Gender Prediction for Japanese Authors

Cybelle Smith
David Edwards
CS224N Winter 2011

Abstract

We compare the performance of several automatic classification systems across a collection of different feature sets in detecting the gender of Japanese authors. The Japanese language is notable for the distinctiveness of the different manners and modes of speech used by the two genders--speakers of one gender often use different verb forms, sentence-final particles, and even personal pronouns than speakers of the other. However, this "gendered speech" predominates in informal conversation, and Japanese formal and narrative texts typically contain little or no direct indication of the author's gender. We investigate the possibility that gender nonetheless has an influence on the writing style of Japanese authors, and that writing style can be used to predict the author of the text. We explored several subtle features for guessing an author's gender as determined by Naïve Bayesian, SVM, and Logistic Regression classification models.

Dataset

We hand-collected two different corpora of Japanese text from online sources: firstly, a collection of 30 essays written by Japanese middle schoolers on a common topic, and secondly, 485 installments of fantasy novels posted online by 40 different authors. These were selected specifically to provide similarity of content across documents, since our intended focus was on lexical and grammatical features rather than choice of content. Further, this lowered the likelihood of mistaking features associated with genre or age-range as indicators of gender.

We chose fictional texts for the second genre in order to lessen the impact of first-person references--in Japanese, there are a variety of first-person pronouns which speakers may opt to use, and their usage is often divided by gender. For this reason, occurrences of gender-based pronouns tended to dominate our classification of first-person essays. Passages of fiction provided a useful environment in which first-person references were less trustworthy (though perhaps not entirely uninformative, since it is plausible that authors may favor writing from the perspective of a character of like gender).

Methodology

Japanese Tokenization

Our first obstacle was to render our raw Japanese data in a tokenized form that would be accessible to classification systems. Since Japanese uses no overt word separators in text like the space used in many European languages, we required a more hands-on approach.

We chose to tokenize our data using the ChaSen morphological parser developed by

the Matsumoto laboratory at the Nara Institute of Science and Technology, available online at <http://chasen.naist.jp/hiki/ChaSen/>. ChaSen tokenization provided not only word separation but other morphological input as well, including:

Stem (an individual morpheme of a Japanese word)
Lemma (the "dictionary-form" or uninflected rendering of a morpheme)
Part of speech
Pronunciation

Feature Set

Personal pronouns: The most obvious feature for gender classification is the occurrence of gender-specific first-person pronouns such as masculine 僕 (*boku*) and 俺 (*ore*) versus feminine-neutral 私 (*watashi*). However, we also explored less overt features that we hoped would be more stable in noisy environments like fiction.

Word choice: Since we restricted our classifiers to texts of similar topic and content, we chose to treat the occurrence of all words as potentially informative features in order to test whether diction and word-choice within a common subject might prove indicative of gender.

Pronunciation and Word shape: The Japanese language utilizes three distinct writing systems: *hiragana*, a phonetic system used for native Japanese words as well as most inflectional suffixes, *katakana*, a phonetic system used for foreign loanwords and onomatopoeia among other functions, and *kanji*, logograms primarily borrowed from Chinese which typically occur as content words and verb stems. A given Japanese word may often be represented in a number of ways: purely in hiragana, in a combination of hiragana and kanji, or even as katakana. Previous research on Japanese gendered speech indicates that males use more words of Chinese origin, and thus more words written in kanji, than females. We thus hypothesized that documents written by male authors would have a larger total number of kanji on average than those written by female authors. In Japanese, there is also a prevalent cultural association between the phonetic script hiragana and femininity. Thus, we hypothesized that, given a word that can be written either in kanji or with hiragana, females would be more likely than males to write the word in hiragana. We designed features to test both of these hypotheses.

Use of quotations: While parsing ChaSen output, our feature extractor tagged words that occurred between quotation marks. Initially we intended to use this information to filter out quoted speech, which might be taken from outside sources and therefore not be indicative of the author. However, we also found that the number of words inside quotation marks and outside quotation marks in the text is a very poor indicator of the author's gender. Thus, we used this as a baseline feature so that we could compare performance of our chosen features both to chance, and to features which were not indicative of gender.

Part of speech: One defining feature of Japanese gendered speech is differences in usage of sentence-final particles, some of which the ChaSen tokenizer glosses as different parts of speech. Therefore, we theorized that male and female authors may produce sufficiently different counts in particular part-of-speech tallies to be indicative of gender.

Lemma: One other possibility was that male and female authors would use different words on average to express themselves. For our purposes, the "lemma" was simply the dictionary form of the stem as parsed by Chasen. Some lemmas that we felt might be particularly indicative of the author's gender include personal pronouns such as "boku," "ore," and "watashi," politeness markers such as "desu" or the "masu" verb stem, and content words that may indicate different author preferences in the content of the story (e.g. "battle"). For our purposes, we counted each lemma as a separate feature, however, since some lemmas are much more predictive than others, this also increased the risk of overfitting our data.

Classifiers

Naïve Bayesian Classifier: We first adopted a Naive Bayes approach to text classification. Naive Bayes is known as the "bag-of-words" approach, in which the probability of each word occurring in each class in the training data is used to compute the probability of the class of a new document given the words in that document. At the heart of Naive Bayes is the "naive" assumption that the features are conditionally independent given the class. This assumption is clearly false in the case of the words contained in a single document, since as mentioned in lecture, wordsense disambiguation can be aided by the presence of words as far as 10,000 words away in the same text. Nonetheless, Naive Bayes Classifiers can be a simple but effective tool for text classfication, and get surprisingly accurate results. In our case also, Naive Bayes proved surprisingly accurate, and was the simplest way to obtain decent results without worrying about parameter estimation.

$$\hat{P}(X \mid Y) = \prod_{i=1}^{m} \hat{P}(X_i \mid Y)$$

Fig 1. The Naive Bayes Assumption

Logistic Regression: The second approach we took to data analysis was logistic regression. Logistic regression, as with other machine learning algorithms, involves training a vector of weights that scale the relative contributions of the features to maximize the probability of the training data. We adapted a logistic regression program that had been written by one of us previously to accept binary input features, so that it would accept real value input features. Logistic regression is biased to overestimate the importance of features given a small sample size. In our case, we found that it was extremely sensitive to the number of training epochs and in many cases, seemed to converge on around a 50% probability of either gender.

Support Vector Model: Next, we analyzed our dataset using the LIBSVM tool produced by Chih-Chung Chang and Chih-Jen Lin at National Taiwan University. The fundamental motivation for the SVM classifier is to consider documents as points in *n* dimensional space, where *n* is the number of feature values, segregated into two (or conceivably more) classes. A hyperplane is then found which divides the segregated sets with the greatest possible margin. Test data is then plotted onto this space and classified according to the side of the hyperplane on which it falls.

However, our specific application of SVM required some optimization, as certain constants used in training may be problem specific--namely, the constant C for penalizing divisions which do not make entirely clean separations of known-class data points, and γ, a parameter of the kernel function K used to map values into a transformed space, in order to accommodate problems for which the class separation is non-linear in the original space.

To select ideal C and γ choices for our application, we used a grid-search method with cross-validation data from our corpus; an initial sweep over possible parameter values was checked for regions of high accuracy in cross-validation data, which was then checked again using finer increments until suitably well-performing parameters were found. Our intent here was to prevent over-fitting of the training data; by selecting a more lenient error-penalty parameter C, we were able to account for the anticipated fuzziness of our dataset, which (in the absence of stark divisions like *boku/watashi* usage in passages written from the perspective of the author) we did not expect to exhibit a clear gulf between the male and female classes. The C parameter therefore needed to permit the possibility of training set points falling on the incorrect side of the hyperplane in order to prevent the classifier from selecting a division which cleanly but narrowly separated the data.

Additionally, to prevent feature values which inherently accumulated much larger counts across all documents from washing out the effects of relevant but low-frequency features, we scaled all feature counts in the training data to the range [-1,1], and applied an identical transformation to the feature counts tabulated for the testing data.

Data Analysis

Bayes Performance: The Naïve Bayesian model was the first we implemented, and we used it to form some initial hypotheses about our feature selections. With the fantasy-novel corpus, in all cases we observed overall moderately better-than-chance accuracy, with much better behavior for identifying male writers than for female writers. This may suggest greater variability in female-indicative language patterns, or a greater tendency for female authors to write from the perspective of male characters.

The best single-feature behavior we found was that of our SPDWS2 metric (for “same pronunciation, different word shape”) which compiled an aggregate count of an author’s usage of a particular type of word shape when another word shape (such as a kanji or hiragana representation) was available.

Single Feature Tests for Fantasy Novel Corpus

Experiment	1	2	3	4	5	6	7	8
Stem	X							
Lemma		X						
Pronunciation			X					
POS				X				
In Quotes					X			
Word Shape						X		
SPDWS1							X	
SPDWS2								X
Male Accuracy	.83537	.84756	.84756	.90244	.61280	.87195	.77134	.92683
Female Acc.	.38393	.38690	.36012	.22619	.16667	.36607	.36012	.34524
Overall Acc.	.60693	.61446	.60090	.56024	.38705	.61596	.56325	.63253

Combining feature values yielded slightly higher accuracy for identifying female writers while male and overall accuracy remained more or less the same. In some ways this was to be expected, since none of the features individually performed very well overall--since Naïve Bayesian models do not weight relevant features over irrelevant features, the relatively high performance of the SPDWS2 metric was dampened by the influence of the other less accurate features.

By a small margin, our best feature combination for identifying female authors was Lemma+SPDWS2, suggesting that choice of word and wordshape are relatively salient features of female-gendered language in Japanese.

Multiple Feature Experiments for Fantasy Novel Corpus

Experiment	9	10	11	12	13
Stem	X	X		X	X
Lemma			X	X	X

Pronunciation				X	X
POS				X	X
In Quotes				X	
Word Shape				X	X
SPDWS1		X		X	X
SPDWS2	X		X	X	X
Male Accuracy	.84756	.83232	.84756	.85061	.84756
Female Acc.	.38095	.38393	.38988	.34821	.36607
Overall Acc.	.61145	.60542	.61596	.59639	.60392

SVM Performance: For testing with the SVM model, we had an additional degree of freedom in how we tuned the feature counts and model parameters. Surprisingly, while we expected that the cross-validation process would set the *C* and *γ* values in a way that prevented overfitting of the training data, we still found large discrepancies between cross-validation accuracy and test-data accuracy with the our best parameter set from training of *C=32, γ* = .00003. In fact, in each case we had notably better accuracy using scaled feature values with the default parameter settings.

	Accuracy			
Features	No Scaling	Scaling	Cross Validation	Cross Validation Parameters (C = 32, γ = .00003) on Test Set
All Features	50.60%	48.49%	79.7386	50.00%
Part of Speech	50.90%	53.01%	67.9739	47.29%
Wordshape	50.60%	63.25%	75.1634	50.60%
Pronunciation	50.60%	64.46%	77.7778	51.81%
mini	55.00%	65.00%	60	50.00%

Influence of Learning Rate and Number of Epochs on Logistic Regression

Learning Rate	Num Epochs	Female Accuracy	Male Accuracy	Overall Accuracy
1.00E-07	11	0	1	0.506024096
1.00E-06	10	0.408536585	0.988095238	0.701807229
1.00E-06	11	1	0	0.493975904
2.00E-04	5	0	1	0.506024096
2.00E-04	9	0	1	0.506024096
2.00E-04	10	0.585365854	0.886904762	0.737951807
2.00E-04	11	1	0	0.493975904
2.00E-04	15	1	0	0.493975904
2.00E-04	100	0	1	0.506024096
0.001	10	0.585365854	0.886904762	0.737951807

Our logistic regression model fluctuated dramatically between classifying all new data as male or female, except when it was trained for exactly 10 epochs. We are still trying to figure out exactly what went wrong, but overfitting of the training data seems likely, given that we were not using cross-validation.

Conclusion: We experimented with three different types of classifiers in order to classify the gender of authors of Japanese text: Naive Bayes, Logistic Regression, and SVM. We were able to get results substantially above chance performance, particularly with the feature "word shape of words with more than one possible word shape for that pronunciation." In future, we would like to examine exactly how much our data can be generalized across types of fictional novel - for example, would a classifier trained on fantasy novels be able to accurately predict the gender of authors of love stories?

References

Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Li. "A Practical Guide to Support Vector Classication". Department of Computer Science, National Taiwan University.

<http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf> Last updated: April 15, 2010

Chih-Chung Chang and Chih-Jen Li. “LIBSVM: a Library for Support Vector Machines”. National Taiwan University, Taipei, Taiwan. <http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf> Last updated: February 10, 2011

Who did what:

David - most of the write-up and background research on SVM. Did much coding and debugging of all of the models. Researched how to web-crawl to gather corpus data off of the web. Also, created the initial testing platform that we used to test our models.

Cybelle - coded up much of the models and did feature analysis based on knowledge of Japanese gendered speech. Performed experiments using various combinations of the features on each of the models. Found and extracted/annotated the two corpora by hand.