ATLANTIS PRESS International Journal of Networked and Distributed Computing, Vol. 5, No. 4 (October 2017) 243–253

Estimating and Visualizing the Time-varying Effects of a Binary Covariate on Longitudinal Big Text Data

Shizue Izumi [1], Tetsuji Tonda [2], Noriyuki Kawano [3], Kenichi Satoh [4]

[1] *Faculty of Data Science, Shiga University,*
1-1-1 Banba,
Hikone, 522-8522, Japan
E-mail: shizue-izumi@biwako.shiga-u.ac.jp

[2] *Faculty of Management and Information System, Prefectural University of Hiroshima,*
1-1-71 Ujina-Higashi,
Minami-ku, Hiroshima, 734-8558, Japan
E-mail: ttetsuji@pu-hiroshima.ac.jp

[3] *Institute for Peace Science, Hiroshima University,*
1-1-89 Higashisenda-machi,
Naka-ku, Hiroshima, 730-0053, Japan
E-mail: nkawano@hiroshima-u.ac.jp

[4] *Research Institute for Radiation Biology and Medicine, Hiroshima University,*
1-2-3 Kasumi,
Minami-ku, Hiroshima, 734-8553, Japan
E-mail: ksatoh@hiroshima-u.ac.jp

Abstract

We propose a method to estimate and visualize effects of a binary covariate on the longitudinally observed text data. Our method consists of series of analytical methods: extracting keywords through a morphological analysis, estimating the time-varying regression coefficient of a binary covariate for keyword's appearance and frequency, classifying summary of estimates, and visualizing the time-varying effects of a binary covariate in animated scatter plots. The procedure is demonstrated with Peace Declaration text data observed for forty years in two cities.

Keywords: Text mining, Summarization, Semiparametric modeling, Classification, Animation, Knowledge discovery.

1. Introduction

In risk communication at disasters written texts to social networking services like Facebook are attracted to attention as big data. And these texts can be treated as longitudinally observed text data. Extraction of the time trends of keyword's frequency and its classification can summarize the changes of characteristics in longitudinal text data. Such technologies are expected to be applicable for various fields including protection against natural disaster, marketing, and social sciences.

A general method of capturing the characteristics of longitudinal text data is to measure keyword's appearance or frequency at each observed time point and examine the changes of measures. Izumi, Satoh, and

Kawano[1] proposed a method to visualize keyword's appearance in the longitudinal text data, using a series of statistical approaches. In following Sec. 2 we propose a method to estimate and visualize time-varying effects of a binary covariate on the longitudinally observed text data, as an extension of the method by Izumi, Tonda, Kawano, and Satoh.[2] Our proposed method is applied to real data in Sec. 3, and practical interpretations of the results are discussed in Sec. 4.

2. Estimation and Visualization Method for Time-varying Effects of a Binary Covariate on Longitudinal Text Data

We propose a method to estimate and visualize effects of a binary covariate on an outcome of interests in the longitudinally observed text data. Our method consists of series of analytical methods: extracting keywords through a morphological analysis, estimating a time-varying regression coefficient of a binary covariate for each keyword, classifying summary estimates, and visualizing the time trends of binary covariate effects. Here we consider two types of outcome: keyword's frequency and appearance. When keyword's frequency is fairly low, one may be rather interested in whether a keyword appears in the text observed at a time point or not. In Sec. 2.1, we explain the extraction of keywords and estimation of a time-varying regression coefficient. In Sec. 2.2, we explain the classification of summary estimates and the visualization of the time trends of binary covariate effects.

2.1. Estimating time-varying effects of covariates on the longitudinal text data

First the m highest frequent words are extracted as keywords through a morphological analysis of text data observed at n time points where $t_1 < t_2 < \cdots < t_{n-1} < t_n$. A $(n \times m)$ matrix, $Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$ is created with a $(n \times 1)$ column vector, $\boldsymbol{y}_i = (y_i(t_1), \ldots, y_i(t_n))'$. When an outcome of our interests is keyword's frequency, let $y_i(t_j)$ be frequency at the j-th time point ($j = 1, \ldots, n$) for the i-th keyword ($i = 1, \ldots, m$). On the other hand, when an outcome of our interests is keyword's appearance, let $y_i(t_j)$ be appearance (1: appear, 0: not appear) at the j-th time point ($j = 1, \ldots, n$) for the i-th keyword ($i = 1, \ldots, m$). In later notations the subscripts i and j are omitted from $y_i(t_j)$ so that the outcome variable is denoted as $y(t)$.

Secondly we apply a statistical model with a time-varying coefficient for longitudinal data to estimate effects of covariates on an outcome at observed time points. While a Poisson regression model is considered to keyword's frequency, a Logistic regression model is considered to keyword's appearance. Statistical modeling is explained in the following sub-subsections: 2.1.1 - 2.1.3.

2.1.1. Keyword's frequency as an outcome

Let $y(t)$ be keyword's frequency at time point t as an outcome variable, and let the time-varying regression coefficients of covariates $a_1(t), \ldots, a_p(t)$ be $\beta_1(t), \ldots, \beta_p(t)$, respectively. When the outcome variable is assumed to be distributed as Poisson distribution, a Poisson regression model with an offset can be described by

$$\log\left(E\left(y(t)\right)\right) = \log\left(w(t)\right) + \sum_{k=1}^{p} \beta_k(t) a_k(t), \tag{1}$$

where the offset is logarithm of observed total word counts $w(t)$. The regression coefficient $\beta_k(t)$ expresses time-varying effects of the corresponding covariate $a_k(t)$. When the covariate $a_k(t)$ is binary, $\exp(\beta_k(t))$ is interpreted as relative keyword's frequency regarding the covariate $a_k(t)$ adjusted for total word counts and other covariates.

2.1.2. Keyword's appearance as an outcome

Let $y(t)$ be keyword's appearance at time point t as an outcome variable. When the outcome variable is assumed to be distributed as Bernoulli distribution, a Logistic regression model can be described by

$$\mathrm{logit}\left(\Pr\left(y(t)=1\right)\right) = \sum_{k=1}^{p} \beta_k(t) a_k(t), \tag{2}$$

where the time-varying regression coefficients of covariates $a_1(t), \ldots, a_p(t)$ are $\beta_1(t), \ldots, \beta_p(t)$, respectively. When the covariate $a_k(t)$ is binary, $\exp(\beta_k(t))$ is interpreted as odds ratio of keyword's appearance regarding the covariate $a_k(t)$ adjusted for other covariates.

2.1.3. Design of a semiparametric coefficient

Regardless the outcome variable is keyword's frequency or appearance, the design of a regression coefficient is common for both models (1) and (2). When the number of time points is large, we can consider a semiparametric coefficient that contains two types of basis covariates, $x(t)$ for linearity and $z(t)$ for

nonlinearity. Combining the two design vectors, the semiparametric coefficient for the k-th covariate can be given by

$$\beta_k(t) = x(t)' \ b_k + z(t)' \ u_k, \quad k = 1, \ldots, p, \tag{3}$$

where the regression coefficient vectors of covariates $x(t)$ and $z(t)$ are b_k and u_k, respectively. In order to make an interpretation easy, we use a simple design vector $x(t) = (1, t)'$ and $z(t) = ((t - \kappa_1)_+, \ldots, (t - \kappa_r)_+)'$ with r ($< n - 2$) knots of $\kappa_1, \ldots, \kappa_r$. Here $(t - \kappa)_+$ is a spline function with knot κ that takes the value $(t - \kappa)$ if $(t - \kappa)$ is positive, otherwise zero. The number of knots, r and arrangement of knots can be decided by the distribution of the observed time points. For example, if the knots are quartiles corresponding to the probabilities 1/4, 2/4, and 3/4, then the number of knots becomes three. We use the method proposed by Brumback, Ruppert, and Wand[3] to estimate these coefficients for a generalized linear mixed effects model assuming the regression coefficient vector u_k to be a random vector. In reality when the number of knots r is much smaller than the number of time points ($r << n$), the computation of the maximum-likelihood method is stable.

We may examine whether the coefficient $\beta_k(t)$ is zero over time (i.e., no effects of the k-th covariate) and whether the coefficient $\beta_k(t)$ is constant over time (i.e., effects of the k-th covariate do not depend on time), in order to evaluate the covariate effects. A hypothesis test on $\beta_k(t)$ can be performed with a significance level of 0.05 or 95% simultaneous confidence interval (CI) of $\beta_k(t)$ can be computed to solve these questions.[4,5] A function of *glmmPQL* in package *MASS* is used to fit linear and nonlinear mixed effects models in the statistical software *R* (See Refs. [6, 7] for more details). It provides the estimates $\hat{\beta}_k(t)$ of regression coefficient $\beta_k(t)$ and their standard errors.

2.2. Visualizing time-varying effects of a binary covariate on the longitudinal text data

Thirdly we extract a time trend of the k-th binary covariate effects using summary of the estimates obtained in the previous Sec. 2.1. A vector of estimates at start, knots, and end of time point, i.e., ($\hat{\beta}_k(t_1)$, $\hat{\beta}_k(\kappa_1)$, . . . , $\hat{\beta}_k(\kappa_r)$, $\hat{\beta}_k(t_n)$) is used as a summary vector of the estimated regression coefficient curve ($\hat{\beta}_k(t_1)$, $\hat{\beta}_k(t_2)$, . . . , $\hat{\beta}_k(t_{n-1})$, $\hat{\beta}_k(t_n)$). The length of the summary vector is $r + 2$ ($r + 2 << n$). Then this summary vector represents a time trend of the k-th binary covariate effects on each keyword.

Fourthly we classify the summary vectors of estimated regression coefficient curve to group keywords with a similar time trend of covariate effects, using k-means method. The number of groups (K) at k-means method should be minimized as long as it is interpretable for a real situation. A scatter plot of group means with calendar time can show a primary time trend of covariate effects in each group.

Finally we create HTML based animations with a series of scatter plots at each time point. When an outcome variable is keyword's frequency, a keyword is plotted with an estimate of binary covariate's regression coefficient (i.e., logarithm of relative frequency) as X-coordinate and maximum of predicted frequency per 100 words between two categories in the binary covariate as Y-coordinate. When an outcome variable is keyword's appearance, a keyword is plotted with an estimate of binary covariate's regression coefficient (i.e. logarithm of odds ratio) as X-coordinate and maximum of predicted probability of keyword's appearance between two categories in the binary covariate as Y-coordinate. Color of keywords in the plot reflects a corresponding group found by k-means method. In this way the animation of scatter plots is more visible. A function of *saveHTML* in the package *animation* is used to insert plot images in *R* programs into an HTML page.[8] Further a function of *saveGIF* in the package *animation* can be used to generate a sequences of plot images in *R* programs into a GIF file. Therefore the animations show the sign and magnitude of covariate effects, a time trend of covariate effects, the similarity of the time trends, and keyword's frequency (or probability of keyword's appearance) simultaneously.

3. Application to Example Data

In this section we demonstrate the application of our proposed method in Sec. 2 to real data. As an example we use the English translated text data observed at forty time points between 1977 and 2016. The original text data written in Japanese is body of Peace Declaration announced by mayors of Hiroshima and Nagasaki cities in an annual Peace Memorial Ceremony held on August 6 in Hiroshima and on August 9 in Nagasaki. During the period of observation four mayors served in each city. The concept of peace has been discussed by Matsuura, Satoh, and Kawano[9,10] that analyzed original Peace Declaration as longitudinal text data. Our interests are to compare a time trend of characteristics between two cities.

A research question in this paper is whether there are city effects on two types of an outcome: keyword's frequency and appearance. In other word whether keyword's frequency differs by city (eg., some keyword's frequency in Nagasaki has been higher than Hiroshima during the observed period) or whether keyword's appearance differs by city (e.g., some keywords in Nagasaki have appeared at more time points than Hiroshima).

Another question is whether keywords can be grouped according city effects. In other words what keywords have a similar time trend (e.g., frequencies of some keywords in Nagasaki are constantly higher than those in Hiroshima during the observed period).

In order to solve the above questions, we create a csv file containing three columns of city, calendar year, and body of peace declaration as longitudinal text data. City is coded as 1 for Nagasaki and 0 for Hiroshima. Calendar year is coded from 1977 to 2016. Original form of words are extracted in a morphological analysis of the former data using KHcoder.[11] Total word counts range from 451 to 966 in Hiroshima and from 479 to 1,253 in Nagasaki. Pearson correlation coefficient between calendar year and total word counts is 0.81 in Hiroshima and 0.48 in Nagasaki.

First frequent 53 nouns (more than 51 appearances in total) are defined as keywords through a morphological analysis of text data. Appearance of 53 keywords at observed time points ranges from 25 to 41 in Hiroshima and from 29 to 42 in Nagasaki. Of these 53 keywords, the top 25 are examined for their frequency, while the rest 28 are examined for their appearance.

3.1. Estimate and visualize time-varying effects of city on keyword's frequency in the Peace Declaration data

Secondly we apply a semi-parametric Poisson regression for keyword's frequency. Let y(t) be frequency of a keyword at time t, $a_1 = 1$ be constant, and a_2 be city expressing 0 for Hiroshima and 1 for Nagasaki. Basis covariates $x(t) = (1, t)'$ and $z(t) = \{(t - 1987)_+, (t - 1997)_+, (t - 2007)_+\}'$ are used to express linearity and nonlinearity, respectively. Years 1987, 1997, and 2007 are the knots selected on the basis of quartiles. Denote an offset be logarithm of total word counts $w(t)$. A Poisson regression model for the frequency $y(t)$ at time t, its semiparametric time-varying coefficients, and estimate of time-varying coefficient for city, $\hat{\beta}_2(t)$ can be individually described by

$$\log\left(E\left(y\left(t \mid a_2\right)\right)\right) = \log\left(w\left(t \mid a_2\right)\right) + \beta_1(t) + \beta_2(t)a_2 \quad (4)$$

$$\beta_k(t) = x(t)'\mathrm{b}_k + z(t)'\mathrm{u}_k, \quad k = 1, 2, \quad (5)$$

$$\hat{\beta}_2(t) = \log\left(\left[\hat{y}(t \mid a_2 = 1) / w(t \mid a_2 = 1)\right] / \left[\hat{y}(t \mid a_2 = 0) / w(t \mid a_2 = 0)\right]\right) \quad (6)$$

where the regression coefficient vectors of covariates $x(t)$ and $z(t)$ are b_k and u_k, respectively, and u_k is assumed to be distributed as multivariate normal distribution MN (0, $\sigma^2\boldsymbol{I}$). Equations (1) and (3) in Sec. 2 are replaced with Eqs. (4) and (5), respectively. The estimates $\hat{\beta}_2(t)$ of regression coefficient $\beta_2(t)$ are obtained by fitting linear and nonlinear mix effects models. When $\beta_2(t)$ is positive, keyword's frequency in Nagasaki is higher than that in Hiroshima. When $\beta_2(t)$ is zero, keyword's frequency in Nagasaki is the same as that in Hiroshima. When $\beta_2(t)$ is negative, keyword's frequency in Nagasaki is lower than that in Hiroshima.

Figures 1-3 show the observed and predicted frequencies of selected keywords by city, where a circle and a triangle are observed, and a solid line and a dash line are predicted from the models. Figures 4-6 show the estimates and their 95% simultaneous CIs of regression coefficient for the corresponding keywords in Figs. 1-3. Estimates and their 95% simultaneous CIs are expressed with a solid line and a dash line, respectively.

Thirdly we extract a time trend of the city effects using a summary of estimates. A vector of estimated regression coefficient curve ($\beta_2(1977)$, $\beta_2(1978)$, $\beta_2(1979)$, … , $\beta_2(2014)$, $\beta_2(2015)$, $\beta_2(2016)$) is summarized with ($\hat{\beta}_2(1977)$, $\hat{\beta}_2(1987)$, $\hat{\beta}_2(1997)$, $\hat{\beta}_2(2007)$, $\hat{\beta}_2(2016)$).

Fourthly we classify the summary vectors of estimated regression coefficient curve to group keywords with a similar time trend of covariate effects, using k-means method (K=3). When K=3, an initial value of random number hardly affects the results of grouping. Figure 7 is a scatter plot of group means of estimates with calendar year. Cluster $C1$ { Hiroshima, hibakusha } is plotted with a red solid line. Cluster $C2$ { weapon, world, peace, year, bombing, government, war, city, abolition, United, victim, disarmament, today, time } is plotted with a black dash line. Cluster $C3$ { people, Nagasaki, bomb, Japan, country, survivor, citizen, state } is plotted with a blue dotted line.

Finally, we create HTML based animations with a series of scatter plot. As a brief version of animations, Figures 8-12 show scatter plots of keywords at years of 1977, 1987, 1997, 2007, and 2016. Colors of keywords: red, black, and blue in Figs. 8-12 reflect those of $C1$-$C3$ in Fig. 7.

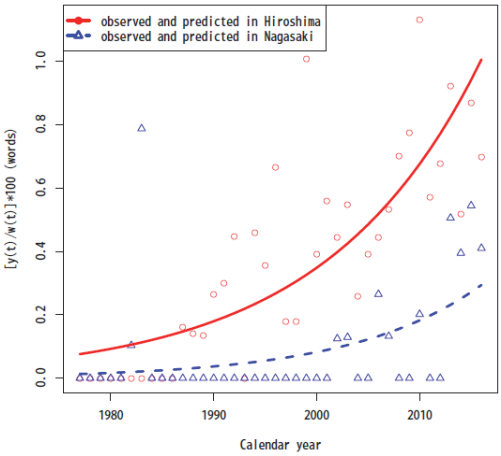

Fig. 1. Observed and predicted frequencies of "hibakusha"

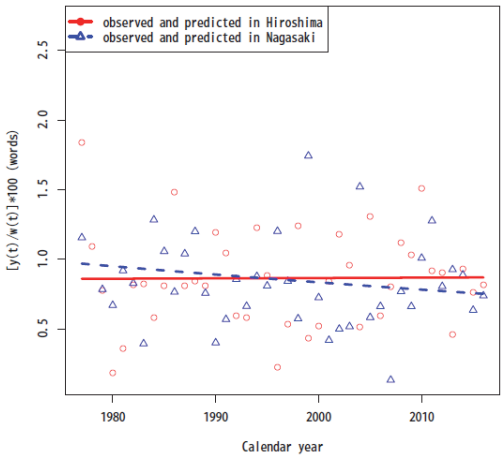

Fig. 2. Observed and predicted frequencies of "world"

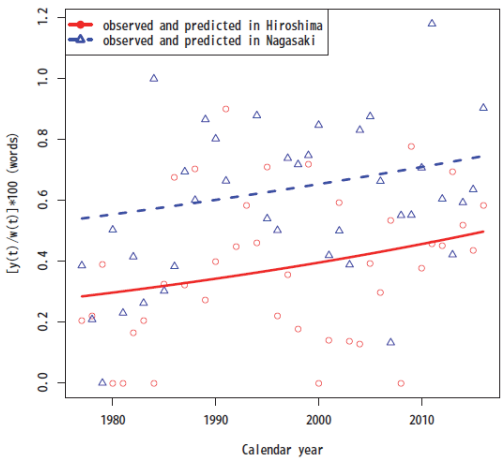

Fig. 3. Observed and predicted frequencies of "people"

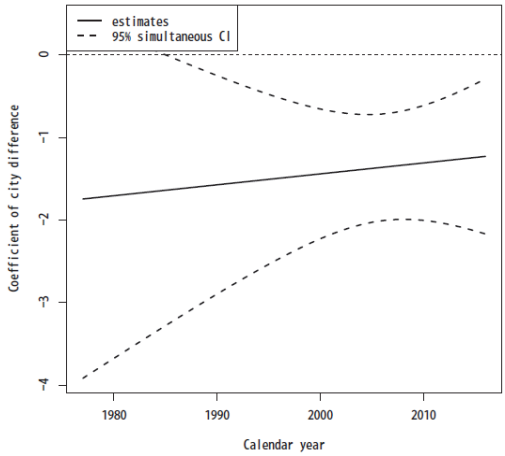

Fig. 4. Estimates and 95% simultaneous CI of regression coefficient in "hibakusha"

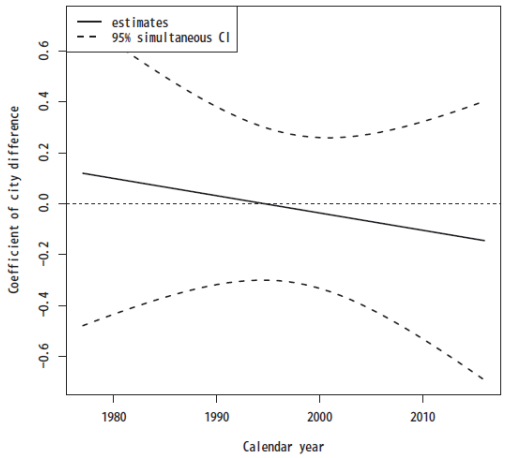

Fig. 5. Estimates and 95% simultaneous CI of regression coefficient in "world"

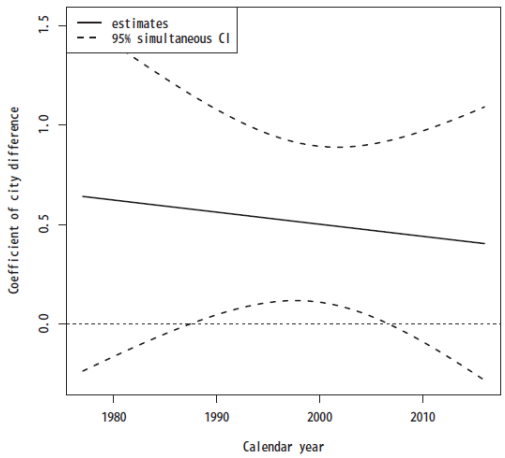

Fig. 6. Estimates and 95% simultaneous CI of regression coefficient in "people"

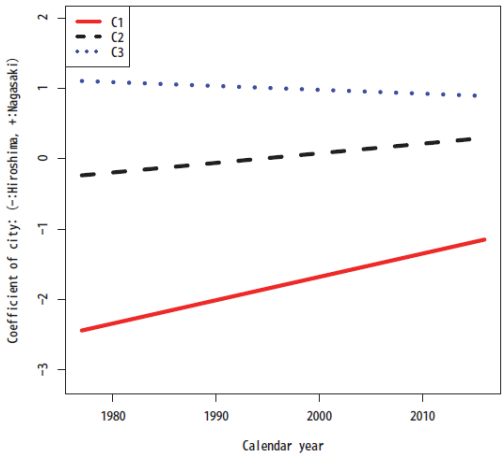

Fig. 7. A grouped time trend of city effects on keyword's frequency when K=3

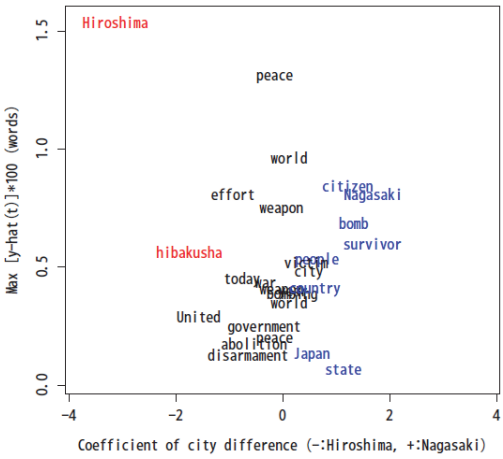

Fig. 8. Scatter plot of keyword's frequency by cluster in 1977

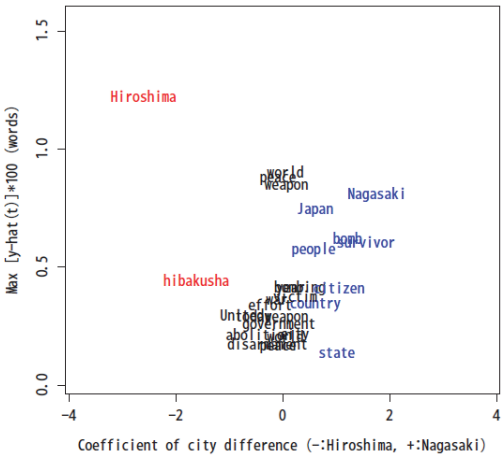

Fig. 9. Scatter plot of keyword's frequency by cluster in 1987

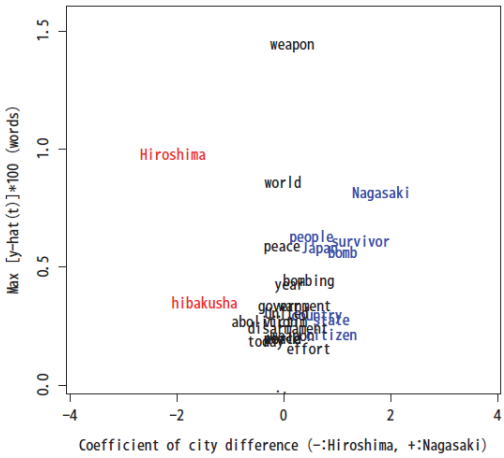

Fig. 10. Scatter plot of keyword's frequency by cluster in 1997

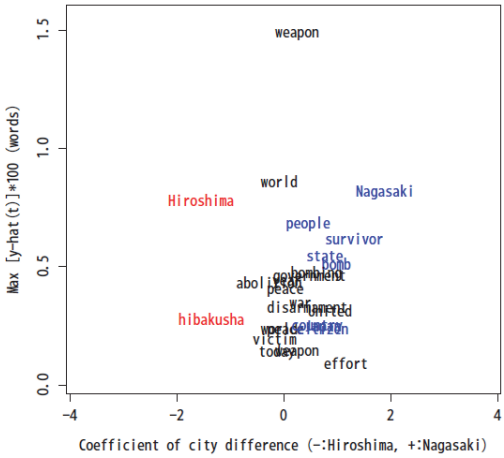

Fig. 11. Scatter plot of keyword's frequency by cluster in 2007

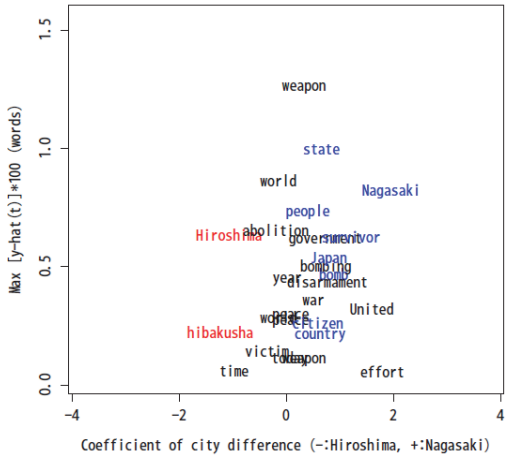

Fig. 12. Scatter plot of keyword's frequency by cluster in 2016

3.2. Estimate and visualize time-varying effects of city on keyword's appearance in the Peace Declaration data

Here we apply a semi-parametric Logistic regression for probability of keyword's appearance. Covariates a_1, a_2, $x(t)$, and $z(t)$ are similarly defined to Subsection 3.1. A Logistic regression model for the probability of keyword's appearance $\Pr(y(t) = 1)$ (y(t) = 1 if keyword appears, otherwise 0) at time t and estimate of time-varying coefficient for city, $\hat{\beta}_2(t)$ can be described by

$$\text{logit}\left[\Pr(y(t)=1\,|\,a_2)\right] = \beta_1(t) + \beta_2(t)a_2\,, \tag{7}$$

$$\hat{\beta}_2(t) = \text{logit}\left[\Pr\left(\hat{y}(t)=1\,|\,a_2=1\right)\right] - \text{logit}\left[\Pr\left(\hat{y}(t)=1\,|\,a_2=0\right)\right]. \tag{8}$$

Equation (2) in Sec. 2 is replaced with Eq. (7). Similar to Subsection 3.1, the estimates $\hat{\beta}_2(t)$ of regression coefficient $\beta_2(t)$ are obtained by fitting linear and nonlinear mix effects models. When $\beta_2(t)$ is positive, probability of keyword's appearance in Nagasaki is higher than that in Hiroshima. When $\beta_2(t)$ is zero, probability of keyword's appearance in Nagasaki is the same as that in Hiroshima. When $\beta_2(t)$ is negative, probability of keyword's appearance in Nagasaki is lower than that in Hiroshima. In addition when $\beta_2(t)$ is constant, city effects do not depend on time.

Figures 13-15 show the observed and predicted probability of appearance, $p(t) = \Pr\left(y(t)=1\right)$ for selected keywords by city, where a circle and a triangle are observed, and a solid line and a dash line are predicted from the models. Figures 16-18 show the estimates and their 95% simultaneous CIs of regression coefficient for the corresponding keywords in Figs. 13-15. Estimates and their 95% simultaneous CIs are expressed with a solid line and a dash line, respectively.

Thirdly we extract a time trend of the binary covariate effects on probability of keyword's appearance using summary of estimates. A vector of estimated regression coefficient curve is summarized with $(\hat{\beta}_2(1977), \hat{\beta}_2(1987), \hat{\beta}_2(1997), \hat{\beta}_2(2007), \hat{\beta}_2(2016))$.

Fourthly we classify the summary vectors of estimated regression coefficient curve to group keywords with a similar time trend of covariate effects, using k-means method (K=3). When K=3, an initial value of random number hardly affects the results of grouping. Figure 19 is a scatter plot of group means of estimates with calendar year. Cluster CC1 { Nuclear, use, Nation, principle, State } is plotted with a red solid line. Cluster CC2 { nation, Peace, experience, power, anniversary, day, soul, century, appeal, age, elimination, leader, August, future, humanity, race} is plotted with a black dash line. Cluster CC3 {life, test, arm, law, radiation, history, repose} is plotted with a blue dotted line.

Finally, we create HTML based animations with a series of scatter plot. As a brief version of animations, Figures 20-24 show scatter plots of keywords at years 1977, 1987, 1997, 2007, and 2016. Colors of keywords: red, black, and blue in the Figs. 20-24 reflect those of CC1-CC3 in Fig. 19.

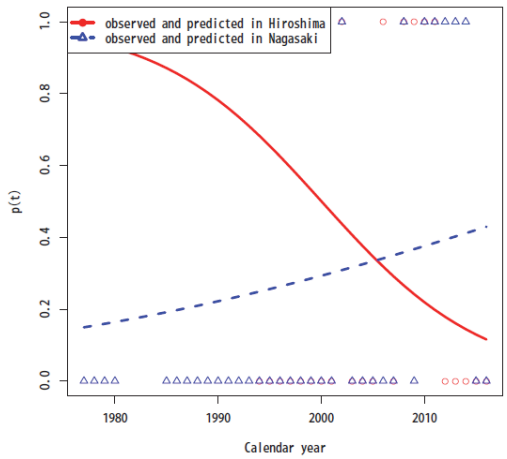

Fig. 13. Observed and predicted probability of appearance for "State"

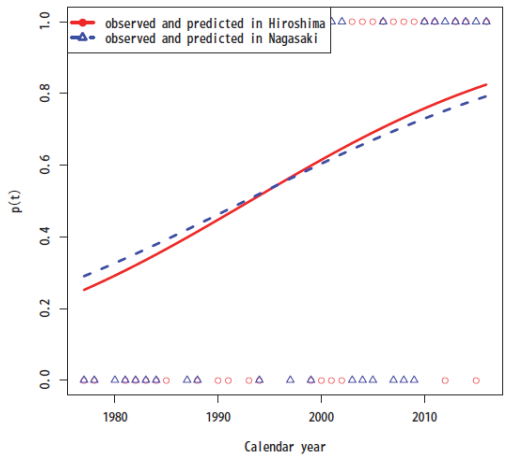

Fig. 14. Observed and predicted probability of appearance for "age"

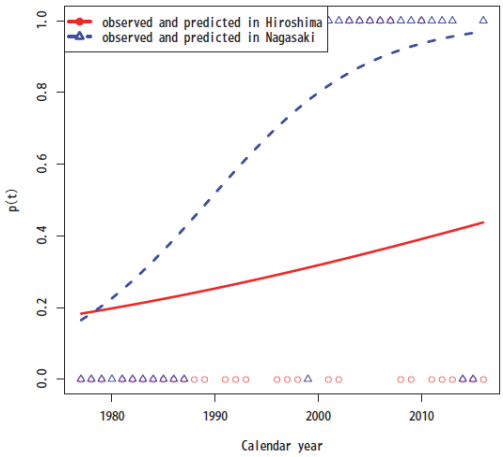

Fig. 15. Observed and predicted probability of appearance for "law"

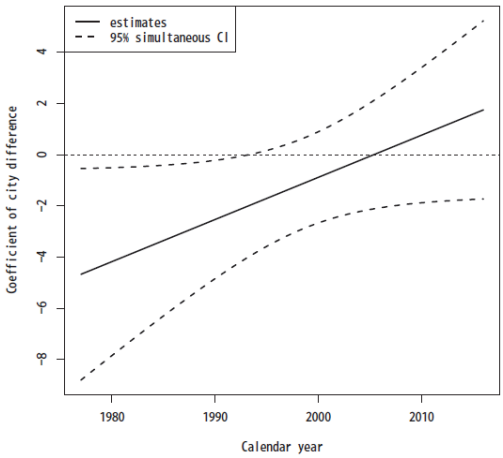

Fig. 16. Estimates and 95% simultaneous CI of regression coefficient in "State"

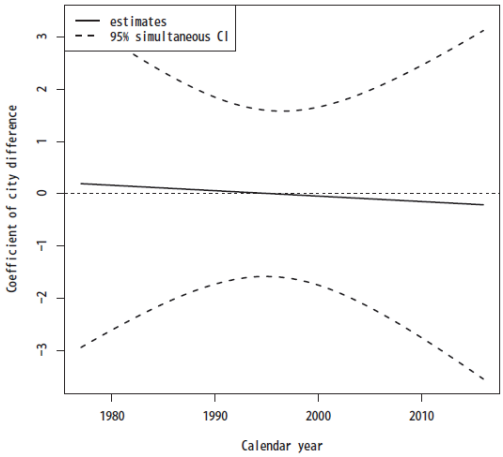

Fig. 17. Estimates and 95% simultaneous CI of regression coefficient in "age"

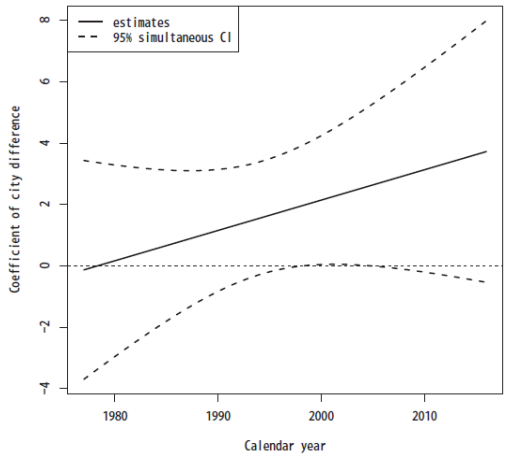

Fig. 18. Estimates and 95% simultaneous CI of regression coefficient in "law"

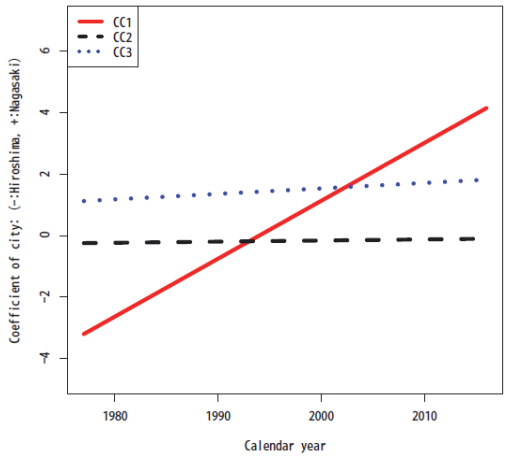

Fig. 19. A grouped time trend of city effects on keyword's appearance when K=3

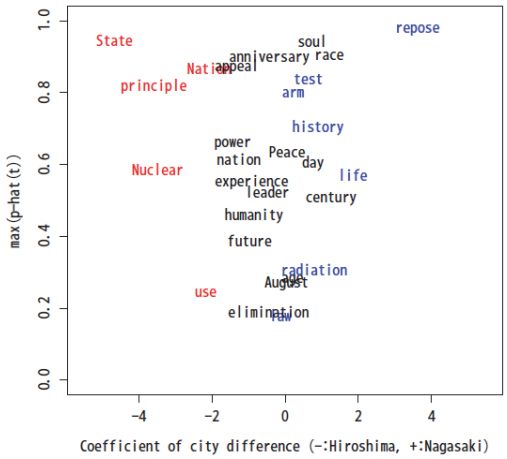

Fig. 20. Scatter plot of keyword's appearance by cluster in 1977

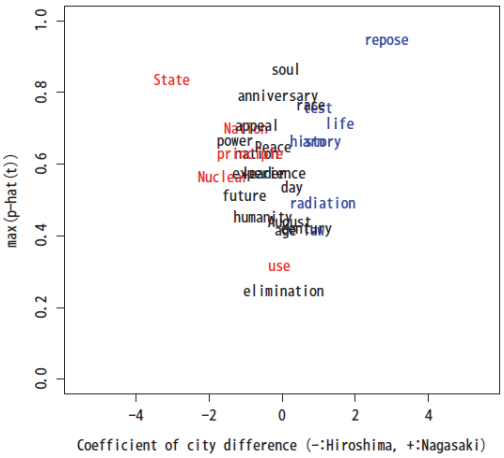

Fig. 21. Scatter plot of keyword's appearance by cluster in 1987

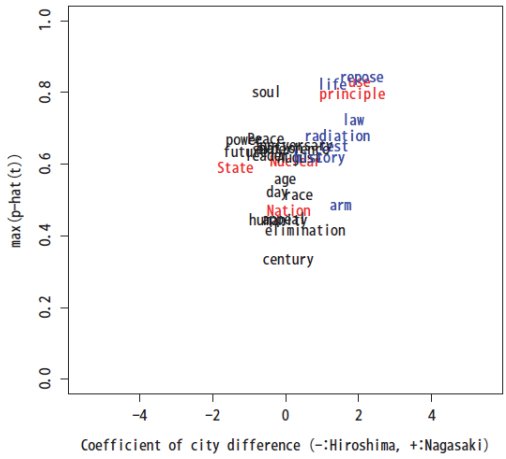

Fig. 22. Scatter plot of keyword's appearance by cluster in 1997

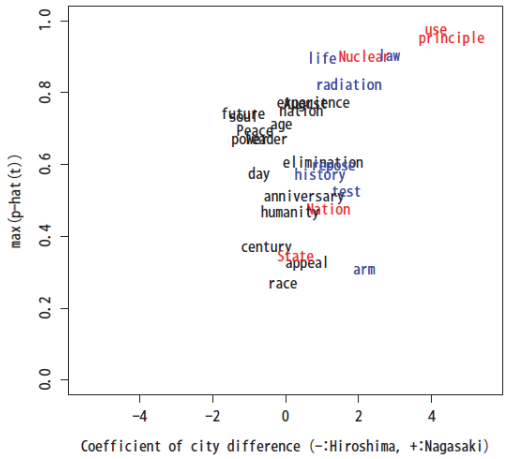

Fig. 23. Scatter plot of keyword's appearance by cluster in 2007

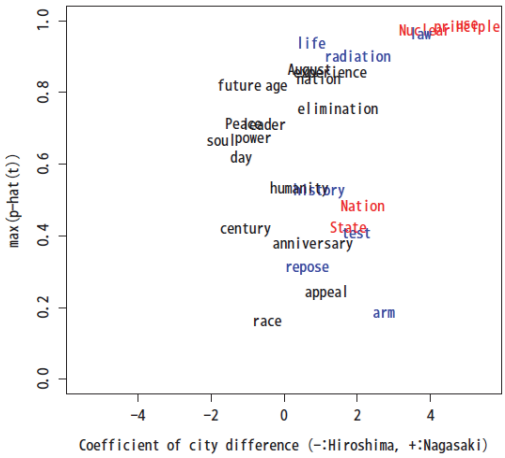

Fig. 24. Scatter plot of keyword's appearance by cluster in 2016

4. Discussion

We propose a method to estimate and visualize effects of a binary covariate on longitudinally observed text data. First we extract keywords through a morphological analysis of the text data. Secondly we estimate a time-varying regression coefficient of the binary covariate with a semiparametric mix effects model. We consider a Poisson regression model with an offset for keyword's frequency and a Logistic model for keyword's appearance. Thirdly a time trend of the covariate effects is extracted using a summary vector of estimates at the start, knots, and end of time points. Fourthly the summary vectors of estimated regression coefficient curve are classified to group keywords with a similar time trend of covariate effects. Finally HTML based animations are created with scatter plots of keywords to visualize the sign and magnitude of covariate effects, a time trend of covariate effects, the similarity of the time trends, and keyword's frequency (or probability of keyword's appearance) simultaneously.

In an application of our proposed method to a real data set, a time trend of city effects on the English translated Peace Declaration observed in forty years is classified into some groups. According to a summary vectors of estimates, selected 25 keywords are classified into three groups such as no effects, positive effects, and negative effects on keyword's frequency. Other selected 28 keywords are also classified into three groups such as no effects, positive effects, and both negative and positive effects on keyword's appearance. Further a

estimate of time-varying regression coefficient for city, keyword's frequency (or appearance) , and groups of a time trend are simultaneously visualized in the animations. As shown in the animations (Figs. 8-12) using keyword's frequency, red keywords in C1 stay left, which implied that these words are more pronounced in Hiroshima in earlier years. Black keywords in C2 stay around zero in coefficient of city, which may represent unchangeable concept of peace. Blue keywords in C3 stay right, which implies that these words are more pronounced in Nagasaki. On the other hand, shown in the animations (Figs. 20-24) using keyword's appearance, red keywords in CC1 move from left to right, which may be influenced by world affairs . Black keywords in CC2 stay around zero in coefficient of city, which may represent unchangeable concept of peace. Blue keywords in CC3 stay right, which implies that these words are more pronounced in Nagasaki.

Practical interpretation of the results from real data is important to examine the appropriateness of our proposed method. For example, it is easy to understand that "Hiroshima" in C1 is more frequently used in Hiroshima and that "Nagasaki" in C3 is more frequently used in Nagasaki. On the other hand "peace" and "world" in C2 is constantly used in both Hiroshima and Nagasaki. Of the top 25 keywords, eight words including "Japan" and "people" are in C3, while only two words are in C1. Words in C3 are more frequently used in Nagasaki and the total word counts are also higher in Nagasaki. So Nagasaki may intend to send their messages to not only the public but also the Japanese government. In Figs. 20-24, State" in CC1 moves from left top to right middle as time goes by, which may imply that "State" along with "United" (i.e. United States) is frequently used in Hiroshima in earlier years because Mayors for Peace was first held in Hiroshima. Similarly "Nuclear" in CC1 moves from left bottom to right upper as time goes by, which may imply that "Nuclear" is more pronounced in later years due to the end of cold war and efforts on abolition of nuclear weapons. Words "soul" and "age" in CC2 is constantly used in Hiroshima and Nagasaki. On the other hand "law" in CC3 is more frequently used in Nagasaki, which is reflected by social movements after special measures for the atomic bomb exposed, under which allowances of various kinds were paid, were enacted, in the form of the law concerning in 1968.

Our proposal is applicable to longitudinal text data in general. Characteristics of text data can be extracted through a series of statistical methods. Concrete visualization of characteristics can provide an understandable image of what big data analysts intend to explore. A joint use of our proposed method with Artificial Intelligence (AI) tools such as IBM Watson may provide for efficiently realizing knowledge discovery.

Acknowledgements

Partially supported by Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology of Japan (17K00047) to the first author (S.I.), and by the Program of the network-type Joint Usage/Research Center for Radiation Disaster Medical Science of Hiroshima University, Nagasaki University, and Fukushima Medical University (67-H) to S.I. Four authors thank Hiroshima city office and Nagasaki city office to provide the documents of the English translated Peace Declaration for public usage. And S.I. thanks her former graduate students: Mr. K. Uchino, Mr. S. Matsuo, Mr. T. Iwakiri, and Mr. K. Go for data check and computational assistance.

References

1. S. Izumi, K. Satoh, and N. Kawano, Statistical classification and visualization based on varying coefficients model for longitudinal text data, *Computational Statistics*, **28**(1) (2015) 81–92. (in Japanese).
2. S. Izumi, K. T. Tonda, N. Kawano, and K. Satoh, Visualize the longitudinal big text data with a binary covariate: an approach based on keyword's frequency, in *Proc. IEEE/ACIT/CSII 2017 International Conference on Big Data, Cloud Computing, and Data Science Engineering (BCD 2017)*, (Japan, Hamamatsu, 2017), 284–289. DOI 10.1109/ACIT-CSII-BCD.2017.47
3. B. A. Brumback, D. Ruppert, and M.P. Wand, Variable selection and function estimation in additive nonparametric regression using a data-based prior: Comment, *Journal of American Statistical Association*, **94** (1999) 794–797.
4. K. Satoh and T. Tonda, Statistical inference of semiparametric varying coefficients using mixed effects model, *Japanese Journal of Applied Statistics*, **42**(1) (2013) 1–10. (in Japanese)

5. K. Uchino, A study of statistical inference on linear varying coefficient on longitudinal data of Poisson distribution, *Oita University Graduate School of Engineering Master Thesis*, (2017) 1–124. (in Japanese)
6. R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/ 2014.
7. W. N. Venables and B. D. Ripley, Modern Applied Statistics with S, 4th ed. (Springer-Verlag, New York, 2002).
8. Y. Xie, animation: An R Package for Creating Animations and Demonstrating Statistical Methods, *Journal of Statistical Software*, **53**(1) (2013) 1–27, URL http://www.jstatsoft.org/v53/i01/.
9. Y. Matsuura, K. Satoh and N. Kawano, Concept of peace in Hiroshima: analysis of Hiroshima Peace Declaration, *Hiroshima Peace Science*, **35** (2013) 67–101. (in Japanese)
10. Y. Matsuura, K. Satoh and N. Kawano, Concept of peace in Nagasaki: analysis of Nagasaki Peace Declaration, *Hiroshima Peace Science*, **36** (2014) 75–100. (in Japanese)
11. K. Higuchi, Quantitative content analysis for social survey. (Nakanishiya Press, Kyoto, Japan, 2014). (in Japanese)