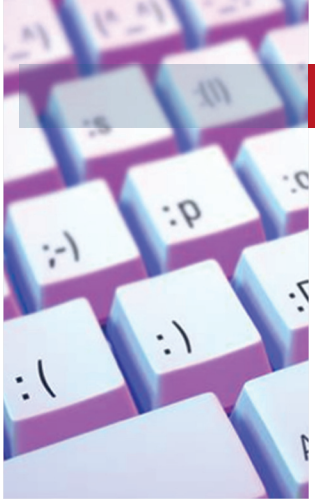

Editor: **Erik Cambria,** Nanyang Technological University, Singapore, cambria@ntu.edu.sg

What Men Say, What Women Hear: Finding Gender-Specific Meaning Shades

Rada Mihalcea and Aparna Garimella, *University of Michigan*

How could one go about uncovering men's and women's different ways of perceiving the surrounding world? We use the personal writings of men and women, as available through a very large collection of weblogs, and attempt to uncover gender differences as expressed in day-to-day accounts of experiences and perceptions. Previous work on understanding gender differences has mainly focused on authorship detection, trying to identify the gender of the author of a certain writing, be that a blog,[1] a tweet,[2] or other works of fiction or nonfiction[3] (see the "Related Work in Computational Studies of Gender" sidebar for more information). In this article, we depart from this earlier research and attempt to move beyond the surface level of word occurrences and counts. We instead use semantic analysis to identify differences that exist between genders in how they use certain words.

Specifically, we address the following question: Can we distinguish between shades of word meanings, as used by the two genders? Do men and women use the word *car* in a similar way, or are there differences between the use of this word in their day-to-day life? What about the words *laugh* or *read*? We answer this question by using a word sense disambiguation framework in which each gender is regarded as a "sense," and we detect the gender corresponding to a given occurrence of a word. Using a large dataset of more than 350 words, we show that gender-based word disambiguation is possible, and that there are indeed differences between the ways men and women use certain words.

Gender-Based Word Disambiguation

Our driving hypothesis is that men and women use some words differently, which we can regard as a reflection of the differences in how they see the world around them. To test this hypothesis, we used examples drawn from men's and women's writings for a large number of words (see the "Data" sidebar for sample blogposts), and we built disambiguation models centered on these target words. We therefore formulated our task as a word sense disambiguation problem, and we attempted to automatically identify the gender of the person using a certain target word.

Target Words

We wanted to investigate the behavior of words in the language of the two genders and verify whether the difference in word behavior comes from changes in sense or changes in wording in the context. Therefore, we chose a mixture of polysemous words and monosemous words (according to WordNet 3.0[4]), and we chose words that frequently appear in both genders' writings as well as words that are frequently used by only one gender.

According to these criteria, for each open-class word (that is, nouns, verbs, adjectives, and adverbs), we selected 100 words, 50 of which have multiple senses and 50 with one sense only. Each of these two sets has a 30-10-10 distribution: 30 words that are frequent in both men's and women's writings, with a distribution in the two genders falling in the 40 to 60 percent range, and 10 words per each gender such that these words are frequent only in one gender (that is, words that

1541-1672/16/$33.00 © 2016 IEEE
Published by the IEEE Computer Society

Related Work in Computational Studies of Gender

One of the earliest studies addressing language differences between men and women found several characteristics of women's language, including words such as *lovely* and *adorable* or phrases such as "it seems to be" or "would you mind."[1] A large body of work also addresses the connection between language and gender in the field of sociology,[2] which we do not address here due to a lack of space.

In computational linguistics, several studies addressed the role of gendered language and the "gender gap" in the blogosphere,[3–6] the significance of gender differences in self-disclosure strategy in teenage blogs,[7] and the validity of author gender predictions based largely on function words (such as pronouns and determiners).[8] Work has also been done on Twitter data, where tweets are used to predict several profile features, including gender.[9–11] Claudia Peersman and colleagues performed age and gender prediction on short messages from social networking sites.[12] The focus in these previous studies has been primarily on investigating the use of automatic classification to distinguish between men's and women's writings, and also on finding words that are specific to each gender by performing statistical analysis on large amounts of data.

Other related work includes recently published research by Dong Nguyen and colleagues,[13] who showed how a person's gender identity can be constructed using various linguistic aspects of male and female speech in language. Also of interest is the work by Vinodkumar Prabhakaran and colleagues,[14] who used topic segments to predict the behavioral patterns of political leaders in election campaigns. In speech, Constantinos Boulis and Mari Ostendorf presented an analysis of the most frequently used words by men and women in telephone conversations.[15]

One exception from the general theme of previous work on surface-level gender classification is the work by Ruchita Sarawgi and colleagues, in which they explicitly avoid topic bias in order to identify stylistic differences between men's and women's writings.[16] The authors use blogs addressing predefined topics (such as education or travel) and scientific publications and show that differences can be found even when the data sources are controlled for topic. In our research, we zoom in even deeper and try to identify the distinctive ways in which men and women use certain words.

References

1. R.T. Lakoff, *Language and Woman's Place*, Cambridge Univ. Press, 1972.
2. P. Eckert and S. McConnell-Ginet, *Language and Gender*, Cambridge Univ. Press, 2003.
3. T.L.M. Kennedy, J.S. Robinson, and K. Trammell, "Does Gender Matter? Examining Conversations in the Blogosphere," *Internet Research 6.0: Internet Generations*, 2005.
4. M. Koppel, S. Argamon, and A. Shimoni, "Automatically Categorizing Written Texts by Author Gender," *Literary and Linguistic Computing*, vol. 4, no. 17, 2002, pp. 401–412.
5. J. Schler et al., "Effects of Age and Gender on Blogging," *Proc. AAAI Spring Symp. Computational Approaches for Analyzing Weblogs*, 2006, pp. 199–204.
6. A. Mukherjee and B. Liu, "Improving Gender Classification of Blog Authors," *Proc. Conf. Empirical Methods in Natural Language Processing*, 2010, pp. 207–217.
7. D. Huffaker and S.L. Calvert, "Gender, Identity and Language Use in Teenage Blogs," *J. Computer-Mediated Communication*, vol. 10, no. 2, 2005; doi:10.1111/j.1083-6101.2005.tb00238.x.
8. S. Herring and J. Paolillo, "Gender and Genre Variation in Weblogs," *J. Sociolinguistics*, vol. 10, no. 4, 2004, pp. 439–459.
9. D. Rao et al., "Classifying Latent User Attributes in Twitter," *Proc. 2nd Int'l Workshop Search and Mining User-Generated Contents*, 2010, pp. 37–44.
10. J. Burger et al., "Discriminating Gender on Twitter," *Proc. Conf. Empirical Methods in Natural Language Processing*, 2011, pp. 1301–1309.
11. S. Volkova and D. Yarowsky, "Improving Gender Prediction of Social Media Users via Weighted Annotator Rationales," *Proc. Workshop Personalization: Methods and Applications*, 2014.
12. C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting Age and Gender in Online Social Networks," *Proc. 3rd Int'l Workshop Search and Mining User-Generated Contents*, 2011, pp. 37–44.
13. D. Nguyen et al., "Why Gender and Age Prediction from Tweets Is Hard: Lessons from a Crowdsourcing Experiment," *Proc. 25th Int'l Conf. Computational Linguistics*, 2014, pp. 1950–1961.
14. V. Prabhakaran, A. Arora, and O. Rambow, "Staying on Topic: An Indicator of Power in Political Debates," *Proc. Conf. Empirical Methods in Natural Language Processing*, 2014, pp. 1481–1486.
15. C. Boulis and M. Ostendorf, "A Quantitative Analysis of Lexical Differences Between Genders in Telephone Conversations," *Proc. 43rd Ann. Conf. Computational Linguistics*, 2005, pp. 435–442.
16. R. Sarawgi, K. Gajulapalli, and Y. Choi, "Gender Attribution: Tracing Stylometric Evidence Beyond Topic and Genre," *Proc. 15th Conf. Computational Natural Language Learning*, 2011, pp. 78–86.

have a frequency for the dominant gender higher than 70 percent).

From the initial set of 400 words, we could not identify enough examples (that is, at least 100) for 28, which left us with a final set of 372 words.

Data Preprocessing

For each target word in our dataset, we collected all the examples found for both genders, for an average number of 5,356 examples per target word. The average number of examples was 492 examples per target word.

We then processed all the extracted snippets: we tokenized and part-of-speech tagged the text using the Stanford tagger,[5] and we removed the contexts that did not include the target word with the specified part of speech. We also identified the target word's position and recorded it as an offset along with the example.

Features and Classification

The classification algorithm we used was inspired by previous work on data-driven word sense disambiguation.[6,7] Specifically, we used a system that integrates local, topical, and sociolinguistic features. The local features include the current word and its part of speech, a local context of three words to the left and right of the ambiguous word, the parts of speech of the surrounding words, the first noun

Data

We use a large corpus of blogposts annotated for gender, which we collected from Blogspot (www.blogspot.com). We chose Blogspot as opposed to other blog communities such as LiveJournal or MSN Spaces because it has richer blogger profile annotations, including gender, age, location, occupation, and others. The kind of writing found in a weblog is ideally suited to what we wish to discover, because weblogs often give an intimate account of personal everyday life and a personal viewpoint of current events.

Starting with the names of approximately 300,000 blogs that were updated with a new entry during the time when the crawling was performed, we collected the bloggers' profile pages and the corresponding profile features. We discarded all the blogs maintained by more than one blogger (collective blogs) and all those that did not include the blogger's gender. Finally, we parsed the entries from the remaining set of blogs and retained only the blogposts written in English and having a 200 to 4,000 character limit. Interestingly, although a large fraction of the blogs listed on Blogspot are spam, our constraints removed almost all the spam—to the point that a random hand-check of 100 blogposts revealed clean spam-free data.

The postprocessing and profile-based filters left us with a total of about 160,000 blog entries annotated for gender, which after balancing between male and female authors, left us with the final set of 75,000 male blog entries and 75,000 female blog entries. Figure A shows two sample entries written by a male and a female writer.

Male-authored blogpost

No word back from the Georges Island people on possible use of their power, so I'm going to proceed with the QRP plans. Even though the QRP stuff is smaller than the 100 watt outfit, there will still be a significant amount of stuff I'll need to wrestle on to the island. I'll bring the Pelican 1510 case outfitted with the Elecraft K 2.

Female-authored blogpost

You could probably tell that I literally enjoy dressing up in costumes and crap. I just don't have the resources nor the skills to make a good costume. But I'm a resource for outlandish ideas. I remember shocking my host dad when I told him that I enjoy dressing up like that.

Figure A. Male- and female-authored blogposts.

Table 1. Results for different parts of speech.

Part of speech	No. words	Average no. examples	Baseline (%)	Disambiguation algorithm (%)
Noun	100	6,144	66.61	73.52*
Verb	92	4,592	65.91	74.37*
Adjective	100	4,295	66.74	74.27*
Adverb	80	6,574	63.81	72.51*
Overall	372	5,356	65.86	73.71*

* Statistical significance measured using a t-test, $p < 0.05$.

before and after the target word, and the first verb before and after the target word. We also determined the topical features from the global context and implemented them through class-specific keywords, which are determined as a list of at most five words occurring at least three times in the contexts defining a certain gender. The sociolinguistic features provide social and psychological insights into the perceptions bloggers have about the words they use. They are calculated as percentages of words belonging to a word class out of the total number of words, wherein the word classes are drawn from Linguistic Inquiry and Word Count,[8] Opinion Finder,[9] Morality Lexicon,[10] and Wordnet Affect.[11] The features are then integrated in an Adaboost ensemble classifier. (The base learner in Adaboost used in our experiments is Decision Stump.)

For evaluation, we calculated the average accuracy obtained in tenfold cross-validations on the data collected for each word. For perspective, we also calculated a simple baseline that assigns the most frequent class by default.

Results and Discussion

Table 1 summarizes the results obtained for the 372 words. Disambiguation results that are significantly better than the baseline are marked with * (statistical significance measured using a t-test, $p < 0.05$). Overall, we found that there are indeed differences between the ways men and women use these target words, with an absolute increase over the baseline of 7.85 percent (which corresponds to a relative error rate reduction of 22.9 percent).

Among the words we considered, some words experienced large improvements over the baseline, such as *knowledge* (with an absolute increase over the baseline of 15.39 percent), *achieve* (15.72 percent), *current* (15.43 percent), or *simply* (13.49 percent). Some words experienced small improvements, including *development* (1.55 percent), *democratic* (0.46 percent), and *convict* (1.50 percent). A few words were dominant in one gender, so

the disambiguation accuracy was below the baseline—for example, *datum* (– 0.71 percent), *fund* (–2.20 percent), *secular* (–7.26 percent), and *effectively* (–5.09 percent).

Per-Gender Word Frequency

To understand to what extent the change in frequency has an impact on gender-based word disambiguation, we report results for words that have high frequency in both genders, or in only one gender at a time (see Table 2). Surprisingly, the words that are used more often by one gender are harder to disambiguate. Although this could be an artifact of the higher baseline, it might also suggest that the words that "belong" to a gender are used in a similar way by both genders (for example, *cozy*), unlike words that are frequent in both genders, which get loaded with gender-specific meaning (for example, *helpful*).

Lexical Meanings

The second analysis that we performed was concerned with the accuracy of polysemous words (words with multiple lexical meanings) as compared to monosemous words (words with only one lexical meaning). Table 3 reports the comparative results. We obtained similar improvements for both monosemous and polysemous words, which indicates that the gender differences we observed were not due to the use of different word meanings, but rather to men and women using the same word meaning in different ways.

To further understand the relationship between lexical meanings and gender-specific word usage, we performed a qualitative analysis: we selected 12 words (adjectives: *young*, *strong*, and *new*; adverbs: *together*, *later*, and *fast*; nouns: *party*, *idea*, and *couple*; and verbs: *heat*, *cause*, and *understand*). We randomly chose 100 examples for each of these words, with an equal split between male and female, and manually annotated their senses with respect to WordNet.[4] From these annotations, we observed that the predominant senses used by each gender were largely the same for most words. For instance, the word *party*, shown in Figure 1a, has a similar distribution over word senses. But there are also a few exceptions: an interesting example is the adverb *together*, which men use more often with the sense of "assembled in one place," whereas women use it with the sense of "in each other's company"; this is in line with the observation made by previous work on gender differences that women are more interested in family and friends, whereas men care more about groups and work.[12]

In general, we find that the distribution of WordNet word senses for men and women for the 12 selected words was mostly similar. For an overall quantification, we used correlation metrics to relate the word sense frequencies of the two genders, resulting in a Pearson score of 0.94 and a Spearman score of 0.88, which represent a high correlation. This suggests once again that the word-centered differences that we observed between men and women are not due to distinct word meanings, but rather to different ways of using a certain word.

Table 2. Results for words that have high frequency in both genders or in one gender at a time.

Part of speech	No. words	Average no. examples	Baseline (%)	Disambiguation algorithm (%)
		High frequency in both genders		
Noun	59	9,141	63.87	72.71
Verb	60	4,516	64.19	73.72
Adjective	60	6,634	62.73	73.10
Adverb	60	8,593	63.31	73.21
Overall	239	7,676	63.52	73.19
		High frequency in one gender		
Noun	41	1,831	70.44	74.62
Verb	32	1,279	69.13	75.57
Adjective	40	786	72.77	76.02
Adverb	20	516	65.30	70.40
Overall	133	1,186	70.05	74.64

Table 3. Results for words that are polysemous or monosemous.

Part of speech	No. words	Average no. examples	Baseline (%)	Disambiguation algorithm (%)
		Polysemous words		
Noun	50	9,612	66.22	72.72
Verb	50	7,697	67.74	74.48
Adjective	50	6,744	68.48	74.56
Adverb	43	8,105	64.80	72.10
Overall	193	8,037	66.88	73.52
		Monosemous words		
Noun	50	2,676	66.91	74.27
Verb	42	897	63.73	74.23
Adjective	50	1,845	65.00	73.98
Adverb	37	4,794	62.66	72.98
Overall	179	2,464	64.75	73.91

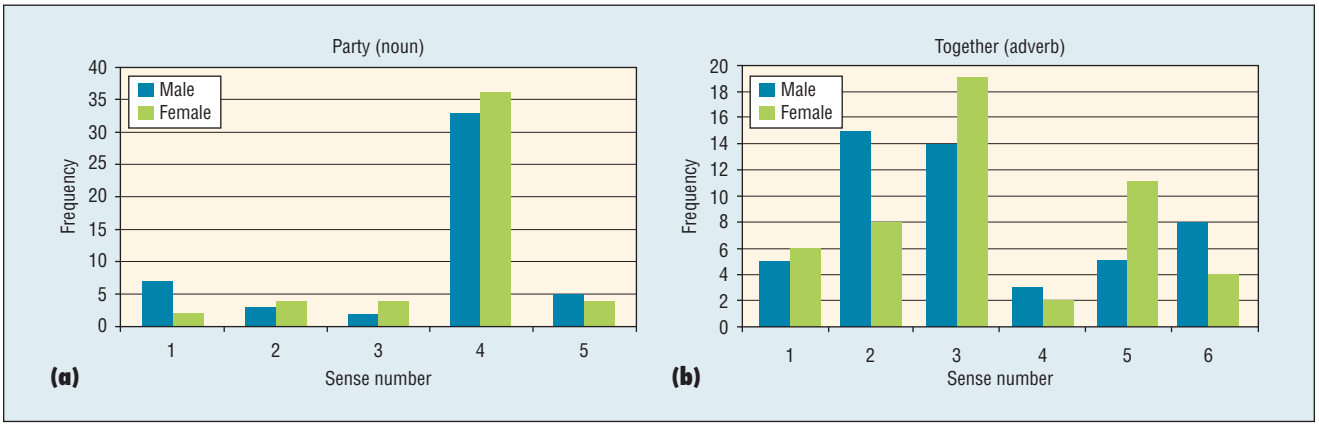

Figure 1. Distribution of WordNet senses for men and women for two words: (a) party (noun) and (b) together (adverb). We chose 100 examples for each word.

Table 4. Results for forward feature ablation for different parts of speech.

Part of speech	Local (%)	Topical (%)	Sociolinguistic (%)	All (%)	Baseline (%)
Noun	66.71	73.14	71.91	73.52	66.56
Verb	66.56	74.07	71.78	74.37	65.91
Adjective	66.61	74.13	72.38	74.27	66.74
Adverb	63.71	72.45	70.05	72.51	63.81
Overall	66.00	73.49	71.60	73.71	65.86

Feature Ablation

We further studied the role of different linguistic features in the disambiguation between word usages by men and women. We retrained our gender classification model for each of the target words, using only one of each of the three feature types (local, topical, and sociolinguistic) at a time. This helped us locate the features that contributed the most to the observed word usage differences between the two genders. Table 4 shows the feature ablation results, averaged over the 372 target words, using the three features types separately. From Table 4, we observe that topical and sociolinguistic features contributed the most to the performance of the word models and led to a significantly greater accuracy compared to that of the baseline.

Topic Modeling

To further identify differences in the usage of words between the two genders, we specifically focused on the top 60 words (top 15 words for each part of speech) with the most significant improvements over the majority baseline. We modeled the different usages of the words in our set of 60 words using topic modeling. Specifically, we used Latent Dirichlet Allocation to find a set of topics for each word and consequently identified the topics specific to either of the two genders.[13] As is typically done in topic modeling, we preprocessed the data by removing a standard list of stop words, words with very high frequency (more than 0.25 percent of the datasize), and words that occur only once. For simplicity, we fixed the number of topics to five in all the topic modeling experiments.

For each data instance, we say that a topic dominates the other topics if its probability is higher than that of the remaining topics. For a given word, we then identified the dominating topic for each gender as the topic that dominates the other topics in a majority of data instances. Consider the noun *team*. Although the dominating topic among men is associated with sports, described by words such as *match*, *league*, *baseball*, *soccer*, and *winning*, the dominating topic among women is associated with *holiday*, described by words such as *trip*, *Sunday*, *sleep*, *pictures*, *ride*, and *dinner*. Another interesting example is the noun *email*. The dominating topic among men is described by the words *website*, *online*, *project*, *design*, and *marketing*. On the other hand, the dominating topic among women is described by the words *birthday*, *miss*, *tomorrow*, *computer*, *care*, and *mother*. Another interesting example is the noun *music*. While the words *film*, *record*, *guitar*, *pop*, *singer*, and *stage* describe the dominating topics among men, the words *evening*, *park*, *coffee*, *beach*, and *drive* dominate among women.

Models of word usage let us move beyond the surface-level text classification approach to gender discrimination and gain insights into the differences between men and women. We believe these distinctions at a deeper semantic level can be regarded as a reflection of the differences between the genders' perception of the world around them. In future work, we plan to improve the disambiguation algorithm by including additional sociolinguistic and psycholinguistic features and perform an in-depth analysis of

the features that best characterize the differences in word usage between men and women. We would also like to perform additional experiments in which we control for potential confounding factors (such as domain, background culture, and age). ▪

Acknowledgments

This material is based in part on work supported by National Science Foundation award #1344257 and by grant #48503 from the John Templeton Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the John Templeton Foundation.

References

1. A. Mukherjee and B. Liu, "Improving Gender Classification of Blog Authors," *Proc. Conf. Empirical Methods in Natural Language Processing*, 2010, pp. 207–217.
2. J. Burger et al., "Discriminating Gender on Twitter," *Proc. Conf. Empirical Methods in Natural Language Processing*, 2011, pp. 1301–1309.
3. M. Koppel, S. Argamon, and A. Shimoni, "Automatically Categorizing Written Texts by Author Gender," *Literary and Linguistic Computing*, vol. 4, no. 17, 2002, pp. 401–412.
4. G. Miller, "WordNet: A Lexical Database," *Comm. ACM*, vol. 38, no. 11, 1995, pp. 39–41.
5. K. Toutanova et al., "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," *Proc. Human Language Technology Conf.*, 2003, pp. 173–180.
6. Y.K. Lee and H.T. Ng, "An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation," *Proc. Conf. Empirical Methods in Natural Language Processing*, 2002, pp. 41–48.
7. B. Dandala, R. Mihalcea, and R. Bunescu, *Word Sense Disambiguation Using Wikipedia*, Springer, 2013.
8. J.W. Pennebaker, M.E. Francis, and R.J. Booth, *Linguistic Inquiry and Word Count*, Lawrence Erlbaum, 2001.
9. T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," *Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 347–354.
10. G. Ignatow and R. Mihalcea, "Injustice Frames in Social Media," Am. Sociological Assoc. Ann. Meeting, 2012.
11. C. Strapparava and A. Valitutti, "WordNet-Affect: An Affective Extension of WordNet," *Proc. 4th Int'l Conf. Language Resources and Evaluation*, 2004, pp. 1083–1086.
12. H. Liu and R. Mihalcea, "Of Men, Women, and Computers: Data-Driven Gender Modeling for Improved User Interfaces," *Proc. Int'l Conf. Weblogs and Social Media*, 2007; www.icwsm.org/papers/2--Liu-Mihalcea.pdf.
13. D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research*, vol. 3, 2003, pp. 993–1022.

Rada Mihalcea is a professor in the Department of Computer Science and Engineering at the University of Michigan. Contact her at mihalcea@umich.edu.

Aparna Garimella is a PhD student in the Department of Computer Science and Engineering at the University of Michigan. Contact her at gaparna@umich.edu.