Author Gender Analysis

Chao-Yue Lai

Abstract

Given an English paragraph of sufficient length, I would like to figure out the gender of the author with sufficiently high accuracy. I wrote a Naïve-Bayes classifier with the assistance of NLTK toolkit, and trained it with frequent words as the main features. The addition of frequent bigrams, trigrams and also part-of-speech tags slightly increased its accuracy. There are some obvious indicators, such as relation-related phrases like “my husband” or “my wife”, or topic-related words like “teaspoon” or “hardware”. However, my goal was building a classifier general enough not to use those. Excluding those salient but biased features, my classifier achieved a sub-optimal accuracy of 69%. This suggests that topic-finding is crucial to author gender analysis. Nevertheless, I still found several mild genders, which may shed light on future research.

I. Introduction

Author gender analysis is an interesting topic with useful applications. The classification of gender is meaningful in machine translation, as some languages employ different grammatical structures for depending on the gender of the author, whereas in English no such thing exists. For example, if we want to translate the English sentence “I am alone” to Italian, it can be either *“Sono solo”*(male) or *“Sono sola”* depending on the gender of the author. It can also be used to verify the claimed gender of the author of blogs, or of the chatters in a chatroom. Also, as Professor Barbara Rosario suggested, it is also interesting to check whether a female author simulating a male talker is authentic enough, or there are some feminine expressions infused.

The genders of authors are easily detected in languages with grammatical genders, such as Romance languages. This task is also trivial in languages where male and female authors generally use different sets of vocabulary, such as Japanese and Korean. In English, however, there is no salient feature for distinguishing male from female authors. Thus, this project served to figure out the cues and hints of the genders of authors.

Considering the descriptive power and the speed of training process, I chose the Naïve-Bayes classifier implemented in NTLK(Bird et al., 2009) in this project. After several trials and errors, the features were set to be the frequency of frequent words, bigrams, trigrams and also part-of-speech tags. The resultant best accuracy on the test

set is 69%. Comparing with the 80%-to-95% accuracy of previous works(Argamon et al., 2003: Herring and Paolillo, 2006; Kågström et al., 2009), this seems to show no improvement. However, the mild author gender indicators found in this project provide insight and may facilitate future research on relevant topics.

II. Related Work

Author gender classification is definitely not a new topic in either natural language processing or linguistics. Here I will list a few pieces of work that perform comparatively well and are most related to this project.

The Argamon et al. (2003) work describes how to use general cues in the texts, such as frequencies of pronouns, to classify author gender of formal texts. They achieved an accuracy of 80%, which is substantially high considering they are only using general features. However, they limited themselves to the analyses of formal texts, such as journal articles and fictions. The explicative power of their results to a broader range of genres is unknown.

Herring and Paolillo (2006) did analysis on the gender analysis of blog authors. They had a rather high accuracy of 95%, but they used some topic-related features in their work, such as diary or non-diary writing style. The topic-related features are exactly what this project tries to avoid.

There is a web-based application of text gender recognition (Kågström et al., 2009). My own experiments on it showed a high accuracy of 95%. Unfortunately, the features and the models they use are unclear to me. Nevertheless, it made mistakes on texts about programming written by female authors. This suggest that it might also utilize some topic-related features.

III. Classifier Selection

There are two criteria for picking a classifier. One is its descriptive power, since the goal of this project is to find legible cues on author genders. The other is its training speed, since this is a term project, which should be finished by the end of semester. I had three types of classifiers in mind: Naïve-Bayes, maximum entropy and support vector machine. The most informative features of the two former classifiers and the support vectors of the latter can serve as indicators of author genders, so all of them passed the first criterion. After experiments, Naïve-Bayes is much faster than the other two classifiers, so I ended up using the Naïve-Bayes classifier implemented in NTLK(Bird et al., 2009).

I tried using LIBSVM(Chang and Lin, 2009) since the format of features in this project, which falls in ranges of numbers instead of Boolean values, is more suitable for SVM training. However, linear kernels of SVM provided the same accuracy as

Naïve-Bayes, while other types of kernels were too slow to train. As a result, Naïve-Bayes still remained as the only classifier of this project.

IV. Data

In order to account for both formal and casual writing styles, I have collected two kinds of literature: books and blogs. They were labeled with the names and genders of the authors, genres and also sources. I had to label all of the data with author genders myself. There were no pre-labeled data in my case. Labeling was rather easily done for books, since the information of the authors could be found via Internet.

However, it became much trickier in the cases of blogs. First I guessed the gender with the authors' names (where name gender detection would come in handy). I also threw a lot of blogs away where I could not identify the genders of the authors at this stage. Then I check the authors' "claimed" genders on their profiles. As a result, the correctness of my labeling depends on the credibility of bloggers.

Books were collected from Project Gutenberg (Hart et al., 2009) and I only adopted English books written in the recent century. Most of them were novels, but there were also letters, poems and biographies. I have collected 47 books with a total of about three million words. Half of them are written by male authors and half written by female authors.

I have also collected 48 blogs by perusing through blogspot.com (Blogger Developers Network, 2009) and wordpress.com (Automattic Inc., 2009). The collected articles from blogs contained about 1.22 million words and the genders of bloggers were evenly distributed. The topics ranged from parenting and education to fashion, politics and even linguistics. In total, about 4.22 million words were gathered. The wide spectrum of topics and the even distribution of author genders made this corpus well-suited for this project.

V. Features

The feature model utilized in this project is the frequency of appearances of target features. In order to account for percentage information, I chopped the data into 100-word chunks. A feature can be a word like "my", a bigram like "of course", or even words with 7 letters. For a target feature *f*, I counted the number of *f* in each 100-word chunk, which is an integer *n* from 0 to an empirical maximum of 20. To make the feature "#*f* = *n*" compatible with the Boolean nature of features in the Naïve-Bayes model, I added "#*f* ≥ *i*" = True for all integers $i \leq n$. Figure 1 provides an example where the target feature is the word "my".

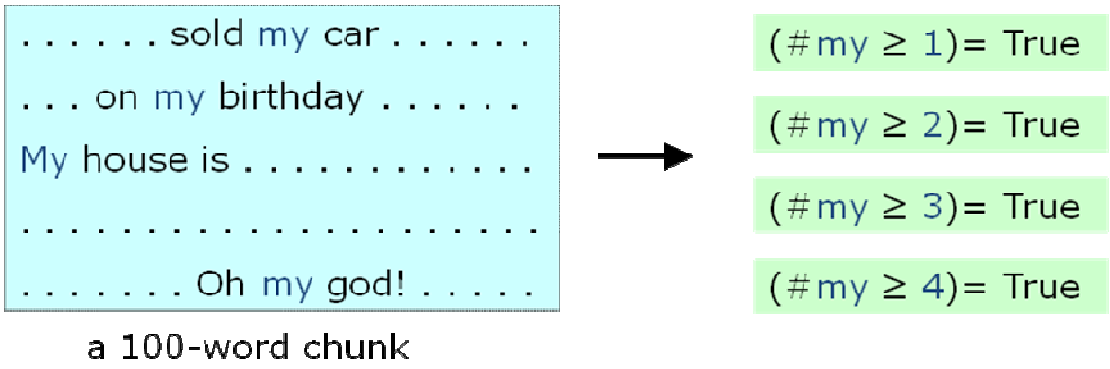

Figure 1: an example of features generated with the target feature as the word "my".

Initially, I treated all the words in the corpus as features to be counted. This scheme yielded an excellent accuracy of 91%. However, after checking the most informative features, the classifier benefited mostly from names, words describing relationships (like "husband" and "boyfriend") and topic-related words (like "teaspoon" and "Mercedes"). Therefore, I narrowed down to frequent words, which is defined as words appearing more than 500 times in all the chunks (approximately 37,000 of them in the training data). I also discovered that accounting for the long-tail distributions of word frequencies not only slowed down the training process, but also over-fitted the training data. Hence, I set a threshold of 5 for the feature counts.

Using only the frequent words gave an accuracy of only 64%, which is much less than 80% claimed in Argamon et al.(2003). I then added frequent bigrams and frequent trigrams, where "frequent" is defined previously. The inclusion of bigrams and trigrams introduced an increase of 4% in accuracy. Finally, counting the frequency of part-of-speech tags brought another 1% of accuracy boost, but it took 3 times longer to train, which was definitely not worth the effort.

Other than words, bigrams, trigrams and POS tags, I tried various kinds of features, but none of them worked. I tried using sentence lengths as features, and it showed that female writers tend to use longer sentences than males. However, introducing these features degraded the classifier instead. Word lengths were also considered and exploited as features, bearing in mind that men prefer shorter words, but the supplement of these features did not help, either. In a nutshell, the features that work best are the combination of words, bigrams, trigrams and POS tags.

VI. Results and Implications

Following the standard procedure in classification tasks, I shuffled the approximately 42,200 100-word chunks and randomly picked 10% of them as the test set, another 10% of them as the development set and the rest were the training set. The following figures are reported by experiments on the test set.

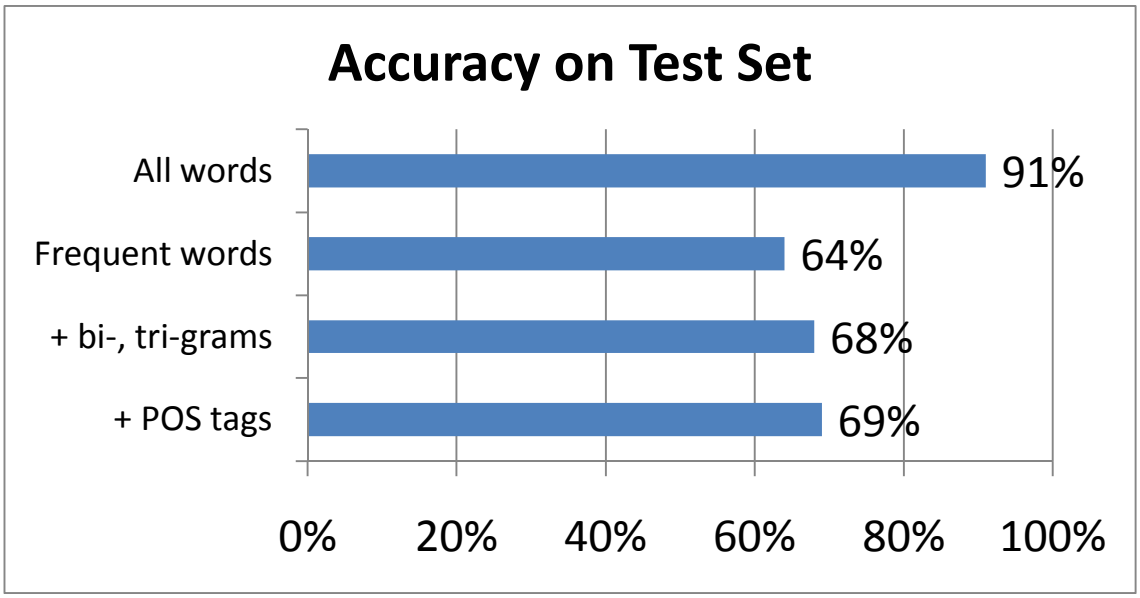

Figure 2: Accuracy with different sets of features.

As shown in Figure 2 and also reported in the last section, using all words performed best, but with a huge portion of over-fitting. Using only the frequent words largely decreased the accuracy in exchange for generality of features. The inclusion of bigrams, trigrams and POS tags pulled the accuracy up to 69%, but it took much longer time to train.

Although the performance was not ideal in terms of accuracy, I still found some mild indicators of author gender by most informative features in the Naïve-Bayes classifier and also extensive data analyses. The frequency of several pronouns, punctuation marks, abbreviations and common verbs hints on the gender of authors. For pronouns, it is not surprising that “he”, “his” and “him” are male indicators, while “she” and “her” suggest female. Nevertheless, other interesting trends are shown: female authors tend to use more of first-person ones, such as “I”, “my”, “our” and “we”. Figure 3 shows the count of “my” in 100-word chunks, and female authors obviously use more than males. Male authors, on the contrary, use more “you” and “it”. This result actually complies with Argamon et al.(2003). According to their analysis, female authors tend to talk more about things related to themselves, while male authors try to address their audience by “you”, and prefer the inanimate pronoun “it”.

For punctuation marks, male authors use more semicolons, which may show men’s analytical nature, while females prefer the expressive exclamation marks. Abbreviations, such as “’s”, “n’t” and “’ll”, are mostly used by men, which coincides with the tendency that men use shorted words. Finally, for common verbs, males like to express uncertainties and assumptions with “could”, “would” and “think”. On the contrary, female authors love to show their affection via the word “love”.

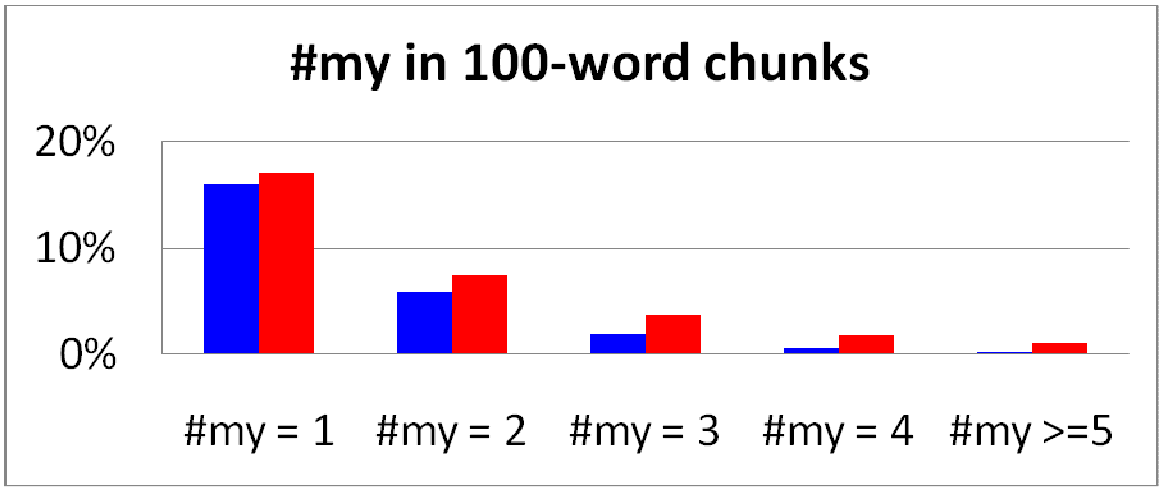

Figure 3: red and blue are for female and male authors, respectively. Female authors collectively use more "my" in their paragraphs.

VII. Conclusions

Author gender analysis is hard by only using general features, as shown by the best-possible 69% accuracy in this project. More data or a suitable model is needed for better performance. This also suggests that a topic-based model is inevitable for accurate author gender classification.

Despite the non-ideal performance, several mild indicators of author gender were still found in this project. These discoveries could be useful information for research on similar or related topics in the future.

References

Argamon, S., Koppel, M., Fine, J. and Shimoni, A. R. 2003. Gender, genre, and writing style in formal written texts. *Text*, 23(3): 321-346.

Automattic Inc. 2009. WordPress.com — Get a Free Blog Here. http://wordpress.com

Bird, S., Klein, E., Loper, E. et al. 2009. NLTK: Natural Language Processing Toolkit. http://www.nltk.org/

Blogger Developers Network. 2009. Blogger: Create your free blog. http://www.blogspot.com

Chang, C.-C. and Lin, C.-J. 2009. LIBSVM – A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Hart, M., et al. 2009. Project Gutenberg. http://www.gutenberg.org

Herring, S. C. and Paolillo, J. C. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4): 439-459.

Kågström, J., Kågström, E. and Karlsson, R. 2009. uClassify Gender Analyzer_v5. http://www.uclassify.com/browse/uClassify/GenderAnalyzer_v5