnr.it

Ester Boldrini **Patricio Martínez-Barco**
DSLI – University of Alicante
Ap. de Correos, 99 – 03080 Alicante
{eboldrini}{patricio}@dlsi.ua.es

Abstract

In this paper we present a method to automatically identify linguistic contexts which contain possible causes of emotions or emotional states from Italian newspaper articles (La Repubblica Corpus). Our methodology is based on the interplay between relevant linguistic patterns and an incremental repository of common sense knowledge on emotional states and emotion eliciting situations. Our approach has been evaluated with respect to manually annotated data. The results obtained so far are satisfying and support the validity of the methodology proposed.

1 Introduction

As it has been demonstrated in Balahur et al. (2010), mining the web to discriminate between objective and subjective content and extract the relevant opinions about a specific target is today a crucial as well as a challenging task due to the growing amount of available information.
Opinions are just a part of the subjective content, which is expressed in texts. Emotions and emotional states are a further set of subjective data. Natural Language is commonly used to express emotions, attribute them and, most importantly, to indicate their cause(s).
Due to the importance of linking the emotion to its cause, a recent subtask of Sentiment Analysis (SA) consists in the detection of the *emotion cause event* (ECE, Lee et al., 2010; Chen et al., 2010) and focuses on the identification of the phrase (if present, as in 1 in bold) mentioning the event that is related to the emotional state (in italics):

(1) Non poteva mancare un accenno alla ***strage* di Bologna**, che costringe l' animo *a infinita vergogna.*
[*There was a mention of Bologna massacre,*
that forces us to feel ashamed.]

This kind of information is extremely interesting, since it can provide pragmatic knowledge about content words and their emotional/subjective polarity and consequently it can be employed for building up useful applications with practical purposes.
The paper focuses on the development of a method for the identification of Italian sentences which contain an emotion cause phrase. Our approach is based on the interplay between linguistic patterns which allow the retrieval of emotion – emotion cause phrase couples and on the exploitation of an associated incremental repository of commonsense knowledge about events which elicit emotions or emotional states. The methodology is only partially language dependent and this approach can be easily extended to other languages such as Spanish. The repository is one of the main results of this work. It allows the discovery of pragmatic knowledge associated with various content

Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pages 153–160,
24 June, 2011, Portland, Oregon, USA

words and can assign them a polarity value which can be further exploited in more complex SA and Opinion Mining tasks.
The present paper is structured as follows. Section 2 shortly describes related work and state of the art on this task. Section 3 focuses on the description of the methodology. Section 4 describes the annotation scheme and the corpus used for the creation of the test set. Section 5 reports on the experiments and their results. Conclusions and future works are described in Section 6.

1 Related Works

Emotional states are often triggered by the perception of external events (pre-events) (Wierzbicka, 1999). In addition to this, emotional states can also be the cause of events (post-events; Chun-Ren, 2010). This suggests to consider emotional states as a pivot and structure the relations between emotional states and related events as a tri-tuple of two pairs:

(2) <<pre-events, emotional state>
<emotional state, post-event>>

This study focuses on the relationship between the first pair of the tri-tuple, namely pre-events (or ECE), and emotional states.
Previous works on this task have been carried out for Chinese (Lee et al., 2009, Chen et al., 2009, Lee et al., 2010). ECE can be explicitly expressed as arguments, events, propositions, nominalizations and nominals. Lee et al (2010) restrict the definition of ECE as the immediate cause of the emotional state which does not necessarily correspond to the actual emotional state trigger or what leads to the emotional state. Their work considers all possible linguistic realization of EKs (nouns, verbs, adjectives, prepositional phrases) and ECEs. On the basis of an annotated corpus, correlations between emotional states and ECEs have been studied in terms of different linguistic cues (e.g. position of the cause events, presence of epistemic markers...) thus identifying seven groups of cues. After that, they have been implemented in a rule-based system, which is able to identify: i.) the EK; ii.) the ECE and its position (same sentence as the EK, previous sentence with respect to the EK, following sentence with respect to the EK) and (iii.) the experiencer of the emotional state(s).The system evaluation has been performed on the annotated corpus in two phases: firstly, identifying those sentences containing a co-occurrence of EK and ECE; secondly, for those contexts where an EK and ECE co-occurs, identifying the correct ECE. Standard Precision, Recall and F-measure have been used. The baseline is computed by assuming that the first verb on the left of the EK is the ECE. The system outperforms the baseline f-score by 0.19. Although the results are not very high, the system accuracy for the detection of ECEs is reported to be three times more accurate than the baseline.

2 Emotional states between linguistic patterns and commonsense knowledge

The work of Lee et al. (2010) represents the starting point for the development of our method. We depart from their approach in the following points: i.) use of data mining techniques (clustering plus a classifier) to automatically induce the rules for sentential contexts in which an event cause phrase is expressed; and ii.) exploitation of a commonsense repository of EK - eliciting ECE noun couples for the identification of the correct ECE noun. The remaining of this section will describe in details the creation of the repository and the methodology we have adopted.

2.1 A source for commonsense knowledge of EKs and ECEs in Italian

Recently crowdsourcing techniques that exploit the functionalities of the Web 2.0 have been used in AI and NLP for reducing the efforts, costs and time for the creation of Language Resources. We have exploited the data from an on-line initiative launched in December 2010 by the Italian newspaper "*Il Corriere della Sera*" which asked its readers to describe the year 2010 with 10 words. 2,378 people participated in the data collection for a total of 22,469 words. We exploited these data to identify preliminary couples of emotional states and cause events, and thus create a repository of affective commonsense knowledge, by extracting all

bigrams realized by nouns for a total of 18,240 couples *noun1-noun2*. After this operation, an adapted Italian version of WN-Affect (Strapparava – Valitutti, 2004) obtained by means of mapping procedures through MultiWordNet (MWN) has been applied to each item of the bigrams. By means of a simple query, we have extracted all bigrams where at least one item has an associated sense corresponding to the "*emotion*" category in WN-Affect. We have applied WN-Affect again to these results and extracted only those bigrams where the unclassified item corresponded to the WN-Affect label of *"emotion eliciting situation*". Finally, two lists of keywords have been obtained: one denoting EKs (133 lemmas) and the other denoting possible ECEs associated with a specific EK. The possible ECEs have been extended by exploiting MWN synsets and lexical relations of `similar-to`, `pertains-to`, `attribute` and `is-value-of`. We have filtered the set of ECE keywords by selecting only those nouns whose top nodes uniquely belongs to the following ontological classes, namely: *event*, *state*, *phenomenon*, and *act*. After this operation we have 161 nominal lemmas of possible ECEs.

2.2 Exploiting the repository for pattern induction

The preliminary version of the repository of EK - ECE couples has been exploited in order to identify relevant syntagmatic patterns for the detection of nominal ECEs. The pattern induction phase has been performed on a parsed version of a large corpus of Italian, the La Repubblica Corpus (Baroni et al., 2004).

We have implemented a pattern extractor that takes as input the couples of the seed words from the commonsense repository and extracted all combinations of EKs and its/their associated ECEs occurring in the same sentence, with a distance ranging from 1 to 8 possible intervening parts-of-speech. We have thus obtained 1,339 possible patterns. This set has been cleaned both on the basis of pattern frequencies and with manual exploration. In total 47 patterns were selected and were settled among the features for the clustering and classifier ensemble which will be exploited for the identification of the sentential contexts which may contain an emotion cause phrase (see Section 5 for details).

3 Developing a gold standard and related annotation scheme

With the purpose of evaluating the validity and reliability of our approach, a reference annotated corpus (*gold standard*) has been created.

The data collection has been performed in a semi-automatic way. In particular, we have extracted from an Italian lexicon, SIMPLE/CLIPS (Ruimy et al., 2003), all nouns marked with semantic type "*Sentiment*" to avoid biases for the evaluation and measure the coverage of the commonsense repository. The keywords have been used to query the La Repubblica Corpus and thus creating the corpus collection. We have restricted the length of the documents to be annotated to a maximum of three sentences, namely the sentence containing the emotion keyword, the one preceding it and the sentence immediately following. As a justification for this choice, we have assumed that causes are a local focus discourse phenomenon and should not be found at a long distance with respect to their effects (i.e. the emotion keyword). Finally, the corpus is composed by 6,000 text snippets for a total of 738,558 tokens.

The corresponding annotation scheme, It-EmoCause, is based on recommendations and previous experience in event annotation (*ISO-TimeML*), emotion event annotation (Lee et al., 2009, Chen et al., 2010), emotion and affective computing annotation (*EARL*[1], the HUMAINE Emotion Annotation and Representation Language, *EmotiBlog*, Boldrini et al, 2010). The scheme applies at two levels: phrase level and token level and it allows nested tags. Figure 1 reports the BNF description of the scheme.

Text consuming markables are `<emotionWord>`, `<causePhrase>` and `<causeEmotion>` tags, which are responsible, respectively, for marking the emotion keyword, the phrase expressing the cause emotion event and the token expressing the cause emotion. The values of the attribute `emotionClass` is derived from Ekman

[1] http://emotion-research.net/earl

(1972)'s classification and extended with the value UNDERSPECIFIED. This value is used as a cover term for all other types of emotion reducing disagreement and allowing further classifications on the basis of more detailed and different lists of emotions that each user can specify. Finally, the non-text consuming <EmLink> link puts in relation the cause emotion event or phrase with the emotion keyword.

```
entry ::= <emotionWord> <causePhrase>+
<ELink>*

<emotionWord>     ::=     ewid     lemma
emotionClass         appraisalDimension,
emotionHolder polarity comment
ewid ::= ew<digit>
lemma ::= CDATA
emotionClass ::= HAPPINESS | ANGER |
FEAR | SURPRISE| SADNESS| DISGUST |
 UNDERSPECIFIED
appraisalDimension ::= CDATA
emotionHolder ::= CDATA
polarity ::= POSITIVE | NEGATIVE
comment ::= CDATA

<causePhrase> ::= epid <causeEmotion>+
epid ::= ep<digit>
<causeEmotion> ::= eid lemma
eid ::= e<digit>
lemma ::= CDATA

<EmLink>      ::=      elid      linkType
emotionInstanceID   causeEventInstanceID
causePhraseID comment
elid ::= el<digit>
linkType ::= POSITIVE | NEGATIVE
relatedToEmotion ::= IDREF
{relatedToEmotion ::= ewid}
causeEventID ::= IDREF
{causeEventID ::= eid}
causePhraseID  ::= IDREF
{causePhraseID ::= epid}
comment ::= CDATA
```

Figure 1 – BNF description of the EmoContext Scheme

The annotation has been performed by two expert linguists and validated by a judge. The tool used for the annotation is the Brandeis Annotation Tool (BAT)[2]. The corpus is currently under annotation and we concentrated mainly on the development of a test set. Not all markables and attributes have been annotated in this phase.

The inter-annotator agreement (IAA)[3] on the detection of the cause event and the cause phrase are not satisfactory. To have reliable data, we have adopted a correction strategy by asking the annotators to assign a common value to disagreements. This has increased the IAA on cause emotion to K=0.45, and P&R= 0.46. A revision procedure of the annotation guidelines is necessary and annotation specifications must be developed so that the disagreement can be further reduced. Table 1 reports the figures about the annotated data so far.

It-EmoContext Corpus	
# of tokens	32,525
# of emotion keyword	356
# of cause emotion	84
# of causePhrase emotion	104
# emotion – cause emotion couples	95
# of emotion – cause phrase couples	121
Agreement on emotion keyword detection	K = 0.91 P&R = 0.91
Agreement on cause emotion detection	K = 0.34 P&R = 0.33
Agreement on causePhrase detection	K = 0.21 P&R = 0.26

Table 1 - It-EmoContext Corpus Figures

4 Emotion cause detection: experiments and results

In order to find out a set of rules for the detection of emotion cause phrase contexts, we experimented a combination of Machine Learning techniques, namely clustering and rule induction classifier algorithms. In particular, we want to exploit the output of a clustering algorithm as input to a rule learner classifier both available in the Weka platform (Witten and Frank, 2005).

The clustering algorithm is the Expectation-Maximization algorithm (EM; Hofmann and Puzicha, 1998). The EM is an unsupervised algorithm, commonly used for model-based

[2] http://www.batcaves.org/bat/tool/

[3] Cohen's Kappa, Precision and Recall have been used for computing the IAA.

clustering and also applied in SA tasks (Takamura et al. 2006). In this work, we equipped the EM clustering model with syntagmatic, lexical and contextual features. The clustering algorithm has been trained on 2,000 corpus instances of the potential EK - ECE couples of the repository from the La Repubblica corpus along with a three sentence context (i.e the sentence immediately preceding and that immediately following the sentence containing the EK).

Four groups of features have been identified: the first set of features corresponds to a re-adaptation of the rules implemented in Lee et al. (2010); the second set of features implements the 47 syntagmatic patterns that specifically codify the relation between the EK and the ECE (see Section 3.2); the last two set of features are composed, respectively, by a list of intra-sentential bigrams, trigrams and fourgrams for a total of 364 different part-of-speech sequences with the EK as the first element and by a list of 6 relevant collocational patterns which express cause-effect relationship between the ECE and the EK, manually identified on the basis of the authors' intuitions. In Table 2 some examples of each group of features are reported[4].

Group of feature	Instance
Re-adaptation of Lee et al., 2010's rules	Presence of an ECE after the EK in the same sentence
Syntagmatic patterns manually identified	S E S | S E RI S | S V RI A S ...
Bigrams, trigrams and fourgrams POS sequences	S EA | S EA AP | S EA AP S
Relevant collocational patterns	S A per RD/RI S ...

Table 2 – Features for the EM cluster.

We expected two data clusters, one which includes cause emotion sentential contexts where the EK and the emotion cause co-occurs in the same sentence and another where either the emotion cause it is not present or it occurs in a different sentence (i.e. the one before the EK or in the one following it).

In order to evaluate the goodness of the cluster configuration created by the Weka version of the EM algorithm, we have run different clustering experiments. The results of each clustering analysis have been passed to the Weka PART rule-induction classifier. The best results were those which confirmed our working hypothesis, i.e. two clusters. The first cluster contains 869 items while the second 1,131 items.

The PART classifier provided a total of 49 detection rules for the detection of EK – ECE contexts. The classifier identifies the occurrence of a cause phrase in the same sentence but is not able to identify the noun which corresponds to the ECE.

The evaluation of the classifier has been performed on the 121 couples of EK – cause phrase of the test set. As we are aiming at spotting nominal causes of EKs, we have computed the baseline by considering as the correct phrase containing the ECE the first noun phrase occurring at the right of the emotion keyword and in the same sentence since this kind of ECEs tends to occur mostly at this position. In this way the baseline has an accuracy of 0.14 (only 33 NPs were correct over a total of 227 NPs at the right of the EKs). By applying the rules of the PART classifier, we have obtained an overall accuracy of 0.71, outperforming the baseline. As for the identification of the EK - cause phrase couples occurring in the same sentence, we computed standard Precision, Recall and F-measure. The results are reported in Table 3. The system tends to have a high precision (0.70) and a low recall (0.58).

	Total	Correct	P	R	F
EK – cause phrase couple	121	85	0.70	0.58	0.63

Table 3 – Evaluation of the classifier in detecting EF – cause phrase couples.

After this, we tried to identify the correct nominal ECE in the cause phrase. Provided the reduced dimensions of the annotated corpus, no training set was available to train a further

[4] The tags S, EA, RI and similar reported for the last three groups of features are abbreviations for the POS used by the parser. The complete list can be found at http://medialab.di.unipi.it/wiki/Tanl_POS_Tagset

classifier. Thus, to perform this task we decided to exploit the commonsense repository. However, the first version of the repository is too small to obtain any relevant results. We enlarged it by applying two set of features (the syntagmatic patterns manually identified and the collocational patterns used for the clustering analysis).

4.1 Incrementing the repository and discovering EK – ECE couples

Our hypothesis is that the identification of the ECE(s) in context could be performed by looking for a plausible set of nouns which are associated with a specific EK and assumed to be its cause. This type of information is exactly the one contained in the repository described in Section 3.1.

In order to work with a larger data set of ECE entries per emotion keyword, we have applied the syntagmatic patterns manually identified and the collocational patterns on two corpora: i.) La Repubblica and ii.) ItWaC[5] (Baroni et al., 2009). For each EK - ECE couple identified we have kept track of the co-occurrence frequencies and computed the Mutual Information (MI). Frequency and MI are extremely relevant because they provide a reliability threshold for each couple of EK and ECE. In Table 4 we report some co-occurrences of the EK "*ansia*" [anxiety] and ECEs.

ECE	Frequency (La Repubblica Corpus)	Mutual Information
crisi [crisis]	119	5,514
angoscia [anguish]	80	8.762
guerra [war]	185	6.609
pianificazione [planning]	1	4.117
ricostruzione [reconstruction]	19	5.630

Table 4- ECEs co-occurrences with EK "*ansia*"[anxiety].

Each ECE has been associated to a probability measure of eventivity derived from MWN top ontological classes, obtained from the ratio between 1 and the sum of all top ontological classes associated to the ECE lemma. The top nodes "*event"*, "*state"*, "*phenomenon"*, and "*act"* have been considered as a unique top class by applying the TimeML definition of event[6]. This measure is useful in case more than one ECEs is occurring in the context in analysis as a disambiguation strategy. In fact, if more than one ECEs is present, that with the higher frequency, MI and eventivity score should be preferred.

Furthermore, to make the repository more effective and also to associate an emotional polarity to the ECEs (i.e. whether they have positive, negative or neutral values) we have further extended the set of information by exploiting WN-Affect 1.1. In particular we have associated each EK to its emotional category (e.g. despondency, resentment, joy) and its emotional superclass (e.g. positive-emotion, negative-emotion, ambiguous-emotion).

This extended version of the repository has been applied to identify the correct ECE noun for the 95 couples of EK – ECE in the test set. We have splitted the whole set of EK – ECE couples into two subgroups: i.) EK – ECE couples occurring in the same sentence (82/95); and ii.) EK – ECE couples occurring in different sentences (13/95). By applying the repository to the first group, we were able to correctly identify 50% (41/82) of the ECE nouns for each specific EK when occurring in the same sentence. Moreover, we applied the repository also to the EK – ECE couples of the second group: a rough 30.76% (4/13) of the ECE occurring in sentences other that the one containing the EK can be correctly retrieved without increasing the number of false positives. This is possible thanks to the probability score computed by means of MWN top ontological classes, even if the number of annotated examples is too small to justify strong conclusions.

[5] http://wacky.sslmit.unibo.it

[6] To clarify, the ECE "*guerra*" [war] has four senses in MWN. Three of them belong to the top ontological class of "*event*" and one to "*state*". This possible ECE has 1 top ontological node, and its eventivity mesure is 1.

5 Conclusions and future works

In this paper we describe a methodology based on the interplay between relevant linguistic patterns and an incremental repository of common sense knowledge of EK – ECE couples, which can be integrated into more complex systems for SA and Opinion Mining.
The experimental results show that clustering techniques (EM clustering model) and a rule learner classifier (the PART classifier) can be efficiently combined to select and induce relevant linguistic patterns for the discovery of EK – ECE couples in the same sentence. The information thus collected has been organized into the repository of commonsense knowledge about emotions and their possible causes. The repository has been extended by using corpora of varying dimensions (la Repubblica and ItWaC) and effectively used to identify ECEs of specific emotion keywords.
One interesting aspect of this approach is represented by the reduced manual effort both for the identification of linguistic patterns for the extraction of reliable information and for the maintenance and extension of specific language resources which can be applied also to domains other than SA. In addition to this, the method can be extended and applied to identify ECE realized by other POS, such as verbs and adjectives.
As future works, we aim to extend the repository by extracting data from the Web and connecting it to SentiWordNet and WN-Affect. In particular, the connection to the existing language resources could be used to spot possible misclassifications and polarity values.

Acknowledgments

The authors want to thank the RSC Media Group. This work has been partially founded by the projects TEXTMESS 2.0 (TIN2009-13391-C04-01), Prometeo (PROMETEO/2009/199), the Generalitat valenciana (ACOMP/2011/001) and the EU FP7 project METANET (grant agreement n° 249119)

References

Baccianella S., A. Esuli and F. Sebastiani. (2010) . SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010), Malta, May 2010

Balahur A., R. Steinberger, M.A. Kabadjov, V. Zavarella, E. van der Goot, M. Halkia, B. Pouliquen, J. Belyaeva. (2010). Sentiment Analysis in the News. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Malta, May 2010.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G.,Mazzoleni, M. (2004). Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper italian. In: Proceedings of the 4th International conference on Language Resources and Evaluation (LREC-04), Lisbon, May 2004.

Boldrini E, A. Balahur, P. Martìnez-Barco and A. Montoyo. (2010). EmotiBlog: a finer-grained and more precise learning of subjectivity expression models. In: Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV '10). Association for Computational Linguistics.

Chen Y., S.Y.M. Lee, S. Li, and C. Huang. (2010) Emotion Cause Detection with Linguistic Constructions. In: Proceeding of the 23rd International Conference on Computational Linguistics (COLING 2010).

Ekman, P. (1972). Universals And Cultural Differences In Facial Expressions Of Emotions.In: J. Cole (ed.), Nebraska Symposium on Motivation, 1971. Lincoln, Neb.: University of Nebraska Press, 1972. pp. 207- 283.3.

Huang, C. (2010). Emotions as Events (and Cause as Pre-Events). Communication at the Chinese Temporal/discourse annotation workshop, Los Angeles, June 2010,.

Lee S.Y.M., Y. Chen, C. Huang. (2010). A Text-driven Rule-based System for Emotion Cause Detection. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.

Pianta, E., Bentivogli, L., Girardi, C. (2002). Multiwordnet: Developing and aligned multilingual database. In: Proceedings of the First International Conference on Global WordNet, Mysore, India, January 2002.

Pustejovsky, J., Castao, J., Saur`ı, R., Ingria, R., Gaizauskas, R., Setzer, A., Katz, G. (2003). TimeML: Robust specification of event and temporal expressions in text. In: Proceedings of

the 5th International Workshop on Computational Semantics (IWCS-5).

Ruimy, N., Monachini, M., Gola, E., Calzolari, N., Fiorentino, M.D., Ulivieri, M., Rossi, S. (2003). A computational semantic lexicon of italian: SIMPLE. In: Linguistica Computazionale XVIII-XIX, Pisa, pp. 821–64

Schroeder M., H. Pirker and M. Lamolle. (2006). First Suggestion for an Emotion Annotation and Representation Language. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, May 2006.

Strapparava C. and A. Valitutti. (2004) WordNet-Affect: an affective extension ofWordNet". In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, May 2004.

Takamura H., I. Takashi, M. Okumura. (2006). Latent Variables Models for Semantic Orientation of Phrases. In: Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006).

Wierzbicka, A. (1999) Emotion Across Languages and Cultures Diversity and Universals. Cambidge.CUP.