LETTER

Using Machine Learning for Automatic Estimation of Emphases in Japanese Documents

Masaki MURATA[†a)], ***Member and*** **Yuki ABE**[†b)], ***Nonmember***

SUMMARY We propose a method for automatic emphasis estimation using conditional random fields. In our experiments, the value of F-measure obtained using our proposed method (0.31) was higher than that obtained using a random emphasis method (0.20), a method using TF-IDF (0.21), and a method based on LexRank (0.26). On the contrary, the value of F-measure of obtained using our proposed method (0.28) was slightly worse as compared with that obtained using manual estimation (0.26–0.40, with an average of 0.35).

key words: *machine learning, automatic estimation, emphasis, bold, conditional random fields*

1. Introduction

Emphasis, using colored or bold fonts and/or underlines, often eases the readability of a document. Including emphasis in a document allows the reader to quickly survey its information arrangement and read it more selectively. If emphases could be automatically generated for a document, it would promote quicker reading and support the creation of a more readable document for authors. Therefore, we aim to apply emphases to documents automatically.

In this study, we propose a method using supervised machine learning with conditional random fields (CRF) [1] for automatic emphasis estimation. In CRF, we can use estimated emphases as features and lay emphasis on important phrases comprising series of words. The term frequency-inverse document frequency (TF-IDF) is useful for extracting important words and titles; moreover, TF-IDFs are also used as features in CRF. We considered automatic emphasis estimation in Japanese documents.

Previous studies have addressed the problem of automatically placing emphases in a document [2]. In addition, nouns and unknown words that appeared in a title have also been emphasized [2]. However, their method cannot emphasize the crucial nouns and unknown words that are used elsewhere in the document and not in the title.

There have also been studies on summarization [3], [4], which are related to emphasis because emphasizing also extracts the important parts from a document. Nomoto used CRF for sentence compression and also used syntactic information to improve the readability of the summarized documents [3]. Shen et al. used CRF to extract the features used in machine learning for summarization [4]. Although summarization is similar to emphasis estimation, a crucial difference between them exists: in summarization, the output should comprise complete sentences. By contrast, in emphasis estimation, the output need not comprise complete sentences and can take the form of phrases or characters.

The important aspects of this study are summarized below.

- This study is novel in that we perform automatic emphasis estimation using machine learning.
- Our method is useful because it has higher performance than other comparable approaches. The value of F-measure (a harmonic average of the recall and precision rates) obtained using our proposed method (0.31) was higher than that obtained using a random emphasis method (0.20), a method using TF-IDF (0.21), and a method based on LexRank (0.26) [5].
- Although one may think that the value of F-measure obtained using our method (0.28) is low, that obtained using manual estimation (0.26 to 0.40, with an average of 0.35) is also low. Therefore, the performance of our method is satisfactory.
- Using supervised machine learning, our proposed method can easily use several features (pieces of information). Further improvements in performance could be achieved by increasing the number of features.
- We noticed a case in which performance improved by using training and test data from documents by the same author. This shows that it may be possible to place different emphases for each author by using documents by the same author as training data.

2. Task

Initially, we took a set of documents with emphasis included. The emphasis was then eliminated from the documents; they were used as an input to our system. The outputs were the documents in which emphasis had been added. Considering the overlaps between the emphases in the output and original documents, an evaluation was performed. An output document that had more emphasis overlaps with its original document was considered to be a better result. Here, we consider the emphases in the original documents to be correct.

Manuscript received December 19, 2016.
Manuscript revised May 9, 2017.
Manuscript publicized July 21, 2017.

[†] The authors are with Tottori University, Tottori-shi, 680–8552 Japan.
a) E-mail: murata@ike.tottori-u.ac.jp
b) E-mail: cobutazura@yahoo.co.jp

DOI: 10.1587/transinf.2016EDL8247

3. Our Proposed Method

In this study, we propose a method for automatically estimating emphasis using supervised machine learning.

We use CRF++, a tool that can create CRF for use in named-entity extraction.

Input documents are divided into words. The method judges whether or not each word should be emphasized, and estimates the necessary emphases using the accumulated results.

3.1 CRF

CRF is a discriminative model that labels series using a maximum entropy method. Because CRF can obtain better results than Hidden Markov models, which perform the same series labeling, CRF has been applied in several fields such as morphological analysis [6] and named-entity extraction [7].

In this study, emphases were estimated for each document using CRF with a standard chain model. Using the following equation, $P(y|x)$ was used to obtain an output sequence, $y = y_1, y_2, \ldots, y_n$, from the input sequence $x = x_1, x_2, \ldots, x_n$.

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{n} \sum_{t=1}^{k} \lambda_t f_t(y_{i-1}, y_i, x)\right) \tag{1}$$

$Z(x)$ is the normalization constant, such that the sum of the probabilities of all the sequences is equal to 1. The feature functions, f_t, depend on the position i, output labels y_i, y_{i-1}, and the input sequence x. λ_i is the weighting factor for feature function f_i and was determined by learning. The output labels y^*, maximizing the probability obtained by Eq. (1), were obtained from the following equation.

$$y^* = \arg\max_{y} p(y|x) \tag{2}$$

In this study, an input sequence x_i corresponds to a sequence of words in a sentence. The corresponding output sequence y_i indicates whether each word should be emphasized.

3.2 Feature Template

Feature templates were used to define the feature functions used in Eq. (1). To define the feature functions in this study, we used feature information from the word currently being processed as well as the two words immediately before and the two words immediately after it. We also used the output label of the word immediately preceding the current word.

3.3 Features

The information required for emphasis estimation was used for the features. Because this is similar to summarization, we used the same TF-IDF and context information (paragraph information and title information) for the features. The features used for machine learning (along with their functions) are as follows.

Word: a word is used as a feature.

Part of speech: the part of speech (POS) of a word is used as a feature.

TF-IDF of a word: the ranges, according to the value of TF-IDF of a word, are used as features.

The average TF-IDF of the words in a sentence: the TF-IDF values calculated for words are averaged per sentence. The ranges, according to the averaged TF-IDF values, are used as features.

Paragraph information: it determines whether the currently targeted word is in the first sentence of a paragraph, the last sentence of a paragraph, or elsewhere.

Title information: it determines whether a currently targeted word is in the title of the document.

TF-IDF is a method for assigning high values to crucial words in a document, and it enables us to use crucial words that do not exist in the title. Because a sentence containing many words with high TF-IDF values is likely to be important, we use the average TF-IDFs of sentences as features. Likewise, because the sentences at the beginning and end of a paragraph are often important, we use this to add additional features.

4. Experiment

4.1 Preparing the Experimental Dataset

We acquired 331 reports (March 2012) from http://www.lifehacker.jp/, which is a lifestyle website. Emphasis was added to the acquired reports, limited to one type (bold font), and the 331 acquired reports were used as a training dataset. The average number of characters per report was 961, and the average number of bold characters was 173.

4.2 Baseline Methods

In this study, we used three types of baseline methods: a random emphasis method, a TF-IDF method, and LexRank.

The random emphasis method judged that sentences selected at random should be emphasized. The TF-IDF method judged that sentences with average TF-IDFs above a given constant value should be emphasized. LexRank [5] is a graph-based summarization method. We used Summpy[†] to implement LexRank. LexRank in our study outputs a number of sentences proportional to the number of characters in the document.

4.3 Automatic Estimation of Emphasis

We conducted experiments using machine learning to add

[†] https://github.com/recruit-tech/summpy

Table 1 Automatic emphasis estimation.

Method	Recall	Precision	F-measure
Our proposed method	0.31	0.31	0.31
Random emphasis method	0.20	0.20	0.20
TF-IDF method	0.21	0.21	0.21
LexRank	0.26	0.26	0.26

emphasis to the documents from which the emphasis had been eliminated.

The total number of characters emphasized by each method across all the documents used in the experiment was adjusted to match the total number of emphasized characters in the correct answers.

Adjustment of the number of emphases output by our proposed method was executed using n-best, a function of CRF++. First, the emphases in the best output were added to the output, followed by the emphases in the second-best output. This was repeated until the number of output emphases was more than or equal to the number of correct answers.

Using the 331 documents obtained in Sect. 4.1, a five-fold cross validation was used for evaluation. We verified the validity of our proposed method by automatically comparing the estimated emphases with the original emphases in the documents and calculating the recall rate, the precision rate, and the F-measure value of the emphases per word[†]. The recall rate is the number of words correctly emphasized by the method divided by the number of words emphasized in the original documents. The precision rate is the number of words correctly emphasized by the method divided by the total number of words emphasized by it. As mentioned above, the F-measure value is the harmonic average of the recall and the precision rates.

The results of automatic emphasis estimation are shown in Table 1. The proposed method outperformed the three baseline methods with an F-measure value of approximately 0.3. Using a one-sided sign test on the F-measures of the 331 training documents, a significant difference between the performance of our proposed method and each of the other methods exists at a significance level of 0.05.

4.4 Effectiveness of Features

We investigated the manner in which each feature contributed to machine learning performance by examining the results when only one feature was used for estimation and when one feature was eliminated. We used the same experimental conditions as in Sect. 4.3 other than features.

The experimental results are shown in Tables 2 and 3. The F-measure values went down most strongly when the POS feature was eliminated. Thus, we find that the POS feature is particularly important to estimate emphases.

Table 2 Performance using only one feature.

Used features	Recall	Precision	F-measure
All features	0.31	0.31	0.31
Word	0.23	0.22	0.23
POS	0.28	0.28	0.28
TF-IDF of a word	0.22	0.22	0.22
TF-IDF in a sentence	0.21	0.21	0.21
Paragraph information	0.20	0.20	0.20
Title information	0.23	0.23	0.23

Table 3 Performance when eliminating one feature.

Eliminated features	Recall	Precision	F-measure
None	0.31	0.31	0.31
Word	0.29	0.29	0.29
POS	0.24	0.25	0.24
TF-IDF of a word	0.30	0.30	0.30
TF-IDF in a sentence	0.28	0.28	0.28
Paragraph information	0.28	0.28	0.28
Title information	0.27	0.27	0.27

Table 4 Number of documents written by each author.

Author	No. of documents
Author 1	327 (326)
Author 2	326 (326)
Multiple authors	1,930 (326)

4.5 Performance When the Author Is Specified

Deciding the portions of a document to emphasize depends on the individual. Therefore, it is possible that machine learning performance will drop when the dataset used for learning contains various authors' documents. Therefore, in this section, we investigate machine learning performance when the same author is used for both the training and test datasets. We selected two authors and used only the documents written by those authors between January and June 2012 for http://www.lifehacker.jp/.

The number of documents written by the authors selected for our experiment is listed in Table 4. When the number of data items used for each method in an experiment differs, it is difficult to judge whether the difference is a true reflection of the performance of each method or whether it should be attributed to the number of data items. Therefore, in this study, we used the same number of data items for each method. For each author in Table 4, 326 documents were selected at random from the documents written by them. For the multiple authors case, 326 documents were selected at random from all the documents written between January and June 2012, and a five-fold cross validation was performed. The table describes the number of documents written by each author between January and June 2012. The numbers enclosed in parentheses are the numbers of data items by each author used in the experiment.

The results are presented in Table 5. The F-measure value for Author 1 (0.37) far exceeds that for multiple authors (0.31), indicating that the performance of this method using the same author for the training and test datasets can

[†] In the evaluation, the emphases obtained by each method were judged to be correct when the location of the emphasis was identical to the location of the emphasis in the original document.

Table 5 Automatic emphasis estimation when the author is specified.

Author	Recall	Precision	F-measure
Author 1	0.37	0.37	0.37
Author 2	0.32	0.32	0.32
Multiple authors	0.31	0.31	0.31

Table 6 Comparison between our proposed method and manual estimation.

Method	Recall	Precision	F-measure
Our proposed method	0.28	0.28	0.28
Manual estimation (Average of the three subjects)	0.35	0.36	0.35
Manual estimation 1	0.40	0.41	0.40
Manual estimation 2	0.39	0.40	0.40
Manual estimation 3	0.26	0.27	0.26

be higher than the case in which multiple authors are used. However, the F-measure value of Author 2 was not as high (0.32). We therefore find that the performance improvement depends on the author as well.

4.6 Comparison of the Proposed Method with Manual Estimation

In this section, we compare the performance of our proposed method with manual estimation.

In manual estimation, the emphases in a document are not known to the subject, who then estimates the emphases. Three subjects[†] were used in the experiment. To study the locations of the emphases occurrence in documents before the experiment began, each subject had the opportunity to read twenty sample documents with emphases included. The sample documents were different from the documents used in the experiment.

Owing to the difficulty in preparing the same number of data items for manual estimation as that used in Sect. 4.3, the number of documents used was set to ten, namely the first ten of the documents used previously in Sect. 4.3. The number of characters output by our method and by manual estimation was set to 2,200, that being the total number of emphasized characters in the correct documents. The average number of characters per report was 1,460, and the average number of bold characters was 220.

For our proposed method, the estimation results for the ten evaluation documents were obtained from the results of the five-fold cross validation discussed in Sect. 4.1.

The results of comparing the proposed method with manual estimation are presented in Table 6. The table shows the manual estimation results of the three subjects, where "Manual estimation X" indicates the result of the Xth subject.

We found that the F-measure value of manual estimation is not that high (0.26 to 0.40, with an average of 0.35). Although one might think that the F-measure value of our method (0.28) is low and that of the manual estimation is also low; therefore, our method actually performs satisfactorily, even if it is not as good as manual estimation.

5. Conclusions

In this study, a method for automatically estimating which parts of a document to emphasize using supervised machine learning was proposed. In an experiment that automatically estimated emphases in terms of word units, the F-measure value (0.31) was higher than that obtained using a random emphasis method (0.20), a TF-IDF method (0.21), and a method based on LexRank (0.26), confirming the efficacy of our proposed method. By conducting experiments to examine the effectiveness of particular features, we found that the POS feature was particularly important in estimating emphases. In an experiment using the same author for the training and test datasets, we observed a case where the F-measure value (Author 1: 0.37) using a single author was higher than obtained using the F-measure (0.31) when several different authors were used. However, comparing our proposed method with manual estimation, the F-measure value obtained using our proposed method (0.28) was worse than that obtained using manual estimation (0.26 to 0.40, with an average of 0.35).

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 26330252.

References

[1] J.D. Lafferty, A. McCallum, and F.C.N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 18th International Conference on Machine Learning, pp.282–289, 2001.

[2] D. Maruyama, M. Murata, M. Tokuhisa, and Q. Ma, "Experiments on reading support techinique highlighting words appearing in titles," 17th Annual Meeting of the Association for Natural Language Processing, pp.639–642, 2011 (in Japanese).

[3] T. Nomoto, "Discriminative sentence compression with conditional random fields," Information Processing and Management, vol.43, no.6, pp.1571–1587, 2007.

[4] D. Shen, J.T. Sun, H. Li, Q. Yang, and Z. Chen, "Document summarization using conditional random fields," 20th International Joint Conference on Artificial Intelligence (IJCAI-07), pp.2862–2867, 2007.

[5] G. Erkan and D.R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," Journal of Artificial Intelligence Research, vol.22, no.1, pp.457–479, 2014.

[6] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," The 2004 Conference on Empirical Methods in Natural Language Processing, pp.230–237, 2004.

[7] R. Sasano and S. Kurohashi, "Japanese named entity recognition using structural natural language processing," The Third International Joint Conference on Natural Language Processing, pp.607–612, 2008.

[†] The subjects were university students and are not authors of this paper.