Verbal Expression of Time Series with Global Trend and Local Features

Motohide Umano and Naoyuki Koizumi
Department of Mathematics and Information Sciences
Graduate School of Science, Osaka Prefecture University
1–1, Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531, Japan
umano@mi.s.osakafu-u.ac.jp

Abstract

We have many kinds of time series such as stock prices. We understand them via their verbal expressions in a natural language rather than conventional stochastic models. We propose a method to have a verbal expression for a global trend and local features of time-series data. A global trend is extracted via representative values, e.g. weighted averages, on the fuzzy intervals in the temporal axis and local features are specified as the positions of large differences between the original data and the data representing the global trend. We apply the method to the data of Multimodal Summarization for Trend Information (MuST).

Keywords: *Verbal Expression, Time Series, Global Trend, Local Feature, Fuzzy Sets.*

1. Introduction

We have various kinds of time-series data such as everyday's temperatures and stock prices. Such time-series data are usually analyzed using stochastic models [1]. In the human information processing, however, we never use such an analysis but express them in verbal expressions, for example, "It is a little increasing globally and there is a moderately larger point in the final term," where the phrase "a little increasing" is a global trend and "a moderately larger point" a local feature.

Using a global trend and local features, we verbally express time-series data for average prices of regular gasoline of all over Japan in the years 1998 and 1999 in a corpus of Multimodal Summarization of Trends (MuST) [2] [3].

2. Verbal Expression of Time Series

We propose a method to express time-series data in a natural language.

Table 1: Time Series Data from the MuST Corpus

months	year/month	price in yen
3	1998/03	95
4	1998/04	94
6	1998/06	92
7	1998/07	92
8	1998/08	92
9	1998/09	92
10	1998/10	92
11	1998/11	92
15	1999/03	91
16	1999/04	91
17	1999/05	90
18	1999/06	92
19	1999/07	93
20	1999/08	94
21	1999/09	95

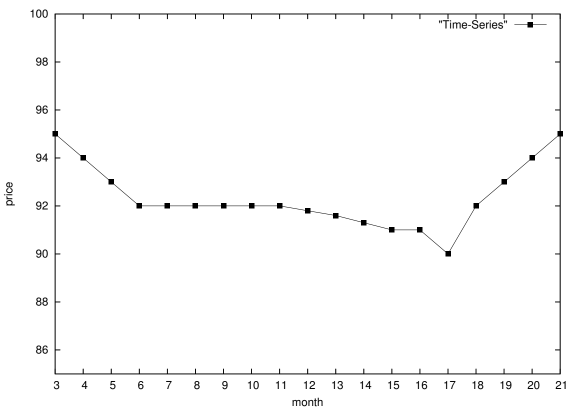

Fig. 1: Plotting of Time-Series Data from the MuST Corpus

2.1 Time-Series from the MuST Corpus

MuST (Multimodal Summarization of Trends [2] [3]) Corpus collects newspaper items on several topics, which are marked up with several tags. We choose the topic on the average prices of regular gasoline of all over Japan, which includes newspaper items in the years 1998 and 1999. In this research, we edit manually time-series data from them. The time-series data $\{(t_1, x_1), (t_2, x_2), \ldots, (x_n, x_n)\}$ are enumerated in **Table 1** and plotted in **Fig. 1**, where the first month is corresponding to January, 1998 and so $t_1 = 3$ and $t_n = 21$.

2.2 Global Trend for Time-Series Data

We first explain how to express a global trend for the time-series data.

(1) Fuzzy Sets of Terms

We divide a whole period into three terms, *the first term*, *the second term* and *the third term*, which are represented by fuzzy sets shown in **Fig. 2** since the boundaries of terms can not be strictly determined. These three fuzzy sets are defined by the following membership functions:

$$\mu_1(t) = Z\left(t; \frac{5t_1 + t_n}{6}, \frac{t_1 + t_n}{2}\right)$$

$$\mu_2(t) = \pi\left(t; \frac{5t_1 + t_n}{6}, \frac{t_1 + t_n}{2}, \frac{t_1 + 5t_n}{6}\right)$$

$$\mu_3(t) = S\left(t; \frac{t_1 + t_n}{2}, \frac{t_1 + 5t_n}{6}\right)$$

where $\mu_1(t)$, $\mu_2(t)$ and $\mu_3(t)$ are for fuzzy sets of *the first term*, *the second term* and *the third term*, respectively, and the functions Z, π and S are standard functions on the real numbers with a piecewise quadratic expression by Zadeh [4].

For the time-series data from the MuST Corpus, we have $t_1 = 3$ and $t_n = 21$ and

$$\mu_1(t) = Z(t; 6, 12) \quad (1)$$

$$\mu_2(t) = \pi(t; 6, 12, 18) \quad (2)$$

$$\mu_3(t) = S(t; 12, 18) \quad (3)$$

These fuzzy sets depends on time-series data. It is, however, very difficult to have appropriate term definitions (we have some discussion in Section 4). It may be better to define them in another way for another time-series data.

(2) Representative Values in Terms

To calculate the differences between adjacent terms, we have the representative value for each term. In this paper, we adopt the weighted average by the fuzzy sets, defined for the k-th term as follows:

$$m_k = \frac{\sum x(t)\mu_k(t)}{\sum \mu_k(t)} \quad (4)$$

For the time-series data from the MuST Corpus, we have $m_1 = 92.91$ for *the first term*, $m_2 = 91.75$ for *the second term* and $m_3 = 92.46$ for *the third term*.

(3) Fuzzy Sets for Difference between Terms

We calculate the differences from *the first term* to *the second term* and from *the second term* to *the third term* as follows:

$$d_{1\to2} = m_2 - m_1$$

$$d_{2\to3} = m_3 - m_2$$

For the time-series data from the MuST Corpus, we have $d_{1\to2} = -1.16$ and $d_{2\to3} = 0.71$.

Next, we express the differences between adjacent terms in fuzzy sets.

We have fuzzy sets, *much decreasing* (*much-dec*, for short), *moderately decreasing* (*mod-dec*), *a little decreasing* (*a-little-dec*), *approximately zero* (*appr-zero*), *a little increasing* (*a-little-inc*), *moderately increasing* (*mod-inc*) and *much increasing* (*much-inc*), which are shown in **Fig. 3** and defined as follows:

$$\mu_{much\text{-}dec}(d) = Z\left(d; \frac{x_a - x_b}{3}, \frac{x_a - x_b}{6}\right)$$

$$\mu_{mod\text{-}dec}(d) = \pi\left(d; \frac{x_a - x_b}{3}, \frac{x_a - x_b}{6}, \frac{x_a - x_b}{10}\right)$$

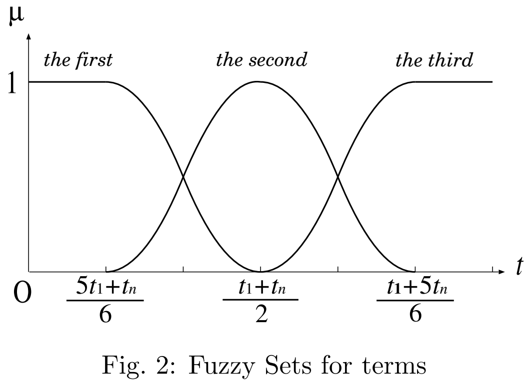

Fig. 2: Fuzzy Sets for terms

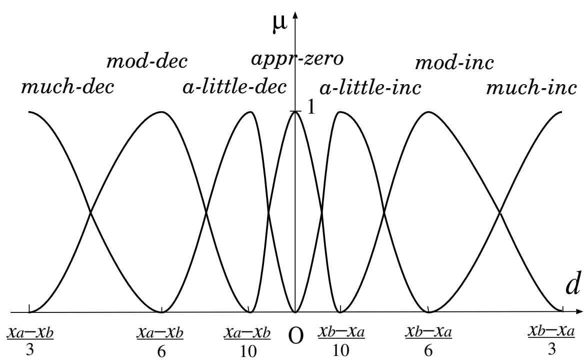

Fig. 3: Fuzzy Sets for Differences between Adjacent Terms

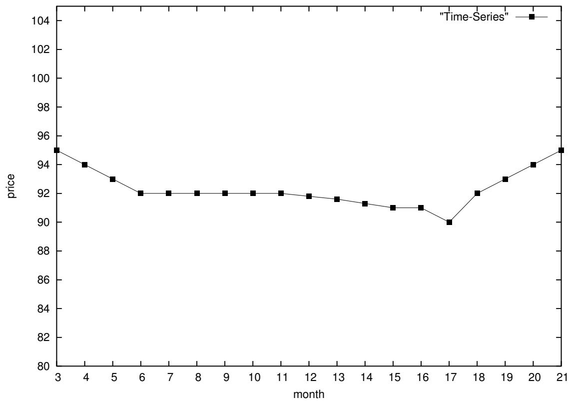

Fig. 4: Another Plotting of the Time-Series Data from the MuST Corpus

$$\mu_{a\text{-}little\text{-}dec}(d) = \pi\left(d; \frac{x_a - x_b}{6}, \frac{x_a - x_b}{10}, 0\right)$$
$$\mu_{appr\text{-}zero}(d) = \pi\left(d; \frac{x_a - x_b}{10}, 0, \frac{x_b - x_a}{10}\right)$$
$$\mu_{a\text{-}little\text{-}inc}(d) = \pi\left(d; 0, \frac{x_b - x_a}{10}, \frac{x_b - x_a}{6}\right)$$
$$\mu_{mod\text{-}dec}(d) = \pi\left(d; \frac{x_b - x_a}{10}, \frac{x_b - x_a}{6}, \frac{x_b - x_a}{3}\right)$$
$$\mu_{much\text{-}dec}(d) = S\left(d; \frac{x_b - x_a}{6}, \frac{x_b - x_a}{3}\right)$$

where x_a and x_b are the floor and ceiling values of x for the plotting area, respectively. It should be noted that these fuzzy sets depend on the plotting area rather than the data themselves. We can compare the plotting in Fig. 1 ($x_a = 85$ and $x_b = 100$) to another plotting in **Fig. 4** ($x_a = 80$ and $x_b = 105$) for the same time-series data in the MuST Corpus.

The expressions for the differences from *the first term* to *the second term* and from *the second term* to *the third term* are fuzzy sets D whose membership degrees $\mu_D(d_{1\to 2})$ and $\mu_D(d_{2\to 3})$ are positive (or greater than some threshold value).

For the time-series data from the MuST Corpus, we have fuzzy sets of differences for Fig. 1 as follows:

$$\mu_{much\text{-}dec}(d) = Z(d; -5, -2.5)$$
$$\mu_{mod\text{-}dec}(d) = \pi(d; -5, -2.5, -1.5)$$
$$\mu_{a\text{-}little\text{-}dec}(d) = \pi(d; -2.5, -1.5, 0)$$
$$\mu_{appr\text{-}zero}(d) = \pi(d; -1.5, 0, 1.5)$$
$$\mu_{a\text{-}little\text{-}inc}(d) = \pi(d; 0, 1.5, 2.5)$$
$$\mu_{mod\text{-}inc}(d) = \pi(d; 1.5, 2.5, 5)$$
$$\mu_{much\text{-}inc}(d) = S(d; 2.5, 5)$$

The expression for the difference $d_{1\to 2} = -1.16$ is *a-little-dec* with the degree 0.944 and *appr-zero* with 0.056, that is, $\{0.944/$*a-little-dec*, $0.056/$*appr-zero*$\}$ and $d_{2\to 3} = 0.71$ is $\{0.520/$*appr-zero*, $0.480/$*a-little-inc*$\}$.

For Fig. 4, we have fuzzy sets of differences as follows:

$$\mu_{much\text{-}dec}(d) = Z(d; -8.3, -4.1)$$
$$\mu_{mod\text{-}dec}(d) = \pi(d; -8.3, -4.1, -2.5)$$
$$\mu_{a\text{-}little\text{-}dec}(d) = \pi(d; -4.1, -2.5, 0)$$
$$\mu_{appr\text{-}zero}(d) = \pi(d; -2.5, 0, 2.5)$$
$$\mu_{a\text{-}little\text{-}inc}(d) = \pi(d; 0, 2.5, 4.1)$$
$$\mu_{mod\text{-}inc}(d) = \pi(d; 2.5, 4.1, 8.3)$$
$$\mu_{much\text{-}inc}(d) = S(d; 4.1, 8.3)$$

and the expressions for the differences are $\{0.499/$*a-little-dec*, $0.501/$*appr-zero*$\}$ from *the first term* to *the second term* and $\{0.827/$*appr-zero*, $0.173/$*a-little-inc*$\}$ from *the second term* to *the third term*. The degrees of *appr-zero* get greater than the ones for Fig. 1.

These fuzzy sets are somewhat common but can be defined in another way for another time-series data if necessary.

(4) Verbal Expression for Global Trend

Now we have the expressions for the differences from *the first term* to *the second term* and from *the second term* to *the third term*, we can have an expression for global trend. We use the following phrases to express a global trend: *much decreasing* (*much-dec*, for short), *moderately decreasing* (*mod-dec*), *a little decreasing* (*a-little-dec*), *approximately constant* (*appr-const*), *a little increasing* (*a-little-inc*), *moderately increasing* (*mod-inc*) and *much increasing* (*much-inc*), *convex* (concave up), *a little convex*, *a little concave*, *concave* (concave down). We have **Table 2** to get a global trend for the differences from *the first term* to *the second term* and from *the second term* to *the third term*.

We get global trends for all combinations of differences in the fuzzy sets, with the degrees of global trends calculated by the multiplication of two degrees of the differences involved. When we have the same global trend from the different combinations, we unify them with the sum of degrees.

For the time-series data from the MuST Corpus in Fig. 1, the differences are $\{0.944/$*a-little-dec*, $0.056/$*appr-zero*$\}$ from *the first term* to *the second term* and $\{0.520/$*appr-zero*, $0.480/$*a-little-inc*$\}$ from *the second term* to *the third term*. We consider all combinations, that is, for the combination *a-little-dec* and *appr-zero* we have a global trend *appr-const* with the degree $0.944 \times 0.520 = 0.491$, for *a-little-dec* and *a-little-inc* we have *a-little-convex* with $0.944 \times 0.480 = 0.453$, for *appr-zero*

Table 2: Table for Global Trend

$1 \to 2$ \ $2 \to 3$	*much-dec*	*mod-dec*	*a-little-dec*	*appr-zero*	*a-little-inc*	*mod-inc*	*much-inc*
much-dec	*much-dec*	*much-dec*	*mod-dec*	*mod-dec*	*a-little-convex*	*convex*	*convex*
mod-dec	*much-dec*	*mod-dec*	*mod-dec*	*a-little-dec*	*a-little-convex*	*convex*	*convex*
a-little-dec	*mod-dec*	*mod-dec*	*a-little-dec*	*appr-const*	*a-little-convex*	*a-little-convex*	*a-little-convex*
appr-zero	*mod-dec*	*a-little-dec*	*appr-const*	*appr-const*	*appr-const*	*a-little-inc*	*mod-inc*
a-little-inc	*a-little-concave*	*a-little-concave*	*a-little-concave*	*appr-const*	*a-little-inc*	*mod-inc*	*mod-inc*
mod-inc	*concave*	*concave*	*a-little-concave*	*a-little-inc*	*mod-inc*	*mod-inc*	*much-inc*
much-inc	*concave*	*concave*	*a-little-concave*	*mod-inc*	*mod-inc*	*much-inc*	*much-inc*

and *appr-zero* we have *appr-const* with $0.056 \times 0.520 = 0.029$ and for *appr-zero* and *a-little-inc* we have *appr-const* with $0.056 \times 0.480 = 0.027$.

We have three *appr-const* with the degrees 0.491, 0.029 and 0.027. We unify them with the degree $0.491 + 0.029 + 0.027 = 0.547$. Thus, we have a global trend $\{0.547/$*appr-const*, $0.453/$*a-little-convex*$\}$ for the time-series data from the MuST Corpus plotted in Fig. 1.

As for those in Fig. 4, the differences are $\{0.499/$*a-little-dec*, $0.501/$*appr-zero*$\}$ from *the first term* to *the second term* and $\{0.827/$*appr-zero*, $0.173/$*a-little-inc*$\}$ from *the second term* to *the third term*. We consider all combinations and unify the same. We have a global trend $\{0.914/$*appr-const*, $0.086/$*a-little-convex*$\}$.

Note that this method performs the ×-+ fuzzy reasoning from the differences from *the first term* to *the second term* and from *the second term* to *the third term* using the fuzzy rules in the form of Table 2 without no defuzzification (as the consequent part is not number or fuzzy set on number).

2.3 Local Features of Time-Series Data

The local features are specified as the positions of large differences between the original data and the data representing the global trend.

(1) Time-Series Representing Global Trend

To find the difference between the time-series data and the global trend, we generate time-series data representing the global trend.

The time-series data are a set of pairs of the time t_i and the value x_i. We propose that the time-series data representing the global trend is a set of the term and its representative value, that is, $\{($*the-first-term*, $m_1)$, $($*the-second-term*, $m_2)$, $($*the-third-term*, $m_3)\}$. Note that a label of fuzz set is concatenated with the hyphen(-) in an equation for easy understanding.

This means the global trend is formulated by the following fuzzy rules with 1 input variable:

rule1: if $t =$ *the-first-term* then $x = m_1$
rule2: if $t =$ *the-second-term* then $x = m_2$
rule3: if $t =$ *the-third-term* then $x = m_3$

where *the-first-term*, *the-second-term* and *the-third-term* are fuzzy sets defined in Equations (1)–(3).

The value of x are easily calculated with the membership functions $\mu_1(t)$, $\mu_2(t)$ and $\mu_3(t)$ of fuzzy sets *the first term*, *the second term* and *the third term*, respectively, as follows:

$$x(t) = \mu_1(t) \cdot m_1 + \mu_2(t) \cdot m_2 + \mu_3(t) \cdot m_3$$

where we assume that these fuzzy sets are defined as the sum of the membership values is 1 for all t. If this does not hold, we must divide the result by the sum of membership values.

When we vary t from t_1 to t_n, $x(t)$ is m_1 for t_1 to $(5t_1 + t_n)/6$ where $\mu_1(t) = 1$, the weighted average of m_1 and m_2 by $\mu_1(t)$ and $\mu_2(t)$ until $(t_1 + t_n)/2$, the weighted average of m_2 and m_3 by $\mu_2(t)$ and $\mu_3(t)$ until $(t_1 + 5t_n)/6$, and m_3 for $(t_1 + 5t_n)/6$ to t_n where $\mu_3(t) = 1$.

For the time-series data from the MuST Corpus in Fig. 1, we have fuzzy rules as follows:

rule1: if $t =$ *the-first-term* then $x = 92.91$
rule2: if $t =$ *the-second-term* then $x = 91.75$
rule3: if $t =$ *the-third-term* then $x = 92.46$

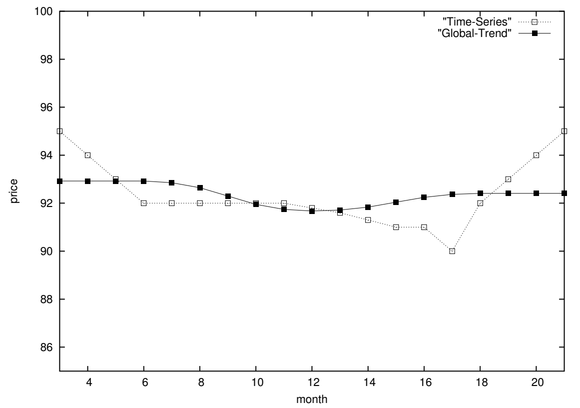

Fig. 5: Plotting of the Data Representing the Global Trend

and the time-series data representing the global trend are generated from these fuzzy rules and shown in **Fig. 5**, where the real line stands for the time-series data representing the global trend and the dotted line the original time-series data. You can find the real line well representing the global trend $\{0.547/$*appr-const*, $0.453/$*a-little-convex*$\}$. Note that for Fig. 4 we have the same the time-series data representing the global trend but a different plotting area.

(2) Local Features

The local features are specified as the positions of large differences between the original data and the data representing the global trend. We find positions of large differences and express the positions and the scale of differences with fuzzy sets. We can have n local features by the n maximum positions.

More specifically, let the original time-series data to be $\{(t_i, x_i)\}$ and the time-series data representing the global trend $\{(t_i, x'_i)\}$. Then the position of local feature is the time t_ℓ such that the absolute value of the difference $|x'_i - x_i|$ gets the maximum value with ranging from t_1 to t_n and expressed by the term that contains t_ℓ with the greatest degree.

The scale of the local feature is the value $x'_\ell - x_\ell$ and expressed with fuzzy sets L for the local feature with the degrees $\mu_L(x'(t_\ell) - x(t_\ell))$. The labels and definitions of fuzzy sets L are similar to those for the difference of adjacent terms. We have fuzzy sets, *much smaller*, *moderately smaller* (*mod-smaller*, for short), *a little smaller*, *approximately zero* (*appr-zero*), *a little larger*, *moderately larger* (*mod-larger*) and *much larger*, which are shown in **Fig. 6** and defined as follows:

$$\mu_{\textit{much-smaller}}(d) = Z\left(d; \frac{x_a - x_b}{2}, \frac{x_a - x_b}{4}\right)$$

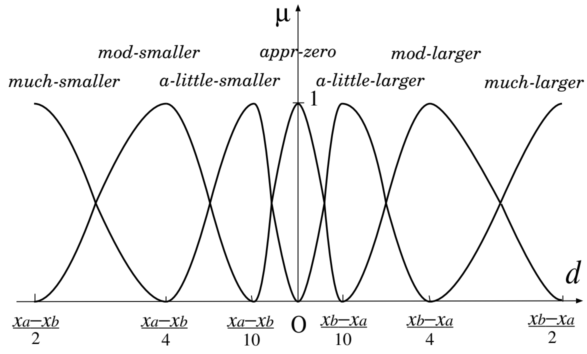

Fig. 6: Fuzzy Sets for Local Feature

$$\mu_{\textit{mod-smaller}}(d) = \pi\left(d; \frac{x_a - x_b}{2}, \frac{x_a - x_b}{4}, \frac{x_a - x_b}{10}\right)$$

$$\mu_{\textit{a-little-smaller}}(d) = \pi\left(d; \frac{x_a - x_b}{4}, \frac{x_a - x_b}{10}, 0\right)$$

$$\mu_{\textit{appr-zero}}(d) = \pi\left(d; \frac{x_a - x_b}{10}, 0, \frac{x_b - x_a}{10}\right)$$

$$\mu_{\textit{a-little-larger}}(d) = \pi\left(d; 0, \frac{x_b - x_a}{10}, \frac{x_b - x_a}{4}\right)$$

$$\mu_{\textit{mod-larger}}(d) = \pi\left(d; \frac{x_b - x_a}{10}, \frac{x_b - x_a}{4}, \frac{x_b - x_a}{2}\right)$$

$$\mu_{\textit{much-larger}}(d) = S\left(d; \frac{x_b - x_a}{4}, \frac{x_b - x_a}{2}\right)$$

Note that these fuzzy sets also depend on the plotting area rather than the data themselves.

For the time-series data from the MuST Corpus in Fig. 1, $|x'_i - x_i|$ gets the maximum at the time 21 and 17, where $x'_i - x_i$ are 2.591 and −2.368, respectively, when we assume two local features. For the largest local feature (21, 2.591), the time 21 belongs to *the third term* with the degree 1 and the value 2.591 belongs to *a-little-larger* with the degree 0.529 and *mod-larger* with 0.471, that is, $\{0.529$*a-little-larger*, $0.471/$*mod-larger*$\}$. For the second largest local feature (17, −2.368), the time 17 is in $\{0.944/$*the-third-term*, 0.056*the-second-term*$\}$ and the value −2.368 is $\{0.703/$*a-little-smaller*, $0.297/$*mod-larger*$\}$.

For Fig. 4, the maximum of $x'_i - x_i$ is the same as that for Fig. 1 since $|x'_i - x_i|$ is not depend on the plotting area but just the data themselves. For the largest local feature (21, 2.591), the time 21 is in $\{1/$*the-third-term*$\}$ and the value 2.591 is $\{0.999/$*a-little-larger*, $0.001/$*mod-larger*$\}$. For the second largest local feature (17, −2.368), the time 17 is in $\{0.944/$*the-third-term*, $0.056/$*the-second-term*$\}$ and the value −2.368 is $\{0.994/$*a-little-smaller*, $0.006/$*mod-larger*$\}$.

(3) Verbal Expression for Time Series

Now we have a fuzzy set of global trend and two fuzzy sets of local features. For a verbal expression, we select the word with the greatest degree.

For the time-series data from the MuST Corpus in Fig. 1, we have a global trend {0.547/*appr-const*, 0.453/*a-little-convex*} and two local features {0.529/*a-little-larger*, 0.471/*mod-larger*} for {1/*the third term*} and {0.703/*a-little-smaller*, 0.297/*mod-larger*} for {0.944/*the-third-term*, 0.056*the-second-term*}. We select the words with the greatest degree and have an expression "It is approximately constant globally and there are a little larger and a little smaller points in the third term." In this case, we can have another expression "It is a little convex globally and there are a little larger and a little smaller points in the third term" because the second greatest degree of the global trend is so close to the the greatest one.

For Fig. 4, we have a global trend {0.914/*appr-const*, 0.086/*a-little-convex*} and two local features {0.999/*a-little-larger*, 0.001/*mod-larger*} for {1/*the-third-term*} and {0.994/*a-little-smaller*, 0.006/*mod-larger*} for {0.944/*the-third-term*, 0.056/*the-second-term*}. We have an expression "It is approximately constant globally and there are a little larger and a little smaller points in the third term."

We feel strange to the expression "a little larger and a little smaller points in the third term," i.e., two opposite words in the term, because this time-series data has two local features in the same term as you can find them in the plotting in Fig. 1 and Fig. 4. We can distinguish two different times 21 and 17 by the expression, for example, "the beginning of the term" and "the end of the term".

3. Future Works

We have several very interesting future works.

(1) Definition of Terms

The definition of appropriate terms is very difficult problem. It depends on time-series data, especially the change of local patterns or trends. Human identifies several points where local patterns change and divides a whole period into terms. To implement this, we must find local patterns and then identify the points to change the pattern. The local pattern is, however, considered to be a global trend in the local period of time. It may be possible to apply the method recursively.

(2) Introduction of Oscillation

Time-series data usually go up and down. We must introduce oscillation to express time-series data well. We have already started consideration on oscillation of time-series data. We use the standard deviation and the extended number of oscillations.

(3) Uncertainty in Time-Series Data

The natural time-series data have various kinds of uncertainty. Some comes from the granularity (e.g., months, weeks, days and hours) and exactness (directly or indirectly written in newspaper items). We have a plan to represent and manipulate the uncertainty of the data itself by the possibility distributions [5].

(4) Retrieval of Time-Series

We store many time-series data into the database with the tag of global trend and local features by this method. Then we can retrieve time-series data from the database by a query with some verbal expression of global trend and local features or some time-series data themselves.

4. Conclusions

In this paper, we proposed a verbal expression for a global trend and local features of time-series data. A global trend is extracted via representative values on the fuzzy intervals in the temporal axis and local features are specified as the positions of large differences between the original data and the data representing the global trend. We apply the method to the data of Multimodal Summarization for Trend Information (MuST).

Future works include the definition of appropriate terms, introduction of oscillation, uncertainty in time-series data, retrieval of time-series.

References

[1] J.D. Hamilton (1994): *Time Series Analysis*, Princeton University Press.

[2] T. Kato, M. Matsushita, Y. Kando (2005): "MuST: A Workshop on Multimodal Summarization for Trend Information," *Proc. of NTCIR-5 Workshop Meeting*, pp.556–563.

[3] http://must.c.u-tokyo.ac.jp/

[4] L.A. Zadeh (1975): "Fuzzy Logic and Approximate Reasoning," *Synthese*, Vol.30, pp.407–428.

[5] L.A. Zadeh (1978) : "PRUF—A Meaning Representation Language for Natural Languages," *International Journal of Man-Machine Studies*, Vol.10, pp.395–460.