Automatic Detection and Analysis of Impressive Japanese Sentences Using Supervised Machine Learning

Daiki Hazure, Masaki Murata, Masato Tokuhisa
Department of Information and Electronics
Tottori University
4-101 Koyama-Minami, Tottori 680-8552, Japan
{s082042,murata,tokuhisa}@ike.tottori-u.ac.jp

Abstract

It is important to write sentences that impress the listener or reader ("impressive sentences") in many cases, such as when drafting political speeches. The study reported here provides useful information for writing such sentences in Japanese. Impressive sentences in Japanese are collected and examined for characteristic words. A number of such words are identified that often appear in impressive sentences, including *jinsei* (human life), *hitobito* (people), *koufuku* (happiness), *yujou* (friendliness), *seishun* (youth), and *ren'ai* (love). Sentences using these words are likely to impress the listener or reader. Machine learning (SVM) is also used to automatically extract impressive sentences. It is found that the use of machine learning enables impressive sentences to be extracted from a large amount of Web documents with higher precision than that obtained with a baseline method, which extracts all sentences as impressive sentences.

1 Introduction

People are always willing to be impressed, and the things that most impress them are liable to be things they need to live, such as food. On the other hand, the wisdom of human beings is recorded in writing, saved in the form of sentences, and inherited by future generations. In this study, we therefore focused on "impressions" and "sentences" and studied sentences that tend to impress the listener or reader. Hereafter for brevity we will refer to these as "impressive sentences". There were two main topics in this study: collecting impressive sentences and analyzing them.

1. Collecting impressive sentences

   We manually collect impressive sentences as well as sentences that are not particularly impressive. By using these sentences and supervised machine learning, we collect more impressive sentences from the Web.

2. Analyzing impressive sentences

   We examine and analyze the impressive sentences. By identifying and collecting words that were often used in them, we clarify the linguistic characteristics of the sentences.

The focus of our study is Japanese sentences.

The study we report in this paper provides useful information for constructing a system that supports the writing of impressive sentences. Such a system would be useful for writing drafts of politicians' speeches or for writing project plan documents where the use of impressive sentences would make the documents more likely to be accepted. In this study, we use natural language processing in an attempt to support persons in their efforts to write impressive sentences.

The main points of the study are as follows:

- This study is the first attempt to use natural language processing for automatic collection and analysis of impressive sentences.

- By collecting sentences automatically and examining the collected data, we identified *jinsei* (human life), *hitobito* (people), *koufuku* (happiness), *yujou* (friendliness), *seishun* (youth), *ren'ai* (love), etc. as words that often appear in impressive sentences. Sentences containing one or more of these words are likely to be impressive sentences. These results should prove to be useful for generating impressive sentences.

- We used machine learning to obtain impressive sentences from a large amount of Web documents at a 0.4 precision rate. This is much higher than the 0.07 rate obtained with a baseline method.

2 Collecting impressive sentences

We first use the Google search engine to collect impressive sentences and sentences that are not particularly impressive. We then use these sentences as supervised data with machine learning to collect more impressive and non-impressive sentences from Web documents.[1]

Hereafter, we will refer to impressive sentences as *positive examples* and non-impressive sentences as *negative examples*.

2.1 Manual collection of impressive sentences

We extract sentences that are obtained by using retrieval words like "... *toiu kotoba ni kando shita*" (I was impressed by the words...) as positive example candidates. We extract sentences that are obtained by using retrieval words like "... *toiu bun*" (the sentences...) as negative example candidates.

Example sentences containing "... *toiu kotoba ni kando shita*" and "... *toiu bun*" are shown below.

Example sentences containing "... *toiu kotoba ni kando shita*":

"*mainichi ga*	*mirai*"	*toiu*	*kotoba ni*	*kando-shita.*
(every day)	(future)	(of)	(word)	(was impressed)

(I was impressed by the words "Every day is the future.")

Example sentences containing "... *toiu bun*":

kanojo wa	*supoutsu wo*	*suru noga*	*suki desu*	*toiu*	*bun*
(she)	(sport)	(play)	(like)	(of)	(sentence)

(The sentence "She likes playing sports")

In the above examples, the sentences *mainichi ga mirai* (Every day is the future) and *michi wa hito to hito no kakehashi desu* (A road is a bridge connecting people with other people) are used as positive example candidates. The sentences *yomitori senyou* (Read only) and *kanojo wa supoutsu wo suru noga suki desu* (She likes playing sports) are used as negative example candidates.

We also use the Google search engine to retrieve famous sentences and use them as positive example candidates.[2] We collect sentences from sources such as Yahoo! News and use them as negative example candidates.

We manually judge whether candidates are positive and negative, and in so doing obtain accurate positive and negative examples.

Our judgment criterion is that sentences that received the comment "*kando shita*" (was impressed by) and famous sentences are judged to be positive. Sentences that do not have emphatic punctuation such as exclamation marks or that describe objective facts only are judged to be negative.

We performed the above procedure and obtained 1,018 positive examples and 406 negative examples.

2.2 Using supervised machine learning to collect impressive sentences

We conduct machine learning using the positive and negative examples obtained as described in Section 2.1 as supervised data. We use sentences in Web documents as inputs for machine learning. Machine learning is used to judge whether the sentences are impressive. In this way we collect impressive sentences from Web documents.

The specific procedure is as follows:

[1] We used the Web documents that Kawahara et al. collected (Kawahara and Kurohashi, 2006).

[2] Some famous sentences are obtained from http://www.meigensyu.com/.

Table 1: Words with high appearance probabilities in positive examples

Word	Ratio of positive	Freq. of positive	Freq. of negative
koufuku (happiness)	1.00	83	0
yujou (friendliness)	1.00	29	0
seishun (youth)	1.00	18	0
kanashimi (sadness)	1.00	12	0
sonzai (existence)	1.00	10	0
...	...	...	...
wareware (we)	0.97	37	1
fukou (unhappiness)	0.97	32	1
aisa (love)	0.96	23	1
ren'ai (love)	0.96	44	2
koi (love)	0.95	122	7
kodoku (loneliness)	0.94	32	2
konoyo (this world)	0.94	16	1
aishi (love)	0.94	31	2
aisuru (love)	0.94	30	2
arayuru (every)	0.93	14	1
omae (you)	0.93	13	1
shunkan (moment)	0.92	11	1
jinsei (life)	0.91	145	14
mirai (future)	0.91	20	2
shiawase (happiness)	0.91	20	2
yorokobi (delight)	0.91	10	1
onna (woman)	0.91	115	12
unmei (destiny)	0.90	19	2
shinu (die)	0.90	37	4
...	...	...	...
hitobito (people)	0.81	17	4
kandou (impression)	0.80	8	2

1. The 1,018 positive and 406 negative examples obtained as described in Section 2.1 are used as supervised data.

2. We use the supervised data to conduct machine learning. The machine learning is used to judge whether 10,000 sentences newly obtained from Web documents are positive or negative. We manually check sentences judged to be positive and construct new positive and negative examples. We add the new examples to the supervised data.

3. We repeat the above step 2 procedure ten times.

We use a support vector machine (SVM) for machine learning (Cristianini and Shawe-Taylor, 2000; Kudoh and Matsumoto, 2000; Isozaki and Kazawa, 2002; Murata et al., 2002; Takeuchi and Collier, 2003; Mitsumori et al., 2005; Chen and Wen, 2006; Murata et al., 2011).[3] We use unigram words whose parts of speech (POSs) are nouns, verbs, adjectives, adjectival verbs, adnominals, and interjections as features used in machine learning.

The judgment criteria for positive and negative examples in this section are as follows: Sentences for which a judge can spontaneously produce certain comments are judged to be positive examples. Sentences that describe objective facts only are judged to be negative examples.

We repeated the procedure ten times. In total, 275 positive and 3,006 negative examples were obtained. When we add these examples to the original ones, the totals become 1,293 positive and 3,412 negative examples. In this case we repeated the learning procedure ten times, but more positive and negative examples could be obtained by repeating it more than ten times.

A subject (Subject A) judged whether the examples were positive or negative. Three other subjects evaluated 20 examples that were judged positive and 20 that were judged negative by Subject A. We compared Subject A's judgments and the majority voting results of the other three subjects' judgments and obtained 0.58 (moderate agreement) as a kappa value.

3 Analysis of collected impressive sentences

In our analysis, we used the abovementioned 1,293 positive and 3,412 negative examples. We used certain words to examine the impressive sentences.

We extracted a number of words from the positive and negative examples. For each word, we calculated its appearance frequency in positive and negative examples and the ratio of its frequency in positive examples to its frequency in negative ones. We extracted words for which the ratio was higher than 0.8 and words that were at least four times likelier to appear in positive examples than in negative ones. Some of the extracted words are shown in Table 1. In the table, "Ratio appearing in positive" indicates the ratio

[3] In this study, we use a quadratic polynomial kernel as a kernel function of SVM. We confirmed that the kernel produced good performance in preliminary experiments.

Table 2: Impressive sentence extraction performance of various methods

Method	Precision	Recall	F measure
ML method (0th)	0.06	0.25	0.10
ML method (first)	0.26	0.08	0.12
ML method (second)	0.29	0.07	0.11
ML method (fifth)	0.31	0.05	0.09
ML method (10th)	0.40	0.05	0.09
Baseline method	0.07	1.00	0.12
Pattern method 1	0.11	0.08	0.09
Pattern method 2	1.00	0.002	0.003

of the word's frequency in positive examples to its frequency in the data. "Frequency in positive" and "Frequency in negative" respectively show the number of times the word appears in positive and negative examples.

As the table shows, the words obtaining the highest ratios included *jinsei* (human life), *hitobito* (people), *koufuku* (happiness), *yujou* (friendliness), *seishun* (youth), and *ren'ai* (love). Sentences in which one or more of these words are used are likely to be impressive sentences.

These results are the most important and interesting points in this paper. We found that using the words shown in the table is a good approach to use if we would like to generate impressive sentences.

Shown below are example sentences containing *jinsei* (human life), *hitobito* (people), and *koufuku* (happiness).

Example sentences containing *jinsei* (life):

jinsei wa	*douro no*	*youna mono da.*	*ichibanno*	*chikamichi wa*	*taitei*	*ichiban warui.*	*michi da.*
(life)	(road)	(be like)	(first)	(shortcut)	(usually)	(worst)	(road)

(Life is like a road. The first shortcut is usually the worst road.)

Example sentences containing *hitobito* (people):

hitobito wa	*kanashimi wo*	*wakachi attekureru*	*tomodachi*	*sae ireba*	*kanashimi wo*	*yawaragerareru.*
(people)	(sadness)	(share)	(friend)	(if only they have)	(sadness)	(can soften)

(People can soften their sadness, if only they have a friend with whom they can share it.)

Example sentences containing *koufuku* (happiness):

fukouna	*hito wa*	*kibou wo*	*mote.*	*koufukuna*	*hito wa*	*youjin seyo.*
(unhappy)	(person)	(hope)	(should have)	(happy)	(person)	(should be on one's guard)

(Unhappy people should have hope. Happy people should be on their guard.)

4 Automatic impressive sentence extraction performance

The method we describe in this paper is a useful one for automatically extracting impressive sentences. In this section, we evaluate the extraction performance of this and other methods.

The evaluation results are shown in Table 2. The data set for evaluation consists of 10,000 new sentences from Web documents. We use each method to extract positive sentences from the set for evaluation. We then randomly extract 100 data items (200 for the baseline method only) from the sentences extracted by each method and manually evaluate them. From the evaluation results we approximately calculate the precision rates, the recall rates, and the F-measures.

We estimate the denominator of the recall rate from the number of positive examples detected by the baseline method. The baseline method judges that all the inputs are positive.

In the "ML method (xth)" we use supervised data for machine learning after adding the xth positive and negative examples to the supervised data (by the method in Section 2.2). In "Pattern method 1" we extract sentences that contain words whose positive appearance ratio is at least 0.8 and that appear at least four times as positive examples. In "Pattern method 2" we extract sentences that contain the word "*kando*" (impression) as positive examples.

With machine learning we obtain a precision rate of 0.40 after we add the 10th positive and negative examples to the supervised data. This precision rate is much higher than the 0.07 rate we obtain with the baseline method.

Some may think that the 0.40 precision rate obtained with machine learning is low. However, since the task of extracting impressive sentences is a very difficult one, and since the rate is much higher than the baseline method rate, we can say that the machine learning results are at least adequate.

5 Related studies

Many methods have been reported that estimated the orientation (positive or negative contents) or the emotion of a sentence (Turney and Littman, 2003; Pang and Lee, 2008; Kim and Hovy, 2004; Alm et al., 2005; Aman and Szpakowicz, 2007; Strapparava and Mihalcea, 2008; Inkpen et al., 2009; Neviarouskaya et al., 2009). However, the studies did not address the task of collecting and analyzing impressive sentences to support the generation of such sentences.

There have been studies that addressed the task of automatically evaluating sentences to support sentence generation (Bangalore and Whittaker, 2000; Mutton and Dale, 2007). However, the studies did not address the task of generating impressive sentences.

In our study, we used machine learning to extract impressive sentences. There have been other studies as well in which machine learning was used to extract information (Murata et al., 2011; Stijn De Saeger and Hashimoto, 2009). Murata et al. extracted articles describing problems, their solutions, and their causes (Murata et al., 2011). Saeger et al. extracted several types of words from a large scale of Web documents by using machine learning (Stijn De Saeger and Hashimoto, 2009). In their method, they manually make supervised data sets for extracted words and extract more words from Web documents using supervised methods. Their study is similar to ours in that both use the same framework of manually making a small scale supervised data set and then extracting more data items from Web documents.

6 Conclusion

We collected sentences in Japanese that impressed readers ("impressive sentences") and examined them through the use of characteristic words in order to support the generation of impressive sentences. In our examination, we obtained *jinsei* (human life), *hitobito* (people), *koufuku* (happiness), *yujou* (friendliness), *seishun* (youth), *ren'ai* (love), etc. as words that often appear in impressive sentences. Sentences in which one or more of these words are used would be likely to impress the listener or reader. The results we obtained should provide useful information for generating impressive sentences.

In this study, we used machine learning to extract impressive sentences and found that with this method we could extract them from a large amount of Web documents with a precision rate of 0.40.

In future work, we intend to use this method to collect more impressive sentences. We also plan to analyze the sentences by using not only words but also parameters such as syntax patterns and rhetorical expressions.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 23500178.

References

Cecilia O. Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language (HLT/EMNLP 2005)*, pages 579–586.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue (TSD'07)*, pages 196–205.

Srinivas Bangalore and Owen Rambow and Steve Whittaker. 2000. Evaluation metrics for generation. In *Proceedings of the first international conference on Natural language generation (INLG '00)*, pages 1–8.

Peng Chen and Tao Wen. 2006. Margin maximization model of text classification based on support vector machines. In *Machine Learning and Cybernetics*, pages 3514–3518.

Nello Cristianini and John Shawe-Taylor. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Diana Inkpen, Fazel Keshtkar, and Diman Ghazi. 2009. Analysis and generation of emotion in texts. In *Knowledge Engineering: Principle and Technique, KEPT 2009*, pages 3–14.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, pages 1–7.

Daisuke Kawahara and Sadao Kurohashi. 2006. Case frame compilation from the web using high-performance computing. *Proceedings of the 5th International Conference on Language Resources and Evaluation,*, pages 1–4.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of Coling 2004*, pages 1367–1373.

Taku Kudoh and Yuji Matsumoto. 2000. Use of support vector learning for chunk identification. *CoNLL-2000*, pages 142–144.

Tomohiro Mitsumori, Sevrani Fation, Masaki Murata, Kouichi Doi, and Hirohumi Doi. 2005. Gene/protein name recognition based on support vector machine using dictionary as features. *BMC Bioinformatics*, 6(Suppl 1)(S8):1–10.

Masaki Murata, Qing Ma, and Hitoshi Isahara. 2002. Comparison of three machine-learning methods for Thai part-of-speech tagging. *ACM Transactions on Asian Language Information Processing*, 1(2):145–158.

Masaki Murata, Hiroki Tanji, Kazuhide Yamamoto, Stijn De Saeger, Yasunori Kakizawa, and Kentaro Torisawa. 2011. Extraction from the web of articles describing problems, their solutions, and their causes. *IEICE Transactions on Information and Systems*, E94–D(3):734–737.

Andrew Mutton and Mark Dras and Stephen Wan and Robert Dale. 2007. Gleu: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 344–351.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2009. Compositionality principle in recognition of fine-grained emotions from text. In *Proceedings of 4th International AAAI Conference on Weblogs and Social Media (ICWSM 2009)*, pages 278–281.

Bo Pang and Lillian Lee. 2008. Opinion minding and sentiment analysis. *Foundation and Trend in Information Retrieval*, 2(1-2):1–135.

Kentaro Torisawa Masaki Murata Ichiro Yamada Kow Kuroda Stijn De Saeger, Jun'ichi Kazama and Chikara Hashimoto. 2009. A web service for automatic word class acquisition. In *Proceedings of 3rd International Universal Communication Symposium (IUCS 2009)*, pages 132–137.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC '08)*, page 1556?1560.

Koichi Takeuchi and Nigel Collier. 2003. Bio-medical entity extraction using support vector machine. *Proceedings of the ACL 2003 Workshop on NLP in Biomedicine*, pages 57–64.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.