Gender in Shakespeare:
Automatic Stylistic Analysis of Shakespeare's Characters

Sobhan Raj Hota
IIT Laboratory of Linguistic Cognition
CS Department, Illinois Institute of Technology
Chicago, IL, 60616
hotasob@iit.edu

Shlomo Argamon
IIT Laboratory of Linguistic Cognition
CS Department, Illinois Institute of Technology
Chicago, IL, 60616
argamon@iit.edu

Abstract

The problem of automatically determining the gender of the author of a literary document (i.e. fiction/non fiction), an electronic document (i.e. text collected from e-mail), an informal text (i.e. essays) etc. has been studied and have achieved reasonable accuracy in the range of 70-80%. These works have proven that noticeable differences exist, which discriminates author gender using simple lexical and syntactic features. Our aim here is, automatically determining gender of literary characters (i.e. from plays) from the playwright's word use. The earlier mentioned works and this work is different because real men and real women are involved as the authors of various forms of documents where as in plays, literary character's gender is characterized by playwright's word usage. Thus we ask questions: *Can the gender of Shakespeare's characters be determined from his word usage? If we are able to find such word use, can we glean any insight into how Shakespeare portrays gender? Are the differences (if any) between male and female language in Shakespeare's characters similar to those found in modern texts by male and female authors? Is there any differences exist in Early and Late Shakespeare on Gender? Is this research going to give a new direction to those who are understanding Shakespeare and plays?* In this work, we have proved Shakespeare used language differently for his male and female characters, and also we have studied what kind of lexical features are used in males and those are used in female characters. We used Sequential Minimal Optimization (SMO) for classification of gender character by importing their speeches. We used function words (FWs), Bag of Words (BoWs) as feature selectors and achieved 67% and 74% accuracy respectively. We have also captured the top features of male and female characters and observed that style has played the role in discriminating gender of a literary character. We observed that, there is similarity with the previous research on classifying gender author of literary books (Argamon et al.2003). Currently we are focusing on Lemma and Part-of-Speech (POS) features in Shakespeare for getting more insight in the classification problem. We will be presenting those features in the conference.

1 Introduction

A recent development in the study of language and gender is the use of automated text classification methods to examine how men and women might use language differently. Such work on classifying texts by gender has achieved accuracy rates of 70-80% for texts of different types (e-mail, novels, non-fiction articles), indicating that noticeable differences exist (de Vel et al. 2002; Argamon et al. 2003).

More to the point, though, is the fact that the distinguishing language features that emerge from these studies are consistent, both with each other, as well as with other studies on language and gender. De Vel et al. (2002) point out that men prefer 'report talk', which signifies more independence and proactivity, while women tend to prefer 'rapport talk' which means agreeing, understanding and supporting attitudes in situations. Work on more formal texts from the British National Corpus (Argamon et al. 03) similarly shows that the male indicators are mainly noun specifiers (determiners, numbers, adjectives, prepositions, and post-modifiers) indicating an 'informational style', while female indicators are a

variety of features indicating an 'involved' style (explicit negation, first- and second-person pronouns, present tense verbs, and the prepositions "for" and "with").

Our goal is to extend this research for analyzing the relation of language use and gender for literary characters. To the best of our knowledge, there has been little work on understanding how novelists and playwrights portray (if they do) differential language use by literary characters of different genders. To apply automated analysis techniques, we need a clean separation of the speech of different characters in a literary work. In novels, such speech is integrated into the text and difficult to extract automatically. To carry out such research, we prefer source texts which give easy access to such structural information; hence, we focus on analyzing characters in plays. The natural choice for a starting point is the corpus of Shakespeare's plays.

We thus ask the following questions. Can the gender of Shakespeare's characters be determined from their word usage? If we are able to find such word use, can we glean any insight into how Shakespeare portrays maleness and femaleness? Are the differences (if any) between male and female language in Shakespeare's characters similar to those found in modern texts by male and female authors? Can we expect the same kind of analysis in understanding Shakespeare's characters' gender, to the ones we discussed above? Keep in mind that here we examine text written by one individual (Shakespeare) meant to express words of different individuals with differing genders, as opposed to texts actually by individuals of different genders.

To address these questions, we applied text classification methods using machine learning. High classification accuracy, if achieved, will show that Shakespeare used different language for his male and for his female characters. If this is the case, then examination of the most important discriminating features should give some insight into such differences and to relate them to previous work on male/female language. The general approach of our work is to achieve a reasonable accuracy using different lexical features of the characters' speeches as input to machine learning and then to study those features that are most important for discriminating character gender.

2 Corpus Construction

We constructed a corpus of characters' speeches from 34 of Shakespearean plays, starting with the texts from the Moby Shakespeare[11]. The reason behind choosing this edition is that it is readily available on the web and has a convenient hierarchical form of acts and scenes for every play, while we do not expect editorial influence to unduly affect our differential analysis. The files collected from this web resource were converted into text files from hypertext media and then we cleaned the text files by removing stage directions. The gender of each character was entered manually. A text file for each character in each play was constructed by concatenating all of that character's speeches in the play. We only considered characters with 200 or more words. From that collection, all female characters were chosen. Then we took the same number of male characters as female characters from a play, restricted to those not longer than the longest female character from that particular play. In this way, we balanced the corpus for gender, giving a total of 83 female characters and 83 male characters, with equal numbers of males and females from each play. This corpus is termed the 'First Corpus'. We also built a second corpus based on the reviewer's comments, in which we equalized the number of words in male and female characters by taking every female character with more than 200 words and an equal number of the longest male characters from each play. The longest male and female characters were then matched for length by keeping a prefix of the longer part (male or female) of the same length (in words) as the shorter part. This procedure ensured that the numbers of words per play for both genders are exactly the same. This corpus is termed the 'Second Corpus'. We also split each corpus (somewhat arbitrarily) into 'early' and 'late' characters. We used the term early to those plays which were written in 16th century and late to those in 17th century. This chronology in plays as

[1] http://www-tech.mit.edu/Shakespeare/

captured from Wikipedia[22]. The numbers of characters from each play for 'First Corpus' and 'Second Corpus' are shown in Table 1[3].

3 Feature Extraction

We processed the text using the ATMan system, a text processing system in Java that we have developed[4]. The text is tokenized and the system produces a sequence of tokens, each corresponds to a word in the input text file. We use two sets of words as features. A stylistic feature set (FW) is a list of more-or-less content-independent words comprising mainly function words, numbers, prepositions, and some common contractions (e.g., "you'll", "he'll"). A content-based feature set comprises all words that occur more than ten times in a corpus, termed Bag of Words (BoW).

We calculate the frequencies of these FWs and BoWs and turn them into numeric values by computing their relative frequencies, computed as follows. We first count the number of times two different features occurring together; then we divide this number to the count of the feature in reference. In this way we calculate the relative frequency for each feature and a collection forms a feature vector, which represents a document (i.e. a character's speech). The FW set has 645 features including contractions; the BoW set has 2129 features collected from the first corpus and 2002 BoW features collected from the second corpus. The numeric vectors collected for each document is used as an input for machine learning.

4 Text Classification

The classification learning phase of this task is carried out by Weka's (Frank & Witten 1999) implementation of Sequential Minimal Optimization (Platt 1998) (SMO) using a linear kernel and default parameters. The output of SMO is a model linearly weighting the various text features (FW or BoW). Testing was done via 10 fold cross validation. This provides an estimation of generalization accuracy by dividing the corpus into 10 different subsets. The learning is then run ten times, each time using a different subset as a test set and combining the other nine subsets for training. In this way we ensure that each character is tested on at least once with training that does not include it. Tables 3 and 4 present the results obtained by running various experiments. It is clear that BoW has performed better than the FW in both selection criteria, as expected, since it has more features on which to operate. This shows that both style and content differ between male and female characters. As expected, the FWs have proven the stylistic evidence and not the content, which are visible from the Table 4[3]. BoW gives a high 74.09 on over all corpuses with the equalizing on number of words selection strategy. Interestingly, FW gives highest accuracy of 74.28 in Late plays with only 63 training samples. This indicates that there is a greater stylistic difference between the genders in late Shakespeare than in early Shakespeare.

5 Discussion

The feature analysis phase is carried out by taking the results obtained from Weka's implementation of SMO. SMO provides weights to the features corresponding to both class labels. After sorting the features based on their weights, we collected the top twenty features from both character genders. Tables 5-10[3] lists the top 20 features from male and female characters and is shown with their assigned weights given by the SMO, for FWs and BoWs respectively. Tables 11-16[3] lists the same for the Second Corpus. These tables also show the 'Average frequency of 100 words', which finds the frequency of a particular feature divided by total gender characters, and then for easy readability this figure is scaled by 100 times. To discriminate binary class labels, SMO uses positive and negative weight values in Weka's implementation. We see from the Tables 5-10[3], male features are designated as negative weights and female characters are given as positive weights. In top 20

[2] http://en.wikipedia.org/wiki/Chronology_of_Shakespeare_plays

[3] http://www.iit.edu/~hotasob/MCLC06-HTML/Hota-MCLC06-Tables.htm

[4] http://lingcog.iit.edu/download.xml

male features, this can be observed that 'Average Frequency of 100 Words' value of male is more than the corresponding value for female. This hold same in the case of the top 20 female features where female 'Average Frequency of 100 Words' value is more than the male for the same feature.

Feature Analysis: BoW

We can see cardinal number usage is found in male characters. Plural and mass nouns ('swords', 'dogs', 'water') are used more in males than females. On the other hand, there is strong evidence for singular noun ('woman', 'mother', 'heart') usage in females. The use of 'prithee' as an interjection is found in female character. This may represent a politeness aspect in their attitude. The past participle form is generally found in females ('gone', 'named', 'known'). Present tense verb forms ('pour', 'praise', 'pray', 'love', 'dispatch', 'despair') are used in female characters. In the case of male characters, Shakespeare used these verb forms ('avoid', 'fight', 'wrought'). Male characters seem to be aggressive while female characters seem to be projected as supporters of relationships.

Feature Analysis: FW

We observed that Shakespeare's female characters used more adverbs and adjectives, as well as auxiliary verbs and pronouns. On the other hand, cardinal numbers, determiners, and some prepositions are generally indicative of male characters. These observations are in line with previous work (Argamon et al. 2003) on discriminating author gender in modern texts, supporting the idea that the playwright projects characters' gender in a manner consistent with authorial gender projection. We did observe some contrasting results in the FW features from the second corpus. Number (i.e. twice) is found in female characters. Certain prepositions are used for females, while negation only appears distinctive for early females. Determiner 'the' which is a strong male character indicator in first corpus is found only in early part of second corpus. Some negation ('cannot') is found in late males as well. Clearly, more and deeper analysis is needed.

6 Conclusion

This is the first work, to our knowledge, in analyzing literary character's gender from plays. It seems clear that male and female language in Shakespeare's characters is similar to that found in modern texts by male and female authors (Argamon et.al 2003), but more work is needed in understanding character gender. We have also observed possible differences between early and late Shakespeare in gender character classification. In particular, the later Shakespeare plays appear to show a greater stylistic discrimination between male and female characters than the earlier plays. We are particularly interested in collaborating with literary scholars on this research to explore these issues further.

References

Koppel M., Argamon S., Shimoni A. (2004). Automatically Categorizing Written Texts by Author Gender : *Literary and Linguistic Computing* 17(4).

Argamon S., **Koppel M.**, **Fine J.**, **Shimoni A**. (2003). Gender, Genre and Writing Style in Formal Written Texts : *Text* 23(3), pp. 321–346

Argamon S., Whitelaw C., Chase P., Hota S., Dhawle S., Garg N., Levitan S. (2005) *Stylistic Text Classification using Functional Lexical Features, Journal of the Association for Information Sciences and Technology*, to appear.

Corney M., Vel O., Anderson A., Mohay G. (2002). *Gender Preferential Text Mining of E-mail Discourse* : In *Proceedings of 18th Annual Computer Security Applications Conference ACSAC*

Corney M., Vel O., Anderson A. (2001). Mining E-mail Content for Author Identification Forensics *: ACM SIGMOD Record* Volume 30, Issue 4 December

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with many relevant features. *ECML-98, Tenth European Conference on Machine Learning.*

Platt, J. (1998). *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector*

Machines. Microsoft Research Technical Report MSR-TR-98-14,

Mitchell, T. (1997) *Machine Learning*. (McGraw-Hill)

Witten I., Frank E. (1999). *Weka3: Data Mining Software in Java* http://www.cs.waikato.ac.nz/ml/weka/