Generating a Set of Rules to Determine The Gender of a Speaker of a Japanese Sentence

KANAKO KOMIYA, CHIKARA IGARASHI, KAZUTOMO SHIBAHARA, KOJI FUJIMOTO*,
YASUHIRO TAJIMA and YOSHIYUKI KOTANI
Department of Computer, Information and Communication Sciences
Tokyo University of Agriculture and Technology
*Tensor Consulting Co.Ltd.
2-24-16, Nakacho, Koganei, Tokyo, 184-8588
*Ichigaya KT bldg. 4-7-16 Kudanminami, ChiyodaKu, Tokyo, 102-0074
JAPAN
{komiya, k-shiba, igaras-c}@fairy.ei.tuat.ac.jp, {ytajima, kotani}@cc.tuat.ac.jp
*koji.fujimoto@tensor.co.jp
http://shouchan.ei.tuat.ac.jp/wiki/index.php?english%2Ffrontpage#d4c92ea8

Abstract: - Some work has been reported on the problem of automatically determining the gender of a document's author as a part of researches to extract features of a document's author. Japanese language has expressions called masculine/feminine expression, and they can often indicate the gender of a speaker of a conversational sentence. The computer system needs this mechanism in order to make or understand natural Japanese conversational sentences.

The authors made a system that determines the suitable gender of a speaker of a single conversational sentence and named it gender-determining system (GDS). It generates a set of rules to determine the more suitable gender of a speaker of a sentence automatically, by decision tree learning. The authors employed six linguistic features for each of two morphemes at the end of a sentence and presence or absence of morphemes whose part of speech is a miscellaneous pronoun or a particle for ending as features of decision tree learning. The authors calculated the accuracy of GDS using the cross validation method and it was approximately 69.3% when human could answer the same problem with approximately 71.7%. The authors showed decision tree learning is more suitable than multiple regression analysis or Bayesian estimation in order to classify the gender of the speaker of Japanese sentences and generate a set of rules to determine them, and selected the suitable features as the inputs of GDS. The set of rules GDS generated indicates, for example, women speak more politely than men in Japan.

Key-Words: - Natural language processing, rule generation, decision tree learning, gender-determining, knowledge acquisition, data mining.

1 Introduction

1.1 Related Work

Some work has been reported on not only categorizing based on speaker's voice [1] but also categorizing written texts by author gender [2~5]. More work has been reported on authorship identification [6, 7] and some of them are about Japanese texts [2, 6, 7]. The methods are extensive from heuristic analysis [7] to SVM [6]. The target texts are also rich in variety for example blogs [2], E-mail texts [3], books [4] and so on.

In this study, we focus on the genders of speakers of single Japanese conversational sentences and determine the suitable gender of them. We generated a set of rules to determine the suitable gender of a speaker using decision tree learning depending on the linguistic features. The system simulates the people's recognizing suitable gender of a speaker of a Japanese sentence and the generated rules can be helpful to know the mechanism of it.

1.2 The Masculine/Feminine Expression in Japanese

Japanese language has expressions called masculine / feminine expression, and they can often indicate the gender of a speaker of a conversational sentence. They are not grammatical rules in the languages such as French, German and so on, but conventional usage trends. For example “ore, I” is a pronoun only for men and a particle for ending “wa” is mostly used by women.

The computer system should imitate this mechanism in order to make or understand natural Japanese sentences.

Hence we developed a system that determines the suitable gender and named it gender-determining system (GDS). It generates a set of rules to determine the suitable gender of a speaker of a single sentence automatically, by decision tree learning from example sentences, and gives us the suitable result for a set of inputs, based on the rules that the system generated. It is an artificial intelligence system that can simulate the people’s recognizing suitable gender of the speaker of Japanese sentence. We can examine the rules that are GDS generated explicitly and it can help us examine those of humans.

We describe the set of selecting rules and the GDS in this paper.

2 Gender Determining System (GDS)

2.1 The Features and Their Values for GDS

For preparation to use GDS, 1230 sentences were gathered from 11 novels and morphological analysis for them was conducted using ChaSen [8] (Matsumoto et al (2000)). Then the linguistic features acquiring system (LFAS) that we developed made linguistic features.

We input the result file of morphological analysis into the LFAS. The LFAS outputs two kinds of linguistic features automatically for each sentence. They are 1) Six features for each of two morphemes at the end of a sentence (:a morpheme itself, a pronunciation, a prototype of a morpheme, a part of speech (POS), a conjugation of a morpheme, a form of a morpheme) and 2) Presence or absence of morphemes whose POS is a miscellaneous pronoun or a particle for ending. We employed these features because the first personal pronouns and the ending of sentences including the particles for ending well indicate the gender of a speaker. In this experiment, GDS used 64 features about presence or absence of a morpheme because the data we gathered included 45 pronouns and 19 particles for ending (cf. Table 1).

These features are used for the inputs of GDS. For example in the case the sentence is “dat tara kekkou da yo”, we will use 12 features: six features for da and yo and 64 more features. (cf. Fig. 1 in appendix).

Table 1 The Number of Features

The Features	Number
The linguistic features of the second last morpheme of the sentence	6
The linguistic features of the last morpheme of the sentence	6
Presence or absence of morphemes whose POS is a miscellaneous pronoun	45
Presence or absence of morphemes whose POS is a particle for ending	19
Total	76

2.2 Results for GDS

We prepared the same number of example sentences for each gender. Table 2 shows the frequencies of appearance of the genders of speakers of Japanese sentences.

Table 2 The Frequencies of Appearance of the Genders

The Gender	The Frequencies of Appearance
Male	615
Female	615
Total	1230

2.3 The System GDS

There are two stages to use GDS: generation of decision tree and performance. In the first stage: the generation of the decision tree, the user inputs the contexts: the linguistic features to determine the gender of a speaker of a sentence and the gender itself into GDS.

In the first stage, GDS performs decision tree learning and outputs a set of rules to determine the suitable gender of a speaker of a sentence.

In the second stage: the performance, the user inputs a set of linguistic features to determine the gender of a speaker of a sentence and GDS determines a suitable gender according to the rules which are obtained in the first stage. (cf. Fig. 2).

Finally, Fig. 3 (cf. appendix) shows the outline of all the processes.

GDS generates a set of rules to determine the suitable gender of a speaker of a sentence automatically, by decision tree learning from many example sentences, and gives us the suitable result for a set of features for a Japanese sentence, based on the rules the system generated. Therefore GDS can simulate determining the suitable gender of a speaker of a sentence, and the set of rules, which is generated by GDS, can help us examine those of humans.

In addition, GDS can teach us the usages that Japanese people can hardly decide in detail, based on the rules that we generated.

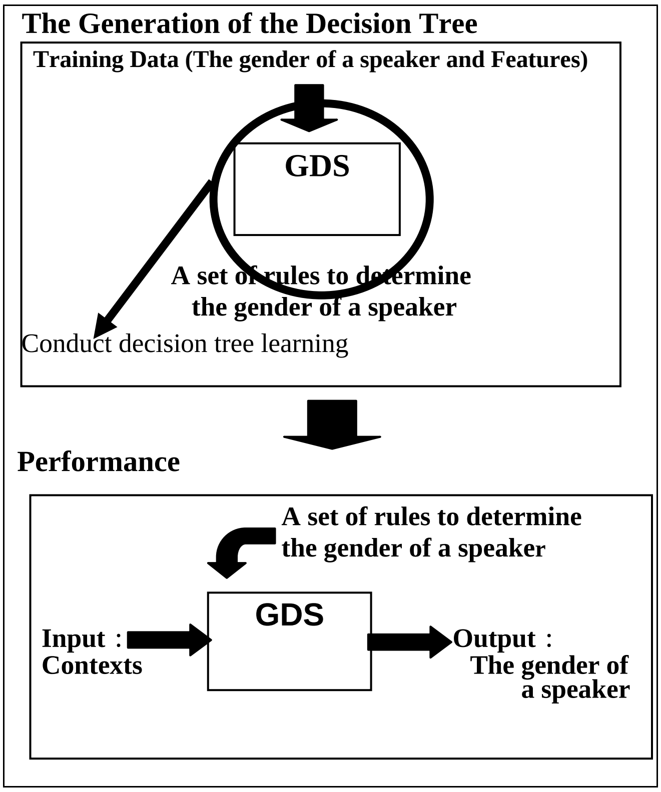
The Generation of the Decision Tree
Training Data (The gender of a speaker and Features)
GDS
A set of rules to determine the gender of a speaker
Conduct decision tree learning
Performance
A set of rules to determine the gender of a speaker
GDS
Input：Contexts
Output：The gender of a speaker

Fig.2 The System GDS (Gender determining system)

3 Decision Tree Learning Based on the Gender of a Speaker

The decision tree is a way to describe some classifications of data, and consists of query nodes. (cf. Fig. 6 in appendix). Each node in the tree classifies the inputs into a few classes according to the feature values of the datasets. The decision tree learning is a machine learning using this property. It generates a tree automatically from leaning data (many examples), and branches lead to leaves that have the suitable result from the root node.

The linguistic features were used in order to get the most type of gender as the output. Yes/no classification (: whether the value of the datum is the same as a certain value or not) was tried. C4.5 (Quinlan (1993) [9]) was used and binary decision tree was generated, which gives us a result for an input that has features, which are not appeared in learning data.

Many papers about classification in WSEAS [10-13] and many are about decision tree [10, 11 and 13].

4 Experiment

1230 sentences were gathered from 11 novels, the sets of features were selected in order to determine suitable gender of a speaker of a sentence using LFAS and the correct meanings were determined manually.

Following termination conditions were used: 1) Whether the information gain is zero or not and 2) threshold values. Two kinds of threshold value were tried: 1) A value of entropy of a node and 2) A value that multiplies a value of entropy of a node and numbers of data in the node together.

5 Evaluation

The accuracies of GDS were calculated many times according to the threshold values, using the five-fold cross validation method, and their highest value was 69.3%. Fig. 4 and fig. 5 show the accuracies of GDS according to the threshold values. The value reached record high when threshold value: a value that multiplies a value of entropy of a node and numbers of data in the node together was 500.

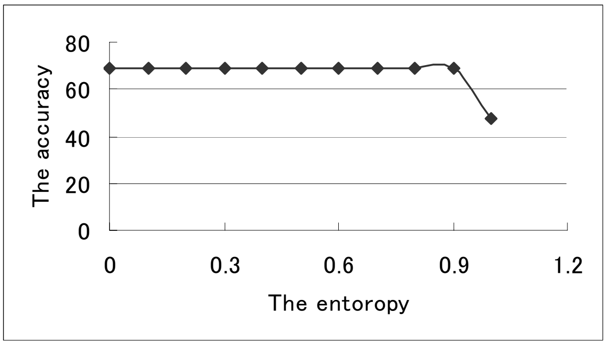

Fig. 4 The accuracies of GDS according to the threshold values; the entropy of a node

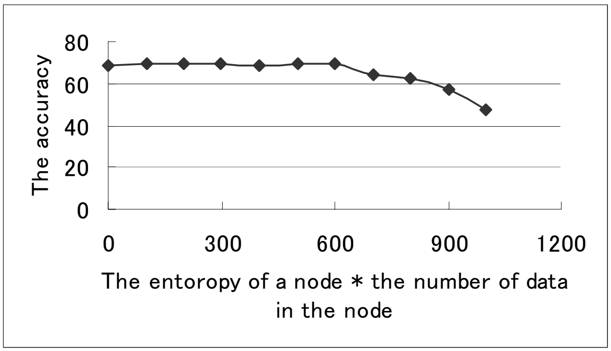

Fig. 5 The accuracies of GDS according to the threshold values; the value the entropy of a node multiplied by the numbers of data in the node

An evaluation test was run in order to evaluate GDS. 20 men and 20 women classified the gender of the speaker of all the questions that GDS answered. (Each man and each woman answered 61 or 62 questions.) The accuracy of men was 71.1% and that of woman was 72.4%. This is because it is difficult to decide the gender of a speaker when they talk formal. Table 3 and table 4 show the breakdown lists of men's answer and women's answers. These tables show that GDS provides less performance than humans, but the difference between GDS and the average of human (71.7%) is only 2.4 points. These tables also show both men and women tend to give their own gender as the answers. Women answered more correctly because they have a less tendency to do so than men.

Table 3 The breakdown list of men's answers

The men's answers	Number	Rate [%]
Correct answer	874	71.06
Answered "male" when correct answer is "female"	202	16.42
Answered "female" when correct answer is "male"	154	12.52
Total	1230	100

Table 4 The breakdown list of women's answers

The women's answers	Number	Rate [%]
Correct answer	890	72.36
Answered "male" when correct answer is "female"	154	12.52
Answered "female" when correct answer is "male"	186	15.12
Total	1230	100

In addition, both men and women answered correctly only 710 problems (57.7 %) and GDS mistook 154 questions in them. It indicates that GDS can still be improved though it is a difficult problem even for humans to determine the gender of a speaker of a single Japanese sentence. For example not only morphemes at the end of the sentence but also morphemes at the end of the clause can indicate the gender.

Moreover, men, women and GDS gave the same answer, which is not correct to 81 questions. These cases are the cases that GDS mistook like humans. For instance, little children's utterance or dialects are difficult to decide the gender of a speaker.

Finally, another experiment was run in order to compare GDS to the other systems. GDS is targeted at a single sentence, and it makes difficult to compare the other systems that are targeted at a document. Therefore we combined sentences that are spoken by the same speaker together into a group, and calculated the accuracy for each group using maximum-likelihood method. The average of the sentences in a group is 7.3 sentences and it was 80.4%. (Minimum is one sentence and maximum is 68.)

6 Comparison with Other Methods

6.1 Multiple Regression Analysis Method

Multiple regression analysis [14] was conducted in order to compare with decision tree learning, as preliminary experiments.

The following four kinds of methods were tried. 983 sentences were gathered from 11 novels and morphological analysis for them was conducted using ChaSen. Table 5 shows the frequencies of appearance of the genders of preliminary experiments.

Table 5 The Frequencies of Appearance of the Genders of Preliminary Experiments

The Gender	The Frequencies of Appearance
Male	607
Female	376
Total	983

Then two types of linguistic features were made by LFAS. Table 6 shows the number of features of the first type and table 7 shows that of the second type.

Table 6 The Number of Features of Preliminary Experiments, the First Type

The Features	Number
Presence or absence of morphemes whose POS is miscellaneous pronoun	36
Presence or absence of morphemes whose POS is particle for ending	18
Presence or absence of morphemes whose POS is miscellaneous noun affix	18
Presence or absence of morphemes whose POS is verb affix	100
Total	172

Table 7 The Number of Features of Preliminary Experiments, the Second Type and The Number of Features of the Second Preliminary Experiments, the First Type

The Features	Number
The linguistic features of the second last morpheme of the sentence	6
The linguistic features of the last morpheme of the sentence	6
Presence or absence of morphemes whose POS is a miscellaneous pronoun	36
Presence or absence of morphemes whose POS is a particle for ending	18
Total	66

The first type of features consists of four kinds of features: presence or absence of morphemes whose POS is a miscellaneous pronoun, a particle for ending, a miscellaneous noun affix or a verb affix. Multiple regression analysis was conducted using these features as independent variables, and features decreased to 15 features by t test (P <0.05 by t-test). Table 8 shows these 15 features. Decision tree learning was also performed using these 15 features and 172 features before conducted t test.

The second type of features also consists of four kinds of features: the linguistic features of the second last morpheme and the last morpheme of the sentence, and presence or absence of morphemes whose POS is a miscellaneous pronoun or a particle for ending. This type of features is the same type of features as the experiments in the chapter five but the number of them is different because the number of data was also different.

Table 8 The 15 features of Preliminary Experiments, the First Type

The Morpheme	POS
Kimi	Miscellaneous pronoun
Watashi	Miscellaneous pronoun
Boku	Miscellaneous pronoun
Minna	Miscellaneous pronoun
Kochira	Miscellaneous pronoun
Achira	Miscellaneous pronoun
Wai	Miscellaneous pronoun
Kanojo	Miscellaneous pronoun
No	Particle for ending
Wa	Particle for ending
Koto	Miscellaneous noun affix
Iru	Verb affix
Teru	Verb affix
Shimawa	Verb affix
Kudasai	Verb affix

In other words, the following four preliminary experiments were conducted.

1) Multiple regression analysis using 15 features of the first type of linguistic features as independent variables.

2) Decision tree learning using 15 features of the first type of linguistic features (P < 0.05 by t-test).

3) Decision tree learning using 172 features of the first type of linguistic features.

4) Decision tree learning using the second type of linguistic features

The way to generate decision trees and the experimental conditions were the same as these of chapter three and chapter four expect the features and the number of the sentences. Table 9 shows the accuracies of these preliminary experiments.

The accuracies of GDS were also calculated many times according to the threshold values, using the five-fold cross validation method.

Table 9 The Accuracies of Preliminary Experiments, Comparison between Multiple Regression Analysis and Decision Tree Learning

Method	The Number of Features	Accuracy [%]
Multiple regression analysis	15	69.48
Decision tree learning	15	63.88
Decision tree learning	172	62.50
Decision tree learning	66	71.82

Table 9 shows the accuracy of multiple regression analysis is higher than that of decision tree learning, if the number of features is the same: 15.

The accuracy of the multiple regression analysis using 15 features of first type of linguistic features as independent variables was about 69.5%, that of decision tree leaning using the same features was about 63.9%, (The value reached record high when threshold value: a value that multiplies a value of entropy of a node and numbers of data in the node together was 700.) and the difference is about 5.6 points. The accuracy of multiple regression analysis was higher because the t test was conducted for multiple regression analysis. The accuracies suggest t test (P <0.05) was too strict to make features of decision tree learning. I think the difference is because decision tree generates a set of rules using combinations of features but multiple regression analysis uses independent variables.

However table 9 also shows the accuracy of decision tree learning can be higher than that of multiple regression analysis. The accuracy of decision tree learning using 15 features of the first type of linguistic features was about 71.8% and the difference between this case and the last case; the number of features is 15 was about 7.9 points. Also, the difference between this case and the multiple regression analysis is still 2.3 points.

Moreover the accuracy of decision tree learning using 172 linguistic features of the first type of linguistic features is 62.5% and it is the worst. (The value reached record high when threshold value: a value that multiplies a value of entropy of a node and numbers of data in the node together was 600.)

The accuracies among them suggest the 66 linguistic features were more suitable than the 15 linguistic features or the 172 features in order to conduct decision tree learning.

In conclusion, table 9 shows decision tree learning is more suitable than multiple regression analysis in order to determine the gender of the speaker of Japanese sentences.

6.2 Bayesian Estimation

Bayesian estimation [15] was also conducted in order to compare with decision tree learning. In this time, 1230 sentences gathered from 11 novels shown in table 2 were used. The accuracies of Bayesian estimation were also calculated using the five-fold cross validation method. The occurrence probabilities of morphemes are used.

Table 12 shows the accuracies of the experiments. It shows the accuracy of Bayesian estimation and that of decision tree are almost the same; the difference is only 0.49 points though Bayesian estimation is a little higher than decision tree learning.

We employed decision tree learning because it can also generate a set of rules.

Table 12 The Accuracies of Experiments, Comparison between Bayesian Estimation and Decision Tree Learning

Method	Accuracy [%]
Decision tree learning	69.27
Bayesian estimation	69.76

6.3 Feature Selection

Additional experiments have done in order to select suitable features to generate a decision tree. Once again, 983 sentences were gathered from 11 novels and morphological analysis for them was conducted using ChaSen. Table 5 shows the frequencies of appearance of the genders of preliminary experiments. The following three kinds of features were tried. All these linguistic features were made by LFAS.

1) The linguistic features of the second last morpheme and the last morpheme of the sentence.

2) The linguistic features in the first trial and presence or absence of morphemes whose POS is a miscellaneous pronoun or a particle for ending.

3) The linguistic features in the second trial and the linguistic features of the second last morpheme and the last morpheme of the first clause of the sentence.

The second type of features is the same as the experiments in the chapter 7.1 and it is shown in table 7. Also table 11 shows the number of features of the first type and table 12 shows that of the third type.

In the third type, all linguistic information of the last and the second last morpheme of the first clause would be "same as the last and the second last morpheme of the sentence" if the sentence consists of only one clause.

The way to generate decision trees and the experimental conditions were the same as these of chapter three and chapter four expect the features and the number of the sentences. The accuracies of GDS were also calculated many times according to the threshold values, using the five-fold cross validation method.

Table 13 shows the accuracies of these preliminary experiments. It shows these three accuracies are almost the same; though the first one is a little lower than the other two. (In the first trial the value reached record high when threshold value: entropy of a node was 0.74, and in the second and the third, the threshold value: a value that multiplies a value of entropy of a node and numbers of data in the node together was 400.) In addition, the new added features in the third features: the linguistic features of the second last morpheme and the last morpheme of the first clause of the sentence have not appeared in the upper part of the decision tree. Therefore it seems these features are not necessary. However, the new added features in the second trial: presence or absence of morphemes whose POS is a miscellaneous pronoun or a particle for ending have appeared in the upper part, for instance, the root node of the decision tree. Therefore we decided to adopt the second type linguistic features to generate a decision tree in GDS.

Table 11 The Number of Features of the Second Preliminary Experiments, the First Type

The Features	Number
The linguistic features of the second last morpheme of the sentence	6
The linguistic features of the last morpheme of the sentence	6
Total	12

Table 12 The Number of Features of the Second Preliminary Experiments, the Third Type

The Features	Number
The linguistic features of the second last morpheme of the sentence	6
The linguistic features of the last morpheme of the sentence	6
Presence or absence of morphemes whose POS is a miscellaneous pronoun	36
Presence or absence of morphemes whose POS is a particle for ending	18
The linguistic features of the second last morpheme of the first clause	6
The linguistic features of the last morpheme of the first clause	6
Total	78

Table 13 The Accuracies of the Second Preliminary Experiments, Feature Selection

The Number of Features	Accuracy [%]
12	71.62
66	71.82
72	71.82

7 A Set of Rules to Determine the Gender

The following questions were selected in the upper part of the decision tree.

(Q1) Whether or not there is the particle for ending "wa" in the sentence.
(Q2) Whether or not "wa" is the last morpheme of the sentence.
(Q3) Whether or not there is the first person pronoun "wai , I" in the sentence.
(Q4) Whether or not "de, please don't" is the last morpheme prototype of the sentence.
(Q5) Whether or not "gozai, be" is the second last morpheme of the sentence.
(Q6) Whether or not "no" is the last morpheme of the sentence.
(Q7) Whether or not there is the particle for ending "cho-dai, please" in the sentence.

Fig. 6 shows the upper part of the best decision tree derived from experiments. (cf. appendix) The question in the root node is whether or not there is the particle for ending "wa" in the sentence. Fig 6 shows, if it is true, the sentence is an utterance of a woman. This is because the particle for ending "wa" is mostly used by women. The second question: whether or not "wa" is the last morpheme of the sentence is selected because morphological analysis sometimes fails. (Q6) is Whether or not "no" is the last morpheme of the sentence. This morpheme is also a particle for ending that mostly used by women. The morphemes "de, please don't" and "cho-dai, please" are particles for ending used for asking politely and "gozai, be" is that for describing politely. These rules shows women talk more politely than men do in Japan.

Meanwhile, "wai, I" in (Q3) is a first person pronoun in dialect. It is not a common feature to determine the gender of a speaker, but GDS selected the feature because women mostly use the morpheme in the training data. GDS generates a set of rules depending on the characteristics of training data because GDS generates them from training data. In one hand the training data should be selected carefully. On the other hand GDS is useful to know the features for texts of specialized area because of this property.

Finally, the features are not appeared in Fig.6 which indicate that the gender of a speaker is a man are, for instance, first person pronouns such as "ore, I" and "boku, I" and morphemes "da, is" and imperative form. These pronouns are always used by men and "da, is" is used for answering without hesitation. It shows men talk with less hesitation than women do.

8 Conclusion

We developed a system that determines the suitable gender of a speaker of a single Japanese sentence in order to examine why Japanese people know the gender of a speaker from written sentences and named it gender-determining system (GDS). GDS generates a set of rules to determine the suitable meanings automatically, by decision tree learning. The inputs of GDS were selected automatically using linguistic feature acquiring system (LFAS). We determined the correct gender manually. The accuracy of GDS was 69.3% when human could answer the same problem with approximately 71.7%. The accuracy is 80.4% when we combined sentences that are spoken by the same speaker together into a group, and calculated the accuracy for each groups using maximum-likelihood method [16]. We showed decision tree learning is more suitable than multiple regression analysis or Bayesian estimation in order to classify the gender of the speaker of Japanese sentences and generate a set of rules to determine them, and selected the suitable features as the inputs of GDS. The rules GDS generated indicate, for example, women talk more politely then men do and men talk with less hesitation than women do.

References:

[1] Dat Tran and Dharmendra Sharma, Automatic Gender Recognition, *WSEAS Transactions on Computers*, ISSN 1109-2750, issue 1, vol. 3, 2004, pp.162-166.

[2] Tsutomu Ohkura, Nobuyuki Shimizu and Hiroshi Nakagawa. Scalable and general method to estimate blogger profile. *IPSJ SIG Technical Report, 2007-NL-181*, 2007, pp.1-5.

[3]Malcolm Corney, Olivier de Vel, Alison Anderson and Georoge mohay. Gender-preferential text mining of e-mail discourse. *In 18th Annual Computer Security Applications Conference*, Las Vegas, 2002.

[4]Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and Liguistic Computing, Vol. 17, No. 4,* 2003, pp. 401-412.

[5] Daisuke Ikeda, Tomoyuki Nannno and Manabu Okumura, Blog Chosya no seibetsu suitei, *NLP2006-C2-3*, http://www.lr.pi.titech.ac.jp/~ikeda/NLP2006-C2-3.pd f (2006).

[6] Yuta Tsuboi, and Yuji Matsumonto. Authorship Identification for Heterogeneous Documents. *IPSJ SIG Technical Report, 2001-NL-148*, 2002, pp. 17-24.

[7]Kazue Kaneko, Tsuyoshi Yagisawa and Minoru Fujita. A sentence generator which reflects the teller's gender and generation, *IPSJ SIG Technical Report, 1996-NL-116,* 1996, pp.129-136.

[8]Yuji Matsumoto, Akira Kitauchi Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda Kazuma Takaoka and Masayuki Asahara, Japanese Morphological Analysis System ChaSen version 2.2.1 , http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.2.1.pdf, (2000).

[9] J. R. Quinlan, C4.5: *Programs for machine learning*, Morgan Kaufmann Series in Machine Learning, 1993.

[10]M. Ikonomakis, S. Kotsiantis and V. Tampakas. Text Classification: A Recent Overview.

Proceedings of the 9th WSEAS International Conference on Computers, 2005, pp. 1-6.

[11]Cheng-Jung Tsai, Chien-I Lee, Chiu-Ting Chen, and Wei-Pang Yang, A Mmultivariate Decision Tree Algorithm to Mine Imbalanced, *Data. WSEAS Transactions on Information Science and applications*. v4 i1., 2007, pp. 50-58.

[12]P. Povalej and P. Kokol. End User Friendly Data Mining with Decision Trees: a Reality or a Wish?, *Proceedings of the 2007 annual Conference on International Conference on Computer Engineering and Applications*, 2007, pp. 35-40.

[13] Tien-Chin Wang and Hsien-Da Lee, Constructing a Fuzzy Decision Tree by Integrating Fuzzy Sets and Entropy, *WSEAS Transactions on Information Science and Applications* , vol.3, no.8, 2006, pp.1547-1552.

[14]Richard A. Berk, *Regression Analysis: A Constructive Critique*, Sage Publications, 2004.

[15] Berger, J.O, *Statistical Decision Theory and Bayesian Analysis,* Springer Verlag, New York, Second Edition, 1985.

[16] A.W. van der Vaart, *Asymptotic Statistics* , Cambridge Series in Statistical and Probabilistic Mathematics, 1998.

APPENDIX

Morphemes, Pronunciation, Prototypes, Parts of Speech, Conjugations, Forms
The second last morpheme of the sentence: da, da, da, auxiliary verb, special/da, end-form or adnominal form
The last morpheme of the sentence: yo, yo, yo, postposition for ending, no information, no information

Fig. 1 The 24 features of «dat tara kekkou da yo»

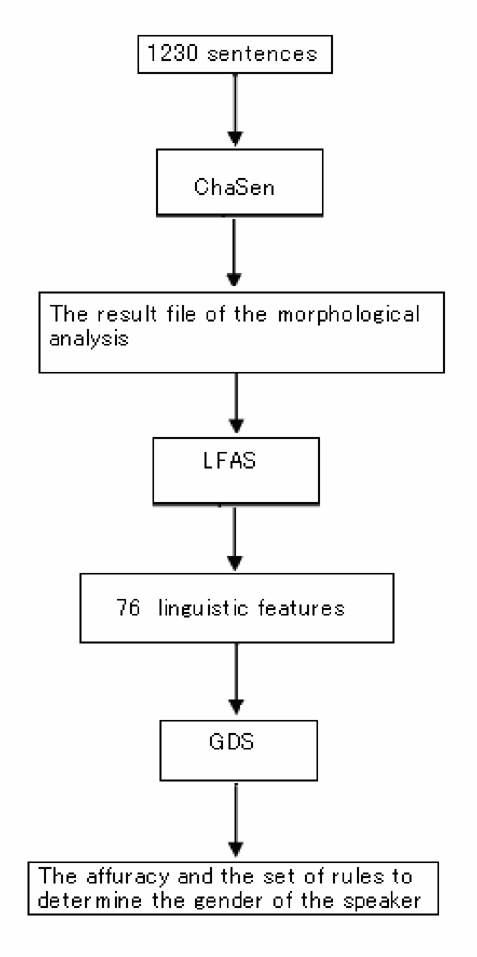
1230 sentences
ChaSen
The result file of the morphological analysis
LFAS
76 linguistic features
GDS
The affuracy and the set of rules to determine the gender of the speaker

Fig. 3 The Outline of All the Processes

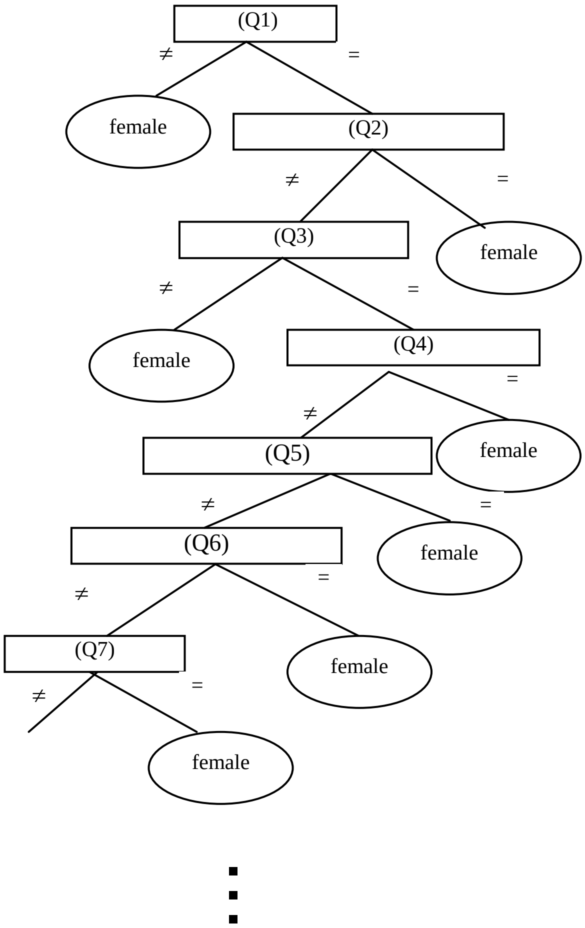

Fig. 6 The Upper Part of the Best Decision Tree Derived from Experiments

Squares mean nodes and questions are in them. These questions consist of a feature and a value, for example "Whether or not the feature is the value". Circles mean the leaves and selected genders are in them.