Predicting the Occurrence of Life Events from User's Tweet History

Shun Abe*, Masumi Shirakawa*†, Tatsuya Nakamura*, Takahiro Hara*, Kazushi Ikeda‡, Keiichiro Hoashi‡
*Graduate School of Information Science and Technology, Osaka University
E-mail : {abe.shun,nakamura.tatsuya,hara}@ist.osaka-u.ac.jp
†hapicom Inc.
E-mail : shirakawa@hapicom.jp
‡KDDI Research, Intelligent Media Laboratory
E-mail : {kz-ikeda,hoashi}@kddi-research.jp

***Abstract*—This paper addresses the new problem of predicting the occurrence of Twitter user's life events using word occurrence tendencies in user's tweet histories. Many previous studies have addressed public event prediction and life event extraction on Twitter. Most of the methods for these two problems use tweets that explicitly refer to event occurrence. However, users who will experience a life event are unlikely to post tweets that reference the occurrence of the life event explicitly; thus, existing methods to find tweets that refer to event occurrence explicitly are not applicable to life event prediction. Therefore, we assume that users who will experience a life event tend to post tweets that refer to life event occurrence implicitly and propose a method to identify such a tendency to predict life events. First, we extract users who experienced a specific life event and collect their past tweets to identify features that implicitly indicate the life event occurrence. We use the word occurrence tendency in such tweets as training data to construct a life event prediction model. We chose five life events, that is, "Giving birth," "Getting a job offer," "Leaving the hospital," "Pregnancy," and "Marriage," and assessed the prediction performance of the proposed method for each event. Experimental results demonstrate that the proposed method outperformed a baseline method for all selected life events except "Leaving the hospital" and achieved the highest prediction accuracy for "Giving birth." We suppose that this event resulted in the highest prediction accuracy because all users who gave birth experienced pregnancy and common features appeared in their tweets over a long period.**

I. Introduction

A. Background and Motivation

In Twitter, which is a representative microblogging service, many users post short messages (i.e. *tweets* restricted to 140 characters) about various real-world events related to their interests and living areas. Many studies have been conducted on Twitter-based real-world event extraction [1], [2] and event prediction [3], [4], [5].

Recently, life events, such as "Giving birth" and "Getting a job offer," have attracted attention as a new type of real-world events [6], [7], [8], [9], [10]. Life event extraction is a problem of extracting the occurrence of Twitter user's life events, i.e., identifying users who experienced a given life event. Extracted user's life events can be used for personalized recommendation and targeted online advertising. For example, we can recommend baby care products to users who experienced the "Giving birth" event, or honeymoon packages to users who experienced the "Marriage" event. Moreover, if we can predict the occurrence of user's life events beforehand, this facilitates applications that use predicted life events to trigger some actions, e.g., recommending various products to users before they experience their life events.

In this paper, we address the problem of predicting the occurrence of Twitter user's life events. To the best of our knowledge, research on predicting Twitter user's life events has not been conducted. The problem can be defined as follows: Given a Twitter user's tweet history and the type of a life event, we predict whether the user will experience the life event in the near future or not. Compared to the problems of event prediction and life event extraction, the problem has the following challenges.

1) **Life events are experienced by individual users**: Studies on event prediction focused on public events, such as sports games and concerts. To predict such events, they assumed that many users post tweets related to the events and the tweets contain mentions of the date. However, a life event tends to be mentioned by a single user who will experience the event. Therefore, the previous event prediction methods relying on a number of tweets cannot be applied to the prediction of user's life events.
2) **Users who will experience a life event hardly post explicit tweets before they experience the life event**: Studies on extracting life events assumed the existence of tweets that contain keywords mentioning the occurrence of a target life event explicitly. Such tweets are generally posted by users after the occurrence of their life events, i.e., users are less likely to post such tweets before experiencing the life events. Thus, previous approaches intended for life event extraction do not work well for life event prediction.

To predict the occurrence of user's life events, it is necessary to consider not only keywords explicitly related to life events, but also the difference of behaviors between users who will experience life events (which we call *event user*) and those who will not. Here, we consider that the most reliable feature is users' word usages. For example, comparing users who post tweets like "*I'm eight months pregnant*" and those who post tweets like "*I heard that celebrity gave birth!*", the former users are more likely to experience the "Giving birth" event regardless of the occurrence of word "birth." Therefore, we hypothesize that life events can be predicted by comparing the word occurrence tendencies in tweets posted by users who will experience the life events with those by another users. However, obtaining training data by identifying whether a user will experience a life event in the near future is difficult. Even if a user posts a tweet contains explicit expressions about a life event, e.g., "*I wish the baby would be born soon,*" it is not obvious that the user will actually experience the "Giving birth" event. Thus, obtaining reliable labeled data for learning the difference of word occurrence tendencies is a nontrivial problem.

B. Contributions

In this paper, we addressed the new problem of predicting the occurrence of Twitter user's life events. Since previous methods proposed for event prediction and life event extraction are not suited for our problem, we proposed a method that builds prediction models by learning the difference of word occurrence tendencies between users who will experience life events and those who will not. To build the prediction models, identifying users who will experience life events for obtaining labeled data is a key challenge. We solved this by regarding the tweet histories of users who have already experienced life events as those of users who will experience life events at some point in the past. Specifically, we leveraged the life event extraction method [10] to identify users who have experienced life events and retrieved their past tweets to obtain tweet histories.

The proposed method extracts the word occurrence tendencies from tweet histories of the users who experienced the life events and randomly sampled users. We define event users' word occurrence tendencies as positive examples and other sampled users' word occurrence tendencies as negative examples. Then, it constructs a prediciton model by using Support Vector Machine (SVM) [11] with the extracted word occurrence tendencies as features. The model identifies whether a given user will experience a given life event in the future. We evaluated the prediction performance of the proposed method experimentally with five life events, that is, "Giving birth," "Getting a job offer," "Leaving the hospital," "Pregnancy," and "Marriage." The experimental results demonstrate that the proposed method is effective for predicting life events.

II. Related Work

Related work can be divided into two categories: (1) event prediction and (2) life event extraction. We introduce related work in these categories in the following subsections.

A. Event Prediction

To the best of our knowledge, research on predicting Twitter user's life events has not been conducted. However, some research on predicting events that many users watch has been conducted. Some previous studies [3], [4] focused on the prediction of soccer games. These studies predict events by analyzing time representations in tweets because many users post tweets that contain events' date. Becker et al. [5] proposed a method that collects information about planned events, such as concerts and conferences. As with soccer games, many users post tweets that contain events' date, location and description. This method analyzes texts collected from Twitter, Flickr and Youtube to identify event information about dates and places. Aside from Twitter, *ChronoSeeker* [12] is a search engine where users can search past and future events that are relevant to user queries. To predict events, this system extracts web pages containing representations of the event dates, and then finds web pages that match a given user query. These studies assume that there are many documents containing explicit information including event dates. However, this assumption does not hold for life events. Tweets that mention the date of a life event occurrence are unlikely to be posted because life events are experienced by individual users. Thus, existing methods that use time representations cannot be applied to life event prediction.

Radinsky et al. [13] proposed a method to predict future events using causal reasoning. Causal reasoning can be used to derive causality patterns from past information. For example, this method automatically derives the fact that a tsunami is likely to occur after a large earthquake and infers that a tsunami would occur in the Pacific Ocean after a magnitude 6.5 earthquake struck the Solomon Islands. It is difficult to apply existing causal reasoning methods to life event prediction due to the lack of data for causal reasoning. Causal reasoning might be an effective approach for life event prediction if sufficient data is available.

Tsuboi et al. [14] proposed a method for Twitter users' product purchase prediction. In this paper, they hypothesized that users who will purchase an expensive product tend to post tweets related to the product and such tweets increase as the purchase date approaches. Based on the hypothesis, they construct a prediction model with various features, such as the content of tweets and the increasing rate of this related tweets. They demonstrated the effectiveness of their proposed method for product purchase prediction. If tweets related to a life event increase as the life event occurrence approaches, their features may also be useful for improving the performance of life event prediction.

B. Life Event Extraction

Life event extraction is a developing area in Twitter mining research. Some previous studies [6], [7] focused on tweets that mention the occurrence of a life event. Choudhury et al. [6] extracted five life events, i.e., "Graduation," "Marriage," "New Job," "New Born," and "Surgery," by using text and interaction features (e.g., replies, retweets, and favorites). In this study, text features were more effective than interaction features. Dickinson et al. [7] extracted five life events, i.e., "Having children," "Beginning school," "Marriage," "Parent's death," and "Falling in love," by using various features, such as user's social position, the frequency of posting, and the pattern of interactions with other users. They examined which features contribute to improvement of the performance of extraction. They revealed that the effectiveness of features to the performance depends on the type of life events. For example, in the "Parent's death" event, n-grams and interaction features contributed to the performence. Akbari et al. [8] focused on diabetic and classified their tweets into 14 types of wellness events, such as "Diet," "Excercise," and "Health." Lin et al. [9] proposed a model to extract events that cause stress and measure the user's stress level. Li et al. [10] discovered interaction patterns for accurately extracting life events. When users post tweets reporting their life events, their followers highly tend to respond to these tweets with fixed expressions, such as *congratulations.*

These studies demonstrated the effectiveness of Twitter as an information source to extract the occurrence of user's life events. However, these methods cannot be applied to predict user's life events because they assume that tweets reporting the occurrence of life events have explicit mentions and reactions from other users. On the other hand, our proposed method utilizes Li et al's method to collect tweet histories of users who have experienced a given life events.

III. Preliminaries

In this section, we define the problem of life event prediction as follows.

Definition 1 (Life event): A life event in this paper is a major personal event that significantly changes user's life due to its occurrence.

Definition 2 (The occurrence date of a life event): We define that the occurrence date of a life event of a user is when the user posts a tweet reporting the experience of the life event. For example, when a user posts a tweet "*I gave birth to my child today*" on October 8th, 2017, we consider the user experienced the "Giving birth" event on that date. Note that some users report their life events after a while, e.g., "*I'm eight months pregnant now.*" In that case, we consider the life event occurred on the reported date because it is most obvious that the user has already experienced the life event at the point of that date. Considering time representations in tweets and adjusting the occurrence date of the event are future work. In this paper, we utilized the existing method [10] to extract the occurrence of life events (i.e. tweets) precisely. The details of this extraction method are described in Section IV.B.

Definition 3 (Life event prediction): For the sake of simplicity, we define life event prediction as a binary classification problem in this paper. Given a user, his/her tweet history for a month, the type of a life event, and prediction period K (months), we predict whether the user will experience the life event within K months from present. Here, the occurrence date of a life event is defined as Definition 2. In the next section, we introduce our proposed method to address the problem of life event prediction.

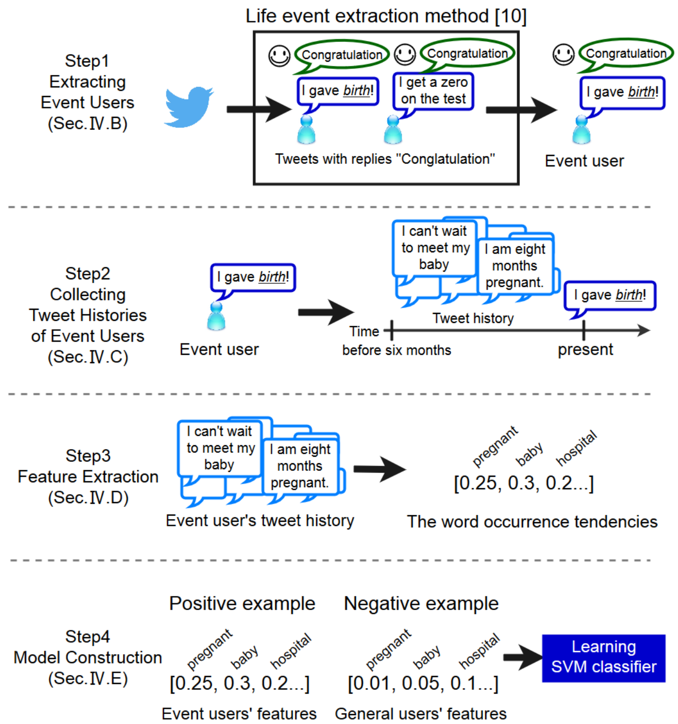

Fig. 1. Overview of the proposed method

IV. Construction of a Life Event Prediction Model

A. Overview

The purpose of this study is to construct a prediction model for the life event prediction problem defined in Section III. To achieve this, we propose a method to construct the prediction model using word occurrence tendencies of user's tweet histories. The proposed method considers tweet histories of users who experienced life events (event users) as those of users who will experience life events in the near future. It learns the difference of word occurrence tendencies in the tweet histories of event users and randomly extracted Twitter users (hereafter *general users*). Fig. 1 shows the procedure of the proposed method. We describe each step of the proposed method in the following subsections:

1) **Extracting Event Users (Section IV.B)**: We extract event users by using the existing method for life event extraction [10]. After that inappropriate data are manually removed.
2) **Collecting Tweet Histories of Event Users (Section IV.C)**: We collect each event user's tweet history (six months) starting from the time when the event user experienced a life event. Note that general users' tweet histories are collected for the same period as the event users.
3) **Feature Extraction (Section IV.D)**: We extract word occurrence tendencies from tweet histories as features.

TABLE I
CANDIDATE OF LIFE EVENTS EXTRACTED USING THE FIXED FORM EXPRESSION OMEDETOU

Birthday	Anniversary	Tournament
Stage performance	Giving birth	Marriage
Leaving the hospital	Winning the lottery	Completion of writing
Passing the exam	Pregnancy	Getting a job offer

4) **Model Construction (Section IV.E)**: We define event users' and general user's word occurrence tendencies as positive and negative examples respectively, and construct a prediction model using an SVM classifier with extracted features.

Note that, in this paper, we evaluated the proposed method using Japanese tweets. Therefore, we explain the details of the proposed method using examples in Japanese and denote Japanese terms in sans-serif in the following sections.

B. Extracting Event Users

In the proposed method, we need to collect tweet histories of event users. To extract event users, we employed the existing method of life event extraction [10]. This method uses the tendency that, when and only when a user posts a tweet mentioning his/her life event occurrence, followers respond to it with certain fixed expressions, e.g. *congratulations*. By using this method, we collected event users as follows: (1) collect replies, which contain the expression omedetou (congratulations in Japanese) ; (2) retrieve tweets to which the collected replies referred; (3) extract tweets that contain predefined keywords representing life events; (4) manually remove tweets that do not mention the users' life events.

It is assumed that the types of life events extracted with the method using the expression omedetou are limited. Therefore, we investigated what kinds of life events could be extracted using the expression omedetou. We collected Japanese replies that contained omedetou from June 10 to August 8, 2016. We then retrieved Japanese tweets to which the collected replies referred and divided them into topics using Latent Dirichlet Allocation [15]. We manually assigned labels to each topic. As a result, we extracted the candidates of life events listed in Table I. Using replies that include the word omedetou, we can construct prediction models for these life events. Note that, Table I does not include seasonal life events, such as *admission* and *graduation,* due to the date range when we collected the replies.

C. Collecting Tweet Histories of Event Users

We collected tweet histories of event users starting from the time when the event users experienced life events. As described in Section IV.A, we regard the tweet histories of event users as the tweet histories of users who will experience the life events in the near future. With respect to general users, we randomly sampled Twitter users and collect their tweet histories starting from their latest tweets.

We used the Twitter REST API[1] to collect tweet histories of event users. However, this API can only collect 3,200 tweets from the latest tweet for each user. Therefore, if the users have posted more than 3,200 tweets, we can collect the latest 3,200 tweets only. To evaluate the peformance of the prediction in a short period as well as a long period, we discarded the users whose tweet histories cannot be collected over six months. Moreover, we discarded event users who have posted tweets rarely because it is difficult to extract features from tweets of such users. Specifically, we discarded event users who posted fewer than 60 tweets over a six-month period. Note that the tweet histories of general users were also collected in the same manner.

D. Feature Extraction

To construct a prediction model, it is important to design features that can be used to discriminate event users (i.e., users who will experience a life event) from general users from their tweet histories. We hypothesize that event users tend to post tweets related to life events more than general users, i.e., event users tend to post tweets using different words from general users. Based on the hypothesis, we extracted word occurrence tendencies as the features.

The previous studies on life event extraction [6], [7] defined keywords that directly identify a life event in advance. For example, the word *birth* was defined as a keyword of the "Giving birth" event. Using such keyword occurrences in some fixed expressions are effective to extract the occurrence of user's life events after they occurred. However, this approach is not effective for life event prediction. If a user does not post tweets containing such keywords until the occurrence of a life event, this approach cannot predict life events. In addition, the keywords are also used by general users, e.g., "*My syster will give birth in the next year.*" Thus, predefined keywords are insufficient in our problem. Therefore, we propose an approach that automatically extracts discriminative words for life event prediction from user's tweet histories.

Given the collected tweet histories of event and general users, the proposed method first replaces a user's tweet history into a vector of discriminative words weighted with their occurrence probabilities. The word occurrence probability represents the percentage of the number of tweets that contain a discriminative word among the total number of tweets in the tweet history. Here, how to determine discriminative words and their weights is a nontrivial problem. The high probability of a word in an event user's tweet history does not necessarily mean that the word is discriminative because the probability may also be high in general users' tweet histories. To find discriminative words, the proposed method recalculates the weight of vectors. In particular, the proposed method subtracts the word occurrence probabilities calculated using randomly sampled tweets (using Twitter Streaming API) from the probabilities of the words from user's tweet histories. This process effectively extracts discriminative words that tend to

[1] https://dev.twitter.com/rest/reference/get/statuses/user_timeline

TABLE II
LIFE EVENTS AND CORRESPONDING KEYWORDS USED IN THE EVALUATION

Life Event	Keyword in Japanese
Giving birth	Syussan
Getting a job offer	Naitei
Leaving the hospital	Taiin
Pregnancy	Ninshin
Marriage	Nyuuseki

TABLE III
NUMBER OF EXAMPLES IN TRAINING AND TEST DATA (#POSITIVE:#NEGATIVE)

Life Event	Training data		Test data	
	$K=1$	$K=6$	$K=1$	$K=6$
Giving birth	191:12415	1146:11460	48:3120	288:2880
Getting a job offer	271:17615	1626:16260	68:4420	408:4080
Leaving the hospital	184:11960	1104:11040	47:3055	282:2820
Pregnancy	164:10660	984:9840	42:2730	252:2520
Marriage	241:15665	1446:14460	61:3965	366:3660

appear in tweet histories of event users. We use top 300 words with recalculated weights as features. We confirmed that more than 300 words hardly contributed to the performance.

E. Model Construction

Using the extracted features, we constructed a model to predict whether a user will experience a given life event within a given period K. We used an SVM classfier [11] as the prediction model. SVM is a type of pattern discriminative model and frequently used for binary classification.

To construct a prediction model, we divided a collected tweet history (for six months) into six monthly histories. Each divided tweet history corresponds to tweets that were posted in each month in the original tweet history. We assigned a positive label if an input monthly tweet history comes from event user's K-month tweet history. In other words, if an input monthly tweet history is older than K months, it is labeled as negative even if it comes from event user's tweet history. All monthly tweet histories of general users are labeled as negative.

V. EXPERIMENT

A. Evaluation Environment

Dataset: To evalute the performance of the proposed method, we selected five life events from Table I, i.e., "Giving birth," "Getting a job offer," "Leaving the hospital," "Pregnancy," and "Marriage." We created the datasets of each life event using Japanese tweets collected from June 10 to August 8, 2016 as follows. First, we extracted event users of each life event using the life event extraction method [10] explained in Section IV.B. In this process, we used Japanese words shown in Table II as keywords, which represent each life event name. Next, we collected tweet histories of event users and general users according to the process described in Section IV.C. Note that, some of general users might experience a

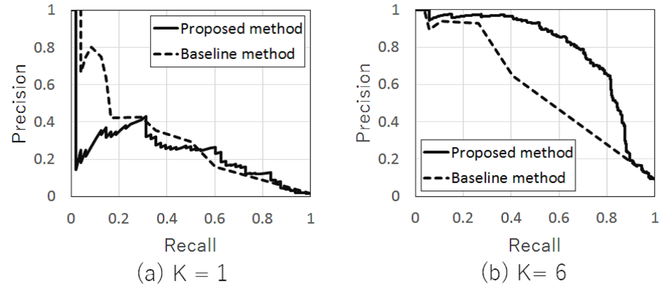

Fig. 2. Precision-recall curve for the "Giving birth" event

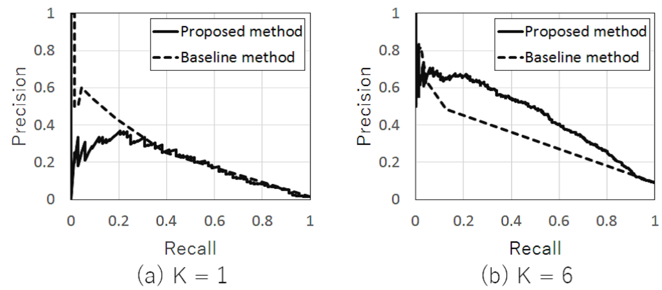

Fig. 3. Precision-recall curve for the "Getting a job offer" event

given life event within K months. However, the proportion of such users is considered to be very small. Therefore, we ignored the existence of such users in this evaluation. Finally, we divided collected tweet histories into monthly histories and labeled each monthly history based on the process described in Section IV.E to create positive and negative examples. Note that, we removed users who can be considered as spam users from the dataset. Specifically, we removed a user who meets either of the following conditions as a spam user: (1) all tweets in the history contain a URL; (2) there are over 20% duplicated tweets in the history; (3) the number of unique words used in the history is less than the number of tweets.

The appropriate prediction period K (months) differs depending on life events. Therefore, we set the prediction period K to 1 for the prediction in a short period and 6 for the prediction in a long period. Tables III shows the number of positive and negative examples in the training and test data for each life event in both K value.

Metric and Comparative Method: We used a precision-recall curve as the metric to evaluate the performance. If the position of the precision-recall curve of a method is closer to the upper right (i.e., both precision and recall are close to 1.0), the method is supposed to achieve better performance. As a comparative method, we employed the baseline method that uses the frequency of a given keyword shown in Table II. The baseline method sorts monthly tweet histories in the test data in order of the frequency of the keyword. Then, we plot a precision-recall curve by changing the threshold of the frequency.

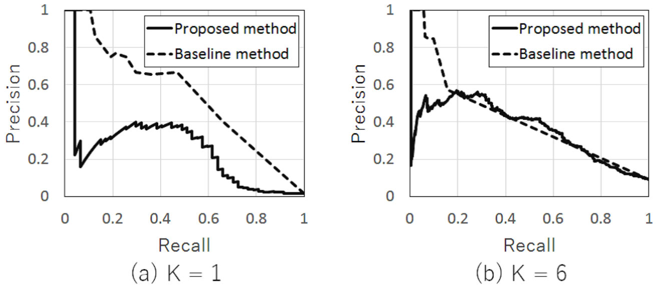

Fig. 4. Precision-recall curve for the “Leaving the hospital” event

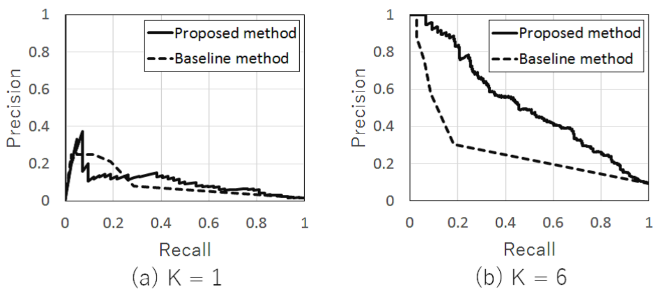

Fig. 5. Precision-recall curve for the “Pregnancy” event

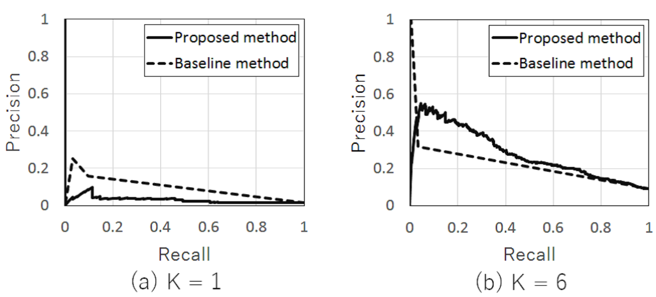

Fig. 6. Precision-recall curve for the “Marriage” event

B. Evaluation Results

Figures 2-6 show the evaluation results (precision-recall curves) for the five life events. In the following, we first compare the proposed method with the baseline method. We then discuss the evaluation results of the proposed method for the five life events.

1) Comparison with the Baseline Method: When $K = 6$, the precision-recall curves of the proposed method were positioned above those of the baseline method for all life events except the “Leaving the hospital” event. Since the baseline method predicts life events using only the frequency of given keywords, it cannot predict event users who do not use the keywords before the occurrence of the life events. On the other hand, the proposed method predicts life events using the word occurrence tendencies, which can capture the differences of the usage of various words between the event users and general users. As a result, the proposed method was able to predict the occurrence of user’s life events more accurately than the baseline method.

With respect to the “Leaving the hospital” event, the curve of the proposed method was not positioned above that of the baseline method (Figure 4 (b)). This result indicates that the word occurrence tendencies are not effective to discriminate event users from general users in this event. To achieve better performance, it is necessary to use other information, such as communication behavior between users and their followers.

For $K = 1$, the prediction performance of both the proposed and baseline methods were lower than the results in $K = 6$. In addition, the precision-recall curves of the proposed method were positioned on or below those of the baseline method. When $K = 1$, event user’s tweet histories before more than 1 month from the occurrence of the life events were labeled as negative examples. However, the proposed method does not consider the differences of the word usage of event users in different periods. Therefore, the proposed method does not work well if word occurrence tendencies of event users before and after 1 month are not different. In fact, the results of the proposed method for the “Giving birth” event contained 112 false positives, out of which 85 false positives were negative examples coming from event users’ tweet histories, at the point (0.625, 0.211) on the curve shown in Figure 2 (a). With respect to the “Leaving the hospital” event, the baseline method marked better performance than the proposed method. This result indicates that the event users tend to use the predefined keyword taiin frequently just before the occurrence of this event. Thus, we should take account of temporal changes of the word occurrence tendencies to achieve better performance.

From these result, we can confirm that the proposed method can effectively predict user’s life events when the prediction period K is long. This is because it captures the differences of word occurrence tendencies for any words. However, if the period K is short, the proposed method cannot work well because it is difficult to discriminate event users’ tweet histories before and after K months from the occurrence of life events. Thus, in the following, we discuss the result of each life event prediction for $K = 6$.

2) Discussion of Each Life Event Prediction: **For the “Giving birth” event (Figure 2)**, the precision-recall curve was highest among all life events, i.e., the proposed method achieved the best performance. This result indicates that the word occurrence tendencies of users who will give birth were very different from those of general users over the long period. This is because all users are generally in pregnancy for several months before they give birth. This result shows that the proposed method works well if users are in the same situation for a long time until they experience a given life event.

For the “Getting a job offer” event (Figure 3), the precision-recall curve was below the “Giving birth” curve. It can be considered that the period of job hunting differs among users who will get a job offer. In this paper, we defined the life event prediction as a binary classification problem with a boundary of K. Due to this definition, if the period in which event user-specific features appear differs among event users, the performance of the proposed method decreases. In fact, when we analyzed tweet histories of users who got a job offer, there was variation of time users spent for job hunting, i.e., from one month to several months. Therefore, it is considered that the proposed method could not extract significant features of event users from their tweet histories. As a result, the performance for this event was decreased compared to that for the “Giving birth” event. In addition, this result indicates that the appropriate period of K differs depending on life events.

For the “Leaving the hospital” event (Figure 4), the precision-recall curve was below that of the “Giving birth” event. It can be considered that the hospitalization period differs depending on event users, similar to the case of the “Getting a job offer” event. As a result, the performance of the “Leaving the hospital” event was lower than that of the “Giving birth” event. The precision-recall curve of the “Leaving the hospital” event is also below that of the “Getting a job offer” event. This result indicates that the hospitalization period tends to be shorter than the job hunting period.

For the “Pregnancy” event (Figure 5) , the precision-recall curve was below that of the “Giving birth” event as well as the “Getting a job offer” and “Leaving the hospital” events. This is because the time at which event users posted tweets about the occurrcence of their pregnancy varied for each user. In fact, some users posted tweets about the event occurrences when they were three months pregnant while others did when they were eight months pregnant. It is considered that the differences of the time when event users report their life event occurrences cause the dispersion of the word occurrence tendencies as well as the “Getting a job offer” and “Leaving the hospital” events. As a result, the performance became lower than the “Giving birth” event.

For the “Marriage” event (Figure 6), the proposed method demonstrated the lowest performance in all life events. Compared to other life events, we observed that the differences of the word occurrence tendencies between event users and general users were small. For example, the number of characteristic words, which were effective to discriminate the “Marriage” event users from general users, was obviously fewer than other events. We assume that users are unlikely to report, or even imply, their “Marriage” events before the event occurrences. Thus, it is difficult to predict the “Marriage” event using only word occurrence tendencies.

3) Revealed problems: Here, we discuss the problems of the proposed method revealed in our experiments. While the proposed method achieved high prediction accuracy for the “Giving birth” event, predicting the other life events seems to remain major challenges. For the “Getting a job offer,” “Leaving the hospital,” and “Pregnancy” events, the period in which discriminative features are likely to appear differs depending on event users. That is, the appropriate period K depends on life events as well as individual event users. We have to consider how to determine appropriate period K for each event and user. Otherwise, we should devise better problem formulation so that manual parameters such as K do not significantly affect the performance.

For the “Marriage” event, we found that using the word occurrence tendencies was not effective since discriminative features were hardly detected in them. In addition, when $K = 1$, the proposed method did not contribute to performance improvements for all life events because the proposed method did not consider the differences of the word usages of event users in different periods. We assume that the feature of word occurrences alone has limitations for predicting life events. Thus, we must consider other features. For example, the transition of the number of tweets that event users posted in each period, or the changes in the patterns of interactions with other users can be considered. If we consider such features, the performance might be improved because previous studies in life event extraction [6], [7] improved their performance by using these features.

VI. Conclution

In this study, we proposed a method that predicts the occurrence of Twitter user’s life events using word occurrence tendencies of tweet histories. We experimentally evaluated the proposed method for five life events (“Giving birth,” “Getting a job offer,” “Leaving the hospital,” “Pregnancy,” and “Marriage”). The results showed that when the prediction period K was set to six months, the proposed method predicted “Giving birth,” “Getting a job offer,” “Pregnancy,” and “Marriage” events more precisely than a baseline method that uses a predefined keyword related to the life events. Especially, the proposed mehod achieved the best performance for the “Giving birth” event. We found that the proposed method works well when users are in the same situation for a long time until they experience the life events.

In future, we plan to utilize other information as features. The proposed method only uses word occurrence tendencies to predict user’s life events. However, in some events, such as the “Marriage” event, event users do not use characteristic words compared to general users. For such events, communication patterns between users can be considered as an effective feature because this feature was effectively used in the life event extractions [6], [7]. In addition, we plan to utilize time series features, e.g., the number of tweets and communication behaviors in each period. Using time series features may improve the performance of life event prediction in a short period. Another challenge is to predict when a given user will experience a given life event using user’s tweet histories. Predicting directly the time when users will experience a given life event is a more difficult problem than the life event prediction we addressed in this paper because we need to consider the relationships between the time and temporal changes of the features.

Acknowledgments

This work was partially supported by JSPS Grant-in-Aid for Scientific Research (A)(JP26240013), JSPS Grant-in-Aid for Scientific Research (B)(17KT0082), and JST, Strategic International Collaborative Research Program, SICORP.

REFERENCES

[1] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," in *Proc. World Wide Web Conf. (WWW)*, 2010, pp. 851–860.

[2] J. Li and C. Cardie, "Early Stage Influenza Detection from Twitter," *arXiv preprint arXiv:1309.7340*, 2013.

[3] A. Hurriyetoglu, N. Oostdijk, and A. van den Bosch, "Estimating Time to Event from Tweets Using Temporal Expressions," in *Proc. Workshop on Language Analysis for Social Media (LASM)*, 2014.

[4] H. Tops, A. van den Bosch, and F. Kunneman, "Predicting Time-to-event from Twitter Messages," in *Proc. Benelux Conf. on Artificial Intelligence (BNAIC)*, 2013.

[5] H. Becker, D. Iter, M. Naaman, and L. Gravano, "Identifying Content for Planned Events Across Social Media Sites," in *Proc. Int'l Conf. on Web Search and Data Mining (WSDM)*, 2012, pp. 533–542.

[6] S. Choudhury and H. Alani, "Personal Life Event Detection from Social Media," *in Proc. Workshop on Social Personalisation (SP)*, 2014.

[7] T. Dickinson, M. Fernandez, L. A. Thomas, P. Mulholland, P. Briggs, and H. Alani, "Identifying Prominent Life Events on Twitter," in *Proc. Int'l Conf. on Knowledge Capture (K-CAP)*, 2015, pp. 4:1–4:8.

[8] M. Akbari, X. Hu, L. Nie, and T. Chua, "From Tweets to Wellness: Wellness Event Detection from Twitter Streams," in *Proc. AAAI Conf. on Artificial Intelligence (AAAI)*, 2016, pp. 87–93.

[9] H. Lin, J. Jia, L. Nie, G. Shen, and T. Chua, "What Does Social Media Say about Your Stress?" in *Proc. Int'l Joint Conf. on Artificial Intelligence (IJCAI)*, 2016, pp. 3775–3781.

[10] J. Li, A. Ritter, C. Cardie, and E. H. Hovy, "Major Life Event Extraction from Twitter based on Congratulations/Condolences Speech Acts," in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1997–2007.

[11] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in *Proc. European Conference on Machine Learning*, 1998, pp. 137–142.

[12] H. Kawai, A. Jatowt, K. Tanaka, K. Kunieda, and K. Yamada, "Chronoseeker: Search Engine for Future and Past Events," in *Proc. Int'l Conf. on Uniqiutous Information Management and Communication (ICUIMC)*, 2010, pp. 166–175.

[13] K. Radinsky, S. Davidovich, and S. Markovitch, "Learning Causality for News Events Prediction," in *Proc. Int'l World Wide Web Conf. (WWW)*, 2012, pp. 909–918.

[14] Y. Tsuboi, A. Jatowt, and K. Tanaka, "Product Purchase Prediction Based on Time Series Data Analysis in Social Media," in *Proc. Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2015, pp. 219–224.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.