SemEval-2015 Task 9: CLIPEval Implicit Polarity of Events

Irene Russo
ILC-CNR "A. Zampolli"
Via G. Moruzzi, 1
56124 Pisa
irene.russo@ilc.cnr.it

Tommaso Caselli
Vrije Universiteit Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
t.caselli@gmail.com

Carlo Strapparava
Fondazione Bruno Kessler
Via Sommarive, 18
38123 Povo (TN)
strappa@fbk.eu

Abstract

Sentiment analysis tends to focus on the polarity of words, combining their values to detect which portion of a text is opinionated. CLIPEval wants to promote a more holistic approach, looking at psychological researches that frame the connotations of words as the emotional values activated by them. The implicit polarity of events is just one aspect of connotative meaning and we address it with a task that is based on a dataset of sentences annotated as instantiations of *pleasant* and *unpleasant* events previously collected in psychological research as the ones on which human judgments converge.

1 Introduction

Current research in sentiment analysis (SA, henceforth) is mostly focused on lexical resources that store polarity values. For bag-of-words approaches the polarity of a text depends on the presence/absence of a set of lexical items. This methodology is successful to detect opinions about entities (such as reviews) but it shows mixed results when complex opinions about events - involving perspectives and points of view - are expressed.

In terms of parts of speech involved, SA approaches tend to focus on lexical items that explicitly convey opinions - mainly adjectives, adverbs and several nouns - leaving verbs on the foreground. Improvements have been proposed by taking into account syntax (Greene and Resnik 2009) and by investigating the connotative polarity of words (Cambria et al., 2009; Akkaya et al., 2009, Balhaur et al., 2011; Russo et al. 2011; Cambria et al., 2012, Deng et al., 2013 among others). One of the key aspects of sentiment analysis, which has been only marginally tackled so far, is the identification of implicit polarity. By implicit polarity we refer to the recognition of subjective textual units where no polarity markers are present but still people are able to state whether the text portion under analysis expresses a positive or negative sentiment. Recently, methodologies trying to address this aspect have been developed, incorporating ideas from linguistic and psychological studies on the subjective aspects of linguistic expressions.

Aiming at promoting a more holistic approach to sentiment analysis, combining the detection of implicit polarity with the expression of opinions on events, we propose CLIPEval, a task based on a dataset of events annotated as instantiations of *pleasant* and *unpleasant* events (PE/UPEs henceforth) previously collected in psychological research as the ones that correlate with mood (both good and bad feelings) (Lewinsohn and Amenson, 1978; MacPhillamy and Lewinsohn, 1982).

2 Measuring Emotional Connotations: Psychological Studies

For a long time research in psychology has been interested in a subjective cultural and/or emotional coloration in addition to the explicit or denotative meaning of any specific word or phrase. Starting with the work of Charles E. Osgood, who in the 50s developed a technique for measuring the connotative meaning of concepts and analyzed human attitudes (Osgood et al., 1957), psychologists have experi-

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 443–450,
Denver, Colorado, June 4-5, 2015. ©2015 Association for Computational Linguistics

mented with emotional values activated by words, often through the evaluation of their pleasantness. Osgood and his colleagues proposed a factor analysis based on semantic differential scales measuring three basic attitudes that people display cross-culturally: evaluation (along the scale of adjectives "good-bad"), potency (along "strong-weak") and activity ("active-passive").

This line of research continued with studies evaluating Osgood's findings with different population and the pleasantness of words became also a dimension to correlate with other dimensions reported in semantic norms studies, such as familiarity and imagery. We know today that pleasantness is a semantic factor influencing short and long term memory (Monnier et al., 2008); similarly, (Hadley and MacKay, 2006) showed that STM for certain unpleasant emotional words (i.e., taboo words) was better than that for neutral words. Emotional words are better recalled because they are related to long-term representations of autobiographical and self-reference units (Ochsner, 2000). Other factors have a role: depressed subjects, for example, recalled more unpleasant words than pleasant words.

Osgood's studies were revised for the production of the Affective Norms for English Words (ANEW) (Bradley et al, 1999), a set of normative emotional ratings for 1034 words in American English. This set of verbal materials have been rated in terms of *pleasure*, *arousal*, and *dominance* in order to create a standard for use in studies of emotion and attention (the same three basic dimensions used by Osgood). Affective valence (or pleasure, ranging from pleasant to unpleasant) and arousal (ranging from calm to excited) were the two primary dimensions. A third, less strongly-related dimension, was called "dominance" or "control".

Connotative meaning emerges as a complex and stratified concept and only psychological studies can guide in this maze, especially when they are supported by significant experimental outputs such as list of words evaluated by human subjects.

All these studies are relevant for NLP because connotative meanings of words can help to refine automatic sentiment analysis on social media, where shared contents are often just short reports on pleasant or unpleasant events and activities. For example, (Fenf et al., 2013) report that connotation lexicon guarantees better performance than other sentiment analysis lexicons that do not encode connotations on Twitter data.

That said, when psychological experiments ask for judgments about single words they oversimplify: we experience the meanings of single words as arising from compositionality, in expressions and sentences. Even neutral words in specific contexts can acquire a polarity as effect of semantic prosody (Louw 1993).

When subjects are asked for the pleasantness of an event they need to evaluate not just single words but complete sentences; for this reason (Lewinsohn and Amenson, 1978; MacPhillamy and Lewinsohn, 1982) developed two psychometric instruments, the Pleasant Events Schedule and the Unpleasant Events Schedule, by sampling events that were reported to be source of pleasure or distress by highly diverse samples of people that rated the frequency of event's occurrence during past month plus a complete mood ratings.

3 CLIPEval Annotation

The CLIPEval exercise provides the NLP community with a newly developed dataset grounded on psychological studies about the pleasantness of events. Dedicated annotation specifications and guidelines for the release of the dataset have been developed.

The starting point for the development of the annotation guidelines was the PE/UPEs lists, the set of 640 pleasant and unpleasant events (320 pleasant events and 320 unpleasant events, respectively) collected by (Lewinsohn and Amenson, 1978) and (MacPhillamy and Lewinsohn, 1982). The dataset could not be used as it is since it is a list of generic sentences describing either states or actions which are labeled as pleasant or unpleasant events. To clarify this, we report two examples extracted from the original dataset. Example 1.) is a pleasant event while example 2.) is an an unpleasant event. The numbers in brackets at the beginning of the sentence refer to the PE/UPEs number in the original dataset.

1.) *(9) Planning trips or vacations.*

2.) *(10) Getting separated or divorced from my spouse.*

Furthermore, a closer examination of PE/UPEs has shown that ambiguity occurs, with the same events considered both as a pleasant and an unpleasant one (e.g. *Being alone*), since this is plausible from a psychological point of view. To overcome these issues and to make the task relevant for sentences from news articles, we have applied the following strategies:

- all ambiguous PE/UPEs have been removed from the original dataset;
- PE/UPEs have been grouped into classes whose labels describe and aggregate different PE/UPEs, referring often to a more general event class with respect to the one the single instance of a PE/UPE event describes. This choice has been necessary because the event instances in the original psychological dataset are conceptually similar but using the original descriptions would result either in too generic cases (e.g. *Being with children*) or too simple (e.g. *Washing my hair*).

The grouping of PE/UPEs in classes has been conducted in two phases by two annotators. In the first phase, both annotators have worked independently: for each PE/UPE the annotators had to decide which of them could be clustered in a more generic class and which were to be excluded, either because it describes a too specific (or a too generic) event or because it explicitly express the pleasanteness of the event (e.g. *(25) Driving skillfully*). As a measure of agreement for this task, we preferred not to use kappa score, because it's not a standard classification task, but we computed the percentage of agreement. The first evaluation shown a relatively low agreement, only 59.06% of the 640 events were considered as belonging to a cluster. An analysis on the cases of disagreement has highlighted some inconsistencies. Thus, a second clusterization task has been performed by asking to the same annotators to go over the same data following new additional rules that were developed during the analysis. The evaluation of this second phase shown a clear improvement with a percentage agreement of 68.25%. As a result of these annotation phases, we had a set of clusters that the annotators were allowed to discuss, finding a joint solution in cases of disagreements and identifying the best labels for the PE/UPEs clusters. The final output of these two phases resulted in 8 classes of PE/UPEs (see Table 1 column "Event Class"). It is important to point out that most of these classes contain PE/UPEs both from the 320 pleasant events and the 320 unpleasant events and as a consequence the polarities of their occurrences in the training data are mixed(see Table 1). Due to the novelty of the task, we could not re-use available datasets for SA. For this reason, the second step concerns the identification and manual annotation of real sentences from the Annotated English Gigaword corpus (Napoles et al., 2012), an automatically-generated syntactic and discourse structure annotated version of the English Gigaword corpus Fifth Edition, which contains a large English corpus of newspaper articles (four billion words ca.). To facilitate the sentence extraction phase, we manually identified the verbal and the nominal keywords from the event mentions composing the classes. We used WN30 and the Oxford Dictionary to extract all verb and noun synonyms of the PE/UPEs in each class. We then queried the Gigaword corpus with this extended set of keywords to extract sentences which contain self-reported events by means of following patterns:

- "I|we + [verbal_keyword]"
- "I|we + [nominal_keyword]"
- "I|we + [verbal_keyword] + [nominal_keyword]".

The sentences thus extracted were manually filtered and annotated with respect to the 8 classes and to their polarity. The annotation has been conducted at sentence level. To provide homogeneous data and annotations, the following guidelines have been developed for the assignment of the class label:

- the class label and the polarity value must be assigned on the basis of the event that correspond syntactically to the main verb in the sentence;
- in case of coordinated main clauses, only the first main clause is taken into account to assign the class label and the polarity value;

Table 1: CLIPEval corpus: Training data.

Event Class	POSITIVE	NEGATIVE	NEUTRAL	Tot. Instances
(FEAR_OF)_PHYSICAL_PAIN	19	131	10	160
ATTENDING_EVENT	83	35	42	160
COMMUNICATION_ISSUE	21	120	19	160
GOING_TO_PLACES	55	72	33	160
LEGAL_ISSUE	24	115	21	160
MONEY_ISSUE	20	109	31	160
OUTDOOR_ACTIVITY	125	18	17	160
PERSONAL_CARE	88	40	32	160

Table 2: CLIPEval corpus: Test data.

Event Class	POSITIVE	NEGATIVE	NEUTRAL	Tot. Instances
(FEAR_OF)_PHYSICAL_PAIN	10	30	5	45
ATTENDING_EVENT	29	5	11	45
COMMUNICATION_ISSUE	8	29	7	44
GOING_TO_PLACES	22	23	3	48
LEGAL_ISSUE	5	27	13	45
MONEY_ISSUE	12	27	12	51
OUTDOOR_ACTIVITY	34	4	8	46
PERSONAL_CARE	24	10	13	43

- subordinated clauses are not annotated with class labels and polarity values.

Although all event mentions in the selected clusters have either a positive (pleasant events) or negative (unpleasant events) polarity that could be reversed by negation, during the annotation phase a third value, namely neutral, has been introduced to cope with those sentences containing self-reporting events whose occurrence is uncertain

We are referring here to the notion of event factuality (Sauríand Pustejovsky, 2009), i.e. the degrees of certainty (e.g. possible, probable, certain) associated to an event description along the category of epistemic modality. In the annotation we focused on the syntactic information between target events instances and factuality markers, such as modal auxiliaries and negation cues (including adverbs, adjectives, prepositions, pronouns and determiners). Events which are in the scope of factuality markers signaling uncertainty or improbability have been marked as neutral.

4 CLIPEval Tasks

The CLIPEval evaluation exercise is composed of two tasks described as follows:

- Task A: identification of the polarity value associated to the event instance. Participants are required to associate each sentence with a polarity value (POSITIVE, NEGATIVE or NEUTRAL);

- Task B: identification of the event mentions with respect to one of the 8 event class labels plus identification of the polarity value. The class labels used are: ATTENDING_EVENT, COMMUNICATION_ISSUE, GOING_TO_PLACES; LEGAL_ISSUE, MONEY_ISSUE, OUTDOOR_ACTIVITIES, PERSONAL_CARE, (FEAR_OF)_PHYSICAL_PAIN. As in Task A the polarity values are (POSITIVE, NEGATIVE or NEUTRAL).

5 Dataset Description

The CLIPEval evaluation exercise is based on the CLIPEval dataset, which consists of two parts: a training set and a test set. The final size of the dataset is 1,651 sentences, divided in 1,280 sentences for the training and 371 for the test. Each event class in the training data contains 160 sentences.

Each class in the training set is available in a separate file composed of four tab separated fields: a sentence id, the sentence extracted from the Gigaword corpus, the polarity value and the class label. Each file is named with the class label. Some examples of the training data are provided in the examples below (examples from 3.) to 5.)):

3.) 8 *I had just gone to a concert with my parents and I identified with the conductor a lot Dudamel said in Spanish during a recent interview in Caracas.* POSITIVE ATTENDING_EVENT

4.) 14 *"It's too cold and I can't ride my bike" he lamented.* NEGATIVE OUTDOOR_ACTIVITY

5.) 4 *"I could take the boys to the sports museum' says James.* NEUTRAL GOING_TO_PLACES

The test data has been provided in a single file with only two fields: the sentence id and the sentence extracted from the Gigaword corpus:

6.) 12 *After having given a friend a lift home I was stopped by police.*

7.) 23 *And then we went to a library.*

Table 1 and Table 2 report the figures for polarity values per class in the training and in the test set, respectively.

The division of the training data for the three polarity values is not balanced due to the event mentions composing the clusters. Only three clusters, namely GOING_TO_PLACES, PERSONAL_CARE and ATTENDING_EVENT, present a relatively balanced distribution for the polarity values. This lack of balance reflects real language data: the prevalence of positive or negative values is due to the classes which may have more PEs or UPEs (e.g. OUTDOOR_ACTIVITY and COMMUNICATION_ISSUE, respectively). Including more sentences which reverse the polarity of the PEs or UPEs to balance the occurrences per polarity value would mean to force the data from real language toward an artificial equilibrium.

6 Evaluation

Since both Task A and Task B of CLIPEval are essentially classification tasks (classification of the polarity value for Task A and classification of the event instance and the polarity value for Task B), we have used Precision, Recall and F1-measure to evaluate the system results against the test set. Furthermore, since this is a multi-classification task (3 possible values for Task A and 24 possible values for Task B), we have computed micro average Precision, Recall and F1-measure per class. This latter measure has been used for the final ranking of the systems. We have adopted standard definitions for these measures, namely:

- Precision: the number of correctly classified positive examples, tp_i per class C_i, divided by number of examples labeled by the system as positive (tp_i plus false positive fp_i): $\frac{\sum_{i=1}^{l} tp_i}{tp_i + fp_i}$

- Recall: the number of correctly classified positive examples tp_i per class C_i divided by the number of positive examples in the data (tp_i plus false negatives fn_i) : $\frac{\sum_{i=1}^{l} tp_i}{tp_i + fn_i}$

- F-measure: the mean of Precision and Recall calculated as follows: $\frac{(\beta^2+1) Precision Recall}{\beta^2 Precision + Recall}$

To better evaluate systems' performances, we have developed three baselines, one per Task A and two per Task B. In particular:

- Task A baseline has been obtained by assigning to each sentence in the test set the most frequent polarity value on the basis of the data in the training set. This resulted in marking all 371 sentences in the test set with NEGATIVE polarity;

- Task B baseline_1 has been obtained in two steps: first, for each class in the training data we have selected the most frequent nouns and verbs lemmas. This has provided us with a list of keywords representing each class. We have then compared each sentence in the test set with each group of keywords and assigned as correct the class which scored the higher number of matches. In case of a draw, a random class between the classes with the highest scores is assigned. If no match is found, a random class is assigned. As for the polarity, we have used the absolute most frequent polarity values, like in task A (i.e. all test set entries have been assigned to NEGATIVE value).

- Task B baseline_2 has been obtained following the approach in Task B baseline_1 for the class assignment and we have assigned the most frequent polarity value per class according to training data (e.g. for items classified as ATTENDING_EVENTS the assigned polarity value is POSITIVE).

6.1 Participant Systems

Overall 26 different teams registered for the task, only two submitted the output of their system for a total of 3 runs: SHELLFBK (Fondazione Bruno Kessler) and SIGMA2320 (Peking Univeristy). Only SHELLFBK submitted results for both tasks. Furthermore, we can provide a short description just for SHELLFBK since the SIGMA2320 team has not submitted a system description paper.

SHELLFBK system implements a supervised approach based on information retrieval techniques for representing polarized information. During the training phase, each sentence is analyzed by applying the parser contained in the Stanford NLP Library. From the results of the parsing activity, both the list of the dependency relations and the parsed trees are used for populating an inverted index data structure containing the relationships between each relation extracted from the sentences and the corresponding information about its polarization. The result of the training phase is a set of three indexes containing, respectively, the positive, negative, and neutral information analyzed in the training set. When the polarity of a new sentence has to be computed, the new sentence is given as input to the Stanford NLP Library by obtaining the list of its dependency relations, as well as, the corresponding parsed tree. Such information are built together for composing a query that is afterwards performed on the indexes built during the training phase. For each of the built indexes, a retrieval score value is retrieved by the system and, based on this, the polarity of the new sentence is assigned.

6.2 Evaluation Results

We report in Table 3 the results of both systems for Task A and the Task A baseline. In Table 4 we report the results for Task B and both baseline for Task B (baseline_1 and baseline_2, respectively).

Table 3: Evaluation for Task A : polarity identification.

System	Precision	Recall	F1-measure
SIGMA2320	0.41	0.42	0.38
SHELLFBK	0.56	0.56	**0.54**
baseline	0.17	0.42	0.25

Table 4: Evaluation for Task B : event instance and polarity identification.

System	Precision	Recall	F1-measure
SHELLFBK	0.36	0.27	**0.29**
baseline_1	0.02	0.04	0.02
baseline_2	0.03	0.05	0.04

SHELLFBK outperforms SIGMA2320 for the Task A; both systems improve the baseline. The results are not as good as in classification tasks concerning the polarities of tweets (Rosental et al., 2014) or reviews (Pontiki et al., 2014) but since this is a novel task about implicit polarity we think they are promising.

For task B SHELLFBK has a better performance both in terms of precision and recall if compared with the two baselines. At the moment we do not know if the results are due to SHELLFBK methodology or if data sparseness in the classes has an influence on the classification task: maybe classes more cohesive from conceptual and lexical point of view could be easier to detect.

7 Conclusions and Future Work

The implicit polarity of words concerns the arising of occasional polarized meanings in specific expressions/linguistic contexts. Labeled as semantic prosody in corpus studies and part of what psychologists call connotative meanings, the implicit polarity is a quite marginal concept in sentiment analysis. It requires a dynamic representation for the polarity of words (i.e. a verb can be neutral in the vast majority of case but can be clearly positive in some contexts) and a compositional approach to sentiment values that goes beyond the oversimplifying assumptions of bag-of-words approaches.

With the CLIPEval task we asked the NLP community to look at these complexities, considering the detection of a set of events as relevant for SA analyses because they have been judged as pleasant or unpleasant by subjects in psychological experiments conducted by (Lewinsohn and Amenson, 1978; MacPhillamy and Lewinsohn, 1982). As future work we plan to extend the dataset, including new classes of events and annotating instances from blogs and tweets. Also, we want to integrate the detection of polarized events with the work on stance and perspectives in news, going toward a theoretical model for SA that takes into account the interplay of linguistic means used by humans to express opinions and feelings.

Acknowledgments

One of the author wants to thanks the NWO Spinoza Prize project Understanding Language by Machines (sub-track 3) for partially supporting this work.

References

Cem Akkaya, Janyce Wiebe and Mihalcea Rada. 2010. Subjectivity Word Sense Disambiguation. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 190 - 199.

Alexandra Balahur, Jesús M. Hermida and Andrés Montoyo. 2011. Detecting implicit expressions of sentiment in text based on commonsense knowledge. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis.*, pages 53 - 60.

Margaret M. Bradley and Peter J. Lang. 1999. Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. *Tech. Rep. No. C-1*

Erik Cambria, Robert Speer, Catherine Havasi and Amir Hussain. 2010. SenticNet: A Publicly Available Semantic Resource for Opinion Mining. *AAAI fall symposium: commonsense knowledge*, pages 14 - 18.

Erik Cambria, Catherine Havasi and Amir Hussain. 2012. SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. *FLAIRS Conference*, pages 202 - 207.

Ling Deng, Yoonjung Choi and Janyce Wiebe. 2012. Benefactive/Malefactive Event and Writer Attitude Annotation. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120 - 125.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1774 - 1784.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. *Proceedings of human language technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503 - 511.

Christopher B. Hadley and Donald G. MacKay. 2006. Does emotion help or hinder immediate memory? Arousal versus priority-binding mechanism. *Journal of Abnormal Psychology*, 87(6), pages 644 - 654.

Peter M. Lewinsohn and Christopher S Amenson. 1978. Some Relations between Pleasant and Unpleasant Events and Depression. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, pages 79 - 88.

Douglas J. MacPhillamy and Peter M. Lewinsohn. 1982. The Pleasant Event Schedule: Studies on Reliability, Validity, and Scale Intercorrelation. *Journal of Counseling and Clinical Psychology*, 50(3), pages 363 - 380.

Catherine Monnier and Arielle SyssauMonnier. 2008. Semantic contribution to verbal short-term memory: Are pleasant words easier to remember than neutral words in serial recall and serial recognition? *Memory and Cognition*, 36, pages 35 - 42.

Courtney Napoles, Matthew Gormley and Benjamin Van Durme. 2012. Annotated gigaword. *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95 - 100.

Kevin N. Ochsner. 2000. Are affective events richly recollected or simply familiar? The experience and process of recognizing feelings past. *Journal of Experimental Psychology. General*, 129 (2), pages 242 261.

Charles E. Osgood, George Suci and Percy Tannenbaum. 1957. *The Measurement of Meaning.* University of Illinois Press,

Roser Saurí and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language resources and evaluation*, 3, pages 227 - 268.

Maria Pontiki, Dimitris Galanis, Pavlopoulos Ioannis, Harris Papageorgiou, Androutsopoulos Ion and Manandhar Suresh. 2014. SemEval 2014 Task 4: Aspect Based Sentiment Analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27 - 35.

Sara Rosenthal, Alan Ritter, Preslav Nakov and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment Analysis in Twitter. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*

Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini and Patricio Barco Martínez. 2011. EMOCause: An Easy-adaptable Approach to Extract Emotion Cause Contexts. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 152 - 160.