Web Intelligence and Agent Systems: An International Journal 9 (2011) 1–13 1
DOI 10.3233/WIA-2011-0209
IOS Press

Mining personal experiences and opinions from Web documents

Shuya Abe, Kentaro Inui *, Kazuo Hara, Hiraku Morita, Chitose Sao, Megumi Eguchi, Asuka Sumita, Koji Murakami and Suguru Matsuyoshi
Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0192, Japan
E-mail: {shuya-a,inui,kazuo-h,hiraku-m,chotose,megumi-e,asuka-s,kmurakami,matuyosi}@is.naist.jp

Abstract. This paper proposes a new UGC-oriented language technology application, which we call experience mining. Experience mining aims at automatically collecting instances of personal experiences as well as opinions from vast amounts of user generated content (UGC) such as weblog and forum posts and storing them in an experience database with semantically rich indices. After discussing the technical issues relating to this new task, we focus on the central problem of factuality analysis, formulate a task definition, and propose a machine learning-based solution. Our empirical evaluation indicates that our factuality analysis defintion is sufficiently well-defined to achieve a high inter-annotator agreement and our Factorial CRF-based model considerably outperforms the baseline. We also present an application system, which currently stores over 50M experience instances extracted from 150M Japanese blog posts with semantic indices and serves an experience search engine for unrestricted users and report on our empirical evaluation of the system's accuracy.

Keywords: Natural language processing, factuality analysis, experience mining, weblog, opinion mining

1. Introduction

The explosive spread of communication media on the Web, such as message forums and weblogs, allows Web users access to a rapidly increasing and massive amount personal experiences and opinions — a potential treasury of wisdom useful for making decisions, resolving troubles and avoiding problems, if only it were all indexed into well-organized user-friendly indices enabling users to easily find what they seek.

This potential is rapidly increasing interest in technologies to extract and analyze automatically personal opinions from such user generated content (UGCs) as customer reviews and weblog posts. Hence, a new field of natural language processing called sentiment analysis or opinion mining is appearing [4,8,9,14,19,20,29]. As indicated by the term *sentiment*, this trend of research has been focused on subjective statements such as *I like* and *is fabulous*.

Subjective information in sentiment analysis, however, is only half of the possible harvest from UGCs. UGCs contain not only subjective material but also a vast range of factual, objective statements describing such personal experiences as in (1).

(1) *On my way home, (in a wheelchair) I could not find my way out of Totsuka Station because all the elevators in the station building stop running at 11pm.*

Such information can indicate the concrete and objective reasons for sentiments or opinions, which are often crucial for decision making and problem solving.

In light of these newly emerging insights, we have been developing a language processing technology for fully automatic extraction of personal experiences as well as opinions from weblog and message forum posts, indexing them with semantically organized indices. In this paper, we use the term *experience* in a very broad sense that includes holding an opinion as well as hearing of an experience of others. So, to restate, our goal is to:

*Corresponding author.

a. collect personal experiences relevant to a broad range of topics including consumer products (automobiles, cellular phones, etc.), public places (tourist sites, hospitals, etc.), social systems (administrative services, welfare systems, etc.), and
b. store them all together in a large database, called an *experience database*, where each experience is represented as a piece of structured information comprising such slots as *topic*, *experiencer*, *event type*, *factuality* and *source pointer* as in (2) below.

(2) a. **Topic object**: What the experience is about (e.g. *Totsuka Station* in the case of example (1) above)
b. **Experiencer**: Who experiences (the *author* of the text)
c. **Event expression**: What event is experienced (*could find my way out*)
d. **Event type**: The semantic type (and sentiment orientation if applicable) of the experienced event (*could find my way out* is a positive/desirable happening)
e. **Factuality**: Whether the event indeed took place or not i.e. the temporal and modal status of the event (*I couldn't find my way out* is an *affirmatively negated past event*)
f. **Source pointer**: A pointer to the source text

The key idea is to index experiences not just by topic keywords and authorship but by a combination of semantic indices such as *event types* and *factuality*. The event types categorize the main predicate of an experience into semantic categories such as *buying*, *using* and *positive/negative happening*. The factuality slot specifies the temporal and modal status of the event referred to by the main predicate of an experience, which indicates, for example, whether the event did indeed take place in the past or is just a hypothetical situation. In the above example, the occurrence of a positive/desirable event is affirmatively negated, from which we can identify this experience as something undesirable, i.e. *trouble*.

Once available, a DB of this type offers a wide range of applications. Semantic indices such as event types and temporal and modal attributes allow retrieval of, for example, *troubles experienced using a particular consumer product* or *complaints and requests regarding a particular local welfare system*. Furthermore, experiences collected from weblog posts, where authorship is identifiable, can also be used to profile an author (the blogger) and enable retrieval of *authors* by such complex queries as *those who have not bought a particular product model while expressing interest in it* or *those who had been using a particular service regularly but have recently stopped using it*. Such retrieval possibilities turn the vast amount of UGCs into a valuable resource useful in evaluating public services and social systems as well as for corporate marketing and risk management (Fig. 1).

In this paper, we call this new UGC-oriented language technology application *experience mining* and discuss the technical issues relating to this new task (Section 3). Then we focus on the central problem of factuality analysis, formulate a task definition, and propose a machine learning-based solution (Section 4). Our empirical evaluation indicates that our factuality analysis defintion is sufficiently well-defined to achieve a high inter-annotator agreement and our Fac-

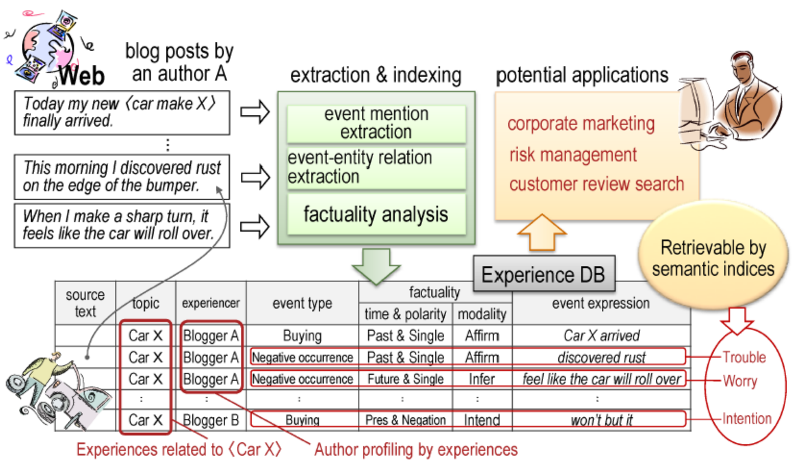

Fig. 1. An overview of experience mining.

torial CRF-based model considerably outperforms the baseline. We also present an application system, which currently stores over 50M experience instances extracted from 150M Japanese blog posts with semantic indices and serves an experience search engine for unrestricted users and report on our empirical evaluation of the system's accuracy.

2. Related work

As stated in the previous section, experience mining is motivated to be an extension of opinion mining. Opinion mining has so far tended to aim at extracting sentiment information mainly from explicit evaluative or emotional expressions such as *useful* (positive) or *disturbing* (negative) [2,3,5,10,13,15,25,30,31]. On the other hand, experience mining covers all the descriptions of events that are related to any use of a wide variety of topic objects including objective descriptions (i.e. facts) with *implicit sentiment* such as *My son passed the exam* or *I discovered rust on the edge of the bumper*.

The task of extracting experiences we consider here is related also to such template-based information (event) extraction as the one driven by the MUC[1] and ACE[2] research funding programs. For example, extraction of event descriptions of a given set of event types and subtypes in the ACE task bears some resemblance to our task in the sense that both aim at extracting event instances from a document collection and structuring them with semantic index labels. What we present in this paper, however, differs from such conventional template-based event extraction in the following two respects.

First, while conventional information extraction tasks are defined on the basis of domain-specific event/relation templates, our task setting is highly domain-independent and our system works for open domains. In this sense, our task may seem closely related also to recently emerging work on open-domain information extraction [1,23,24]. This work, however, primarily considers named entities and heads of proper noun phrases rather than event expressions and the relations extracted are those commonly held between NPs (e.g. city-of-state) rather than a more general relevance relation between a topic and event.

Second, extraction of a wide variety of events motivates us to explore fine-grained analysis of temporal and modal attributes of each event description, which has attracted little attention in the opinion mining or information extraction literature. For example, in the ACE (Automatic Content Extraction) research program, each event mention is supposed to be annotated with temporal and modal markers as in (3).

(3) a. **TENSE**: *Past*, *Present*, *Future*, *Unspecified*
b. **POLARITY**: *Positive*, *Negative*
c. **MODALITY**: *Asserted*, *Other*

This markup scheme, however, is too simple for our purpose. For example, ACE has only two labels for modality, Asserted and Others, while we need finer-grained distinct labels, as described below.

Another effort we should refer to is TimeML [21], a specification language for events and temporal expressions, which annotates event mentions with tense, aspect, polarity and modality information as in (4).

(4) a. **TENSE**: *Past*, *Present*, *Future*, *None*, *Infinitive*, *Present-Perfect*, *Past-Perfect*
b. **ASPECT**: *Progressive*, *Perfective*, *Perfective-Progressive*, *None*, *Initiation*, *Culmination*, *Termination*, *Continuation*, *Reinitiation*
c. **POLARITY**: *Positive*, *Negative*
d. **MODALITY**: *must*, *may*, *should*, *would*, *could*
e. **S-LINK**: *Modal*, *Factive*, *Counter-factive*, *Evidential*, *Negative-evidential*, *Conditional*

While the labels are more fine-grained than those of ACE, the markup scheme of TimeML is, however, highly dependent on the syntax of the target language (currently only English and Chinese are supported) and, more importantly, is too shallow to capture such factuality information as we require. In fact, researchers engaged in the TimeML project are currently developing a more semantic-oriented level of representation of factuality for the purpose of reasoning textual entailment [22]. This work in an extension of [11].

3. Technical challenges

Our task can be decomposed into the following sequence of subtasks:

[1] Message Understanding Conference
http://www-nlpir.nist.gov/related_projects/muc/
[2] Automatic Content Extraction
http://www.nist.gov/speech/tests/ace/

- **Event mention extraction**: Given an input text, we first identify event mentions which may constitute an experience by simple dictionary look-up. For this purpose, we build a typologized lexicon of expressions of experiences as we briefly describe below. When a text is given, (1) for example, we identify *can find (my way out)* as a positive/desirable state and *(elevators) stop running* as a negative/undesirable happening.
- **Entity-event relation extraction**: For each identified event mention, we next seek from the local context an entity about which the event is an experience (i.e., *can be interpreted* as an experience). *could not find* in (1), for example, can be considered to be an experience about *Totsuka Station* but not *home*.
- **Factuality analysis**: If any appropriate entity mention is found, we then carry out factuality analysis to identify the factuality status of the event. By doing this, we can distinguish, for example, between events which actually took place and those merely surmised or desired by the author.
- **Experiencer identification**: Finally, we identify the experiencer of each experience.

Each of the steps represents an interesting technical challenge. Entity-event relation extraction and experiencer identification have already been addressed in the context of opinion mining [3,13,15,25, etc.]. Entity-event relation extraction is the task of identifying the relation instances between an evaluative expression and its subject in opinion mining whereas experiencer identification can be taken as an extension of the task of identifying opinion holders. The other two steps, on the other hand, involve new challenges so far given paid little attention in opinion mining.

One major issue in event mention extraction is how to create a lexicon of event expressions with a sufficiently broad coverage. For the event typology, we currently assume that the following distinctions are useful for characterizing experiences:

(5) a. **Sentiment**: Predicative expressions of an emotion or subjective evaluation. Each has a sentiment orientation (i.e. *positive* or *negative*).

   - **Emotion**: *enjoy*, *disappointed*
   - **Evaluation**: *tasty*, *inconvenient*
   - **Reputation**: *popular*, *criticised*

b. **Happening**: Predicative expressions referring to a non-volitional event or state which is related to the use of a topic object and has a sentiment orientation

   - **General**: *pass (an exam)*, *get slim*, *do (something) on time*, *cheated*, *broken*, *run out*
   - **Availability**: *released*, *(system) go into effect*
   - **Usability**: *get used to*, *prohibited*

c. **Action**: Predicative expressions referring to experiencers' volitional actions related to the use of a topic object. Sentiment orientations are not necessarily involved.

   - **Buying/Selecting**: *buy*, *get*, *apply to (a social system)*, *choose*
   - **Using**: *use*, *drive (a car)*
   - **Stopping**: *cancel*

Expressions of Emotion, Evaluation and Reputation can be largely imported from existing sentiment lexicons such as SentiWordNet for English and Kobayashi's sentiment lexicon [15] for Japanese. For Action expressions, on the other hand, our preliminary exploration into weblog posts reveals that most expressions can be covered by a relatively small list of predicates. To obtain those predicates, WordNet-like general purpose thesauri can be employed. In contrast to the above two classes, collecting Happening expressions with a sentiment orientation is a new challenge given their wide variety. To this challenge, we approach by exploring a method of combining large-scale acquisition of sentiment-bearing expressions from a Web corpus and pattern-based composition of acquired expressions. As a result, we have so far obtained over 50M sentiment-bearing experience/event expressions at an affordable cost for manual cleaning and fed them to our demonstrative experience mining system presented in Section 5. The details of the acquisition method is out of scope of this paper as it will be presented elsewhere in a paper under submission.

The last, but very important, subtask of experience mining is factuality analysis. We believe this task could serve as an important semantic component across a wide range of language technology applications. However, it has so far attracted surprisingly little attention in the literature. One major technical contribution of our present work is that we designed the task and gave a machine learning-based solution to it as we describe in the next section.

4. Factuality analysis

4.1. Aims and background

For each event mention, we want to identify the temporal and modal status of the event entity referred to by the event mention. Namely, we want to know, for example:

- whether the event indeed took place, is intended to take place, or just hypothetical,
- whether the happening of the event is desired by the author or not, and
- whether the event is a single event, a series of repeated events, or a state.

To this end, one might consider adopting a highly formal representation like temporal logic. However, introducing such a logic-based representation would require extremely sophisticated language understating and the state-of-the-art technology has not reached that level.

4.2. Factuality markup scheme

Given this context, we have created a new markup scheme for annotating event mentions with factuality information. We annotate each event mention in a given text with a triplet ⟨ *Event-time, Modality, Modality-time* ⟩.

The *Event-time* slot represents the tense, aspect and polarity status of the event in question, consisting of three sub-slots *Past-Present-Future*. Each sub-slot is to be filled with one of the following ASPECT-POLARITY labels, denoting the aspect and polarity (negation) information:

(6) **ASPECT-POLARITY**: *Punctual* (*Pnc*), *State-Continuation* (*StC*), *Repetition* (*Rpt*), *Initiation* (*Int*), *Termination* (*Trm*), *Negation* (*Ngt*), *Uncommitted* (*Unc*)

where all but *Negation* and *Uncommitted* implicitly denote *Positive* in terms of polarity. *Uncommitted* denotes that the author does not say anything about whether the event takes place in the corresponding slot of time. An example is given in (7), where the *Event-time* of the event mention *using* is annotated with *Int-Rpt-Unc*.

(7) a. *I started using FireFox recently.*
b. ⟨ *Int-Rpt-Unc, Affirm, Unc-Pnc-Unc* ⟩

In experience mining, it is often meaningful to distinguish between repeatedly happening events and single punctual events. For example, corporate marketers may seek customers who use their product repeatedly; and troubles which recur may well be more serious than single occurrence. It is also important to capture the initiation and termination of a repetitive or continuous event, for this Will enable a search, for example, *those who recently stopped using a particular social welfare system*.

The *Modality* slot specifies the author's mental or communicative attitude toward the event in question. As a set of possible values of this slot, we have so far identified the following classes based on several reference books on Japanese modality [17, etc.]:

(8) **MODALITY**: *Affirm, Infer, Doubt, Hear, Intend, Ask, Recommend, Hypothesize, Other*

For example, while the *Modality* of the event *Using* in (7a) is *Affirm*, the *Modality* of the event *possess* in the next sentence (9a) is interpreted as *Hear*.

(9) a. *I watched a TV program reporting isoflavone-rich foods possessed activity against cancer.*
b. ⟨ *Unc-StC-Unc, Hear, Pnc-Unc-Unc* ⟩

An important point to note here is that unlike the modality labels defined in TimeML (see 3.1 above), our modality labels are defined at the semantic level. More specifically, in TimeML, each modality label simply corresponds to an auxiliary verb and each S-LINK label is also strictly associated with a small set of modality verbs; for example, Factive is associated with verbs such as *forget* and *regret*. However, on the other hand, what we want to do in factuality analysis is to identity the temporal and modal status of each event mention at a semantic level. For example, in Japanese, a modality value *Doubt* may be linguistically realized by such a verb as *utagau* (*doubt*) or an interrogative particle *ka*. There is also a range of adverbs and adverbial functional expressions that can be used to express a doubt. Some of them are highly context-dependent and are thus apparently ambiguous. To make a factuality analysis component applicable to experience mining, we need to handle these phenomena.

4.3. Training factuality analysis models

To automate the above factuality analysis task, we created a manually annotated corpus and trained a statistical machine learning-based model.

Table 1

The results of the experiments (label accuracy)

Model	Domain	Past	Pres	Fut	Mod
Baseline	all	.61	.61	.76	.66
SVM	beverage	.49	.52	.72	.82
SVM	automobile	.38	.48	.74	.84
SVM	shampoo	.53	.63	.80	.84
Fact. CRF	beverage	.66	.61	.90	.83
Fact. CRF	automobile	.75	.59	.88	.85
Fact. CRF	shampoo	.68	.58	.90	.85

To create an annotated corpus, we first randomly sampled from our weblog corpus (see Section 5.1) sentences including any one of the three chosen topic keywords (beverage name, automobile name, shampoo name). We then asked two annotators to annotate with factuality tuples all the event mentions included in the sampled sentences. After rehearsing several times, the annotators came to exhibit a remarkable agreement on unseen data — the κ statistics citekappa was 0.68, where they were considered to agree for an event mention only if *all its slots agreed*. This figure indicates that our annotation scheme is reliable enough. We then re-sampled sentences for the same topic keywords, obtaining 2,646 sentences in total, and asked one of the above two annotators to annotate all 4,417 event mentions included in the obtained sentences.

As easily imagined, the distribution of the value of each slot is highly skewed. Therefore, a simple baseline is given by choosing the most common values for each slot (*Unc* for all the three sub-slots of *Event-time* and *Assert* for *Modality*). The results are shown in the baseline of Table 1. The *Modality-time* slot was neglected in the experiment because its value was *Unc-Pnc-Unc* (i.e. the present tense) over 95 percent of the time.

Our task is now restated as one of determining the values of the four slots $\langle\, Et_1{-}Et_2{-}Et_3,\ Mdl\,\rangle$. We have so far examined two machine learning models.

First, the three *Event-time* sub-slots Et_1, $-Et_2$ and $-Et_3$ may well be highly dependent on their neighbors. We therefore employed the SVM-HMM algorithm [28] to train an *Event-time* model so that it could optimize the labels of those three slots simultaneously and we used the SVM-Multiclass package [28] to train a *Modality* model, which took care of the *Modality* slot independently of the *Event-time* slots.

The second model we examined is more sophisticated. Besides the inter-dependency between the

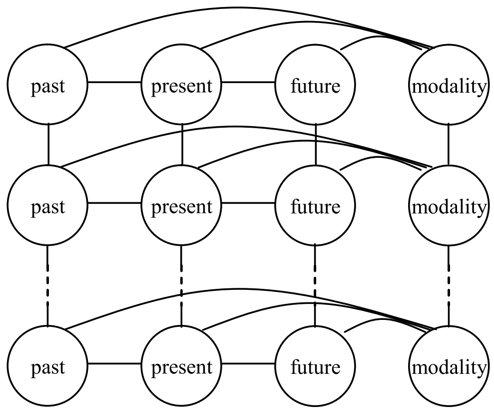

Fig. 2. A graphical model representing the interdependencies between the factuality labels of neighboring event mentions.

Event-time slots, each slot may well be dependent also on the *Modality* slot. Furthermore, the factuality of an event mention is also likely to interact with that of any neighboring event mentions appearing in the same sentence. Such interdependencies led us to consider the graphical model illustrated in Fig. 2. To train this mesh model, we employed the state-of-the-art GRMM toolkit, designed for the paradigm of conditional likelihood maximization [27]. This toolkit can deal with graph structures which include loops as in Fig. 2. Training this type of mesh model on the basis of conditional likelihood maximization is also called *Factorial CRFs (Conditional Random Fields)*.

The feature set we used for both models included bag-of-words features with part-of-speech tags and lexemes extracted from neighboring base-NP/VP phrases and from the head phrase of the sentence.

4.4. Empirical evaluation

Finally, we conducted a three-fold cross-validation using our annotated corpus, where, for each fold, a model was trained on the data of two of the three domains (beverage, automobile, shampoo) and tested on the third domain. The results are shown in Table 1.

The tendency we observe from these figures is clear. First, the SVM-based model did not particularly outperform the baseline. This indicates the difficulty of the task, which is partly due to the skewness of the labels (i.e., the baseline is already quite high). Second, on the other hand, the Factorial CRF-based model substantially improved the accuracy for all the slots, which shows the importance of considering the interdependency between neighboring labels in this task.

Our error analysis revealed considerable room for improvement. Concretely, feature engineering is expected to be of great help — the present bag-of-words-based features set is doubtlessly too simple to represent complex combinations of Japanese auxiliary verbs and particles. While our factorial CRF-based model worked well across domains, for practical use, it would also be effective to extend the training data to a wide variety of other domains. We are planning to employ an active learning schema for efficient collection of informative training data.

5. An application system

Employing these components just described, we have developed an application system, and evaluated its overall performance.

5.1. System overview

The system is designed for users of topic objects (consumer products, public places, social systems, etc.). Given one or more topic objects specified by a user of the system, the system provides the user with facilities for browsing bloggers' experiences related to those topic objects. Each experience instance is automatically classified into about ten *experience classes*. Each experience class is defined in terms of event types and factuality labels. For example, the experience class *Experienced troubles* is defined as a small number of combinations of event types and factuality labels including:

(10) a. *negative happening* and ⟨*Pnc-Unc-Unc, Affirm, Unc-Pnc-Unc*⟩ and
  b. *positive happening* and ⟨*Trm-Ngt-Unc, Affirm, Unc-Pnc-Unc*⟩.

By this classification, a user of the system can restrict a search to, for example, only *troubles experienced by the users of a specific topic object in question*.

To build the system, we first collected recent 18-month worth of Japanese weblog posts, which amounted to about 150M posts. We next collected a set of potential topic objects from Wikipedia[3]. From the categories under the *technology*, *culture* and *society* super categories in Wikipedia, we obtained about 200K keywords (i.e. topic objects) each corresponding to a Wikipedia article.

[3] http://ja.wikipedia.org/

We should employed a fast parser in order to handled the large amount of corpus. Therefore, we replaced *Factrial CRFs* with *SVM* in the factuality analysis component. The replacement increased the speed of parser, however the precision was declined slightly.

The lexicon of experience/event expressions was built in a way as follows:

- Expressions of Emotion, Evaluation and Reputation were imported from Kobayashi's sentiment lexicon [15].
- Action expressions were collected from an existing general purpose thesaurus, *Bunruigoihyo* [6].
- For Happening expressions, a newly devised knowledge acquisition method, which is going to be present elsewhere in a paper under submission, was first used to obtain about 25K candidate (compound) nouns with positive sentiment economic recovery) and 10K candidate (compound) nouns with negative sentiment from a large-scale treebank of Web documents [12]. Here, positive nouns are those which are commonly desired to appear, increase, or take place (e.g. *profit, motivation, economic recovery*), while negative nouns are those which are commonly undesired to appear, increase or take place (e.g. *wrinkle, spam, domestic violence*). The obtained candidate nouns were then cleaned manually, which filtered out about 20% of the candidates. Then each remaining noun was combined to each from two distinctive sets of predicative expressions:

  * *Increasing verbs*: verbs and adjectives meaning *to exist, to appear, to increase, to strengthen, to take place, to continue, to see, to gain*, etc., and
  * *Decreasing verbs*: verbs and adjectives meaning *not to exist, to disappear, to decrease, to weaken, to stop, to loose*, etc.

  The sentiment orientation of each combined expression can be calculated based on a small set of simple composition patterns. For example, combining a positive noun with an increasing verb generates a positive event expression (e.g. *get a profit*), while combining with a decreasing verb generates a negative event expression (e.g. *loose motivation*). Filtering out meaningless combinations based on their frequency counts in our corpus, we finally obtained over 550K event expressions with a sentiment orientation.

Table 2

Examples of extracted experiences from blog entries

Blog url	Entry url	Sentence	Topic	Event expression	Event type	Time	Polarity	Modality
http://.../hai	.../080110	My friend said that when he used the ABC Spray his hair became **difficult to get messed up**.	ABC Spray	difficult to get messed up	Positive	Past/Present	non-Negation	Hear
http://.../bal	.../080110	We could not **win** at volleyball even once.	volleyball	win	Positive	Present	Negation	Affirm
http://.../atm	.../080113	At the end of the month I waited in line at a **crowded** ATM of ABC bank.	ATM of ABC bank	crowded	Negative	Present	non-Negation	Affirm
http://.../tmt	.../080112	I **drink** tomato juice every night.	tomato juice	drink	Using	Past/Present	non-Negation	Affirm
http://.../bkk	.../080112	I **eat too much** at yesterday's barbeque.	barbeque	eat too much	Using	Past/Present	non-Negation	Affirm
http://.../car	.../080113	This morning I **discovered rust** on my <CAR ABC>.	<CAR ABC>	discovered rust	Negative	Past/Present	non-Negation	Affirm
http://.../mac	.../070609	I will **buy** the iPhone on launch day.	iPhone	buy	Buying	Future	non-Negation	Intend
http://.../mac	.../070609	I think that the iPhone will be hardest hit.	iPhone	be hardest hit	Positive	Future	non-Negation	Infer
http://.../mac	.../070611	I went to the shop, but I could not **got** the iPhone.	iPhone	get	Buying	Past/Present	Negation	Affirm
http://.../mac	.../070620	At last, I **buy** the iPhone.	iPhone	buy	Buying	Past/Present	non-Negation	Affirm
http://.../mac	.../070620	The iPhone **is good**.	iPhone	is good	Positive	Past/Present	non-Negation	Affirm

The whole lexicon is available from our Web site[4].

For identifying topic-experience relations (i.e. the task of entity-event relation extraction described in Section 3), we devised proximity-based heuristic rules. Namely, we extracted only experience/event expressions that met the all the following conditions:

- The experience expression must appear in the same sentence as the one where the corresponding topic word appears.
- The experience expression must appear in the subtree (i.e. a descendant position) of the corresponding topic word in a dependency parse tree.
- The number of the base phrases (i.e. so called *bunsetsu* phrases in Japanese) intervening the topic word and experience expression mush be smaller than eight.

Obviously there is much room for refinement. We believe we must eventually incorporate state-of-the-art technologies of, for example, ellipsis and coreference resolution into our system. This issue is definitely included in our future directions.

[4]http://cl.naist.jp/~inui/research/EM/sentiment-lexicon.html

We next automatically extracted sentence-chunked texts from the weblog post set, and conducted tokenization and POS tagging with ChaSen[5] and dependency parsing with CaboCha[6]. We then carried out experience mining on the parsed texts and obtained over 50M experience instances related to one of our keywords and stored all of them in a relational database (Table 2).

Figure 3 shows a snapshot of the system's view, where a summary of the search results for a query *Dogo Onsen* (hot spring), *Ikaho Onsen* and *Kurokawa Onsen*, is presented. For each given topic object, the system presents the number of bloggers who have described one or more experiences related to that topic object, where the bloggers are counted separately for each experience class. Furthermore, for each experience class, several major experience expressions are presented. Given this view, the user can overview the reported experiences for each topic object and compare them across different topic objects.

By clicking one of the experience classes (e.g. *trouble*), the user is led to another view, as shown in Fig. 4,

[5]http://chasen-legacy.sourceforge.jp/

[6]http://chasen.org/~taku/software/cabocha/

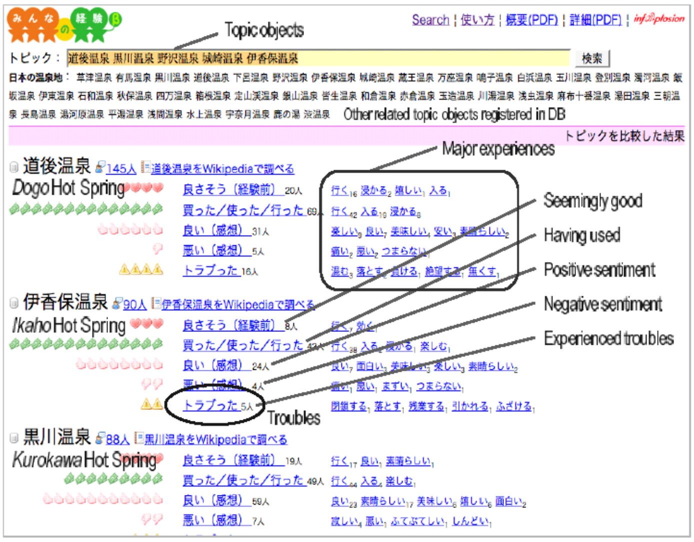

Fig. 3. A snapshot of the summary view of the experience search engine.

Fig. 4. A snapshot of the experience view showing troubles experienced while using *iPod*.

and can browse there all the mentions about troubles experienced during the use of the topic object. In addition, by clicking one of the link of the blogger, the user is led to another view, as shown in Fig. 5, and can browse there all the mentions about the history of the blogger during the use of the topic object. Characteristic of our system is that it presents experience-mentions blogger by blogger and ranks bloggers according to the number of their experience mentions about the queried topic object. We are assuming that the more experienced a person is with a given topic, the more he/she knows about it and the more important his/her mentions about it are. Based on this assumption, the system also allows a user to browse a blogger's experiences with a topic object in chronological order, possibly a clue regarding the blogger's background (expert, confederate, etc.).

A demonstration site was released to unrestricted users at our Web site[7] in December 2008.

We also extended the above system by enhancing the user interface as shown in Fig. 6, which demonstrates how a user can specify complex queries comprising event type and factuality configurations. Collaborating with a major Internet service provider (the leading UGC-based marketing research business in Japan), we designed the user interface and defined the

[7] http://minna.naist.jp/

Fig. 5. A snapshot of the blogger view showing a single author's history of experiences with *electric piano*.

Fig. 6. A snapshot of the system view customized for corporate marketing research.

default set of experience classes based on a marketing theory [7,26]: *Attention*, *Interest*, *Desire*, *Experience*, *Enthusiasm*, and *Share*. Those classes were straightforwardly defined in terms of our event types and factuality labels. Using those notions, a user of the system can seek, for example, *those who have not bought a particular product model while expressing interest in it* or *those who had been using a particular service regularly but recently stopped using it*.

5.2. Evaluation for an experience search

To evaluate the overall performance of our system, we first created gold-standard date set in the following steps so that we could estimate the system's recall as well as the precision.

We first randomly chose 80 topic keywords from our experience database and manually filtered out over-generic words from them, obtaining 53 words. For each of the 53 keyword, we randomly sampled two or three blog posts including it from our 150M-post data set and manually filtered out those suspected to be a spam from them, obtaining 86 posts (3154 sentences in total). As a result, each post was associated with one of the 53 keywords, which we call the *document topic* of the post in question. Finally, for each post, we manually identified as many experience instances related to its document topic as possible. All the tasks were

Table 3
The precision/accuracy of the system's output

(a) Precision for topic-experience relation extraction	0.76
(b) Accuracy of event type classification	0.96
(c) Accuracy of factuality analysis (Polarity)	0.92
(d) Accuracy of factuality analysis (Modality)	0.81

done by a linguist who was familiar with our event typology and factuality labels but was not involved in the development of our experience mining system.

The precision/accuracy of the system's output is summarized in Table 3. The *precision for topic-experience relation extraction* (a) shows how many of the experience instances identified by the system as one related to the document topic were indeed related to that topic. Since we simply devised several heuristic rules for this subtask in the current implementation as mentioned in Section 5.1, there is still much room for improvement, which we consider as one of the important issues we should address in our future work. The accuracy figures in (b), (c) and (d) are calculated only for the experience instances whose topic-experience relation was judged correct (i.e. 76% of all the system's outputs). These figures indicate that our dictionary lookup-based event type classifier and factuality analysis component both worked reasonably well. The accuracy of modality classification is slightly lower than those shown in Table 1. We consider this is within a reasonable deviation given that we used the multi-class SVM model instead of the factorial CRF model in this experience.

To measure inter-annotator agreement, another annotator evaluated 313 examples that were judged to satisfy criterion (a). In other words, the annotator does not evaluate examples judged to be labeled as non-experiences. The other judged examples using criteria (b), (c) and (d). The κ statistics of criteria (b), (c) and (d) were 0.91, 0.83 and 0.71, respectively. There are high level of agreement between two annotators. The reason for the high level of agreement is that criteria (b), (c) and (d) are binary classifications, and also natural criteria for a human. Annotators consistently gave the same evaluation. As a result, two annotators could easily evaluate perfectly using the criteria.

Furthermore, we also need to consider the coverage of our event/experience lexicon. Our gold-standard data contained 1,605 experience instances but only 45% of them were actually covered by our event/experience lexicon. Our error analysis revealed that we still needed to scale-up the semi-automatic acquisition of sentiment-bearing words, while devising a robust mechanism for open-domain word-sense disambiguation so as to maintain the precision.

6. Conclusion

In this paper we have proposed a new UGC-oriented language technology application called experience mining. Experience mining aims at automatically collecting instances of personal experiences as well as opinions from an explosive number of UGCs such as weblog and forum posts and storing them in an experience database with semantically rich indices. Experience mining can be regarded as a substantial extension of opinion mining. Opinion mining has so far tended to aim at extracting sentiment information mainly from explicit evaluative or emotional expressions such as *useful* (positive) or *disturbing* (negative) [3,5,15, etc.]. On the other hand, experience mining covers all the descriptions of events that are related to any use of a wide variety of topic objects including so-called *implicit evaluative descriptions*.

We have also argued the technical issues of this new task. Focusing on factuality analysis, we have designed the task anew and given a machine learning-based solution to it. Our empirical evaluation indicates that the task is sufficiently well-defined to achieve a high inter-annotator agreement, and our factorial CRF-based model considerably outperforms the baseline. Furthermore, our technology will also benefit other types of applications. In the biomedical domain, for example, recognizing the factuality of each event mentioned in research papers is crucial, though very few researchers have addressed this issue [16,18,32].

We have also presented an application system, which currently stores over 50M experience instances with semantic indices — published an experience search engine for unrestricted users. Although we empirically evaluated the factuality analysis component, the experience search system as a whole is still to be evaluated from various angles, such as accuracy, utility and usability. An extrinsic evaluation of the whole system is included in our future work.

Our application system employed the SVM-based model instead of the Factrial CRF-based model for improving scalability[8]. The SVM-based model is scalable. On the other hand, the Factorial CRF-based

[8]We compared elapsed time between our SVM-based model and Factorial CRF-based one. The SVM-based model was faster than the Factorial CRF-based one by one to three magnitudes.

model does not scale. However it outperforms the SVM-based model. It is clearly suitable to combine the good scalability of the SVM-based model and the performance of the Factorial CRF-based model for an application system. We therefore suggest that a system employ the SVM-based model for building an experience database from weblogs and creating summaries of experiences, and employ the Factorial CRF-based model for showing the details of an experience to a user.

Acknowledgements

This work was partly supported by Japan MEXT Grant-in-Aid for Scientific Research on Priority Areas, Cyber Infrastructure for the Information-explosion Era (No. 19024033), and by Japan National Institute of Information and Communications Technology.

References

[1] M. Banko and O. Etzioni, The tradeoffs between open and traditional relation extraction, in: *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics (ACL) with the Human Language Technology Conference (HLT) of the North American Chapter of the ACL (ACL-08: HLT)*, Association for Computational Linguistics, June 2008, pp. 28–36.

[2] E. Breck, Y. Choi and C. Cardie, Identifying expressions of opinion in context, in: *Proc. of the 20th International Joint Conference on Artificial Intelligence (IJCAI-2007)*, Morgan Kaufmann Publishers Inc., 2007, pp. 2683–2688.

[3] Y. Choi, E. Breck and C. Cardie, Joint extraction of entities and relations for opinion recognition, in: *Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, July 2006, pp. 431–439.

[4] K. Dave, S. Lawrence and D.M. Pennock, Mining the peanut gallery: opinion extraction and semantic classification of product reviews, in: *Proc. of the 12th International World Wide Web Conference* (*WWW*), ACM, 2003, pp. 519–528.

[5] A. Esuli and F. Sebastiani, Determining term subjectivity and term orientation for opinion mining, in: *Proc. of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, The Association for Computer Linguistics, 2006, pp. 193–200.

[6] The National Institute for Japanese Language, *Bunrui Goi Hyo*, Dainihon Shuppan, 1964, (in Japanese).

[7] H.M. Goldman, *How to Win Customers*, Pan Books, 1958.

[8] K. Hiroshi, N. Tetsuya and W. Hideo, Deeper sentiment analysis using machine translation technology, in: *Proc. of the 20th International Conference on Computational Linguistics (COLING)*, Association for Computational Linguistics, 2004, pp. 494–500.

[9] M. Hu and B. Liu, Mining and summarizing customer reviews, in: *Proc. of the 10th International Conference on Knowledge Discovery and Data Mining* (*KDD*), ACM, 2004, pp. 168–177.

[10] M. Hu and B. Liu, Mining opinion features in customer reviews, in: *Proc. of the 19th National Conference on Artificial Intelligence* (*AAAI*), AAAI Press, 2004, pp. 755–760.

[11] K. Inui, S. Abe, H. Morita, M. Eguchi, A. Sumida, C. Sao, K. Hara, K. Murakami and S. Matsuyoshi, Experience mining: Building a large-scale database of personal experiences and opinions from web documents, in: *Proc. of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, IEEE Computer Society, 2008, pp. 314–321.

[12] D. Kawahara and S. Kurohashi, A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis, in: *Proc. of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 2006, pp. 176–183.

[13] S.-M. Kim and E. Hovy, Extracting opinions, opinion holders, and topics expressed in online news media text, in: *Proc. of the Workshop on Sentiment and Subjectivity in Text*, Association for Computational Linguistics, July 2006, pp. 1–8.

[14] N. Kobayashi, R. Iida, K. Inui and Y. Matsumoto, Opinion extraction using a learning-based anaphora resolution technique, in: *Proc. of the 2nd International Joint Conference on Natural Language Processing (IJCNLP)*, 2005, pp. 175–180.

[15] N. Kobayashi, K. Inui and Y. Matsumoto, Opinion mining from web documents: Extraction and structurization, *Journal of the Japanese Society for Artificial Intelligence* **22**(2) (2007), 227–238.

[16] M. Light, X.Y. Qiu and P. Srinivasan, The language of bioscience: Facts, speculations, and statements in between, in: *Proc. of the HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*, Association for Computational Linguistics, May 6 2004, pp. 17–24.

[17] T. Masuoka and Y. Takubo, *Fundamental Japanese Grammar*, Kuroshio, 1992, (in Japanese).

[18] B. Medlock and T. Briscoe, Weakly supervised learning for hedge classification in scientific literature, in: *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, Association for Computational Linguistics, June 2007, pp. 992–999.

[19] B. Pang and L. Lee, A sentiment education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics* (*ACL*) , Association for Computational Linguistics, 2004, pp. 271–278.

[20] A.-M. Popescu and O. Etzioni, Extracting product features and opinions from reviews, in: *Proc. of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2005, pp. 339–346.

[21] J. Pustejovsky, J.M. Castano, R. Ingria, R. Sauri, R.J. Gaizauskas, A. Setzer, G. Katz and D.R. Radev, Timeml: Robust specification of event and temporal expressions in text, in: M.T. Maybury, (ed), *New Directions in Question Answering*, AAAI Press, 2003, pp. 28–34.

[22] R. Sauri and J. Pustejovsky, Determining modality and factuality for text entailment, in: *Proc. of the International Con-*

ference on Semantic Computing (ICSC 2007), IEEE Computer Society, 2007, pp. 509–516.

[23] S. Sekine, On-demand information extraction, in: *Proc. of the COLING/ACL on Main Conference poster sessions*, Association for Computational Linguistics, 2006, pp. 731–738.

[24] Y. Shinyama and S. Sekine, Preemptive information extraction using unrestricted relation discovery, in: *Proc. of the Human Language Technology Conference of the NAACL, Main Conference*, Association for Computational Linguistics, 2006, pp. 304–311.

[25] V. Stoyanov and C. Cardie, Topic identification for fine-grained opinion analysis, in: *Proc. of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Coling 2008 Organizing Committee, 2008, pp. 817–824.

[26] E.K. Strong, Theories of selling, *Journal of Applied Psychology* **9** (1925), 75–86.

[27] C. Sutton, Grmm: A graphical models toolkit, http://mallet.cs.umass.edu, 2006.

[28] I. Tsochantaridis, T. Joachims, T. Hofmann and Y. Altun, Large margin methods for structured and interdependent output variables, *Journal of Machine Learning Research (JMLR)* **6** (2005), 1453–1484.

[29] P.D. Turney, Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews, in: *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics* (*ACL*), Association for Computational Linguistics, 2002, pp. 417–424.

[30] J. Wiebe, T. Wilson and C. Cardie, Annotating expressions of opinions and emotions in language, *Language Resources and Evaluation* **39**(2–3) (2005), 165–210.

[31] T. Wilson, J. Wiebe and P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: *Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Association for Computational Linguistics, 2005, pp. 347–354.

[32] L. Zhou, G.B. Melton, S. Parsons and G. Hripcsak, A temporal constraint structure for extracting temporal information from clinical narrative, *Journal of Biomedical Informatics* **39**(4) (2006), 424–439.