Automatically Inferring Gender Associations from Language

Abstract

In this paper, we pose the question: do people talk about women and men in different ways? We introduce two datasets and a novel integration of approaches for automatically inferring gender associations from language, discovering coherent word clusters, and labeling the clusters for the semantic concepts they represent. The datasets allow us to compare how people write about women and men in two different settings - one set draws from celebrity news and the other from student reviews of computer science professors. We demonstrate that there are large-scale differences in the ways that people talk about women and men and that these differences vary across domains. Human evaluations show that our methods significantly outperform strong baselines.

Introduction

It is well-established that gender bias exists in language – for example, we see evidence of this given the prevalence of sexism in abusive language datasets BIBREF0, BIBREF1. However, these are extreme cases of gender norms in language, and only encompass a small proportion of speakers or texts.

Less studied in NLP is how gender norms manifest in everyday language – do people talk about women and men in different ways? These types of differences are far subtler than abusive language, but they can provide valuable insight into the roots of more extreme acts of discrimination. Subtle differences are difficult to observe because each case on its own could be attributed to circumstance, a passing comment or an accidental word. However, at the level of hundreds of thousands of data points, these patterns, if they do exist, become undeniable. Thus, in this work, we introduce new datasets and methods so that we can study subtle gender associations in language at the large-scale.

Our contributions include:

Two datasets for studying language and gender, each consisting of over 300K sentences.

Methods to infer gender-associated words and labeled clusters in any domain.

Novel findings that demonstrate in both domains that people do talk about women and men in different ways.

Each contribution brings us closer to modeling how gender associations appear in everyday language. In the remainder of the paper, we present related work, our data collection, methods and findings, and human evaluations of our system.

Related Work

The study of gender and language has a rich history in social science. Its roots are often attributed to Robin Lakoff, who argued that language is fundamental to gender inequality, “reflected in both the ways women are expected to speak, and the ways in which women are spoken of” BIBREF2. Prominent scholars following Lakoff have included Deborah Tannen BIBREF3, Mary Bucholtz and Kira Hall BIBREF4, Janet Holmes BIBREF5, Penelope Eckert BIBREF6, and Deborah Cameron BIBREF7, along with many others.

In recent decades, the study of gender and language has also attracted computational researchers. Echoing Lakoff's original claim, a popular strand of computational work focuses on differences in how women and men talk, analyzing key lexical traits BIBREF8, BIBREF9, BIBREF10 and predicting a person's gender from some text they have written BIBREF11, BIBREF12. There is also research studying

how people talk to women and men BIBREF13, as well as how people talk about women and men, typically in specific domains such as sports journalism BIBREF14, fiction writing BIBREF15, movie scripts BIBREF16, and Wikipedia biographies BIBREF17, BIBREF18. Our work builds on this body by diving into two novel domains: celebrity news, which explores gender in pop culture, and student reviews of CS professors, which examines gender in academia and, particularly, the historically male-dominated field of CS. Furthermore, many of these works rely on manually constructed lexicons or topics to pinpoint gendered language, but our methods automatically infer gender-associated words and labeled clusters, thus reducing supervision and increasing the potential to discover subtleties in the data.

Modeling gender associations in language could also be instrumental to other NLP tasks. Abusive language is often founded in sexism BIBREF0, BIBREF1, so models of gender associations could help to improve detection in those cases. Gender bias also manifests in NLP pipelines: prior research has found that word embeddings preserve gender biases BIBREF19, BIBREF20, BIBREF21, and some have developed methods to reduce this bias BIBREF22, BIBREF23. Yet, the problem is far from solved; for example, BIBREF24 showed that it is still possible to recover gender bias from “de-biased” embeddings. These findings further motivate our research, since before we can fully reduce gender bias in embeddings, we need to develop a deeper understanding of how gender permeates through language in the first place.

We also build on methods to cluster words in word embedding space and automatically label clusters. Clustering word embeddings has proven useful for discovering salient patterns in text corpora BIBREF25, BIBREF26. Once clusters are derived, we would like them to be interpretable. Much research simply considers the top-n words from each cluster, but this method can be subjective and time-consuming to interpret. Thus, there are efforts to design methods of automatic cluster labeling BIBREF27. We take a similar approach to BIBREF28, who leverage word embeddings and WordNet during labeling, and we extend their method with additional techniques and evaluations.

Data Collection

Our first dataset contains articles from celebrity magazines People, UsWeekly, and E!News. We labeled each article for whether it was reporting on men, women, or neither/unknown. To do this, we first extracted the article's topic tags. Some of these tags referred to people, but others to non-people entities, such as “Gift Ideas” or “Health.” To distinguish between these types of tags, we queried each tag on Wikipedia and checked whether the top page result contained a “Born” entry in its infobox – if so, we concluded that the tag referred to a person.

Then, from the person's Wikipedia page, we determined their gender by checking whether the introductory paragraphs of the page contained more male or female pronouns. This method was simple but effective, since pronouns in the introduction almost always resolve to the subject of that page. In fact, on a sample of 80 tags that we manually annotated, we found that comparing pronoun counts predicted gender with perfect accuracy. Finally, if an article tagged at least one woman and did not tag any men, we labeled the article as Female; in the opposite case, we labeled it as Male.

Our second dataset contains reviews from RateMyProfessors (RMP), an online platform where students can review their professors. We included all 5,604 U.S. schools on RMP, and collected all reviews for CS professors at those schools. We labeled each review with the gender of the professor whom it was about, which we determined by comparing the count of male versus female pronouns over all reviews for that professor. This method was again effective, because the reviews are expressly written about a certain professor, so the pronouns typically resolve to that professor.

In addition to extracting the text of the articles or reviews, for each dataset we also collected various useful metadata. For the celebrity dataset, we recorded each article's timestamp and the name of the author, if available. Storing author names creates the potential to examine the relationship between the

gender of the author and the gender of the subject, such as asking if there are differences between how women write about men and how men write about men. In this work, we did not yet pursue this direction because we wanted to begin with a simpler question of how gender is discussed: regardless of the gender of the authors, what is the content being put forth and consumed? Furthermore, we were unable to extract author gender in the professor dataset since the RMP reviews are anonymous. However, in future work, we may explore the influence of author gender in the celebrity dataset.

For the professor dataset, we captured metadata such as each review's rating, which indicates how the student feels about the professor on a scale of AWFUL to AWESOME. This additional variable in our data creates the option in future work to factor in sentiment; for example, we could study whether there are differences in language used when criticizing a female versus a male professor.

Inferring Word-Level Associations

Our first goal was to discover words that are significantly associated with men or women in a given domain. We employed an approach used by BIBREF10 in their work to analyze differences in how men and women write on Twitter.

Inferring Word-Level Associations ::: Methods

First, to operationalize, we say that term i is associated with gender j if, when discussing individuals of gender j, i is used with unusual frequency – which we can check with statistical hypothesis tests. Let f_i represent the likelihood of i appearing when discussing women or men. f_i is unknown, but we can model the distribution of all possible f_i using the corpus of texts that we have from the domain. We construct a gender-balanced version of the corpus by randomly undersampling the more prevalent gender until the proportions of each gender are equal. Assuming a non-informative prior distribution on

f_i, the posterior distribution is Beta(k_i, $N - k_i$), where k_i is the count of i in the gender-balanced corpus and N is the total count of words in that corpus.

As BIBREF10 discuss, “the distribution of the gender-specific counts can be described by an integral over all possible f_i. This integral defines the Beta-Binomial distribution BIBREF29, and has a closed form solution.” We say that term i is significantly associated with gender j if the cumulative distribution at k_{ij} (the count of i in the j portion of the gender-balanced corpus) is $p \le 0.05$. As in the original work, we apply the Bonferroni correction BIBREF30 for multiple comparisons because we are computing statistical tests for thousands of hypotheses.

Inferring Word-Level Associations ::: Findings

We applied this method to discover gender-associated words in both domains. In Table TABREF9, we present a sample of the most gender-associated nouns from the celebrity domain. Several themes emerge: for example, female celebrities seem to be more associated with appearance (“gown,” “photo,” “hair,” “look”), while male celebrities are more associated with creating content (“movie,” “film,” “host,” “director”). This echoes real-world trends: for instance, on the red carpet, actresses tend to be asked more questions about their appearance –- what brands they are wearing, how long it took to get ready, etc. –- while actors are asked questions about their careers and creative processes (as an example, see BIBREF31).

Table TABREF9 also includes some of the most gender-associated verbs and adjectives from the professor domain. Female CS professors seem to be praised for being communicative and personal with students (“respond,” “communicate,” “kind,” “caring”), while male CS professors are recognized for being knowledgeable and challenging the students (“teach,”, “challenge,” “brilliant,” “practical”). These trends are well-supported by social science literature, which has found that female teachers are praised for

“personalizing” instruction and interacting extensively with students, while male teachers are praised for using “teacher as expert” styles that showcase mastery of material BIBREF32.

These findings establish that there are clear differences in how people talk about women and men – even with Bonferroni correction, there are still over 500 significantly gender-associated nouns, verbs, and adjectives in the celebrity domain and over 200 in the professor domain. Furthermore, the results in both domains align with prior studies and real world trends, which validates that our methods can capture meaningful patterns and innovatively provide evidence at the large-scale. This analysis also hints that it can be helpful to abstract from words to topics to recognize higher-level patterns of gender associations, which motivates our next section on clustering.

Clustering & Cluster Labeling

With word-level associations in hand, our next goals were to discover coherent clusters among the words and to automatically label those clusters.

Clustering & Cluster Labeling ::: Methods

First, we trained domain-specific word embeddings using the Word2Vec BIBREF33 CBOW model ($w \in R^{100}$). Then, we used k-means clustering to cluster the embeddings of the gender-associated words. Since k-means may converge at local optima, we ran the algorithm 50 times and kept the model with the lowest sum of squared errors.

To automatically label the clusters, we combined the grounded knowledge of WordNet BIBREF34 and context-sensitive strengths of domain-specific word embeddings. Our algorithm is similar to BIBREF28's approach, but we extend their method by introducing domain-specific word embeddings for clustering as

well as a new technique for sense disambiguation. Given a cluster, our algorithm proceeds with the following three steps:

Sense disambiguation: The goal is to assign each cluster word to one of its WordNet synsets; let S represent the collection of chosen synsets. We know that these words have been clustered in domain-specific embedding space, which means that in the context of the domain, these words are very close semantically. Thus, we choose S^* that minimizes the total distance between its synsets.

Candidate label generation: In this step, we generate L, the set of possible cluster labels. Our approach is simple: we take the union of all hypernyms of the synsets in S^*.

Candidate label ranking: Here, we rank the synsets in L. We want labels that are as close to all of the synsets in S^* as possible; thus, we score the candidate labels by the sum of their distances to each synset in S^* and we rank them from least to most distance.

In steps 1 and 3, we use WordNet pathwise distance, but we encourage the exploration of other distance representations as well.

Clustering & Cluster Labeling ::: Findings

Table TABREF11 displays a sample of our results – we find that the clusters are coherent in context and the labels seem reasonable. In the next section, we discuss human evaluations that we conducted to more rigorously evaluate the output, but first we discuss the value of these methods toward analysis.

At the word-level, we hypothesized that in the celebrity domain, women were more associated with appearance and men with creating content. Now, we can validate those hypotheses against labeled

clusters – indeed, there is a cluster labeled clothing that is 100% female (i.e. 100% words are female-associated), and a 80% male cluster labeled movie. Likewise, in the professor domain, we had guessed that women are associated with communication and men with knowledge, and there is a 100% female cluster labeled communication and a 89% male cluster labeled cognition. Thus, cluster labeling proves to be very effective at pulling out the patterns that we believed we saw at the word-level, but could not formally validate.

The clusters we mentioned so far all lean heavily toward one gender association or the other, but some clusters are interesting precisely because they do not lean heavily – this allows us to see where semantic groupings do not align exactly with gender association. For example, in the celebrity domain, there is a cluster labeled lover that has a mix of female-associated words (“boyfriend,” “beau,” “hubby”) and male-associated words (“wife,” “girlfriend”). Jointly leveraging cluster labels and gender associations allows us to see that in the semantic context of having a lover, women are typically associated with male figures and men with female figures, which reflects heteronormativity in society.

Human Evaluations

To test our clusters, we employed the Word Intrusion task BIBREF35. We present the annotator with five words – four drawn from one cluster and one drawn randomly from the domain vocabulary – and we ask them to pick out the intruder. The intuition is that if the cluster is coherent, then an observer should be able to identify the out-of-cluster word as the intruder. For both domains, we report results on all clusters and on the top 8, ranked by ascending normalized sum of squared errors, which can be seen as a prediction of coherence. In the celebrity domain, annotators identified the out-of-cluster word 73% of the time in the top-8 and 53% overall. In the professor domain, annotators identified it 60% of the time in the top-8 and 49% overall. As expected, top-8 performance in both domains does considerably better than overall, but at all levels the precision is significantly above the random baseline of 20%.

To test cluster labels, we present the annotator with a label and a word, and we ask them whether the word falls under the concept. The concept is a potential cluster label and the word is either a word from that cluster or drawn randomly from the domain vocabulary. For a good label, the rate at which in-cluster words fall under the label should be much higher than the rate at which out-of-cluster words fall under. In our experiments, we tested the top 4 predicted labels and the centroid of the cluster as a strong baseline label. The centroid achieved an in-cluster rate of .60 and out-of-cluster rate of .18 (difference of .42). Our best performing predicted label achieved an in-cluster rate of .65 and an out-of-cluster rate of .04 (difference of .61), thus outperforming the centroid on both rates and increasing the gap between rates by nearly 20 points. In the Appendix, we include more detailed results on both tasks.

Conclusion

We have presented two substantial datasets and a novel integration of methods to automatically infer gender associations in language. We have demonstrated that in both datasets, there are clear differences in how people talk about women and men. Furthermore, we have shown that clustering and cluster labeling are effective at identifying higher-level patterns of gender associations, and that our methods outperform strong baselines in human evaluations. In future work, we hope to use our findings to improve performance on tasks such as abusive language detection. We also hope to delve into finer-grained analyses, exploring how language around gender interacts with other variables, such as sexual orientation or profession (e.g. actresses versus female athletes). Finally, we plan to continue widening the scope of our study – for example, expanding our methods to include non-binary gender identities, evaluating changes in gender norms over time, and spreading to more domains, such as the political sphere.