Gender in Movie Dialogue

Anne Huang, Natawut Monaikul
Computer Science Department, University of Illinois at Chicago
ahuang27@uic.edu, natawutmo@gmail.com

Abstract

Existing work suggests that women make linguistic choices to signal friendliness and collaboration. We examine collections of lines from movies in the action, adventure, drama and romance genres. We run Latent Dirichlect Allocation (LDA) and Biterm Topic Modeling (BTM) on movie scripts from each genre to determine if there are differences in the topics men and women discuss. Using the Stanford Politeness API, we also run a politeness detector on requests in these scripts to see if there are gender differences in politeness in any of the four genres. We find no significant consistencies in cohesive topics. We find a statistical significance in politeness for requests from only the adventure genre, in which the female requests are scored as more polite.

1 Introduction

In popular movies, women are often underrepresented and underwritten. Only 32% of all speaking characters in the top 100 movies of 2016 were women according to an annual study by the Center for the Study of Women in Television & Film (Lauzen, 2017).

There is interest in natural language processing into differences in male and female utterances because it is believed that these utterances encode information about underlying social norms and power dynamics. In this project, we select an available corpus of movie scripts in which the gender of the speaker has been labeled (if known). We do our research in 2 parts to explore differences in utterances by male and female characters. In part 1, we run our corpus under topic models. In part 2, we run a politeness detector on random samplings from this corpus.

Researchers into topic models have written about the potential for topic models to scale up the volume of textual data that can be explored to answer traditional social science questions. A topic model is an unsupervised algorithm that takes a collection of documents as input and outputs a list of groupings of words from that collection, each of which the model considers to have a high probability of belonging to one topic (Blei, 2012).

We use the probabilistic topic models Latent Dirichlet Allocation and Biterm Topic Modeling. We run these topic models on documents of female lines and documents of male lines from collections of movies. Because genre is information available about movies, we think of movies from the same genre as data where we expect to find consistent distributions of topics, and run each instance of a topic model on a collection all from the same genre, which we compare to other collections.

Our secondary step is to use the Stanford Politeness API to empirically test social observations about the differences in gender. This API includes a request classifier. We first run the data on the request classifier to detect requests, and then score these requests on politeness.

What differences do these topic models and the politeness detector demonstrate in male and female dialogue in movies? We test known approaches on a different type of corpus than they have previously been applied to, to see if the results support hypotheses about the discursive differences in male and female dialog.

2 Related Work

2.1 Politeness and Hedges

Previous work has evaluated the politeness of requests on Wikipedia edits and StackExchange questions. A **request** is an utterance that asks for something that potentially burdens the addressee. That work shows that the less powerful someone is, based on their relative status along tangible metrics on online communities, the more polite they are. In that research, utterances are scored for politeness based on the presence of a variety of indicators, such as expressions of gratitude ("appreciate") and hedges (Danescu-Niculescu-Mizil and Lee, 2013).

The data used for this research is not annotated for gender, and thus it provides no inferences for whether this politeness, as they've defined it, varies by demographic characteristics such as gender. A **hedge** is a word that softens an utterance, such as "would," "could," "may."

However, other work has observed differences in the frequency of hedges by men and women in academic writing (Yeganeha et al, 2015). Following the intuitions from this and other work linking women's smaller relative authority to linguistic choices, which we discuss below, we investigate whether these gender differences can be generalized to politeness as a whole, of which hedges are a part.

2.2 Gender in Natural Conversation

One area of social science that natural language processing has been applied to is the issue of the uneven distribution of power between genders. Natural language processing has looked at gender dynamics in both naturally occurring and fictional texts. Work on naturally occurring conversational texts suggests that men speak to establish their status within a hierarchy and are more likely to use language to assert their authority; while women speak to maintain rapport, and thus make linguistic choices to signal friendliness and collaboration, and come across as more tentative, such as by using hedges and interrogatives to minimize embarrassment for the listener. This work suggests that women tend to talk about topics to do with self-disclosure that lend their way more easily to hedges. They also show that women differentiate between downplaying their authority at work but asserting it to their kids (Prabhakaran and Rambow, 2017).

This subject is salient because there have been a lot of arguments that women are not asserting themselves confidently at work, which seems to frame their linguistic styles as deficient. However, the natural language processing work described above suggests that women make these linguistic choices for an affirmative social strategy -- to promote collaboration, which should not necessarily be discouraged.

2.3 Gender in Fictional Conversation

In addition to work on naturally occurring conversation, there is precedent for natural language research on fictional texts. Work on coordination between genders in movie dialogue asserts that movie dialogue can encode what screenwriters see as the power dynamic between different demographic groups, and their communicative styles. That is, even though the utterances are fictionally produced by screenwriters, they reflect the screenwriters' mental models of how men and women actually talk (Danescu-Niculescu-Mizil and Lee, 2011).

2.4 Topic Models

Probabilistic topic models offer the potential to explore themes across large text corpora. They have previously been applied to psychology and political science, and have the potential to help answer questions in other fields centered around textual data, such as comparative literature and sociology (Blei, 2012). Given the above work, we also apply two generative probabilistic topic models to conversational texts in movies to see whether they support similar hypotheses to the social science research discussed above.

The first is Latent Dirichlet Allocation. In this model, a **topic** is a multinomial distribution over the vocabulary in a collection (Blei, 2012; Chang et al, 2009). LDA has been used on news and academic articles (Blei, 2003).

Topic models are based on word co-occurrence. However, for short documents, the data may be too sparse for LDA to produce coherent topics. The solution that Biterm Topic Modeling provides is to model biterms for a corpus. A **biterm** $\boldsymbol{b}$ is a word pair (w_i, w_j) that co-occurs. BTM uses the model to determine the topics for each document. BTM has been shown

to be effective for short documents, such as Twitter tweets (Yan et al, 2013).

2.4 Topic Model Evaluations

There are several ways that topic models have traditionally been evaluated, but no definitive one. One method is through a held-out set. Various topic models are run on the rest of the data. The model that produces the set of topics that most fits the held-out set is chosen (Blei, 2012).

Two other metrics that have been developed to evaluate topic models are **word intrusion** and **topic intrusion**.

Word intrusion has to do with the extent to which the words the model identifies as belonging to a coherent topic, matches what humans judge to be a coherent topic. This is tested by putting words with a high probability of belonging to the same topic together, inserting an intruder word with a low probability of being from that topic, and seeing if test subjects identify the same word as being out of place (Chang et al, 2009).

Similarly, topic intrusion has to do with the extent to which the topics that the model labels a document with, coincide with what humans would identify as the topics of that document. This is also tested by placing the topics with a high probability of being associated with a document together, inserting an intruder topic with a low probability of being relevant to this document, and seeing if test subjects identify it as being out of place (Chang et al, 2009).

3 Corpus / Data

Our raw input is from an available corpus, the Cornell Movie-Dialogs Corpus (https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html). This corpus is made available by Cornell researchers for research on gender and movie dialogue.

The compilers of the corpus have inferred and labeled the gender of the utterers by their names. They provide lines of dialogue from movies, labeled by the character, in a standard format; as well as character metadata, which includes gender, and movie metadata, which includes the genre of the movie (Danescu-Niculescu-Mizil and Lee, 2011). This makes it feasible for us to parse out the lines by character, associate those characters with their gender, and create separate documents of male and female dialogue. The genres we have selected from this corpus are action, adventure, drama, and romance, as we expect to see biased gender roles from these genres. All of the films we use are English-language.

In selecting this data, we make the simplifying assumption that a genre is conceptually cohesive enough to find reasonably consistent topics within. We also make the simplifying assumption that gender roles are consistent enough across time that we are not placing restrictions on the time periods that these movies cover, though looking at specific time periods is an interesting future direction.

For the topic models, we treat each individual line from a movie as a document. By **line**, we refer to an entire block of text uttered by one character in a script before the next character's turn in a conversation. Each collection of documents is all the lines for one gender for one genre.

In Table 1 we provide the size of the data, in terms of the number of lines for each genre-gender collection that we use.

Table 1

	Number of Lines	
Genre	**Female**	**Male**
Action	11,407	37,552
Adventure	6,644	25,590
Drama	41,038	100,032
Romance	23,849	41,445
Total	82,938	204,619

For the topic models, we treat each document as a vector in the bag of words model, where each entry in the vector corresponds to the count in that document for that word in the vocabulary. Thus, the different features correspond to the counts for different words, and the number of features is the number of words in the vocabulary for a collection of documents.

For the politeness detector, the features are hedges, positive/negative words, etc. For the politeness part of our project, we select a random subset from the above data to parse (500 for each genre/gender pair).

4 Methods

Whereas previous work has looked at differences between the genders in the use of function words through coordination, we also investigate their differences through probabilistic topic models, which focus on their content. In addition, we run the Stanford Politeness API built for the existing work, on the movie corpus instead.

Much of our work consists of transforming the raw text input into a form expected by the off-the-shelf topic model and politeness interfaces. This includes parsing the scripts and metadata based on their format, and separating the corpus by gender. For both of the topic models, this also includes cleaning the text of tokens irrelevant to topic assignment.

4.1 Topic Model Preprocessing

We conduct standard text preprocessing, such as converting all text to lowercase and cleaning it of digits and punctuation. We also experiment with stop word filtration and stemming.

We make use of Python's Scikit-Learn machine learning libraries. We start with the default list of English stop words provided with their repository, which includes function words such as articles and prepositions, and modal verbs like "would" and "will" (Loginov, 2015). On top of this, we add stop words such as additional pronoun contractions, and verbs such as "know," "like," and "see," which we see coming up commonly across these conversational texts without adding meaningful information to topic clusters. We also exclude some words from the default stop words, such as "perhaps" and "may," that may be hedge words and may differentiate the male and female utterances.

We experiment with stemming so that words that have similar roots can be grouped together. For this, we use the Porter Stemmer, available in NLTK, which heuristically removes endings on words that roughly correspond with inflectional endings, such as "-es," "-ed," and "-ing" (van Rijsbergen et al, 1980).

4.2 Latent Dirichlet Allocation

The first topic model we use is Latent Dirichlet Allocation, because it is a widely used and simple topic model. The main parameter for this algorithm is the number of topics *n*, which we predetermine.

The algorithm assumes that the distribution of these *n* topics is random (Blei, 2012). It assumes that each word in each document is generated as follows: at each iteration, we draw a topic and generate a word based on the likelihood it occurs given this topic (Blei et al, 2003). Like other generative models, it uses this assumption and learns what the topics would have been given the probabilities implied by this assumption. The algorithm starts with a random assignment of each word to a topic. The algorithm iterates for a user-specified number of times, reassigning the words to topics based on the probabilities with the assumption that the topic assignments are steady at that point (Chen, 2011).

In using this algorithm, we treat each line as one document. All of the lines of one gender in one genre then form a collection of documents, which we compare against other collections.

To run this algorithm, we use an off-the-shelf Python module called **lda**. Blei (2012) gives examples of the algorithm being fit to 20 and 100 topics, as LDA is effective when the input number of topics is large. We experiment with 20, 100, and 500 topics.

The lda module initializes a model that is fit with a **document-term matrix** for a collection of documents. A document-term matrix is a 2-dimensional array in which each row corresponds to one document and each column corresponds to a word in the vocabulary. Each entry corresponds to the number of times that word appears in the document (Grisel et al, 2017).

For LDA, we store all the text documents for the same gender in one genre in an array, and convert this to a 2-dimensional array of word counts using Scikit-Learn's CountVectorizer (Pedregosa, 2011). To initialize CountVectorizer, we try both using the default parameters (leaving all parameters as none), and the parameters from the LDA example from the Scikit-Learn documentation: a minimum count of 2, a maximum frequency of 95%, and at most 1000 features (Grisel, Buitinck, Yau, 2017).

We use this 2-dimensional NumPy array as the document-term matrix input into LDA. We display the top 8 words in each topic as in the example provided with the module documentation.

4.3 Biterm Topic Modeling

Because movie lines are short, and because the Biterm Topic Model has been shown to be effective for short documents, we also use the Biterm Topic Model to see if we can get more coherent topics.

In BTM generation, a word distribution is determined for each topic, and a topic distribution is determined for the collection of documents. Each biterm *b* is assigned a topic *z* based on the distribution of topics. *b* is assigned 2 words w_i, w_j based on the probability of each word independently occurring given *z*. Thus,

$$P(b) = \sum_z P(z)\, P(w_i \mid z)\, P(w_j \mid z)\ .$$

We use the GitHub repository **BTM**, from the creators of the model. The input is a text file of lines, where each line is a document, consistent with how we use LDA. We again run BTM for 20, 100, and 500 topics.

4.3 Topic Model Evaluation

Existing methods to evaluate topic models treat what humans would identify as a coherent topic as the ground truth, and compares an output topic to what humans would identify as coherent. A topic model groups together a set of words that have a high probability of being from the same topic. Existing methods evaluate these topics by seeing if humans would consider any of these words as out of place.

In our project, we examine the topics output to evaluate whether we would consider any of the words as out of place with each other.

4.4 Politeness

We use the Stanford Politeness API, which is available as a GitHub repository. This includes a SVM classifier trained on requests from StackExchange and Wikipedia. These requests were annotated for politeness (by crowdsourcing), and the classifier uses features such as hedge words, pronouns, "please", use of apology or gratitude, positive/negative words, etc., that can be automatically extracted. The politeness classifier outputs a politeness score (between 0 and 1) for each request input, where a score closer to 1 is more polite (Danescu-Niculescu-Mizil and Lee, 2013).

The classifier is trained only on requests, so the API includes a tool for heuristically determining if a sentence is a request, using features such as the presence of a question mark and the use of keywords such as "if" and "will" in specific positions.

The input for both the politeness classifier and the request classifier requires a dependency parse tree for each sentence (Suhof and Danescu-Niculescu-Mizil, 2017). For this, we run the Stanford CoreNLP toolkit in Python (Manning et al, 2014).

We randomly select 500 lines from each genre for each gender (a total of 4,000 lines) due to time requirements for parsing. We run the request classifier on the parsed subcorpus to select a subset of the lines as requests, giving an average of 181 female lines and 182 male lines per genre. We then compare the means of the politeness scores output across all female requests against all male requests for each genre. Our evaluation of statistical significance for this is a standard t-test.

5 Results and Discussion

5.1 Topic Model Results

Based on our evaluation, most of the words in all the topics output seem out of place with each other, regardless of which of the parameters we describe above we use (with or without minimum and maximum word count frequencies, with and or without a maximum number of features).

An example of a topic that was output from LDA is:

"curtain cuttings sparkie champs pellet stars warmth"

We would consider this an incoherent topic because none of the words grouped together here have any relation to each other. Some topics contain two words which could be considered related to each other, but this does not provide enough evidence to evaluate the topic as cohesive.

For BTM, we also find generally incoherent topics. One of the topics output that almost seems coherent is:

"transactions banks cayman irs transfers dollars"

However, looking at the source input, all of these words are from the same line: "*Transactions* are offshore. *Dollars* or euros. Secure internet *transfers*. We have lists of *Cayman* and Isle of Man *banks* infiltrated by *IRS.*"

Thus, we would not consider this a meaningful topic.

5.2 Topic Model Discussion

Accordingly, the topics generated from running LDA and BTM on this corpus are incoherent. However, this is somewhat expected. Previous work on LDA has shown that some topics that seem to mostly be about one concept also include other words that are incoherent with this concept. This demonstrates one of the disadvantages of using LDA, as it is an unsupervised learning model (Chen and Liu, 2014).

Though the total number of male lines and the total number of female lines, and the total number of male versus female characters are very disproportionate with each other, we do not think this affects our topic model results because we do not find coherent topics for either gender.[1]

One of the limitations of our work is that the methods used here may not be getting applied in the right contexts. Other work that has been more successful in finding differences in the natural language production of men and women has focused on function words, such as use of coordination, rather than the content. These features may be more relevant to the differences between male and female utterances, as they encode power dynamics, which is a large part of what characterizes the social differences between male and female utterances. It is difficult to discern those types of differences through topics.

Other work that has been more successful in using topic models has used large corpora of news and academic articles in which it may be more reasonable to expect cohesiveness in topics. An entire genre of movies may be too broad to look for coherent topics in. Furthermore, an additional variable that we do not control for is the time periods of different movies.

5.3 Politeness Results

In Table 2 we show the average politeness scores across all male and all female requests for each genre.

Table 2: Mean Politeness Scores

	Male	**Female**	***p***
Action	0.310	0.314	0.70
Adventure	**0.296**	**0.320**	**0.03**
Drama	0.308	0.324	0.16
Romance	0.325	0.326	0.99

Here, the only genre with a statistically significant difference in male and female politeness scores is adventure, where the female requests are on average more polite.

The adventure movies include superhero movies, *Star Trek*, and *Back to the Future*. Examples of requests from this genre and their scores from the politeness classifier include:

Male: "You wanna explain that?" (0.175)
Female: "Would you mind if […]?" (0.770)

In this example, the female request is scored as more polite due to the presence of hedges such as "Would" and "mind."

5.4 Politeness Discussion

As for the lack of statistical significance for the other genres, one of the limitations of our work is that we sometimes speak requests differently from how we write them. Speech includes intonation, which can help differentiate between a polite and less polite request, whereas we need to include more information in the actual words

[1] We also try considering all of the lines for one gender in a movie as a document in case the length affects LDA results, but we also find no substantially more coherent topics. So that our definition of document is consistent in comparing LDA and BTM results, we only report results in which each line is considered a document.

in writing to convey that we are making a polite request. An utterance that sounds impolite based on the words alone may sound softer out loud due to the tone used. The classifier we use in this project has been trained and tested on written requests, rather than oral conversation.

6 Conclusions and Future Work

In this paper, we follow intuitions implied from various previous work to test the differences in the natural language utterances of men and women. We try different aspects from previous computational work on language and gender. We specifically look at topic models, which is distinct from the focus of previous work. Previous work looks at politeness without gender, or some other aspect of movie dialogue with gender, and not movies with topics and politeness, as we have tied together.

Our work can be improved upon by trying other topic models that use some supervision. For example, knowledge-based topic models have been used to make topics converge more (Chen and Liu, 2014).

Another limitation is that the data we use is fictional, whereas some of the related work makes claims based on naturally occurring dialogue, which is a different phenomenon. Thus, our data does not capture how real-life women actually talk, but rather, screenwriters' impressions of how they talk. However, we presuppose that the screenwriters' mental models of how women and men talk are ultimately derived on some level from society's expectations.

We recommend that future work take a different approach, such as by trying supervised over unsupervised models; or exploring functional rather than thematic characteristics of the language. We suggest using other data, such as newspaper articles, which may have more narrow focal points and thus fewer topics, and identifying gender based on author byline.

We also suggest that future work try retraining the politeness classifier on spoken dialogue.

Appendix

Anne researched the presuppositions behind the methods and applications of natural language processing to gender. She wrote Python code to preprocess and adapt the input to the interface expected by lda, ran movie scripts on the topic models with some parameters, and drafted the explanation of the experimental setup for the topic models. Nat also performed preprocessing on the data, wrote code to associate lines from the Cornell corpus to gender and genre, executed the topic models and Stanford parser on this corpus, wrote methods to adapt the input to the dependencies and interface expected by the politeness detector, ran the t-test, and examined the output to explain results.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993-1022.

David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM (Review Articles)*, 55(4):77-84.

Jonathan Chang, Jordan Boyd-Graber, Wang Chong, Sean Gerrish, David M. Blei. 2009. Reading tea leaves: how humans interpret topic models. In *NIPS*. 288– 296.

Edwin Chen. Introduction to latent Dirichlet allocation. 2011. http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/ Accessed 12 Dec. 2017.

Zhiyong Chen and Bing Liu. 2014. Topic modeling using topics from many domains, lifelong learning and big data. In *Proceedings of the 31st International Conference on Machine Learning*, 32:703-711.

Christian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination and linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*. https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, Christopher Potts. 2013 A computational approach to politeness with application to social factors. *Proceedings of ACL*.

Olivier Grisel, Lars Buitinck, and Chyi-Kwei Yau. 2017. Topic extraction with Non-negative Matrix Factorization and Latent Dirichlet Allocation. In *Sci-kit Learn*. http://scikit-learn.org/stable/auto_examples/applications/plot_topics_extraction_with_nmf_lda.html#sphx-glr-auto-examples-applications-plot-topics-extraction-with-nmf-lda-py Accessed 25 Nov. 2017.

Martha M. Lauzen. 2017. It's a man's (celluloid) world: portrayals of female characters in the top 100 films of 2016. http://womenintvfilm.sdsu.edu/wp-content/uploads/2017/02/2016-Its-a-Mans-Celluloid-World-Report.pdf Accessed 9 Oct. 2017.

lda. 2015. http://pythonhosted.org/lda/ Accessed 15 November 2017.

Alex Loginov. stop_words.py. *scikit-learn*. https://github.com/scikit-learn/scikit-learn/blob/master/sklearn/feature_extraction/stop_words.py Accessed 10 Dec. 2017.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60.

NLTK. http://www.nltk.org Accessed 15 November 2017.

Pedregosa et al. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825-2830.

Vinodkumar Prabhakaran and Owen Rambow. 2017. Dialog structure through the lens of gender, gender environment, and power. *Journal for Dialogue & Discourse* 8(2):21-55. arXiv:1706.03441

Moritz Suhof and Christian Danescu-Niculescu-Mizil. 2017. *Stanford Politeness API*. https://github.com/sudhof/politeness Accessed 11 Dec. 2017.

Xiaohui Yan. 2013. BTM. https://github.com/xiaohuiyan/BTM Accessed 15 November 2017.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*. 1445-1456.

Maryam Tafaroji Yeganeh, Issa Mellati Heravi, Abdolrasoul. 2015. Hedge and booster in newspaper articles on Iran's Presidential election: A comparative study of English and Persian articles. In *Procedia - Social and Behavioral Sciences* 192:679-83. https://doi.org/10.1016/j.sbspro.2015.06.103

Cornelis J. van Rijsbergen, Stephen E. Robertson, and Martin F. Porter. 1980. *New models in probabilistic information retrieval*. London: British Library Research and Development Department.