https://doi.org/10.20965/jaciii.2020.p0422

Paper:

Study on Development of Humor Discriminator for Dialogue System

Tomohiro Yoshikawa and Ryosuke Iwakura

Graduate School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8603, Japan
E-mail: yoshi@suzuka-u.ac.jp

[Received October 31, 2019; accepted April 6, 2020]

Studies on automatic dialogue systems, which allow people and computers to communicate with each other using natural language, have been attracting attention. In particular, the main objective of a non-task-oriented dialogue system is not to achieve a specific task but to amuse users through chat and free dialogue. For this type of dialogue system, continuity of the dialogue is important because users can easily get tired if the dialogue is monotonous. On the other hand, preceding studies have shown that speech with humorous expressions is effective in improving the continuity of a dialogue. In this study, we developed a computer-based humor discriminator to perform user- or situation-independent objective discrimination of humor. Using the humor discriminator, we also developed an automatic humor generation system and conducted an evaluation experiment with human subjects to test the generated jokes. A t-test on the evaluation scores revealed a significant difference (P value: 3.5×10^{-5}) between the proposed and existing methods of joke generation.

Keywords: automatic dialogue system, non task oriented, humor discriminator

1. Introduction

Studies on automatic dialogue systems, which allow people and computers to communicate with each other using a natural language, have been attracting attention. Typical examples of practical automatic dialogue systems include Apple's "Siri,"[1] NTT Docomo's "Shabette concierge,"[2] Softbank's "Pepper,"[3] Microsoft's "Rinna"[4] and "Cortana,"[5] Google's "Google Home (Google Nest),"[6] and Amazon's "Amazon Echo (Alexa)."[7]

Such systems can be classified into task-oriented and non-task-oriented systems. A task-oriented dialogue system aims to achieve a specific task by providing answers to a user's question or request, through a dialogue. For example, weather information systems [1] and tourist information systems [2] are typical task-oriented systems. In addition, Siri, Shabette concierge, Cortana, Google Home (Google Nest), and Amazon Echo (Alexa) in the aforementioned list are also task-oriented systems because they are used for management of schedules or home electric appliances, or for support of Web browsing.

On the other hand, the main objective of a non-task-oriented dialogue system is not to achieve a specific task but to amuse users through chats and free dialogue. Rinna, among the given examples, is a non-task-oriented dialogue system because its major aim is chatting with users. This system is sometimes called "chat dialogue system" or "chatting system." For this sort of dialogue system, continuity of the dialogue is important because users can easily get tired if the dialogue is monotonous.

For a user to want to continue the dialogue, speech with humorous expressions [3], frequent responses in the dialogue [4], and conversations about hobbies or preferences [5] are found to be effective. If we focus on humor in particular, we find that there are several studies on the generation of humor: automatic generation of comic scenarios [6, 7], support for browsing funny images to use for presentations [8], and comedy talks by multiple robots [9]. Topics of studies on humor generation associated with dialogue systems include telling riddles [10], making jokes [11, 12], and creating humor suitable for dialogue [13].

However, humor generated from these systems is considered to be insufficiently funny. One of the reasons for this problem is the ambiguity feature of humor, as funniness can change depending on the listeners and situations. In this study, we discriminate the funniness of the generated humor by using a computer. Automatic judgment by the computer is expected to be able to discriminate funniness in a way that does not depend on the people or situations; in other words, the system should be able to circumvent the ambiguity of humor. In addition, with an automatic humor discrimination function, one can generate humor that the computer judges to be funny and use that kind of humor in a dialogue.

In this research, we first develop a humor discrimina-

[1] http://www.apple.com/jp/ios/siri/
[2] https://www.nttdocomo.co.jp/service/shabette_concier/
[3] http://www.softbank.jp/robot/consumer/products/
[4] http://www.rinna.jp/
[5] https://www.microsoft.com/ja-jp/windows/cortana
[6] https://store.google.com/jp/product/google_nest_hub
[7] https://developer.amazon.com/ja/echo

© Fuji Technology Press Ltd. Creative Commons CC BY-ND: This is an Open Access article distributed under the terms of the Creative Commons Attribution-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nd/4.0/).

tor on the basis of manual evaluation and, from the results of automatic discrimination, determine the effective features of humor. We then look for an appropriate combination of the features used for discrimination to complete the humor discriminator. Furthermore, we check the validity of the humor discriminator. To generate humor, we use a method based on dynamic programming (DP) matching [11] as the baseline. We generate jokes via DP matching and output, as humorous expression, only the joke data that the discriminator judges to be funny. We perform an evaluation experiment with human subjects and conduct a t-test for the evaluation scores. As a result, we confirm that our method could enhance funniness significantly (P value: 3.5×10^{-5}).

2. Related Studies

In this study, we assume a simple form of humorous expressions as the basis of the humor for developing automatic humor discrimination and focus on the funniness that arises in a combination of words. For each of these expressions, we consider not only a single word but a combination of multiple words; whereas it is usually difficult for a single word to produce funniness, a combination of two words can produce funniness based on their relationship with each other.

In previous studies, discrimination methods for humor with a story nature [14, 15] and for jokes [12] have been reported. Humor with a story nature is expressed in a prose consisting of several sentences with a punch line. Previous methods [14, 15] have discriminated humor by using the bag-of-words model, which checks the occurrence frequencies of words, and word2vec [16, 17] to find features. However, these methods examine sentences with certain lengths and are not intended to find the funniness of a word or combination of words. Meanwhile, the joke discrimination method [12] examines the funniness of words, but only when these words are in a joke style.

In this research, we study funniness as a whole, generated not only by jokes but also by word combinations. We use combinations of nouns because nouns are important and are the most common words. To create a combination of two nouns in the form “noun+preposition+noun,” we consider “of” as the preposition because “of” appears the most frequently in Twitter data. In other words, we develop a humor discriminator for “noun$_2$ of noun$_1$.” Because we restrict these combinations of words to only two kinds of nouns, we then need to consider the features that capture their funniness.

Few studies have been conducted on humor generation methods that use humor discrimination, because, in the first place, only a few studies have been made on humor discrimination methods. Research has been conducted on a method of speech generation for a dialogue system [18] that uses a joke discrimination method to generate jokes in dialogue, although this is the only study that has focused on a joke discrimination method.

Here, we attempt to generate funnier jokes of the form “noun$_2$ of noun$_1$.” by using a discriminator that we develop. In particular, for joke candidates generated via a conventional joke generation method, we perform humor discrimination and output only the jokes that are judged to be funny.

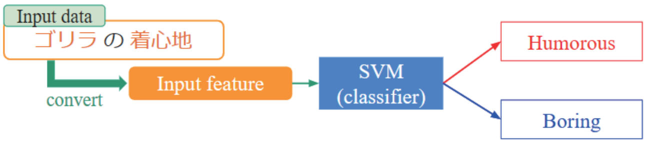


Fig. 1. Overview of humor discriminator.

3. Development of Humor Discriminator

3.1. Manual Collection of Evaluation Data

For the development of a humor discriminator using a support vector machine (SVM), we first collect manual evaluation data. We use humor expressions consisting of two nouns combined by “of,” namely “noun$_2$ of noun$_1$.” The two nouns are chosen randomly. However, combinations of nouns that frequently co-occur in ordinary sentences or that are unpopular and unfamiliar are excluded. This is because we have previously obtained results revealing that nouns that frequently co-occur are usually not funny, whereas unpopular nouns could not be easily imagined by an audience. After the expressions are chosen, multiple subjects evaluate whether each chosen set of “noun$_2$ of noun$_1$.” is funny. We then examine and discuss the results of the evaluation, and use these results to develop the humor discriminator.

3.2. Extraction of Effective Features for Humor Discrimination

Now, we consider and extract features that are capable of capturing humor. Referring to a preceding study [11], we first use as many feature candidates, that could be elements of funniness, as possible. The humor discriminator is then developed with SVM using the feature candidates to extract only the features that are effective as funniness elements. The humor discriminator is overviewed in **Fig. 1**.

The humor discrimination performs binary classification on the “noun$_2$ of noun$_1$” expressions, between funny (positive example) and not funny (negative example) expressions. We consider a used feature to be effective if the F-measure obtained via leave-one-out cross-validation [19] exceeds a threshold. For the creation and verification of the discriminator, we use the data collected in Section 3.1.

For the indices, we use precision (matching ratio), recall (reproduction ratio), F-measure, and accuracy.

Table 1. Classification of stuffed toys.

Part of speech	Class	Mid-category	Category	Category number	Header	Reading
Noun	Product	Tool	Toy, ornament, statue, etc.	1.4570	Stuffed toy	Stuffed toy

3.3. Finding Appropriate Combination of Effective Features for Humor Discrimination

With the features extracted in Section 3.2, we look for a combination of features that is capable of effectively capturing funniness. The procedure is outlined in the following paragraph. To search for an optimal combination, we need to examine all possible combinations of the features, which would consume a tremendously long amount of time for calculation. Therefore, we employ the following search procedure to develop a discriminator, which performs a small number of searches by ignoring unnecessary features if possible.

- BEGIN

  1. A humor discriminator based on SVM for all features extracted in Section 3.2 is created.
  2. The leave-one-out cross-validation is used for classification, and F-measure is calculated for reference and set to be F_{basis}.

- LOOP

  1. The humor discriminator is created again with a single kind of feature excluded.
  2. The leave-one-out cross-validation is used for classification, and F-measure is calculated and compared with F_{basis}.
  3. If the calculated F-measure is smaller than F_{basis}, the excluded feature is returned. If the F-measure is higher than F_{basis}, the new F-measure is set as the new F_{basis}, and the excluded feature is kept excluded.
  4. The previous steps are repeated until all features are processed (excluded or returned).

- END

  1. After the whole process is finished, the remaining features are considered as an appropriate combination. The F-measure calculated with this combination of the features is the final evaluation value.

With the aforementioned procedure, we search for an appropriate combination of the multiple features extracted in Section 3.2. The created discriminator is used as humor discriminator, having the F-measure obtained in the final calculation.

4. Experiment

4.1. Collection of Manual Evaluation Data

We collect data from manual evaluation to develop the humor discriminator using SVM.

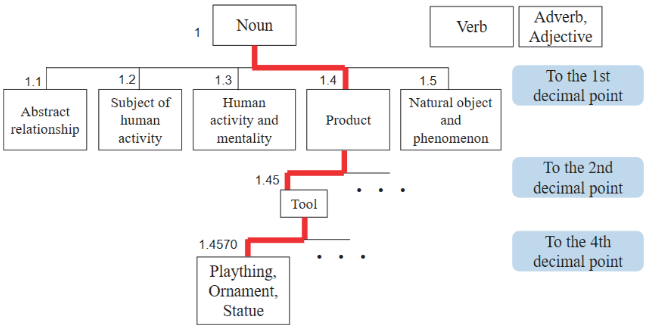


Fig. 2. Classification of stuffed toys in thesaurus.

4.1.1. Experiment Method

We use a thesaurus [20] to generate "noun_2 of noun_1" expressions to be evaluated. The thesaurus is a dictionary of about 100,000 words classified and arranged in terms of their meanings. As an example, the classification of the word "stuffed toy" in the thesaurus is shown in **Table 1** and **Fig. 2**.

Three nouns are selected from each of five classes with category numbers from 1.1 to 1.5 in the thesaurus. These 15 nouns are used as "noun_2." In addition, for each noun_2, ten nouns are randomly selected from each of 43 mid-categories with category numbers from 1.10 to 1.57 to obtain 430 kinds of "noun_1." As a result, we have $450 \times 15 = 6{,}450$ combinations of "noun_1" and "noun_2" to evaluate. To select "noun_2," we use Twitter data. Namely, for each of the classes with category numbers from 1.1 to 1.5, we randomly choose nouns from the top 50 nouns in terms of number of occurrences. The "noun_2" words that we use and their category numbers are listed in **Table 2**.

As mentioned in Section 3.1, the "noun_1" words with difficult meanings and those which are closely related to their counterparts (i.e., "noun_2") are excluded because such combinations are expected to be not funny. For this purpose, we perform a preliminary experiment. On the basis of the results of the preliminary experiment, we exclude the nouns whose search results in Google are less than 300,000, whereas the nouns that co-occur 20 times or more with the counterpart "noun_2" in Twitter data (about 60 million tweets from July 2017 to September 2017) are excluded, with another noun being selected again via the aforementioned procedure.

These 6,450 word combinations are then judged by eleven male university students on whether these expressions are funny. A screen system, as shown in **Fig. 3**, is used for the evaluation. In the experiment, funniness is defined for the subjects as "nature of things that are felt as being funny, including interesting or humorous, from a subjective point of view."

Table 2. Noun$_2$ used and classification number.

None$_2$ (in Japanese)	Classification number
Level	1.1101
Show-up	1.1210
Birthday	1.1633
Oneself	1.2020
World	1.2600
School	1.2629
Information	1.3123
Movie	1.3240
Event	1.3510
Ticket	1.4040
Oden (*Japanese food)	1.4310
Train	1.4650
Typhoon	1.5150
Cat	1.5501
Body	1.5600

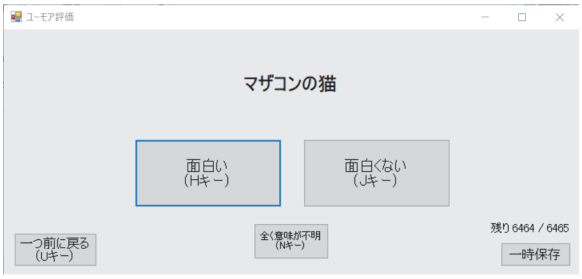


Fig. 3. System screen used in evaluation experiment.

4.1.2. Experiment Results

Table 3 outlines the number of word combinations that each evaluator felt were funny. The number of word combinations evaluated as funny varied largely from 20 to 618, depending on the evaluators. This could be because the evaluators got tired of the evaluation of 6,450 word combinations and because they have different preferences regarding funniness. A questionnaire survey made after the experiment revealed that there were many opinions akin to “I got tired of too many word combinations.”

Table 4 then lists the number of word combinations that multiple evaluators evaluated as being funny. Of these combinations, 73% were evaluated as not being funny by all of the evaluators. In other words, there exist a certain number of word combinations that are universally not funny, regardless of the evaluators’ preferences. The maximum number of evaluators who all evaluated the same word combination as being funny was 5, and the ratio of these expressions to the total was 0.3%. Therefore, there was no word combination that all the evaluators felt was funny and only few word combinations that multiple evaluators felt were funny.

Table 5 shows examples of humor expressions that five evaluators felt were funny. Some of them, such as “train of cold” and “cat of mother-complex,” could be felt as being funny because of the gap between the two words, such as between “cold” and “train” or between “mother-complex” and “cat.” In addition, “myself of the greatest” and “oden of foreign-species” could be evaluated as being funny because of the funniness of the words themselves, such as “the greatest” or “foreign-species.”

Table 3. Number of word combinations that each evaluator felt were funny.

	Number of word combinations evaluated as funny (/6,450 combinations)
Evaluator 1	318
Evaluator 2	117
Evaluator 3	20
Evaluator 4	590
Evaluator 5	66
Evaluator 6	70
Evaluator 7	67
Evaluator 8	161
Evaluator 9	43
Evaluator 10	618
Evaluator 11	476

4.2. Extraction of Features Effective for Humor Discrimination

Afterward, we examine and extract features that are capable of capturing humor. As explained in Section 4.1.2, the simple form of the word combination “noun$_2$ of noun$_1$” could have two elements of funniness: relationship (gap) between the two nouns, and each word’s own funniness (potential).

An example of an expression with funniness due to the “relationship between the two nouns” is “hardtack of first-class.” In this example, the funniness is caused by the gap between the high-grade atmosphere of “first-class” and the frugality of “hardtack.” Namely, images of the two nouns and their relationship cause the funniness.

An example of an expression with funniness due to “each word’s own funniness” is “comfort of gorilla.” In this example, the word “gorilla” itself has funniness, and basically, it can make the word combination funny no matter which word is combined. Namely, the potential of the noun itself causes the funniness.

Features that represent the relationship of the two nouns are listed as follows.

a) word2vec

b) Adjective vector

c) Gap in image

d) Mora number

e) Ratio of occurrence of word connected to a target noun

f) Category number

a) word2vec refers to the distributed representation of words that is learned with no supervision from the large-scale corpus. Because words similar to each other are learned as similar distributed representations, this

Table 4. Number of word combinations that multiple evaluators felt were funny.

Number of evaluators who evaluated the expression as funny (/11 evaluators)	Number of word combinations	Ratio (/6,450)
0 evaluator	4,717	73%
1 evaluator	1,172	18%
2 evaluator	386	6.0%
3 evaluator	116	1.8%
4 evaluator	41	0.6%
5 evaluator	18	0.3%

Table 5. Examples of humor expression that five evaluators felt were funny.

Train of cold
Cat of mother-complex
Myself of the greatest
Oden of foreign-species (*Oden is a Japanese food.)
Birthday of unanimity

Table 6. Example of antonym set.

Good	Bad
Cute	Scary
Tasty	Non-tasty
Wide	Narrow
Hot	Cold

technique is often used as a method for accurately obtaining synonyms. We can then use the obtained distributed representation of words as a feature that represents the meanings of nouns. With this feature, we would be able to grasp an image of the two nouns. We use Japanese Wikipedia for the learning corpus, morphological analyzer MeCab [21, 22] for separating the words, and mecab-ipadic-NEologd [23] as the dictionary. The dimension of the word vector is set to 300.

b) An adjective vector is a word vector that uses co-occurrences with adjectives. The number of co-occurrences of a noun and an adjective is calculated in corpus, and therefore the adjective vector is a feature that reflects an image of the noun. Although it resembles word2vec, the latter uses words of all parts of speech. Because the adjective vector uses the co-occurrence of a noun with an adjective that expresses human feelings, among others, it would reflect an image of the noun more directly. We use Twitter data as corpus. We also use MeCab for separating words and extract only the adjectives and nouns. From the extracted adjectives, we choose 36 adjectives that have many occurrences and whose antonyms are also adjectives. Using the 36 adjectives and their respective antonyms as 36 pairs, we create an adjective antonym set. An example of the antonym set is shown in **Table 6**.

The adjective vectors are created via the following procedure.

- The extracted data are used to calculate the number of co-occurrences of each adjective from the antonym set for each noun (co-occurrence is defined within a single tweet).
- An adjective vector with 72 dimensions ($= 36 \times 2$), having the numbers of co-occurrences as its elements, is created for each noun.
- Because some adjectives have zero co-occurrence, each element value is smoothed (Eq. (1)).
- To eliminate differences in the numbers of occurrences of the adjectives, each element is divided by the total number of occurrences of the adjectives to change it to the ratio of occurrence.

In other words, with the number of co-occurrences $n_{i,j}$ of noun $i(1,2,\cdots,N)$ and adjective $j(1,2,\cdots,M_a)$, the adjective vector $\boldsymbol{w}_i$ and its elements $w_{i,j}$ are defined as follows.

$$\boldsymbol{w}_i = (w_{i,1}, w_{i,2}, \cdots, w_{i,M_a})$$

$$w_{i,j} = \frac{n_{i,j}+1}{\sum_{i=1}^{N} n_{i,j} + N} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (1)$$

The norms of the adjective vectors are all set to 1.

c) Gap in image is the sum of the products of antonym elements between the adjective vectors of two nouns (for example, product of "tasty" of "oden" and "non-tasty" of "chilled-Chinese-noodle"). Through the use of the product of elements that are antonyms to each other, this feature can capture an image of the gap between two nouns.

d) The mora number is the number of characters in kana (which are the syllabaries that form parts of the Japanese writing system) of nouns, with consideration for the contracted sound characters (such as "ゃ," "ゅ," and "ょ"). For example, "kabocha" (meaning: "pumpkin") has three morae, whereas "ocha" (meaning: "tea") has two morae. This number expresses linguistic sense and can indicate affinity of rhythm, which is often seen in the 5-7-5 structure of haiku. When we see a noun, we usually read it silently. Therefore, the mora number would be a more appropriate feature than only the number of characters in kana.

e) In the Japanese language, certain words will sometimes be connected to the end of a noun to change its function to that of a verb or an adjective verb. The ratio of occurrence in corpus of a word that is connected to the end of a target noun is therefore also a feature that is used to represent the relationship between two nouns. This feature should be used because nouns comprise not only simple nouns that represent objects but also those nouns that have verb-like or adjective-verb-like meanings, which is a distinction that is important in Japanese grammar. Examples of simple nouns include "stuffed-toy" and "cat." Examples of verb-like nouns include "research" and "purchase," whereas examples of adjective-verb-like nouns include "flooding" and "pool-of-blood." These nouns contain stems of a verb or adjective verb. The funniness of "noun$_2$ of noun$_1$" could change depending on how the aforementioned three types of nouns are combined in the two-noun expression. One could judge the type of a noun by examining the types of words that can be connected to the end of that noun. For example, the word "suru (verb)" (meaning "do" in Japanese) can be connected to the end of a verb-like noun, and the word "na" (conjugative suffix of adjective verb) to the end of an adjective-verb-like noun. Using the occurrence rates of these words as feature, one would be able to determine the types of the target nouns. Among the words that can be connected to the target nouns, we consider 14 kinds of these words, including "suru/na" and postpositional particles "wa/ga/shi/no/wo/ni/he/to/kara/yori/de/ya" of the Japanese language. We use Twitter data as corpus and calculate the ratios of occurrences of these words connected to each of the target nouns. As shown in Eq. (2), the number of occurrences $n_{i,j}$ of noun i and word $j(1,2,\ldots,M_c;\ M_c = 14)$, which is connected to the noun, is divided by the total number of occurrences to obtain the ratio of occurrence $w_{i,j}$.

$$w_{i,j} = \frac{n_{i,j}}{\sum_{j=1}^{M_c} n_{i,j}} \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \quad (2)$$

f) The category number of each noun in the thesaurus list is also a feature that can represent the relationship between two nouns. In **Fig. 2**, nouns in the thesaurus have been classified manually, and category numbers have been assigned. Using these numbers, we can find a combination of nouns that has a funny meaning. In the experiment, the five categories of nouns with category numbers 1.1 through 1.5 are converted to dummy variables.

On the other hand, features representing a noun's own funniness are listed as follows.

g) Imageability
h) Number of occurrences in corpus
i) Number of search results in Google
j) Number of hiragana/katakana/Chinese characters
k) Number of occurrences in power words
l) Number of occurrences in tweets with 10,000 likes or more
m) Number of occurrences in tweets with 1,000 likes or more

g) Imageability is a feature used in a preceding study [11]. It is presented in NTT's imageability database [24], and the ease of instinctively imagining an event that a word represents is measured with a positive real value. For a noun to be evaluated as being funny, the image that people have for the noun is important, and the ease of imaging the noun is essential. Because of the limitation of vocabulary in the database [24], the ease of co-occurrence with adjectives in a corpus is used as this feature in the present experiment. We use Twitter data as corpus. As shown in Eq. (3), the total number of occurrences $n_{i,j}$ of noun i and adjective $j(1,2,\ldots,M_a)$ in the corpus is divided by the number of occurrences N_i of the noun i to calculate ease e_i of co-occurrence with adjectives.

$$e_i = \frac{\sum_{j=1}^{M_a} n_{i,j}}{N_i} \quad . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \ . \quad (3)$$

h) The number of occurrences in corpus indicates how familiar the noun is to the general population. For unfamiliar nouns, people may not have a concrete image or know the meanings of such nouns. Because one can consider nouns with higher numbers of occurrences in the corpus as the more general nouns, we use the number of occurrences as a feature. Here, we use Twitter data as corpus.

i) The number of search results in Google, like the number of occurrences in corpus, also signifies how familiar the noun is to the general population. One can consider nouns with higher numbers of search results in Google as the more general nouns. This feature is also considered to be able to circumvent inconsistent spelling.

j) The number of hiragara/katakana/Chinese characters in a noun is a feature that is considered in determining the inherent funniness of the noun. Whereas the mora number represents the number of phonological characters, or characters that we say, this feature denotes the number of characters that we see. Hiragana sounds soft, whereas katakana sounds like foreign words. On the other hand, Chinese characters are formal. Thus, these three types of characters have different characteristics. Another difference is that hiragana and katakana are phonograms, whereas Chinese characters are logograms [25]. The number of characters of each of the three types could affect the visual image that we have of a noun.

k) The number of occurrences of a noun in Twitter "power words," or impressive words that are frequently used in Twitter, is also considered. An example of a tweet containing a power word is "Spicy-curry rice gratin with a hamburger steak on iron plate is a power-word-like product name." In this example, three different names of dishes, "spicy-curry," "rice gratin," and "hamburger steak on iron plate," are contained in a single product name, generating funniness in the words. In other words, power words create funniness from unfamiliar combination of words and often consist of funny words. Therefore, nouns that appear in the power words could be funny. We examine about 75,000 tweets that are extracted via searches for "power words" in Twitter.

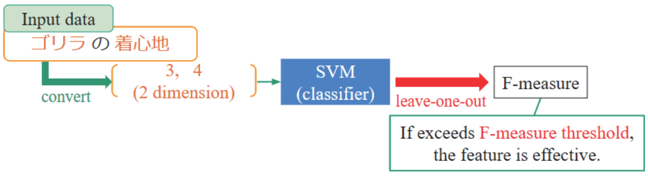


Fig. 4. Overview of development and verification of humor discriminator.

l) One of the features that may determine the funniness of a noun is the number of its occurrences in popular tweets. “Like” is a function in Twitter to register the tweets that we like. A large number of likes indicates that the tweet is popular. In general, a tweet with 10,000 or more likes is particularly popular. In addition, nouns that tend to occur in popular tweets would attract attention from many people.

m) This feature also looks at the number of occurrences of a noun in popular tweets. However, whereas feature “l” uses 10,000 as the threshold for the number of likes, feature “m” uses 1,000. This is a much smaller threshold than 10,000, but we can still consider these target tweets as popular.

4.2.1. Experiment Method

To extract the features that are effective as funny elements, from among the aforementioned features, we develop a humor discriminator using SVM with radial basis function (RBF) kernel for each kind of feature (a) to (m). **Fig. 4** shows an overview of this process. For the hyperparameters of SVM, we set $C = 8$ and $\gamma = 0.3$, according to the result of a simple coarse grid search.

In the humor discrimination, an input of “noun$_2$ of noun$_1$” is binary-classified to either “funny” (positive example) or “not funny” (negative example). For the development and verification of the discriminator, we use the data collected in Section 3.1. In the experiment, we define positive and negative examples of the data collected from eleven people in the following manner.

- Positive example: Word combination that at least one of the evaluators felt was funny
- Negative example: Word combination that no evaluator felt was funny

We evaluate the effectiveness of the six features that represent the relationship between two nouns and the seven features that represent the inherent funniness of a noun. To utilize these thirteen features, we use nouns that appear in the vocabulary of word2vec and whose total numbers of co-occurrences with adjectives in Twitter data (about 11 million tweets in September 2017) are 10 or higher. As a result, we use 3,252 out of the 6,450 word combinations in the collected evaluation data.

A feature is judged as effective if the F-measure value is equal to or higher than the threshold as a result of classification using the leave-one-out cross-validation. Here, we set the threshold of F-measure to 0.311. This value is chosen because the F-measure value is 0.311 in the case of random classification according to the ratio between the positive and negative examples in the learning data (3,252 word combinations of the evaluation data to use, 1,011 positive examples, and 2,241 negative examples).

4.2.2. Experiment Results

Using each of the thirteen features, we developed the humor discriminator with SVM and performed leave-one-out cross-validation. The result is shown in **Fig. 5**.

From the results, we show the features with F-measure values equal to or higher than the threshold and those with values less than the threshold.

- Features whose F-measure values are equal to or higher than the threshold:

  1. word2vec
  2. Adjective vector
  3. Gap in image
  4. Mora number
  5. Ratio of occurrence of word connected to a target noun
  6. Category number
  7. Imageability
  8. Number of occurrences in corpus
  9. Number of search results in Google
  10. Number of hiragana/katakana/Chinese characters

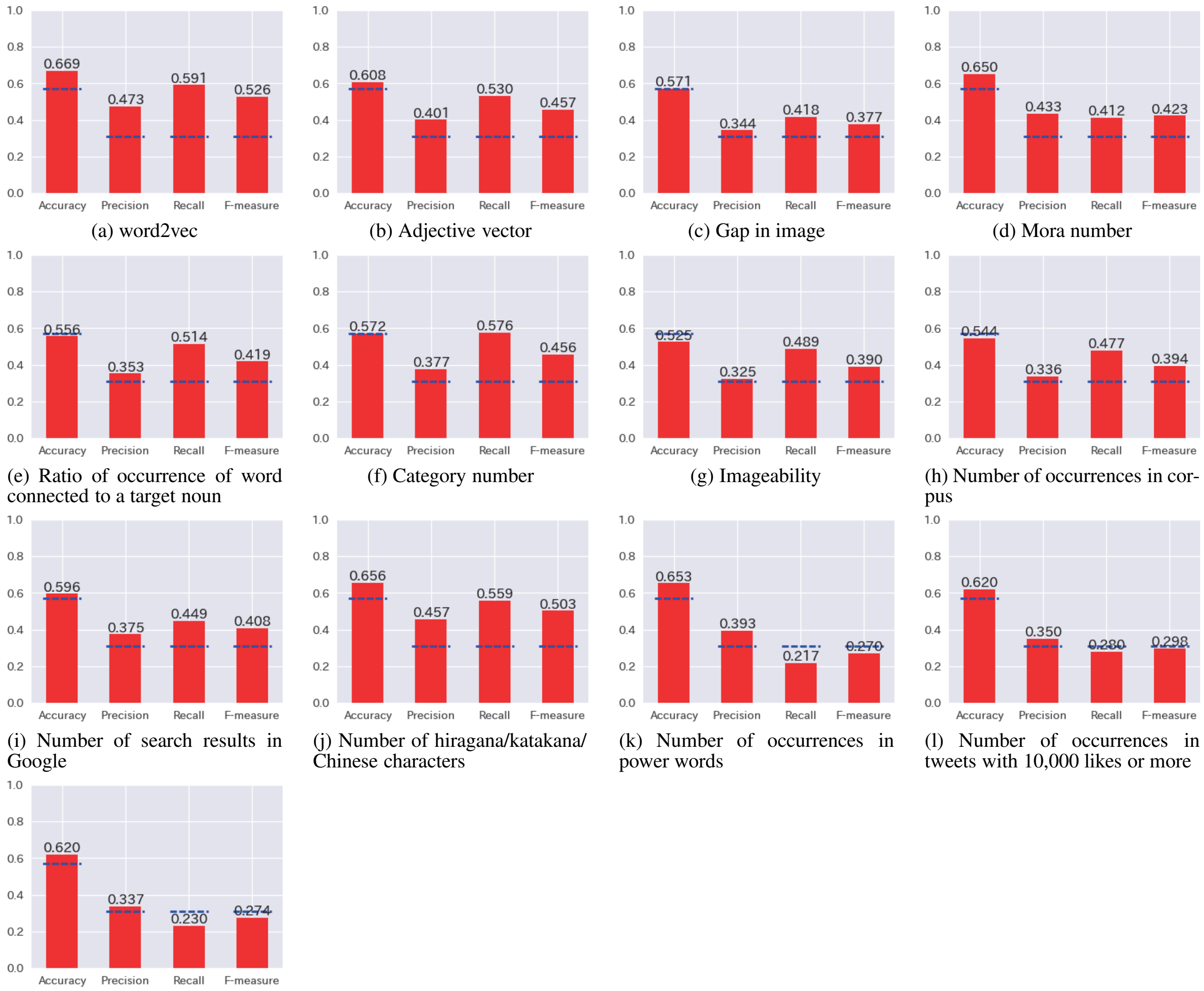
- Features whose F-measure values are lower than threshold:

  1. Number of occurrences in power words
  2. Number of occurrences in tweets with 10,000 likes or more
  3. Number of occurrences in tweets with 1,000 likes or more

The features whose F-measure values are equal to or higher than the threshold would be effective as funny elements. In the next section, we search for appropriate combinations of these features.

4.3. Search for Appropriate Combination of Effective Features for Humor Discrimination

To develop the humor discriminator, we first search for appropriate combinations of the features extracted in Section 4.2 whose F-measure values are equal to or higher than the threshold. However, because the number of occurrences in power words improved the scores if combined with other features in preliminary experiments, it was added to the feature set for searching for an appropriate feature combination. Namely, we perform searches with a total of 11 kinds of features.



(a) word2vec

(b) Adjective vector

(c) Gap in image

(d) Mora number

(e) Ratio of occurrence of word connected to a target noun

(f) Category number

(g) Imageability

(h) Number of occurrences in corpus

(i) Number of search results in Google

(j) Number of hiragana/katakana/Chinese characters

(k) Number of occurrences in power words

(l) Number of occurrences in tweets with 10,000 likes or more

(m) Number of occurrences in tweets with 1,000 likes or more

Fig. 5. Results of classification using different features.

4.3.1. Experiment Method

The procedure to search for an appropriate combination is described in Section 3.3. The order of removal is outlined as follows, with consideration for each feature's dimension and contribution to discrimination.

1. (e) Ratio of occurrence of word connected to a target noun
2. (b) Adjective vector
3. (i) Number of search results in Google
4. (a) word2vec
5. (f) Category number
6. (j) Number of hiragana/katakana/Chinese characters
7. (k) Number of occurrences in power words
8. (g) Imageability
9. (h) Number of occurrences in corpus
10. (d) Mora number
11. (c) Gap in image

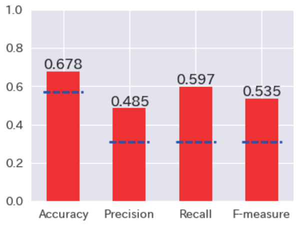
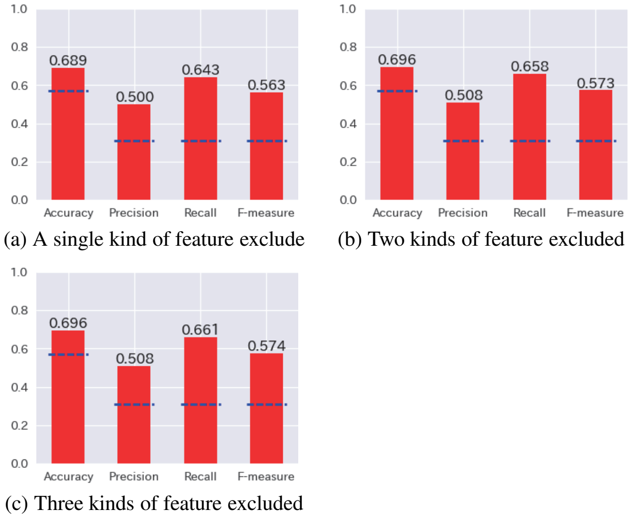


Fig. 6. Results of classification using a combination of all eleven features.

4.3.2. Experiment Result

Figure 6 shows the result of making the humor discriminator when all 11 features were used. On the other hand, **Fig. 7** shows the results of making the discriminator when one, two, or three features were excluded, in the order listed in Section 4.3.1.

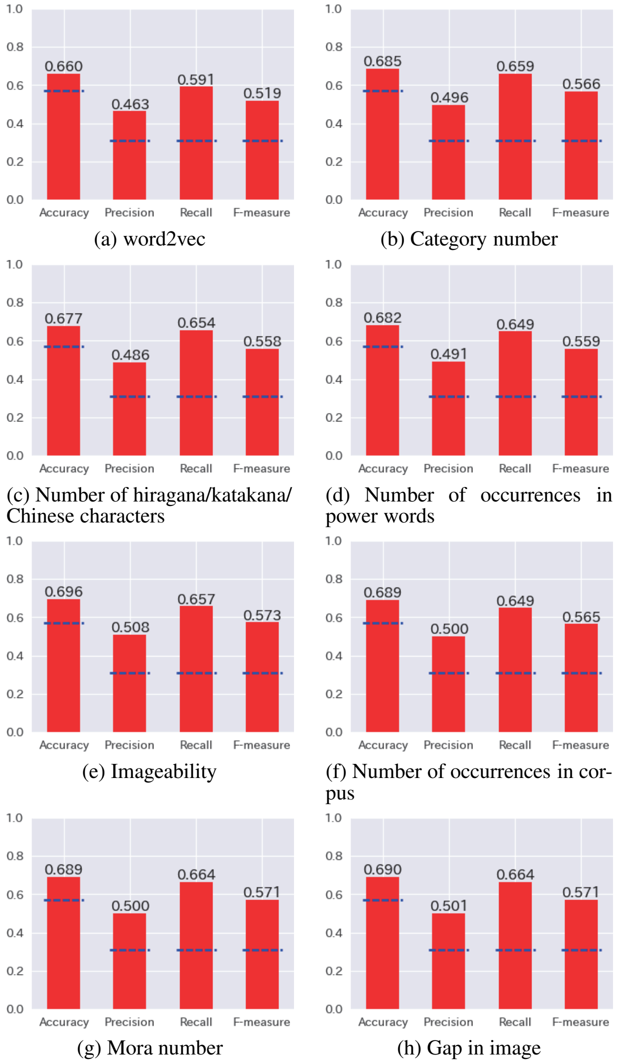
The results in **Figs. 6** and **7** reveal that the F-measure values and other scores are improved if some features are excluded, compared to the case where all eleven features



(a) A single kind of feature exclude (b) Two kinds of feature excluded

(c) Three kinds of feature excluded

Fig. 7. Results of classification with some features excluded.



(a) word2vec (b) Category number

(c) Number of hiragana/katakana/ Chinese characters (d) Number of occurrences in power words

(e) Imageability (f) Number of occurrences in corpus

(g) Mora number (h) Gap in image

Fig. 8. Results of classification with three specific features, plus an additional feature, excluded.

are combined. **Fig. 7** also shows that, among the results for when one/two/three features were excluded, the results for when three features (specifically, (e), (b), and (i)) were excluded have the highest score.

In addition, **Fig. 8** shows the results for when an additional feature was excluded, one at a time, in the order listed in Section 4.3.1. As for the F-measure, all eight results in **Fig. 8** are lower than the 0.574 value in **Fig. 7(c)**. In other words, the eight features used in **Fig. 7(c)** are the appropriate combination for developing the humor discriminator. We again list these eight features as follows.

- word2vec
- Category number
- Number of hiragana/katakana/Chinese characters
- Number of occurrences in power words
- Imageability
- Number of occurrences in corpus
- Mora number
- Gap in image

The combination of these features has 625 dimensions, and the accuracy and precision of the developed humor discriminator (**Fig. 7(c)**) are about 70% and about 50%, respectively. Precision indicates the ratio of the number of actually funny word combinations to the number of word combinations judged funny by the discriminator. In the experiment, a positive example (i.e., “funny”) was defined as a “word combination that at least one evaluator felt was funny,” and therefore, about half of the word combinations that the humor discriminator judged as being funny could actually be funny to someone.

The data that the developed humor discriminator judged as being funny are shown in **Figs. 9** and **10**, which give the numbers and ratios, respectively, of the word combinations that are judged as being funny by the discriminator and by different numbers of evaluators on the basis of the data collected in Section 3.1 from the eleven evaluators. **Fig. 10** shows that the word combinations that many evaluators felt were funny were also judged as being funny by the discriminator. The word combinations that four or five evaluators felt were funny were also correctly discriminated (as funny ones) by the discriminator. In the experiment, a positive example (i.e., “funny”) is defined as a “word combination that at least one evaluator felt was funny,” and therefore, the word combinations that only one evaluator felt were funny and those which multiple evaluators felt were funny are treated similarly. However, **Fig. 10** shows an increasing ratio as a function of the number of evaluators who judged the expression as being funny, indicating that the humor discriminator that we developed may be capable of capturing the characteristics of funniness.

5. Joke Generation with Humor Discriminator

Here, we use the humor discriminator to generate jokes. We generate jokes using the discriminator because, by

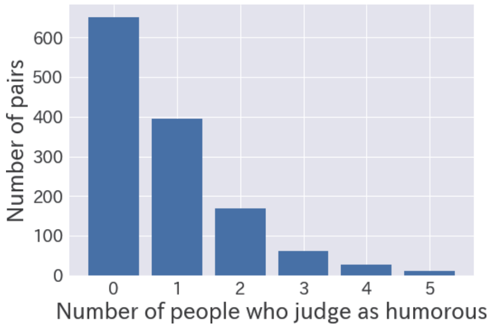


Fig. 9. Number of word combinations judged as being funny.

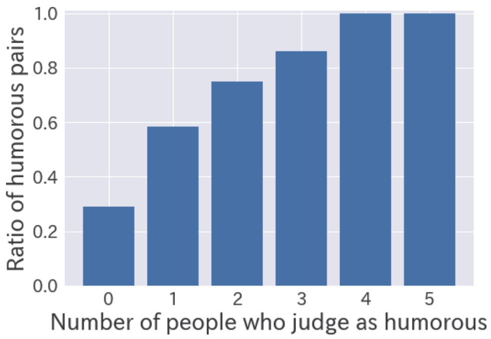


Fig. 10. Ratio of the number of word combinations judged as being funny.

doing so, we can compare our method with an existing one [11] to verify the effectiveness of the discriminator. The discriminator that we developed can be used for the overall funniness that arises in combining words and therefore be used to discriminate the funniness of jokes.

5.1. Joke Generation Method

In the following subsection, we outline a procedure for joke generation using the humor discriminator.

1. Extract ordinary terms in the form “noun$_2$ of noun$_1$.”
2. Calculate the score of DP matching [11] of “noun$_1$” and all possible “noun$_{1after}$.”
3. Replace “noun$_1$” with “noun$_{1after}$” and apply the humor discriminator to “noun$_2$ of noun$_{1after}$.”
4. Randomly select a “noun$_2$ of noun$_{1after}$” from those having minimum DP matching score among those judged as being funny by the discriminator.

In the calculation of the phonetic similarity of words via DP matching [11], a penalty score for discrepancy or difference in phonemes is calculated. A smaller score indicates higher similarity of phenomes. The following demonstrates an example of discrepancy or difference of phonemes.

- “Mikan” (meaning “orange” in Japanese) and “Hikan” (meaning “despair” in Japanese) have a discrepancy between m and h.

Table 7. Top five “noun$_2$ of noun$_1$” in terms of number of occurrences in corpus.

Saying of entertainer
Party of hope
Sideline of adult
Chance of travel
Influence of typhoon

- “Mikan” and “Arumikan” (meaning “aluminum can” in Japanese) have a difference of aru.

Previous baseline methods have used DP matching but with no humor discriminator.

One of the ways of direct use by the dialogue system is that the system converts “noun$_1$,” in the form “noun$_2$ of noun$_1$” from the user’s speech into a humor expression and then outputs the joke. The following is an example of such a dialogue.

User: “Shokuba no senpai (a “senior-staff of company,” in Japanese) scolded me.” System: “Never mind about Shokupan no senpai (a “senior-staff of bread,” in Japanese)!”

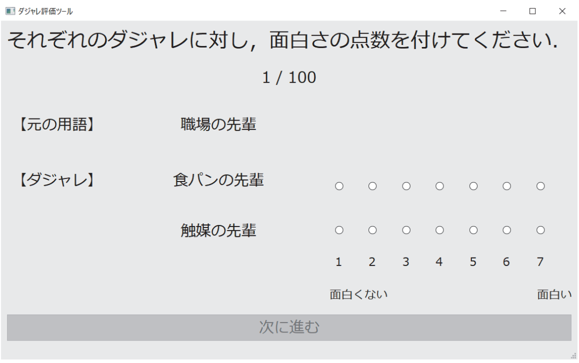
In this example, “Shokupan (“bread” in Japanese)” and “Shokubai (“catalysis” in Japanese),” which have high phonetic similarities to “Shokuba (“company” in Japanese),” are extracted as possible components for a joke in response to the user’s speech “Shokuba no senpai.” In addition, the humor discriminator judges the funniness of “Shokupan no senpai” and “Shokubai no senpai” and decides to output “Shokupan no senpai.”

5.2. Joke Generation Experiment

We generate jokes using the proposed method in Section 5.1 and check the effectiveness of this method via manual evaluation. For the extraction of ordinary terms in the form “noun$_2$ of noun$_1$,” we extract the top 100 “noun$_2$ of noun$_1$” expressions in terms of their numbers of occurrences in Twitter corpus. As examples, the top five of these expressions are listed in **Table 7**.

The DP matching scores of all “noun$_1$” words of the extracted “noun$_2$ of noun$_1$” expressions are calculated from the thesaurus. A new “noun$_2$ of noun$_1$” is generated via replacement of the original “noun$_1$.” New word combinations of the form “noun$_2$ of noun$_1$” that the humor discriminator judges to be funny are then collected, and the word combinations are randomly output in ascending order in terms of score. The output from the previous method are the randomly selected word combinations in ascending order in terms of score. In the thesaurus, there are 22,409 kinds of nouns that appear in the vocabulary of word2vec and whose numbers of occurrences in the corpus are 10 or higher.

Funniness is evaluated on a scale of seven grades: Grade 1 (not funny) to Grade 7 (funny). Each of the thirteen evaluators evaluate 100 word combinations generated using either the proposed method or an already existing method. **Fig. 11** shows a screen of the tool used



ダジャレ評価ツール

それぞれのダジャレに対し，面白さの点数を付けてください．

1 / 100

【元の用語】　職場の先輩

【ダジャレ】　食パンの先輩　○ ○ ○ ○ ○ ○ ○

触媒の先輩　○ ○ ○ ○ ○ ○ ○

1 2 3 4 5 6 7

面白くない　面白い

次に進む

Fig. 11. Tool screen for joke evaluation.

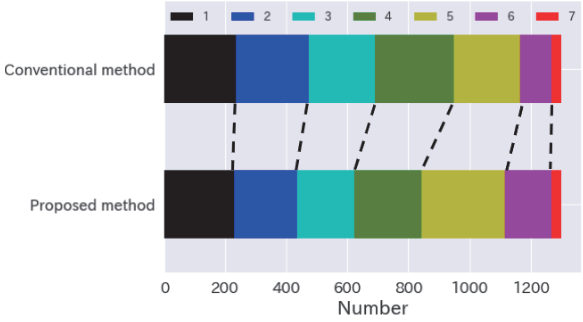


Fig. 12. Score distribution of jokes.

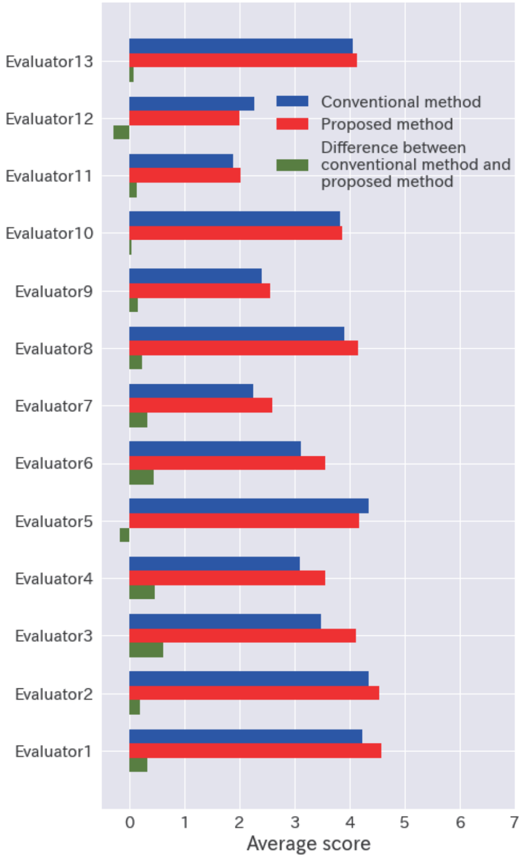


Fig. 13. Average scores by different evaluators.

for the evaluation. As in the figure, the evaluators do not know whether a given joke was generated via the proposed method or an already existing method (jokes are presented randomly).

5.3. Experiment Result

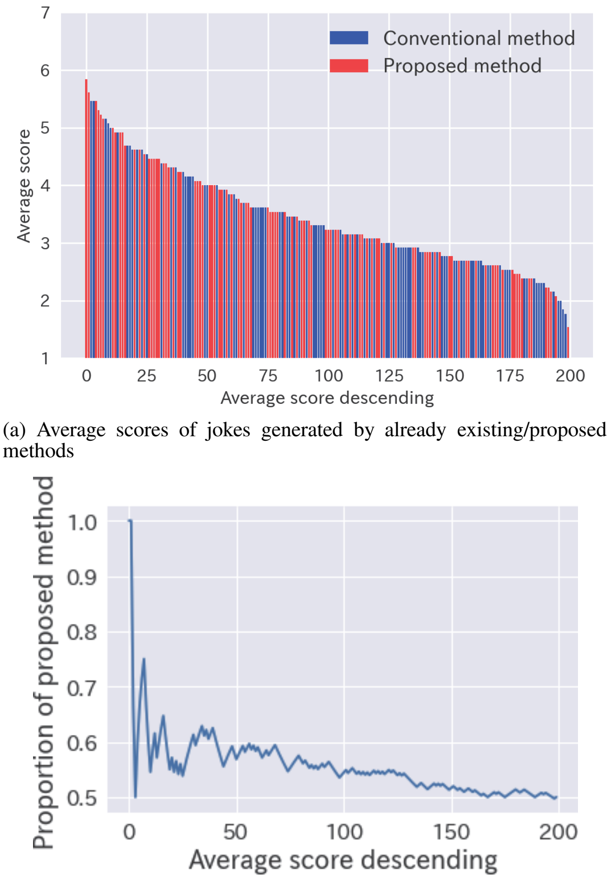
Figure 12 shows the score distribution of the jokes generated via the proposed and already existing methods. From the figure, the jokes generated by the proposed method are observed to have higher scores than those by the already existing method. In particular, the proportion of jokes with scores of 6 or 7, which the evaluators felt were funny, is larger for the proposed method. A t-test (two-sided test with 1,299 degrees of freedom) is conducted for the average difference between the proposed and already existing methods. The resulting t value is 4.2 and P value is 3.5×10^{-5}. This confirms that the joke generation by the proposed method can enhance the funniness of a dialogue. The t-test is conducted for the scores of 1,300 jokes (i.e., 100 jokes multiplied by thirteen evaluators) generated by the proposed and already existing methods.

However, even among the jokes generated by the proposed method, almost half have scores less than 4 and are evaluated as "not funny" by the evaluators. Namely, the jokes generated by the proposed method are not always felt as being funny. This is because the humor discriminator used in the experiment was developed via learning of the data considered as "positive examples," i.e., that at least one of the eleven evaluators felt were funny, in Section 3. Because the discriminator picks up data that at least one of the users might consider as funny, the discriminator is versatile; however, it also often picks up word combinations that some users may feel are not funny. As a result, the jokes are felt as not funny by some of the evaluators. This could be the reason for the aforementioned result.

The average scores of the jokes generated by the proposed and already existing methods and their differences are shown in **Fig. 13** for different evaluators. The figure indicates that the average score changes depending on the evaluators, whereas the differences in the scores between the proposed and already existing methods are larger than 0 in most cases. In other words, although there are slight differences among the scores by the evaluators, the proposed method is confirmed to be effective.

Figure 14(a) shows the average scores of the jokes in descending order. The colors indicate the method, whether the proposed or already existing method, used to generate the jokes. **Fig. 14(b)** shows the proportion of the jokes generated by the proposed method among the top *i* jokes in terms of the average score.

One can see from **Fig. 14(a)** that the proposed method, compared to the already existing method, generates a larger number of jokes with high average score. Because each of the proposed and already existing methods generated 100 jokes, the occupancy of the jokes generated by



(a) Average scores of jokes generated by already existing/proposed methods

(b) Proportion of jokes generated by proposed method

Fig. 14. Average scores of jokes in descending order.

the proposed method should converge to 0.5 as the number i increases, and as can be seen from **Fig. 14(b)**, the proportion does not go below 0.5 almost always until the proportion converges to 0.5. Namely, the jokes generated by the proposed method tend to be judged as being funny in comparison to those generated by the already existing method.

For reference, the top five jokes in terms of average score are listed in **Table 8**. We also had the evaluators answer a questionnaire survey with the question, "What is your evaluation criterion for judging funniness?" Free descriptive answers were allowed in the survey. The answers categorized based on the authors' judgments are shown in **Table 9**.

The results in **Table 9** reveal that "good sound of word" and "ease of imagining" were the most critical factors in the evaluation. The "good sound of word" is generated through DP matching, which both the proposed and already existing methods use. On the other hand, the "ease of imagining" is not considered in the already existing method and is generated only by the proposed method. Therefore, the already existing method can generate "good sound of word" to enhance the funniness, whereas the proposed method can generate both "good sound of word" and "ease of imagining" to enhance the funniness. This could be the reason that the proposed method was evaluated as being more effective.

6. Conclusions

In this study, we developed a humor discriminator, initially based on manual evaluation, and determined its effective features using the discrimination result. The discriminator was then completed via discovery of an appropriate combination of the features. We also proposed a method of using the humor discriminator, which was developed for word combinations of the form "noun$_2$ of noun$_1$." Previously, few studies have been conducted on the use of humor discrimination. Whereas a method of making conversations in a dialogue system via discrimination of jokes does exist, this had previously been the only way of using humor discrimination. On the other hand, our proposed method uses a newly developed humor discriminator for the automatic generation of jokes. To create a cheap joke, we combined the humor discriminator with the conventional method of DP matching and confirmed significant improvement in funniness in comparison to previously developed techniques. The proposed method can be applied to two-word combinations of any form. Therefore, improvement in funniness can also be expected for the automatic generation of riddles, performance of comedy talks by multiple robots, or creation of scripts of comic dialogue.

Acknowledgements

This work was supported by a driving support project using agent, Nagoya University.

References:

[1] V. Zue, S. Seneff, J. R. Glass, J. Polifroni, C. Pao, T. J. Hazen, and L. Hetherington, "JUPITER: a telephone-based conversational interface for weather information," IEEE Trans. on Speech and Audio Processing, Vol.8, No.1, pp. 85-96, 2000.

[2] T. Misu and T. Kawahara, "Speech-Based Interactive Information Guidance System using Question-Answering Technique," Proc. of the 2007 Int. Conf. on Acoustics, Speech and Signal Processing, Volume 4, pp. 145-148, 2007.

[3] K. Miyazawa, T. Tokoyo, Y. Masui, N. Matsuo, and H. Kikuchi, "Factors of Interaction in the Spoken Dialogue System with High Desire of Sustainability," The Trans. of the Institute of Electronics, Information and Communication Engineers A, Vol.95, No.1, pp. 27-36, 2012 (in Japanese).

[4] T. Yamaguchi, K. Inoue, K. Yoshino, K. Takanashi, N. G. Ward, and T. Kawahara, "Analysis and Prediction of Morphological Patterns of Backchannels for Attentive Listening Agents," Proc. of the 7th Int. Workshop on Spoken Dialogue Systems, pp. 1-12, 2016.

[5] S. Kobyashi and M. Hagiwara, "Non-task-oriented dialogue system considering user's preference and human relations," Trans. of the Japanese Society for Artificial Intelligence, Vol.31, No.1, Article No.DSF-A_1-10, 2016 (in Japanese).

[6] Y. Yoshida and M. Hagiwara, "An Automatic Manzai-dialogue Creating System," Trans. of Japan Society of Kansei Engineering, Vol.11, No.2, pp. 265-272, 2012 (in Japanese).

[7] R. Mashimo, T. Umetani, T. Kitamura, and A. Nadamoto, "Automatic generation of Japanese traditional funny scenario from web content based on web intelligence," Proc. of the 17th Int. Conf. on Information Integration and Web-based Applications and Services, Article No.21, 2015.

[8] S. Saiki and K. Nishimoto, "Enigma Image Searcher: A System for Retrieving Funny Images based on Multistage Word Association," IPSJ SIG Technical Report: SIG Human-Computer Interaction (HCI), Vol.2016-HCI-167, No.1, 2016 (in Japanese).

Table 8. Top five jokes in terms of average score.

Original words	Joke	Average score
Shokuba no senpai (“Senior of company,” in Japanese)	Shokupan no senpai (“Senior of bread,” in Japanese)	5.85
Densetsu no hanta (“Hunter of legend”)	Dentetsu no hanta (“Hunter of train”)	5.62
Shokuba no senpai (“Senior of company”)	Shokubai no senpai (“Senior of catalyst”)	5.46
Oparu no miryoku (“Charm of opal”)	Omaru no miryoku (“Charm of bedpan”)	5.46
Gakko no shiken (“Exam of school”)	Dakko no shaken (“Exam of proctoptosis”)	5.31

Table 9. Typical results in the questionnaire survey.

Type	Item	Number of respondents
Regarding jokes	Good sound of word (rhyme, rhythm)	7
	Similarity of the number of characters between expressions before and after word replacement	2
Regarding imagination	Ease of imagining	6
	Familiar word	4
	Appropriate gap in meaning between expressions before and after word replacement	2
	Word combination that actually exists	2
	Dirty joke	2

[9] A. Kobayashi, T. Isezaki, T. Mochizuki, T. Nunobiki, and T. Yamada, “Pirot study about Oogiri System, Japanese Comedy Show, with Some Robots,” IPSJ SIG Technical Report: SIG Consumer Devices and Systems (CDS), Vol.2017-CDS-18, No.6, 2017 (in Japanese).

[10] M. Maeda and T. Onisawa, “Generation of Nazokake Words Considering Funniness and Relation Degrees Among Words,” J. of Japan Society of Kansei Engineering, Vol.5, No.3, pp. 17-22, 2005 (in Japanese).

[11] H. Yamane and M. Hagiwara, “Oxymoron generation using an association word corpus and a large-scale *N*-gram corpus,” Soft Computing, Vol.19, No.4, pp. 919-927, 2015.

[12] M. Yatsu and K. Araki, “Comparison of Pun Detection Methods Using Japanese Pun Corpus,” Proc. of the 11th Int. Conf. on Language Resources and Evaluation (LREC2018), 2018.

[13] T. Matsui and M. Hagiwara, “Non-task-oriented Dialogue System with Humor Considering Utterances Polarity,” Trans. of Japan Society of Kansei Engineering, Vol.14, No.1, pp. 9-16, 2015.

[14] Y. Amaya, R. Rzepka, and K. Araki, “Performance Evaluation of Recognition Method of Narrative Humor Using Words Similarity,” Technical Report: JSAI Special Interest Group on Language Sense Processing Engineering (SIG-LSEB), Vol.43, pp. 63-69, 2013 (in Japanese).

[15] D. Yang, A. Lavie, C. Dyer, and E. Hovy, “Humor recognition and humor anchor extraction,” Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing, pp. 2367-2376, 2015.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint, arXiv:1301.3781, 2013.

[17] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Advances in Neural Information Processing Systems 26 (NIPS2013), pp. 3111-3119, 2013.

[18] M. Yatsu, “A Study on Topic Adaptation and Pun Humor Processing in Integrated Dialogue Systems,” Doctoral thesis, Hokkaido University, 2017 (in Japanese).

[19] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” Proc. of the 14th Int. Joint Conf. on Artificial Intelligence (IJCAI), Volume 2, pp. 1137-1145, 1995.

[20] National Institute for Japanese Language and Linguistics (NINJAL), “Bunrui-goi-hyo,” 2004 (in Japanese).

[21] T. Kudo, “Applying Conditional Random Fields to Japanese Morphological Analysis,” Proc. of the 2004 Conf. on Empirical Methods in Natural Language Processing (EMNLP-2004), pp. 230-237, 2004.

[22] T. Kudo. “Mecab: Yet another part-of-speech and morphological analyzer,” 2005, http://taku910.github.io/mecab/ [accessed October 31, 2019]

[23] S. Toshinori. “Neologism dictionary based on the language resources on the Web for Mecab,” 2015, https://github.com/neologd/mecab-ipadic-neologd [accessed October 31, 2019]

[24] N. Amano and K. Kondo, ”On the NTT Psycholinguistic Databases: Lexical Properties of Japanese,” Sanseido, 1999 (in Japanese).

[25] C. Kano, “A Proposed Syllabus for Kanji Teaching,” J. of Japanese Language Teaching (Nihongo Kyoiku Ronshu), International Student Center, University of Tsukuba, Vol.9, pp. 41-50, 1994 (in Japanese).

Name:
Tomohiro Yoshikawa

Affiliation:
Department of Medical Information Science, Suzuka University of Medical Science

Address:
1001-1 Kishioka, Suzuka-city, Mie 510-0293, Japan

Brief Biographical History:

1997 Ph.D., Department of Information Electronics, Nagoya University
1997-1998 Visiting Researcher, University of California at Berkeley
1998-2005 Assistant Professor, Department of Electrical and Electronic Engineering, Mie University
2005-2006 COE Designated Associate Professor, COE Project "Frontiers of Computational Science" at Nagoya University
2006-2020 Associate Professor, Department of Computational Science and Engineering, Nagoya University
2020- Professor, Department of Medical Information Science, Suzuka University of Medical Science

Main Works:

- "Analysis of Pareto Solutions Based on Non-Correspondence in Spread Between Objective Space and Design Variable Space," J. Adv. Comput. Intell. Intell. Inform., Vol.19, No.5, pp. 681-687, 2015.

Membership in Academic Societies:

- Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT)
- The Institute of Electrical and Electronics Engineers (IEEE)
- The Japanese Society for Evolutionary Computation
- The Japanese Society for Artificial Intelligence (JSAI)
- Information Processing Society of Japan (IPSJ)

Name:
Ryosuke Iwakura

Affiliation:
Department of Information and Communication Engineering, Graduate School of Engineering, Nagoya University

Address:
Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8603, Japan

Brief Biographical History:

2017 Bachelor, Department of Electrical Engineering, Electronics, and Information Engineering, Nagoya University
2019 Master, Department of Information and Communication Engineering, Graduate School of Engineering, Nagoya University
2019 SCSK Corporation

Main Works:

- "A Basic Study on Generating Back-channel Humor Phrases for Chat Dialogue Systems," Proc. of the 2018 Joint 10th Int. Conf. on Soft Computing and Intelligent Systems and 19th Int. Symp. on Advanced Intelligent Systems (SCIS&ISIS 2018), pp. 1275-1278, 2018.