Title	マイクロブログからの典型的使用場面付き辞書の構築
Author(s)	岡，利成
Citation	
Issue Date	2020-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/16386
Rights	
Description	Supervisor: 白井　清昭，先端科学技術研究科，修士（情報科学）

Construction of Lexicon with Typical Situations of Words from Microblog

1810033 Toshinari Oka

A lexicon is a database of words, which contains information of words such as pronunciation, part of speech, and synonym. It is one of the fundamental knowledge for natural language processing. In particular, it is very useful for various researches to construct a lexicon that compiles words with time or place where each word is frequently used or with the information of people who frequently use each word. The goal of this research is to automatically construct such a lexicon, i.e. a lexicon with typical situations of words. This study considers three types of typical situations of words: (1) time and (2) place where a word is frequently used as well as (3) job of a person who frequently uses a word. For example, a typical situation of time of the word "breakfast" is morning. In this study, categories of typical situations are defined as follows: [morning], [noon], [evening], [night], and [midnight] for time, 47 prefectures for place, and 44 representative jobs such as [doctor] and [teacher]. From Twitter, one of the microblogs, we collect texts with information of time, place and job, and identify typical situations where the words frequently appear in the text.

Although several previous studies aimed at identifying typical situations of words and applied them to specific applications of natural language processing, no attempt has been made to construct a lexicon with typical situations of words that can be widely used in general. In addition, time or place was considered as a typical situation in the previous work, but a job has not been considered as a typical situation. This study is the first attempt to consider the job as the typical situations of words. Furthermore, although typical situations of nouns were considered so far, we also consider other types of words, that is, verbs, adjectives, adverbs, and hashtags, as words to be compiled in a lexicon.

In the proposed method, a lexicon with typical situations of words is constructed by the following procedures. First, tweets annotated with a category of time, place, and job are collected using TwitterAPI Tweepy. As for the place, tweets are searched using the place name code of a prefecture as a query, then tweets annotated with the category of the prefecture are retrieved. As for a job, we collect tweets posted by job users. A "job user" is a Twitter user who has a certain occupation (job), which is automatically collected by our proposed method described later. As for the time, since a time stamp is attached to all tweets as metadata in Twitter, we reuse a collection of tweets with place categories as tweet data with time categories. Next, after several preprocessings, word segmentation and part of speech tagging are performed on tweets, then nouns, verbs, adjectives, adverbs, and hashtags are extracted as candidates of words to be compiled in a lexicon with typical situations.

Typical situations of these candidates of words (time, place, or job that are highly associated with a word) are identified by the following three methods. The first is a method using Pointwise Mutual Information (PMI). Correlation (co-occurrence) between words and categories is measured by PMI to identify typical situations of words. The second method is based on Kleinberg's burst detection algorithm. Regarding a sequence of categories as a virtual time series, we identify the time period (corresponding to the category) where the frequency of a word is sharply increased by Kleinberg's method. The detected category is set as the typical situation of the word. The frequency of words in each category is measured by the number of tweets containing a word. Hereafter, we call this method "Kleinberg-tweet method". The third method is also based on Kleinberg's burst detection algorithm, but the frequency of words in each category is measured by the number of users who use the word. Hereafter, we call this method "Kleinberg-user method". Finally, when the score calculated by each method is greater than a predefined threshold, the category is specified as a typical situation of the word, then the words and their identified typical situations are added to the lexicon.

In the above method, a list of job users is required to obtain tweets annotated with jobs. It is obtained semi-automatically as follows. First, for a given job, user profiles in Twitter account are searched using the job name as a query. Then, a user who actually has the job (we call it as a seed job user) is manually selected. Next, the followees of the seed job user is retrieved. When the profile of the followee includes the job name, it is added to the list as a new job user. This procedure is repeated recursively until 100 job users are obtained.

Several experiments are conducted to evaluate the proposed method. First, we evaluated the method of identifying typical situations of words using PMI. For the place category, we selected several pairs of a word and its identified typical situation (place category) and manually evaluated whether they were appropriate. However, the accuracy was insufficient, i.e. 0.26. Next, we evaluated the Kleinberg-tweet method. The top 20 words of the score calculated by the Kleinberg-tweet method for each category were manually evaluated whether they are appropriate by two subjects. Then the accuracy, the ratio of the words that are judged as appropriate to the total number of words, is measured. The accuracy was 0.44, 0.79, and 0.52 for the time, place, and job category, respectively. On the other hand, the κ coefficients of judgement of two subjects were 0.76, 0.77, and 0.53 for time, place, and job, respectively. Similarly, the Kleinberg-user method was evaluated. The accuracy was 0.59, 0.90, and 0.92 for the time, place, and job category, respectively, which outperformed the Kleinberg-tweet method. The κ coefficients of judgement of two subjects were 0.83, 0.41, and 0.49 for time, place, and job, respectively. In the Kleinberg-tweet method, when one user use the same word many times, the number of tweets of that word increases, and the word is wrongly detected as it is highly related to a specific category. In contrast, since the Kleinberg-user method detects words that are used by many users, such errors are suppressed. Finally, using the Kleinberg-user method, we constructed a lexicon consisting of 1,152 words with time typical situations, 6,793 words with place typical situations, and 199,475 words with job typical situations.