# Web Reviews and Events Matching Based on Event Feature Segments and Semi-Markov Conditional Random Fields

Yuanzi Xu
School of Computer Science and Technology, Shandong University, Jinan, China
xu_yuanzi@163.com

Qingzhong Li, Zhongmin Yan and Wei Wang
School of Computer Science and Technology, Shandong University, Jinan, China
lqz@sdu.edu.cn

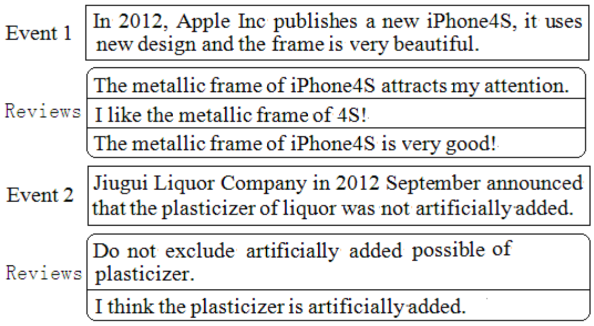
***Abstract*—To establish links between a large number of reviews and events, we propose a web reviews and events matching approach by event feature segments and semi-Markov conditional random fields (CRFs). We extract named entities and verb phrases from reviews as event feature segments. We use semi-Markov CRFs to label the reviews and to recognize event feature segments at the segment level. This approach uses event feature segments to match reviews and events. Therefore, it is more accurate than other approaches which use only named entities to match. We use several feature rules to recognize the variants of named entities, such as abbreviation and acronym. In addition, we use phrase dependency parsing tree to recognize verb phrases. A compositive similarity measurement function is presented to combine similarity results of event feature segments. Experimental results demonstrate that this method can accurately match reviews and events.**

***Index Terms*—event feature segments, semi-Markov CRFs, feature rules, phrase dependency parsing tree**

## I. Introduction

Millions of people publish reviews via microblogs and forums, thereby providing useful information about web events. An event is an activity that occurs at a special time and involves participants. For example, product reviews provide vast important information for enterprise policymakers. Matching online reviews and events is an effective way to track products. Online reviews have the characteristics of free description and publication. Thus, matching them is difficult. However, artificial intelligence and machine learning present an opportunity to match reviews and events automatically.

This study aims to automatically match web reviews and events. In the web, a news report contains many independent events, and a news headline is the generalization of multiple independent sentence-level events [1]. In this study, events from news are extracted, and a sentence is considered an event. Some reviews which match associated sentence-level events are showed in Fig. 1. The first event in Fig. 1 reports a new phone called iPhone4S and emphasizes that its frame is very beautiful. Three reviews are associated with the metallic frame of iPhone4S. The second event reports that Jiugui Liquor Company announced that plasticizer is not artificially added in their products. The following two reviews match the second event and showed consumer opinions that distrust the addition of plasticizer. Reviews show the opinions of users regarding an event, and these reviews provide important information for enterprise decision.



Event 1

In 2012, Apple Inc publishes a new iPhone4S, it uses new design and the frame is very beautiful.

Reviews

The metallic frame of iPhone4S attracts my attention.

I like the metallic frame of 4S!

The metallic frame of iPhone4S is very good!

Event 2

Jiugui Liquor Company in 2012 September announced that the plasticizer of liquor was not artificially added.

Reviews

Do not exclude artificially added possible of plasticizer.

I think the plasticizer is artificially added.

Figure 1. Matched Reviews and Events.

The automatic matching of a large number of reviews and events is a new research in machine learning and information integration. We represent an event as an eight-dimension vector to accurately match sentence-level events and reviews. An event is denoted as {agent, activity, object, time, location, cause, purpose, manner}. In this paper, the contents of agent, activity, and object dimensions are considered event feature contents. In a review, some segments containing event feature contents are considered event feature segments. Reviews and events matching research aims to establish a link between an event and its associated reviews.

Matching reviews and existing events pose interesting technical challenges. First, a review contains entities or entity attributes that are consistent with some dimension contents of an event. However, the descriptions of entities and entity attributes in reviews are optional. Many variants of named entity, such as abbreviations, acronyms, and nicknames, appear in reviews. Thus, identifying the variants and their associated named entities in reviews and events is difficult. Second, some reviews show user

doi:10.4304/jsw.9.9.2401-2408

opinions about the activity of an event. For example, two reviews show the skeptical attitude of users about the activity of “artificially added” plasticizer on the second event in Fig. 1. According to integrated events and prior knowledge, verbs or verb phrases are the usual content components of activity dimension. Therefore, effectively identifying verbs and verb phrases is important in a review. This paper presents the event feature segment that contains named entities and verb phrases. This approach is more effective than other methods that match using only named entities [2]. Identifying the event feature segment in a review has a crucial role in reviews and events matching research.

To match reviews and events automatically and accurately, we proposed an approach by using event feature segment and semi-Markov conditional random fields (CRFs). An event feature segment is a text segment that contains named entities and verbs or verb phrases in a review. We use semi-Markov CRFs to identify event feature segments of reviews and use a compositive similarity measurement function to calculate the matching degree of reviews and events. Semi-Markov CRFs is an extended CRF that can perform sequence labeling tasks at the segment level. This method is perfect for labeling and identifying event feature segments composed of multiple word fragments. Compared with single word matching, the event feature segment comprises semantic information and syntactic structure, and event feature segment recognition has higher accuracy for matching. In the event feature segment recognition process, we use some feature rules to identify the variants of named entities and use a phrase dependency parsing tree to identify verb phrases. In the matching process, we use multiple matchers to compute the similarity of event feature segments. A compositive similarity measurement function is presented to calculate the matching result.

This paper is organized as follows. The background is reviewed in Section 2. Reviews and event matching problem and the solving process are described in Section 3. The identifying and matching methods are presented in Section 4. The experiments are described and the results are evaluated in Section 5. The conclusion is given in Section 6.

## II. Background

Web reviews and events matching research belongs to association research of unstructured and structured data. Yahoo research institute proposed a language production model [2] to match reviews and entities without entity recognition. This approach can match reviews and structured entity data from database, but has low accuracy. Most reviews and entities matching research employ the named entity recognition [3] technology to identify named entities in a review. Some entity recognition research utilizes dictionaries [4] to learn the schema character of named entities for recognition. However, constructing a dictionary is time consuming, and existing dictionaries cannot be directly used.

An event includes elements such as time, participant, and activity [5, 6]. Researchers labeling these elements can extract the structured data of an event. The participant is the named entity; therefore, some technology of named entity recognition can be used for reviews and events matching. Monge [7] used the recursive field matching algorithm to find the abbreviations of names of persons for identifying the same semantic entity. In this paper, named entities of reviews and events not only include person names but also contain much product information; therefore, other feature rules were added for matching.

With respect to mining opinions from reviews [8], researchers employ shallow semantic analysis to comprehensively mine opinions in product reviews [9-12]. Wu [13] utilized the phrase dependency parsing method to extract the expressions of opinions. Reviews can provide opinions to different elements in an event and most of these elements are participants and activities in an event. These elements can be summed up as named entity and verb phrase. Matching reviews and events with event feature segments, namely, named entity and verb phrase, has higher accuracy compared with matching reviews and events only with named entity [14].

The CRF model is a statistical learning model widely used in named entity recognition [15], text categorization [16], and information extraction. The semi-Markov CRF model [17, 18] is an extended CRF model that relaxes the Markovian assumptions to allow sequence labeling at the segment level.

We represent an event as an eight-dimension vector and use event feature segments to match reviews and events. In an identifying process, we use semi-Markov CRFs to label event feature segments in reviews. We create some feature rules to identify the variants of named entities and use phrase dependency parsing tree to extract the verb phrases in reviews. In a matching process, we use a compositive similarity measurement function to combine the similarity of event feature segments as the matching result.

## III. Problem Definition

In this paper, event feature segments contained in reviews was used to match web reviews and events. To make a clear presentation, reviews and events matching problem was described and some concepts of our approach were explained in this section.

- **Event:** An event is an activity that occurs at a special time and involves participants. Eight dimensions were used to represent an event and an event can be denoted as {agent, activity, object, time, location, cause, purpose, manner}.
- **Event feature segment:** Some event dimension content segments are contained in a review. The content of event feature segment is summed up in two parts, namely, named entity and verb phrase, and formalized as follows: Event Features={<named entity>,<activity phrase>}.
- **Reviews and events match problem:** A link between an event and its associated reviews is

established. For example, three reviews matched the first event about the "metallic frame of iPhone4S" in Fig. 1. Matching reviews and events can be denoted as $e_{match}$= {($e$ , $r_i$); $i$=1, 2,…, n}. The event and review matching set contains all the reviews associated an event $e$, and $r_i$ represents the $i$th matched review. Event feature segment is the basis of reviews and events matching. Reviews and events matching problem is according to the matching degree of the event feature segments and the event dimension contents.

In this paper, we proposed a review and event matching approach that uses the semi-Markov CRFs to identify the event feature segments in reviews and use a compositive similarity measurement function to calculate the result of reviews and events matching. The overall identifying and matching process is shown in Fig. 2.

The process inputs in Fig. 2 are the web event and review sets, whereas the output is the result of matching reviews and events. First, we preprocess these reviews, which are extracted from microblogs, news, and forums. We then merge repetitive reviews and use the word segmentation tool ICTCLAS 3.1 to divide sentences into words.

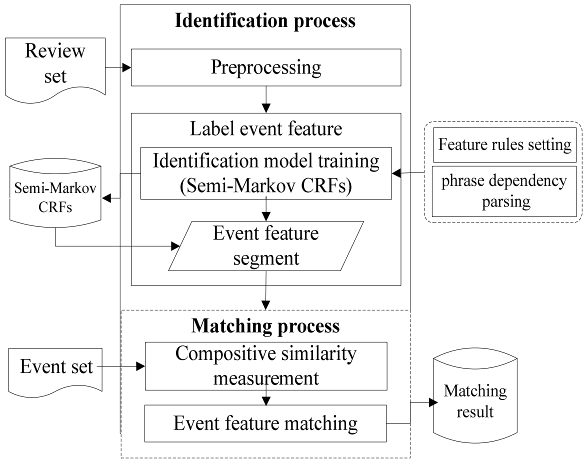


Figure 2. Reviews and Events Matching Process.

After preprocessing, we obtain a review item set for every review in the review set, which can be denoted as $\forall r \in R$ , r={ $r_{w1}$, $r_{w2}$,…, $r_{wn}$}. Second, we use the semi-Markov CRFs to identify event feature segments in reviews in the identifying process. To identify the variants of named entities, we propose some feature rules and identify the abbreviation and acronym of a named entity. We use phrase dependency parsing tree to analyze the review and to find verb phrases as a part of the event feature segment. Third, in the matching process, we use multiple matchers to compute the similarity of event feature segments and events and use a compositive similarity measurement function to calculate the final matching result.

## IV. Matching Approach of Reviews and Events

In this section, we explain our approach in detail. We also define the semi-Markov CRF model and propose feature rules, phrase dependency parsing algorithm, and compositive similarity measurement function.

### A. Semi-Markov CRFs

An extended version of CRFs [19] named semi-Markov CRFs is introduced in this study. This model can be used to label opinion expression and parse information in the segment level. In the semi-Markov CRFs, each observed sequence $X$={$x_1$, $x_2$,…, $x_m$} corresponds to a sequence of consecutive segments $S$ ={$s_1$, $s_2$,…, $s_j$,…, $s_n$}, where $s_j$ is a triple, $s_j$=($t_j$ ,$u_j$ , $y_j$); $t_j$ denotes the start position of segment $s_j$; $u_j$ denotes the end position; and $y_j$ denotes the label of the segment. Segments are restricted to have positive length and adjacent segment touch, that is, $1\leq t_j \leq u_j \leq |s|$, $t_{j+1} = u_j +1$.Given an observed sequence $X$, the conditional probability of a segmentation $s$ is defined as

$$\Pr(s \mid x, W) = \frac{\exp(W \cdot G(x,s))}{Z(x)} \tag{1}$$

In this function, $W$ is the weight vector over the components of G; G($x$, $s$) is the feature function; and $G(x,s) = \sum_{j}^{|s|} g(j,x,s)$. Z($x$) is the normalized factor, and $Z(x) = \sum_{s'} e^{W \cdot G(x,s')}$ . $s'$ denotes any segment in consecutive segment set $S$. $S$ ={$s_1$, $s_2$,…, $s_j$,…, $s_n$}, s' ∈ $s_i$ .In the semi-Markov model, we consider the usual first-order Markovian assumption, where $g$($j$, $x$, $s$) is a segment level feature function of $x$; $s_j$ is the current segment; and $y_{j-1}$ is the label of the preceding segment $s_{j-1}$. Therefore, $g$($j$, $x$, $s$ ) can be rewritten as $g$($y_j$, $y_{j-1}$, $x$ , $t_j$, $u_j$). In the model training process, the Viterbi algorithm[20] [20] was used to estimate the parameters.

$$\begin{aligned} &arg\,max_s\, Pr(s \mid x, W) = arg\,max_s\, W \cdot G(x,s) \\ &= arg\,max_s\, W \cdot \sum_j g(y_j, y_{j-1}, x, t_j, u_j) \end{aligned} \tag{2}$$

In this formula, the best segmentation corresponds to the path traced by $max_s Pr$($s$ | $x$, $W$). An event feature segment is composed of multiple words and contains useful semantic and syntactic structure information, so the event feature segment has high accuracy for identification.

Named entities and verb phrases are both event feature segments, and we use different methods to identify them.

### B. Named Entity Recognation

Given that an event is represented as an eight-dimension vector in this paper, named entities can act as the content of agent, object, and location dimensions. Named entities have some variants such as full name, abbreviations, and acronyms. For example, "4S" is the abbreviation of the full name "iPhone4S," as shown in Fig. 1. Given these variants, named entities in reviews are inconsistent in events. Identifying these variations of named entities can improve the identification accuracy. The similarity of named entity in reviews and events can be used as a feature to train semi-Markov CRFs, but the similarity of full name and variations is not high. In this case, we proposed some feature rules to effectively identify the variations of named entities.

**(1). Pattern feature**

The pattern of named entity in existing events provides important information for identification. Although some variations make a named entity in reviews become inconsistent with an event, we used some non-overlapping sub-patterns of regular expression to express the pattern feature of a variation. Table I presents the sub-patterns of regular expression about the product name.

TABLE I
SUB-PATTERNS OF REGULAR EXPRESSION.

| id | Pattern |
|---|---|
| 1 | [A–Z] |
| 2 | [a–z] |
| 3 | [A–Z] [a–z]+ |
| 4 | {number}[A–Z] |
| 5 | [a–z]{number} |

In Tab. I, [A-Z] [a-z]+ denotes a string written beginning with capital letters. [a–z]{number} denotes a string that includes two parts, beginning with lowercase letters [a–z] and followed by numbers (0–9). For example, the full name of iPhone4S can be denoted as "2, 3, 4" and "4S," which is an abbreviation that can be denoted as "4". In the same way, the phone "Galaxy Grand" can be denoted as "3, 3".

**(2). Label data feature**

The labeled data of a review provide some useful information, such as the length of attribute value and data type. If the named entity is the name of a person, then the segment is usually composed of two or three words. In addition, some named entities, such as company, community, and region, may show special ending features in the entity. For example, the "company," "commission," and "province" always provide the data feature of some named entities. In reviews, "Samsung Company" may be represented as "Samsung." Nevertheless, the two named entities both point to the Samsung Company.

**(3). Association feature**

When different names of an entity appear in an event at the same time, we consider the different names to have an association feature. The association is shown in some regular forms, such as linked with parentheses. For example, "WWDC (apple worldwide developers conference)" is shown in an event, and many users only write "WWDC" in reviews. This variant can be identified by association feature in reviews. The following is a table of some common symbols which can indicate association feature.

TABLE II
COMMON SYMBOLS OF ASSOCIATION FEATURE

| id | Symbol | Explanation |
|---|---|---|
| 1 | ( ) | Variations appear in the brackets |
| 2 | ‘ ’ | Variations appear in the quotes |
| 3 | —— | Variations appear before or after the dash |
| 4 | : | Variations appear after the colon |
| 5 | [ ] | Variations appear in the square brackets |

### C. Verb Phrase Identification

We define an event feature segment containing a named entity and a verb phrase. Identification using only the named entity cannot reach high accuracy because the same entity participating in different activities can represent different events. Users can present reviews about the activity in an event. Therefore, identifying the verb phrase in an activity is important for event feature segment recognition.

We use phrase dependency parsing technology [13] to build a dependency parsing tree for verb phrase identification. According to the result of word segmentation and the label of a review, we utilize the phrase dependency parsing tree to identify the verb phrase. In Fig. 3, we use the phrase dependency parsing tree to identify the verb phrase of the review "I really like using the iPhone4S battery".

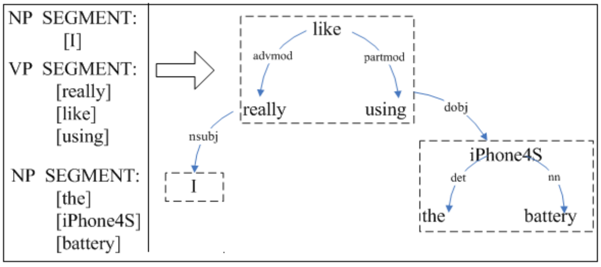


Figure 3. Verb Phrase Identification by Phrase Dependency Parsing Tree.

The left part of Fig.3 lists some segments which can be used to build the phrase dependency parsing tree. The right part of Fig.3 is the tree, and tree nodes are all from listed segments. We can detect the dependency relation of verbs and verb phrases and identify them by using the dependency parsing tree. These labels on edges of the tree are the part-of-speech tagging of words of their children nodes. For instance, the label "nn" means the "battery" is a noun. To accurately recognize verb phrases, we define $VP_{Root}$ and $VP_{leaf}$ to represent different meanings of the verb node in a phrase dependency parsing tree [17].

- **VP$_{\mathbf{Root}}$:** $VP_{Root}$ denotes that the node is a verb node and its parent node is not a verb note.
- **VP$_{\mathbf{leaf}}$:** $VP_{leaf}$ denotes that the head of $VP_{leaf}$ is a verb node and the children nodes of $VP_{leaf}$ are not verb notes.

**Algorithm 1. Verb phrase identification algorithm.**

**Input:** A preprocessed review that is represented as $r=\{s_1, s_2, \ldots, s_n\}$
**Output:** Verb phrase candidate segment $S_V=\{s_{v1}, s_{v2}, \ldots, s_{vm}\}$

```
Obtained the segment units {$s_1, s_2, \ldots, s_n$}, build the dependency parsing tree $T$.
Detecting $VP_{Root}$ and $VP_{Leaf}$ by dependency parsing tree $T$,
  Form verb sequence set $V=\{v_1, v_2, \ldots, v_m\}$
    For $i$=1 to m do
        CurNode = $V_i$
        /* Selection start from the root node */
            getChildrenNode ($T$, CurNode);
            $S_V = S_V \cup$ getChildrenNode ($T$, CurNode);
        CurNode = $V_{i+1}$
    While CurNode = $VP_{Leaf}$
    End While
```

Figure 4. Algorithm for Verb Phrase Identification.

We identify the verb phrase of the event feature segment by using the dependency parsing tree from $VP_{Root}$ to $VP_{leaf}$. Fig. 4 shows the verb phrase identification algorithm. The verb phrase identification algorithm can recognize some candidate verb phrases as a part of an event feature segment.

This algorithm uses phrase dependency parsing tree to identify verb phrases. First, algorithm uses review segment units to build phrase dependency parsing tree. Second, algorithm detects the $VP_{Root}$ and $VP_{Leaf}$ by tree, selects and analyzes children node from the $VP_{Root}$. Finally algorithm obtains verb phrase set $S_v$ which contains verb phrase candidate segments.

### *D. Matching Reviews and Events*

We match reviews and events by calculating the similarity of an event feature segment and the event dimension content. In this paper, we divide an event feature segment into two parts, named entity and verb phrase. We use different similarity matchers to compute for the similarity degree of an event feature segment and use a compositive similarity measurement function to calculate the matching result.

**(1). Named entity similarity matcher**

Given a review $r\{s_1, s_2,..., s_n\}$, $s_1, s_2,..., s_n$ are identified as event feature segments. An event is represented as an eight-dimension vector, and the content of the agent, activity, object, and time dimension are not blank. We use a cosine similarity matcher to compute the similarity of the named entity segment, agent, and object (agent and object are both named entities). The computer function is as follows:

$$Sim(N_e, N_r) = \frac{\overrightarrow{N_e} \cdot \overrightarrow{N_r}}{||N_e|| \cdot ||N_r||} \quad (3)$$

$N_e$ is the named entity in an event, and $N_r$ is the named entity of the event feature segment in a review. For example, $N_e$ is the content of an agent; $(t_2, t_3, t_4)$ is an identified event feature segment that corresponds to $N_r$ in a review, which can be denoted as $N_r = g(y_i , y_{i-1} , 2 ,4)$ and $y_i$ =[agent].

**(2). Verb phrase similarity matcher**

To compute the similarity of a verb phrase and the content of the activity dimension, we use Hownet [21] to compute the semantic similarity. In the Hownet semantic network architecture, a word is composed of primitives. The verb itself may be the primitive or can be deconstructed into primitives. We use a primitive to denote a verb, and a verb phrase can be divided into primitives. For instance, the primitives of "sell at half price" are "sell," "depreciate," and "range is 50%." $V_r$ is a verb phrase and can be denoted as $V_r=\{P_{vr1},P_{vr2},...,P_{vri}\}$; in the same way, $V_e$ is the activity content and can be denoted as $V_e=\{P_{ve1},P_{ve2},...,P_{vej}\}$. $P_{vri}$ is the primitive of a verb phrase in a review, and $P_{vej}$ is the primitive of a verb phrase in an event. The following is a formula of depth-based semantic matcher to measure the similarity of verbs:

$$Sim_{Verb}(P_{Vei}, P_{Vrj}) = \frac{\alpha}{\alpha + dist(P_{Vei}, P_{Vrj})} \quad (4)$$

In this formula, $P_{Vei}$ and $P_{Vrj}$ are primitives of two verb words, *dist* is the shortest path length of two primitives in Hownet; $\alpha$ is an adjustable factor, and $\alpha$=0.5 was set by experiments. The similarity of $V_{review}$ and $V_{event}$ is the maximum similarity of the primitives.

$$Sim(V_e, V_r) = \max_{i=1\cdots m, j=1\cdots n} \left| Sim(P_{Vei}, P_{Vrj}) \right| \quad (5)$$

**(3). Compositive similarity measurement**

An event feature segment is composed of two parts, so we respectively calculate the matching degree and use a compositive similarity measurement to obtain the result.

We obtain the similarity of the named entity and add the measure of an event feature matching rules to compute the matching degree of the named entity.

$$Match(N_e, N_r) = Sim(N_e, N_r) + Pattern(N_e, N_r) + Association(N_e, N_r) \quad (6)$$

In this formula, Sim($N_e$, $N_r$) is the similarity of the named entity, Pattern($N_e$, $N_r$) is the pattern matching degree of the named entity, and Association ($N_e$ , $N_r$) is the association degree of the named entity. Match ($N_e$ , $N_r$) is the combined matching degree of the named entity.

To match reviews and events, we need to combine the matching degree of the two parts of an event feature segment. We use a formula to compute the matching degree of the named entity. We consider the similarity of the verb phrase as the matching degree of the verb phrase and use a compositive similarity measurement to calculate the final matching result of reviews and events.

$$Match(e,r) = \beta_1 Match(N_e, N_r) + \beta_2 Match(V_e, V_r) \quad (7)$$

$\beta_1$ and $\beta_2$ are the proportional factors of the named entity and verb phrase, $\beta_1+\beta_2=1$, Match($V_e$, $V_r$)= Sim($V_e$, $V_r$). We obtain the optimal value of $\beta_1$ and $\beta_2$ by conducting the experiment.

## V. Evaluation and Results

In this paper, we proposed an approach for reviews and events matching on the basis of event feature segment and semi-Markov CRFs. We used real data and conducted a series of experiments on different fields. We randomly selected 7% reviews from a review set as a training set to label. To evaluate the effectiveness of our approach, we tested the approach in three aspects. (1) We evaluate the matching accuracy of the different approaches. (2) We assess the influence of the training set size on matching accuracy. (3) We test the effective influence of interference data on our approach.

### *A. Dataset*

We extracted 600 events from sina.com.cn, news.xinhuanet.com, and finance.qq.com. These events constitute two event datasets, food and phone. We extracted reviews from web pages (i.e., sina microblog, news linked web pages, and forums), which also constitute food and phone datasets. After merging repetitive reviews and filtering short and meaningless reviews, such as "nonsense!" we obtained 8,236 reviews. These reviews are relevant to extracted events. We extracted 2,034 irrelevant reviews as interference data. In the review dataset, we randomly selected 7% reviews as training data, and the remaining reviews are used as testing data.

### *B. Experiment Evaluation*

We adopted the classic evaluation standard of information retrieval and used recall, precision, and

F-measure for assessment. Evaluation standards are defined as follows.

A represents the number of reviews (contain irrelevant reviews as interference data). B represents the number of reviews that accurately matched the associated event. C represents the number of reviews with a mismatch. Precision, recall, and F-measure are calculated as follows:

$$Precision = \frac{B}{B+C} \tag{8}$$

$$Recall = \frac{B}{A} \tag{9}$$

$$F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{10}$$

*C. Experiment Results*

**(1). We compared the accuracy of the proposed matching approach with other approaches in different datasets.**

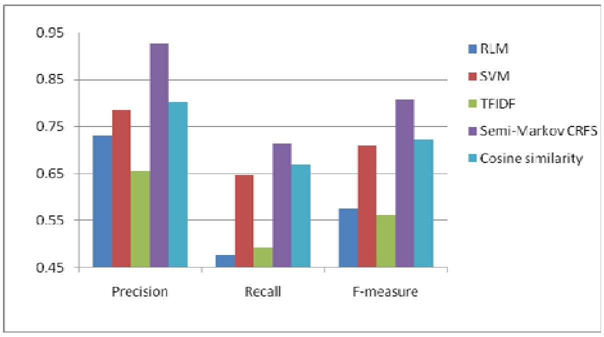


Figure 5. Accuracy of Different Matching Approaches in Food Dataset

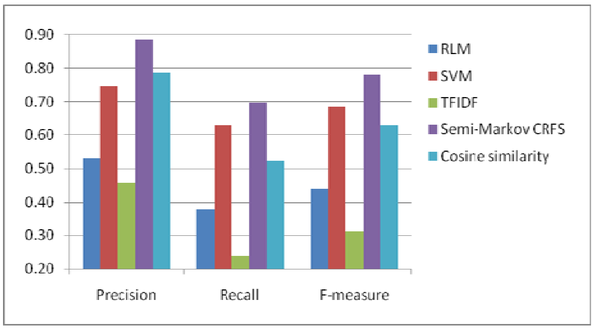


Figure 6. Accuracy of Different Matching Approaches in Phone Dataset

The accuracy of the proposed approach that uses semi-Markov CRFs for matching reviews and events is higher than RLM, SVM, TFIDF, and cosine similarity (Fig.5 and 6). Compared with other methods, semi-Markov CRFs identify the event feature segment with richer semantic information than single words in reviews. We use feature rules to recognize the variants of named entities and use the dependency parsing tree to analyze the verb phrase. Therefore, our proposed method can identify more accurately compared with other methods, which identify only named entities in reviews. In the matching process, we utilize mutiple matches to calculate the similarity of event feature segment and employ a compositive similarity measurement to calculate the final matching result. This approach combines the advantages of multiple matchers. This is another reason that it has high matching accuracy compared with other methods (i.e., cosine similarity, SVM).

**(2). We assessed the influence of the training dataset size on matching accuracy.**

To test the influence of training set size on the accuracy of the approach, we randomly selected 1%, 3%, 5%, 7%, 9%, 11%, and 13% reviews in the review dataset. We labeled event feature segments in these different sizes of training dataset. The remaining reviews are used as test dataset. In the case of larger artificial labeled training dataset, many methods have higher matching accuracy. But if a method needs a large size of training set to keep high accuracy, we consider the automation of the method is low. In order to adapt to large-scale web reviews and events matching, the approach needs keep high accuracy with small training dataset size (i.e., less than 10%). We obtained the changing curve of the F-measure with different sizes of training dataset (size of training set is 1%, 3%, 5%, 7%, 9%, 11%, and 13%).

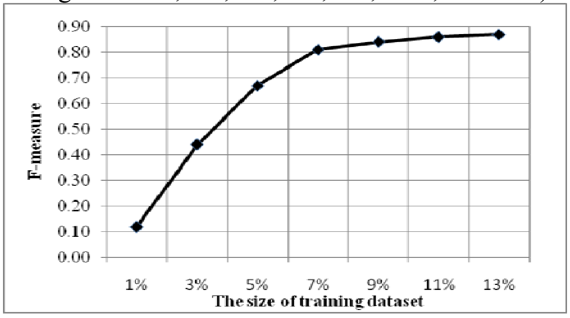


Figure 7. F-measure for Different Training Dataset in Food Dataset.

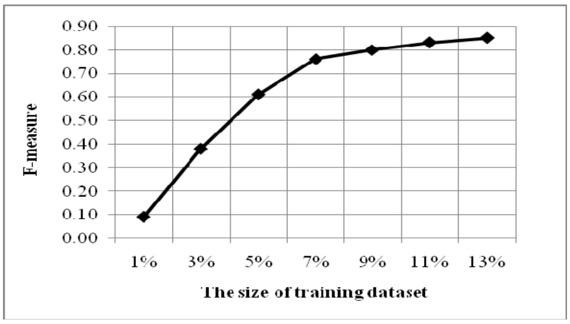


Figure 8. F-measure for Different Training Dataset in Phone Dataset.

Fig.7 and Fig.8 show that the number of training dataset significantly influences the matching accuracy. When the size of manual label is from 1% to 7%, the F-measure significantly increases. The rising trend of the changing curve becomes smooth from 7% to 13%. Thus, with the increase in training data, the effectiveness of adding manual training dataset decreases. Therefore, the number of training dataset can be controlled in a certain range. Our proposed approach needs only 7% training dataset to reach a high F-measure. Figs. 7 and 8 show that the proposed approach is less dependent on the size of the training dataset and is suitable for large-scale web reviews and events matching.

**(3). We tested the effect of interference data on our approach.**

The extracted reviews were obtained from microblogs, forums, and other web pages, as well as many event irrelevant reviews in the review dataset. The antinoise ability of our approach must also be evaluated. In the case of existing interference data, we tested whether this approach can reach a high matching accuracy. In this experiment, we added irrelevant reviews to the review dataset and observed the changes in the curve of the F-measure for different numbers of interference data. First, we chose 300 event relevant reviews and then added 50 irrelevant reviews in each test.

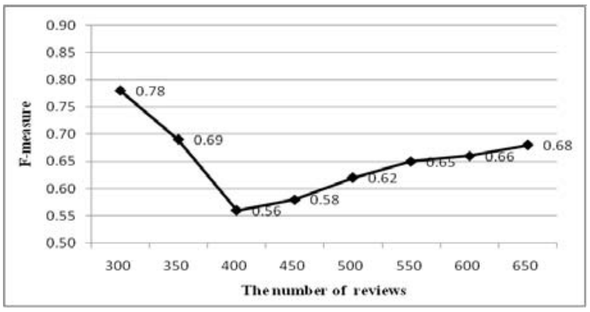


Figure 9. F-measure for Different Number of Irrelevant Reviews.

Fig. 9 shows the changing curve of the F-measure with increasing number of interference data. We added irrelevant reviews as interference data, and the F-measure had an obvious change. Upon adding the first 50 irrelevant reviews, the F-measure exhibited a shape decrease. When the irrelevant reviews reached 100, the F-measure decreased to its minimum. When the number of reviews achieved 400, the changing curve increased slowly. The proposed approach can automatically identity the pattern of named entities and verb phrases, thus, the increased reviews provided rich semantic and structure information for improving identification accuracy. This experiment showed that the effect of interference data on the proposed approach is limited.

## VI. Conclusion

In this paper, we propose a web reviews and events matching method based on event feature segments and semi-Markov CRFs. An event feature segment is composed of named entity and verb phrase. This approach uses the semi-Markov CRFs to identify event feature segments at the segment level, which can use rich semantic information. The approach uses event feature segments to match reviews and events. Therefore, it is more accurate than other approaches that use only named entities to match. We use several feature rules to recognize the variants of named entities, such as abbreviation and acronym, and utilize the phrase dependency tree to recognize verb phrases. In the matching process, we employ different matchers to compute the matching degree of two parts of event feature segments. A compositive similarity measurement function is presented to combine the matching result. A series of experiments is carried out to evaluate the effectiveness of the proposed approach. The results demonstrate that this approach can accurately match large scale of web reviews and events.

## Acknowledgment

This work is supported by the Natural Science Foundation of China (No.61303005) and the Shandong Distinguished Middle-aged and Young Scientist Encouragement and Reward Foundation (No.BS2012DX015).

## References

[1] M. Naughton, et al., "Sentence-level event classification in unstructured texts," Information retrieval, vol. 13, pp. 132-156, 2010.

[2] N. Dalvi, et al., "Matching reviews to objects using a language model," in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2, 2009, pp. 609-618.

[3] J. Nothman, et al., "Learning multilingual named entity recognition from Wikipedia," Artificial Intelligence, vol. 194, pp. 151-175, 2013.

[4] S. Kim, et al., "Multilingual named entity recognition using parallel data and metadata from wikipedia," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, 2012, pp. 694-702.

[5] C. J. Fillmore, "Frames and the semantics of understanding," Quaderni di semantica, vol. 6, pp. 222-254, 1985.

[6] C.-Y. Zhang, et al., "Extracting web entity activities based on SVM and extended conditional random fields," Ruanjian Xuebao(China)/Journal of Software, vol. 23, pp. 2612-2627, 2012.

[7] A. E. Monge and C. Elkan, "The Field Matching Problem: Algorithms and Applications," in KDD, 1996, pp. 267-270.

[8] W. Xu, et al., "Sentiment Recognition of Online Chinese Micro Movie Reviews Using Multiple Probabilistic Reasoning Model," Journal of Computers, vol. 8, pp.1906-1911, 2013.

[9] S. Moghaddam and M. Ester, "Aspect-based opinion mining from product reviews," in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 1184-1184.

[10] R. Kaula, "Business Rules Modeling for Business Process Events: An Oracle Prototype," Journal of Computers, vol. 7, pp.2099-2106, 2012.

[11] S. Huang, et al., "Fine-grained product features extraction and categorization in reviews opinion mining," in Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, 2012, pp. 680-686.

[12] R. Robinson, et al., "Textual factors in online product reviews: a foundation for a more influential approach to opinion mining," Electronic Commerce Research, vol. 12, pp. 301-330, 2012.

[13] Y. Wu, et al., "Phrase dependency parsing for opinion mining," in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3, 2009, pp. 1533-1541.

[14] R. Balasubramanyan and W. W. Cohen, "Block-LDA: Jointly modeling entity-annotated text and entity-entity links," in SDM, 2011, pp. 450-461.

[15] G. A. Seker and G. Eryigit, "Initial Explorations on using CRFs for Turkish Named Entity Recognition," in COLING, 2012, pp. 2459-2474.

[16] A. Balaram, "A Two Stage Approach Using SVM/CRF for Text Categorization," International Journal for Research in Science & Advanced Technologies, vol. 2, pp. 105-108, 2013.

[17] B. Yang and C. Cardie, "Extracting opinion expressions with semi-markov conditional random fields," in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 1335-1345.

[18] S. Sarawagi and W. W. Cohen, "Semi-Markov Conditional Random Fields for Information Extraction," in NIPS, 2004, pp. 1185-1192.

[19] F. Wang and X. Zhang, "CRFs in Music Chord Recognition Algorithm Application Research," Journal of Computers, vol. 8, pp. 1016-1019,2013.

[20] W. W. Cohen and S. Sarawagi, "Exploiting dictionaries in named entity extraction: combining semi-markov extraction processes and data integration methods," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 89-98.

[21] Y. Bin, et al., "Using Information Content to Evaluate Semantic Similarity on HowNet," in Computational Intelligence and Security (CIS), 2012 Eighth International Conference on, 2012, pp. 142-145.

**Yuanzi Xu,** Born in 1983, currently a Ph.D candidate in Shandong University of computer science and technology. Her research interests are web information integration, artificial intelligence and event detection.

**Qingzhong Li,** Ph.D, currently a professor in Shandong University. His research interests are web information integration, artificial intelligence and large-scale network data management.

**Zhongmin Yan,** Ph.D, currently an associate professor in Shandong University. Her research interests are web information integration and artificial intelligence.

**Wei Wang,** currently a master candidate in Shandong University of computer science and technology. His research interests are artificial intelligence and event detection.