Correlating open rating systems and lived experiences extraction from text

Ehab HASSAN
LIPN, Université Paris XIII,
CNRS UMR 7030
93430 Villetaneuse - France
ehab.hassan@lipn.univ-paris13.fr

Davide BUSCALDI
LIPN, Université Paris XIII,
CNRS UMR 7030
93430 Villetaneuse - France
davide.buscaldi@lipn.univ-paris13.fr

Aldo GANGEMI
LIPN, Université Paris XIII,
CNRS UMR 7030
93430 Villetaneuse - France
aldo.gangemi@lipn.univ-paris13.fr

ABSTRACT

Event extraction is a very important task for research textual information. This task can be applied to various types of written text, e.g. news messages, blogs, manuscripts, and user reviews for products or services. In this paper, we report results about an experiment in correlating Event-based lived experience patterns obtained from machine reading, and ranking derived from open rating systems.

Keywords

Lived experiences extraction, event extraction, machine reading, semantic web, user reviews.

1. INTRODUCTION

The web has significantly changed how people express themselves and interact with others. Now they can post reviews of products and services in merchant websites and express their opinions and interact with others through blogs and forums. It is now well agreed that user generated content contains valuable information that can be used for many applications.

Reviews offer (often implicitly) suggestions or opinions on the basis of lived experiences. These reviews are very important in user decisions since they contain non-fictional narrative or stories that people tell about their experiences.

Open rating systems allow to synthetically grasp the opinion of the crowds with reference to specific entities: products, services, statements of ideas, etc.

In this paper, we describe our efforts to study if lived experience events given in reviews correlate with open rating, and if it is possible to extract relevant lived experience event dictionaries from user reviews.

We formulate these tasks as a binary text classification task. We explored machine learning techniques to build a classifier so as to classify two types of reviews (Positives and Negatives) using lived experience features.

We define user lived experiences as textual discourse that describes events, where the author is among the participants. In addition, we consider that lived experiences should tell real lived facts and do not contain generic user opinion.

As an example, let us consider the following hotel review:

- The view from this hotel's rooms is quite stunning. And that's what make it very special, possibly better than the next door 4 star hotel and than many other hotels in Paris. The bedrooms interior decor is extremely nice. I asked for a room overlooking the pantheon and I got it. My deluxe room was number 32, and was tastefully decorated with a classic and beautiful Pierre Frey wallpaper, and an extra day bed. The bath had bathtub-shower combination and was separated from the toilet. If you book directly through the hotel, you'll get a voucher for a free-drink upon arrival. It was a bit cold at night at some point, maybe because it's March and the heating is not constantly on anymore. Each room has its own heating control, though. Strongly recommended.

According to our definition, this review contains lived experiences represented by events where the user is among their participants: (1) *I asked for a room overlooking the pantheon and I got it.* (2) *My deluxe room was number 32, and was tastefully decorated with a classic and beautiful Pierre Frey wallpaper, and an extra day bed.* (3) *If you book directly through the hotel, you'll get a voucher for a free-drink upon arrival.* The user here tell three facts indicating three lived experiences represented by three personal events {*Got, Ask, Decorate*} and their participants.

In order to extract lived experience events from user reviews, we use a *machine reader* to perform a deep semantic parsing of text which allow to obtain a RDF Linked-Data-ready graph representation of the text.

2. RELATED WORK

2.1 Lived Experience Definition

Lived experiences have been studied mainly in the context of anthropological, historical, and health studies [2, 11].

To our knowledge, the only studies devoted specifically to lived experiences, called "personal stories", from the information extraction perspective are [5, 4], which defines personal stories as textual discourse that describes a specific series of causally related events in the past, spanning a period of time of minutes, hours, or days, where the author or a close associate is among the participants. In this research, we subscribe to this definition to identify lived experience events and used them to classify reviews.

2.2 Personal Event Extraction

Event extraction is the field of research to which our work is more related. Previous work in event extraction focused largely on news articles, There has been relatively little

SEMANTICS 2016: Posters and Demos Track
September 13-14, 2016, Leipzig, Germany

previous work in event extraction from user reviews. Van Oorschot et al. [12] extract game events (e.g. goals, fouls) from tweets about football match to automatically generate match summaries. Events were detected and classified using a machine learning approach. Ritter et al. [10] present an open domain event extraction within Twitter. They propose an approach based on latent variable models to categorize events and classifying extracted events in an open-domain text with 0.64 of F-measure. Due to the difference in structure, this work is not suitable for extracting events from user reviews. In addition, tweets typically include a single event while the users reviews include sequence of events, which makes our task different from that discussed in [10]. Ploeger et al. [8] introduce an automatic activist events extraction method from various news sources using NLP tools. They extracted 1829 events with 0.71 precision 0.58 recall and 0.64 F-Measure. Unlike most approaches mentioned above, we do not use a predefined list of potentially interesting events but, we use a system of "machine reading", FRED [9], to automatically identify and extract events. Hassan et al. (2015) [6] studied the correlation between events and open rating system. They found an F-Measure of more about 0.84 for classifying either positive or negative reviews with event types extracted from reviews. In our work, we follow an approach similar to Hassan et al. (2015)[6]. However, in this work, we attempt to use lived experience events for classifying polarity of reviews and to correlate user reviews with their rating.

3. DATASET

Ott et al. [7] have recently created the first publicly available [1] dataset for deceptive opinion spam research containing 800 positive reviews (400 truthful reviews and 400 fake reviews) which have been assigned with 5-stars in the open system ranking and 800 reviews for the negative polarity (400 truthful reviews and 400 fake reviews) which have 1-star in the open system ranking. In this work, we were only interested in the truthful reviews which are collected from the 20 most popular Chicago hotels on TripAdvisor [2]. We selected 600 reviews, 300 user reviews for the positive reviews (15 reviews for each hotel) and 300 user reviews for the negative ones. In our experimentation, 420 user reviews were used as a training set to build the classifier, 210 reviews for the positive class and 210 reviews for the negative one, and 180 user reviews were used to evaluate our classifier (90 for each class).

4. PERSONAL EVENT DICTIONARY CONSTRUCTION

4.1 Event extraction

In this research, we focus on events that can be directly extracted from explicit mentioning and expressed by verbs, propositions, common nouns, and named entities (typically proper nouns).

We employ a deep variety of machine reading [3], as implemented in the FRED tool [3], which extracts knowledge (named entities, senses, taxonomies, relations, events, etc.) from text, resolves it onto the Web of Data, adds data from background knowledge, and represents all that in RDF and OWL.

FRED is event-centric, therefore it natively supports event extraction. It is available as a RESTful API, and as a web application.

Applying SPARQL query to the semantic graph produced by FRED, we can extract all events mentioned in the text with their main participants. From our example in the introduction, we are able to extract eight events {*Get, Overlook, Recommend, Ask, Decorate, Have, Make, Separate*}.

According to our definition in the introduction, we consider events referenced in non-fictional narrative that people tell about their lives as a lived experience. Therefore, we categorize extracted events from user reviews in two types: *(1)* Personal (or lived experience) events: the events which have the narrator among their participants. These events should have the first or the second person pronoun (i.e. I, You, We, Me, Us, My, Mine, Our, Ours, Your, Yours, ...) as a participant. *(2)* General (or non lived experience) events: the events which do not have the user among their participants. In this work, we are interested in personal events to build our dictionary and classify the polarity of user reviews.

4.2 Dictionary Construction

Hassan et al. (2015)[6] found that events may be used to find whether a certain review is negative or positive. Therefore, we assume that lived experience events may be also used to perform this task with some good results: often, users describe personal events that led to an uncomfortable or an enjoyable experience during their test of the products or visit to the hotels. For instance, from our example in the introduction, we are able to identify and extract three events {*Get, Ask, Decorate*}, which represent lived experience events since they have the user among their participants.

Our approach for the construction of the dictionary is the following:

1. Select a set of "positive" and "negative" reviews from TripAdvisor;
2. Extract all personal events contained therein;
3. Consider the events that appear in the highly rated reviews as positive events (agreeable), and the events that appear in the worst rated reviews as negative events (disagreeable);
4. According to event frequencies, select a limited number events to create the dictionary D_E;
5. Use the events in D_E as features for a multinominal Naïve Bayes classifier to check the correlation between personal event types and ranking.

Using our training set, we were able to recognize and extract 759 personal events: 306 for the positive reviews and 453 for the negative ones.

The resulting dictionary could not be used to discriminate between the two types of events since many reviews, positive and negative, may contain common events which

[1]Available by request at: http://www.cs.cornell.edu/~myleott/op_spam

[2]http://www.tripadvisor.com/

[3]http://wit.istc.cnr.it/stlab-tools/fred, see also http://www.semantic-web-journal.net/system/files/swj1297.pdf for a more recent description

are distributed in the two classes in a homogeneous manner. Therefore, We improved our classifier results by removing those events. We considered that an event is representative to a class if the probability $P(c|e) \geq \sigma$ where:
$c \in C = \{+,-\}$;
e: A generic event;
σ: A threshold that we determined empirically between 4 possibilities: {0.6, 0.7, 0.8, 0.9} . The best value of σ was 0.7 [6].

4.3 Event Participants Extraction

Event participants are the semantic arguments associated with this event. Each event argument play a semantic role (e.g. Agent, Patient, Oblique, Theme, etc.).

Our SPARQL query generates an event sub-graph containing personal events, with their direct or indirect participants. Direct participants are the arguments which connect to an event directly. i.e. the direct objects of the events. Indirect participants belong to events that sometimes occur as direct objects of main events. This sub-graph also contain event modifiers such as modality, negation, and adverbial qualities.

We assume that event participants and modifiers could be useful for the polarity classification task and can upgrade our results. Therefore, we extract these arguments for the first, second and third degree and used them with their events to build an event-participant dictionary.

Using our training set, we were able to build a new event-participant dictionary containing 2058 features: 668 uniques for the positive reviews, 76 common events discriminating the positive reviews, 1153 uniques for the negative reviews, and 161 common events, but important to discriminate the negative reviews.

5. EVALUATION

As we mentioned above, we used our dictionaries as a collection of features to train a multinomial Naïve Bayes classifier as our classification model. Naïve Bayes is very well suited for binary classification tasks, and it has the ability to deal with large space features. Since we were looking at entire reviews rather than segments of sentences, it was useful to consider the frequency of events in the review. Therefore, each review is transformed into a feature vector where the $i-$th component value is the frequency of the $i-$th event in the review, and then classified as either "Positive" or "Negative".

Our approach is evaluated in terms of precision, recall, and F-measure. We performed 10-fold cross-validation on the training set and yielded good results Table 1.

To validate the obtained results, we applied our review classifier to the test set which contains 180 user reviews. The achieved results for this set can be shown in Table 1.

As shown in Table (1), the results using personal events when $\sigma = 0.7$ have been particularity good. However, using all personal events without deleting the common ones decrease the performance by about 5% for the training set and 1% for the test set. Further adding participant features to all personal events give good results, but less effective than the results obtained using personal features with $\sigma = 0.7$, which achieves the best results for the training set. Using these features allows to obtain a slight increase, (8%) for the training set and (8%) for the test set, compared to the achieved results using event features.

Table (2) shows the results which were obtained by Hassan et al. (2015)[6] using all event features, personal and general events, which are mentioned in the text for the same training set and test set.

From Table (1) and Table (2), we observe that performance using all event features is more efficient than performance using only personal events for the polarity classification task. However, It is clear that personal events, which constitute (39%) of the total events, achieve results very close to the results that have been obtained using all events. In other word, personal events present the most efficient events in user reviews and could be used to classify the sentiments of user reviews with good results.

To be able to meaningfully evaluate our model, we compared our approach with the systems which participated in the Polarity Detection task [1], the elementary task in the ESWC-14 challenge on Concept Level Sentiment Analysis. The reviews which used in this task were extracted from the Blitzer dataset [4]. To build our classifier for this task, we extract personal events with their participants from the training set which contain 8000 reviews (4000 positives and 4000 negatives). Then, we used them as features for a multinomial Naïve Bayes classifier.

Table 3 shows the results of our approach and the results of the top three participants in this challenge. The evaluation is carried out on the test, which is composed of 2429 sentences constructed in the same way and from the same sources as the Blitzer dataset. Our system using personal events achieved the second best performance on Recall and the third best system in F-measure.

6. CONCLUSION AND FUTURE PLAN

In this paper, we present an approach to detect and classify the polarity for customer reviews. We employ a machine reading system, which implemented in the FRED tool to extract features based-events and use them for a naïve Bayes classifier to predict the classes of reviews and study the correlation, which can be found between the reviews and theirs rating. We were capable to build a personal event dictionary, which can discriminate the two types of reviews. In addition, we were able to use personal events to correlate the reviews to their rating, confirming our initial hypothesis that some events have an influence on the rating scores given by users. In addition, we compared personal events with all events mentioned in reviews. We found that personal events are very important arguments in user reviews and very useful in the polarity classification task.

In the future, we plan to use our personal event dictionary to extract lived experience sentences from user reviews. We assume that personal events can be used to extract all user motivations, which are written in their reviews.

7. ACKNOWLEDGMENTS

This work is partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the program "Investissements d'Avenir" (reference: ANR-10-LABX-0083).

[4]http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

		Training Set			Test Set		
Features	*Nb_Features*	Precision	Recall	F-Measure	Precision	Recall	F-Measure
All_Events	621	70.5%	70.5%	70.5%	65.0%	65.0%	65.0%
$Event_\sigma = 0.7$	520	75.8%	75.5%	75.7%	65.7%	65.6%	65.7%
All_Event_Participant	2486	77.9%	77.4%	77.7%	76.8%	76.7%	76.7%
$(Event\text{-}Participant)\sigma = 0.7$	2058	84.1%	83.8%	84.0%	73.4%	73.3%	73.4%

Table 1: Overall results for review classification using lived experience features and NB method with four configurations. *Nb_Features*: **the number of features,** *All_Events*: **use all personal events as features,** $Event_\sigma = 0.7$: **delete some common events and use the rest as features,** *All_Event_Participant*: **use personal events with their participants as features, and** $(Event\text{-}Participant)\sigma = 0.7$: **remove the common events and participants, and use the rest as features.**

		Training Set			Test Set		
Features	*Nb_Features*	Precision	Recall	F-Measure	Precision	Recall	F-Measure
All_Events	1547	79.8%	79.8%	79.8%	78.3%	77.8%	78%
$Event_\sigma = 0.7$	1340	82.6%	82.6%	82.6%	75.5%	73.3%	74.4%
All_Event_Participant	3821	83.4%	83.1%	83.3%	82.8%	82.8%	82.8%
$(Event\text{-}Participant)\sigma = 0.7$	3197	88.4%	88.1%	88.3%	80.3%	79.4%	79.7%

Table 2: Overall results for review classification using event features and NB method with four configurations. *Nb_Features*: **the number of features,** *All_Events*: **use all events as features,** $Event_\sigma = 0.7$: **delete the common events and using the rest as features,** *All_Events_Participant*: **use events with their participants as features, and** $(Event\text{-}Participant)\sigma = 0.7$: **remove the common events and common participants, and use the rest as features.**

Participant	**Precision**	**Recall**	**F-Measure**	**Final position**
NCU	0.78	0.57	0.66	1
IBM	0.66	0.59	0.62	2
FBK	0.42	0.47	0.44	3
Event Approach	0.68	0.60	0.63	
Personal Event Approach	0.64	0.59	0.61	

Table 3: Results of Polarity Detection Task at ESWC2014

8. REFERENCES

[1] J. K.-C. Chung, C.-E. Wu, and R. T.-H. Tsai. Polarity detection of online reviews using sentiment concepts: Ncu iisr team at eswc-14 challenge on concept-level sentiment analysis. In *Semantic Web Evaluation Challenge*, pages 53–58. Springer, 2014.

[2] M. Eastmond. Stories as lived experience: Narratives in forced migration research. volume 20, pages 248–264, 2007.

[3] O. Etzioni, M. Banko, and M. Cafarella. Machine reading. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 2006.

[4] A. Gordon and R. Swanson. Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA*, 2009.

[5] A. S. Gordon. Story management technologies for organizational learning. In *International Conference on Knowledge Management. In Special Track on Intelligent Assistance for Self-Directed and Organizational Learning, Graz, Austria*, 2008.

[6] E. Hassan, D. Buscaldi, and A. Gangemi. Correlating open rating systems and event extraction from text. In *Neural Information Processing*, pages 367–375. Springer, 2015.

[7] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 309–319, 2011.

[8] T. Ploeger, M. Kruijt, L. Aroyo, F. de Bakker, I. Hellsten, A. Fokkens, J. Hoeksema, and S. ter Braake. Extracting activist events from news articles using existing nlp tools and services. In *Proceedings of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web*, page 30, 2013.

[9] V. Presutti, F. Draicchio, and A. Gangemi. Knowledge extraction based on discourse representation theory and linguistic frames. In *EKAW: Knowledge Engineering and Knowledge Management that matters*, 2012.

[10] A. Ritter, O. Etzioni, S. Clark, et al. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112, 2012.

[11] C. Treloar and T. Rhodes. The lived experience of hepatitis c and its treatment among injecting drug users: qualitative synthesis. volume 19, pages 1321–1334, 2009.

[12] G. Van Oorschot, M. Van Erp, and C. Dijkshoorn. Automatic extraction of soccer game events from twitter. page 15, 2012.