Automatic Bechdel: Paying Attention to Women in Movies.

Naidoo, Alec; Fenton, Zachary
DATASCI 266: Natural Language Processing
UC Berkeley School of Information
{alecarnassi, zacharytfenton}@ischool.berkeley.edu

Figure 1. Alison Bechdel, 1985, "The Rule", *Dykes to Watch Out For*

Abstract

The Bechdel test is a rudimentary test that has been applied in the social sciences for assessing gender bias in popular culture, in particular the underrepresentation of women in movies. Recent advancements in language models open the door to the possibility of automating the Bechdel test. In this paper, we present two NLP approaches to the automation task. We train a transformer based model with chunking on screenplays that are labeled with binary Bechdel scores (pass/fail). We also take a coreference resolution approach to identify characters within screenplay dialogue to evaluate the Bechdel criteria. Our results show that both methods improve on baseline models and give confidence to the idea that automating the Bechdel test, with modern NLP procedures, is possible.

1 Introduction

The Bechdel Test was created in 1985 by Alison Bechdel in their comic titled "The Rule"[1] (Figure 1). The comic strip portrays two women discussing the requirements to see a movie. These rules are ultimately translated into what is now known as the Bechdel test, a simple test to measure the representation of women in movies and literature.

In order to pass the Bechdel Test, three criteria must be met; (1) the piece of media must have at least two named female characters, (2) they must talk to each other, (3) and they must discuss something other than a man. Previous studies have utilized a machine learning approach to apply the Bechdel Test to assess gender equality. For example the work of Scheiner-Fisher and Russell III (2012) have applied ML to asses male bias in history curriculum and Garcia et al. (2014) have applied it to study social media posts (i.e. Twitter, Myspace). With the increasing presence of machine learning, automating the Bechdel Test and applying it to movie scripts has become an achievable reality.

In this paper, we build off of the work of Agarwal et al. (2015) in automating the Bechdel Test, specifically as it is applied to movie scripts. We first built a baseline model by replicating the work of Umairican (2014) and obtained an *accuracy* score of 0.558 and an *F1* score of 0.558. Our goal for this paper is to build an automatic Bechdel test that will achieve higher *accuracy* and *F1* while overcoming the challenges of applying transformer models to long document sizes of screenplays. With the advancement of language models, we approach the goal of creating an automatic Bechdel test by first utilising transformer based models and then a character coreference approach.

2 Background

Gender bias in media is an area in the social sciences that receives much attention and is continuously studied to better understand why and how it occurs (Rattan et al. 2019). The Bechdel Test in particular has gained attention for its simplicity and its strong implications for analyzing mainstream storytelling and the representation of women in culture. Hickey (2014) found that, in the United States, films that pass the Bechdel test have a higher return on investment compared to films that did not pass the test. Appel and Gnambs (2023) also found a positive correlation between movies that pass the test and higher revenues and audience IMDb ratings.

With the advances of natural language processing and their linguistic applications, automating this test using machine learning offers a natural opportunity to improve its accessibility and streamline the process of generating meaningful insights.

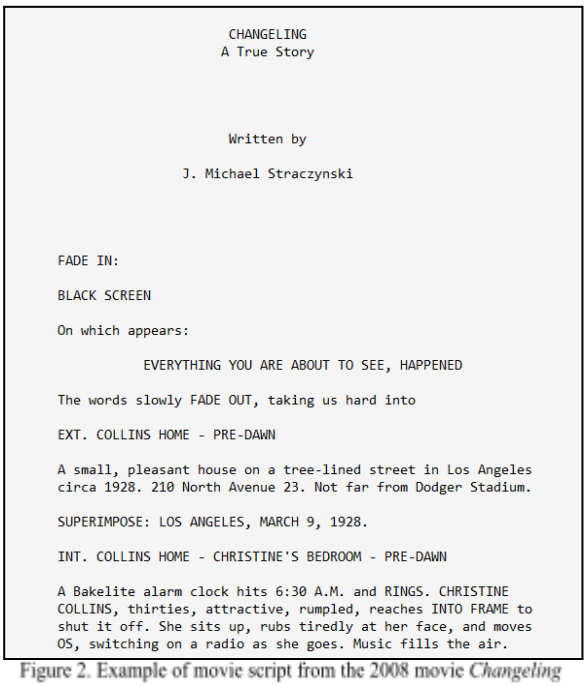
The seminal work of Agarwel et al. (2015) utilizes many statistical and machine learning techniques to approach the problem of automating the Bechdel Test. The paper concludes that the most effective technique is to develop features based on social network analysis (SNA) and a set of complex rules. In order to analyze such a large text, the authors apply their previous work to format screenplays for ease of use. (Agarwal et al., 2014).

[1] The original publication was in a short comic strip titled '*Dykes to Watch Out For*' that ran in the LGBTQ+ friendly newsletter **Funny Times.**

3 Methods

3.1 Data

We used a publicly available dataset from Hugging Face[2] that consists of 426 scripts, each having a corresponding Bechdel rating score. The average script length is 24,000 words with a minimum script length of 571 words and maximum of 52,766. The rating score is drawn directly from bechdeltest.com. This website utilizes user rankings and discussions for deriving the scores. The website is active in its discussion and revision of evaluations, giving confidence to the scores.

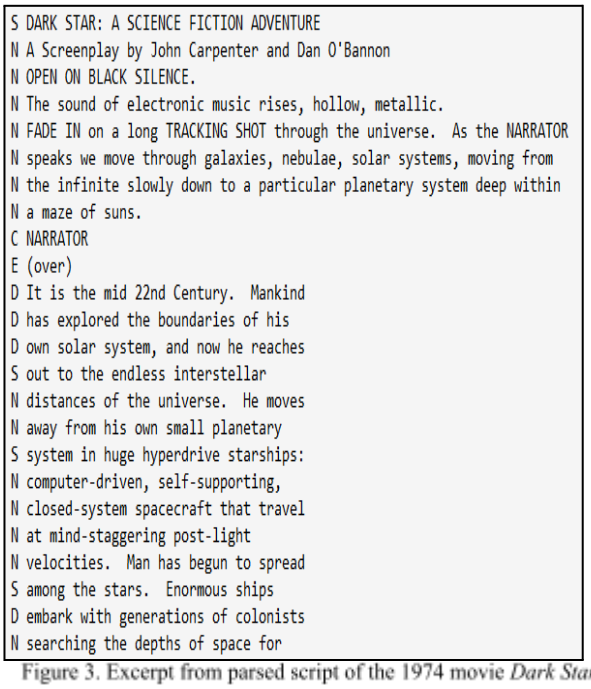


```
                        CHANGELING
                      A True Story



                        Written by

                    J. Michael Straczynski



FADE IN:

BLACK SCREEN

On which appears:

            EVERYTHING YOU ARE ABOUT TO SEE, HAPPENED

The words slowly FADE OUT, taking us hard into

EXT. COLLINS HOME - PRE-DAWN

A small, pleasant house on a tree-lined street in Los Angeles
circa 1928. 210 North Avenue 23. Not far from Dodger Stadium.

SUPERIMPOSE: LOS ANGELES, MARCH 9, 1928.

INT. COLLINS HOME - CHRISTINE'S BEDROOM - PRE-DAWN

A Bakelite alarm clock hits 6:30 A.M. and RINGS. CHRISTINE
COLLINS, thirties, attractive, rumpled, reaches INTO FRAME to
shut it off. She sits up, rubs tiredly at her face, and moves
OS, switching on a radio as she goes. Music fills the air.
```

Figure 2. Example of movie script from the 2008 movie *Changeling*

3.2 Data Wrangling

Movie scripts are created in a semi-standardized format and contain dialogue, setting, description and direction for a movie. Figure 2 shows an example screenplay. One concern of the dataset was the inconsistencies in publicly available scripts. We decided to utilize the robust screenplay parser created by Baruah et. al. (2023). The screenplay parser uses an RNN, taking advantage of the semi-consistent structure of screenplays, and outputs a line-by-line list of tags that identify components of the script. The returned tags are as follows: 'S': scene header, 'N': scene description, 'C': character name, 'D': utterance, 'E': expression, 'T': transition, 'M': metadata, 'O': other. The list and screenplay are then combined into a single element. Figure 3 shows an example of a parsed script with tags. After parsing the dataset, we analyzed the data for missing scripts, duplicates, and outliers. After final cleaning, the dataset totaled 414 unique scripts having an average length of 28,000 words with the minimum length being 648 words and the maximum being 62,384 words. The discrepancies between the raw and parsed data come from the addition of the tags at each line and how the encoder tokenizes them.



```
S DARK STAR: A SCIENCE FICTION ADVENTURE
N A Screenplay by John Carpenter and Dan O'Bannon
N OPEN ON BLACK SILENCE.
N The sound of electronic music rises, hollow, metallic.
N FADE IN on a long TRACKING SHOT through the universe.  As the NARRATOR
N speaks we move through galaxies, nebulae, solar systems, moving from
N the infinite slowly down to a particular planetary system deep within
N a maze of suns.
C NARRATOR
E (over)
D It is the mid 22nd Century.  Mankind
D has explored the boundaries of his
D own solar system, and now he reaches
S out to the endless interstellar
N distances of the universe.  He moves
N away from his own small planetary
S system in huge hyperdrive starships:
N computer-driven, self-supporting,
N closed-system spacecraft that travel
N at mind-staggering post-light
N velocities.  Man has begun to spread
S among the stars.  Enormous ships
D embark with generations of colonists
N searching the depths of space for
```

Figure 3. Excerpt from parsed script of the 1974 movie *Dark Star*

The Bechdel Test has three criteria that are required to be met in order to pass. Each script is scored by a human with a rating from 0 to 3. A `0` score indicates the movie failed all three tests, a `1` indicates it passed one test, `2`, two tests, and a `3` indicates that a script passes all three tests. The value counts are broken down as follows: `0` - 29; `1` - 138; `2` - 57, `3` - 190. The uneven class groups proved to be a problem with application of multiclass labelling via a machine learning model. To better balance the class and improve the accuracy of the model, the classes were converted to a binary format of either fully passing (1) or failing (0). The final result was 224 failing classes (~ 54%) and 190 passing (~ 46%) classes.

2 https://huggingface.co/datasets/mocboch/movie_scripts/tree/main

3.3 Chunking

BERT models typically restrict their token input length to 512. After tokenization, a movie script far exceeds this restriction. To allow for optimum training, it is desirable to encode an entire movie script for input to a model. In work done by Sun et al. (2020), they experiment with 2 methods for handling large texts; truncation or hierarchical. They show that ultimately, the hierarchical method performs the best on large texts. In our work, we build off of the hierarchical method through a cross application with chunking.

Chunking is the application of breaking down large texts into smaller subdocument embeddings that can be fed individually and sequentially into a model, while still retaining necessary contextualization (Jaiswal et al., 2023). The scripts were divided into 100 equal length chunks of 512 tokens per chunk. This decision was made based on maximum script length and resource limitations while still maintaining the greatest amount of information in a script as possible. Each chunk is then fed into a pre-trained BERT model and the hidden state of the [CLS] token is used as the representation for that chunk. The representations are concatenated and fed as the output of the pre-trained BERT model, into the next layer which in our case is an LSTM layer.

3.4 Bert Approach

In the baseline model, an LSTM was used to train a model for automatic Bechdel test. The introduction of transformers, specifically BERT, allows for greater attention across larger documents (Vaswani et al., 2017). We choose 3 pre-trained BERT models.

3.4.1 'bert-base-cased' - BC:

'Bert-based-cased' is trained on cased English text. This was chosen as the first BERT pretrained model due to its standard use as well as reasonable size.

3.4.2 'bert-large-cased' - LC:

We chose to use this model due to its larger parameter set that is able to handle more complex tasks, such as large text span entity recognition.

3.4.3 'bert-base-cased-finetuned-mrpc' - MRPC:

This model was chosen due to the fact that it was fine tuned on the GLUE-MRPC dataset. This fine tuning is specific to determining whether paraphrasing or semantic similarities between sentences exist. The hope was this could be applied to dialogue understanding.

All BERT models were trained for 5 epochs with a batch size of 5, dropout rate of 0.1, and learning rate of 0.00002. We decided to use 331 movies for training and evaluation and 83 movies for testing (~20%). To preserve experiment integrity, we only varied the pre-trained model for each run. The standard functional model had the final layer unfrozen and output the pooler-token to an LSTM layer. Due to the overall complexity of the model, we used Adam for optimization. We utilized sparse categorical cross-entropy as our loss function and chose binary accuracy as the model metric.

3.5 Coreference Approach

In the second approach, we adopt the coreference resolution research of Baruah et al. (2023) to our task. Our objective is to identify character references within each conversation of a scene and apply a set of rules to score a conversation according to the Bechdel Test. Combining coreference resolution modeling with scoring heuristics for each scene aims to capture nuanced details of conversations within a scene that would improve the performance of our Bechdel Test automation.

Baruah et al. (2023) applies the word-level coreference resolution model of Dobrovolskii (2021) to the screenplay format while also addressing the challenge of long document sizes for screenplays. We adapt a fusion-based approach, according to their research, that divides a given screenplay and infers based on smaller, overlapping documents.

The model works by encoding tokens into word representations using a pretrained RoBERTa transformer model (Liu et al. 2019). These representations are combined with embedded word features like POS, named entities, and structural screenplay tags from the parser. A bi-directional RNN is applied to obtain a hidden layer vector, and feeds into a feed-forward Neural Network to output a character score for each word representation. Each antecedent word representation is then given an antecedent score for a given character

word representation, and the antecedent with the max score is paired. Words with no antecedent candidates are negatively scored. The model infers this scoring on the smaller, overlapping documents of each screenplay with a set of hyperparameters. Each subdocument has 5,120 tokens maximum, and the overlap between subdocuments can contain a maximum of 2,048 tokens. Character scores and antecedent pairing scores are calculated for each subdocument. If a character reference and an antecedent pair lie between two adjacent subdocuments within the overlap region, it is possible that we calculate two separate coreference scores but, in that case, these scores are averaged for the same antecedent. These overlapping regions allow coreference clusters to link across subdocuments and overcome long document sizes of screenplays.

For each screenplay, we apply the parser to label the structural tags of the screenplay by each line, and then we apply the fusion-based coreference resolution model to identify instances of men in dialogue-tag lines in conversations between women. Due to resource constraints, we use a pre-trained weights file from Baruah et al. (2023) to make coreference predictions in our dataset. The weights were initially trained on Onto Notes and fine tuned to six individual scripts using 48 GB A40 NVIDIA GPUs. Each entity represents a character, and we maximize the count of the number of pronouns associated with that character to label whether it is a man, a woman, or neither.

Next, we apply a set of heuristics to determine criteria Two (are there two women that have a conversation together?) and Three (is the conversation about a topic other than a man?). Based on the character-tags, their dialogue, and the sequential rotation of who is talking and who is being spoken to, we parse conversations from each scene and identify those that are between coreference-identified character entities that are women. In their dialogue, we find instances of coreference-identified character entities that are men and, if so, label a failure to pass the final stage of the Bechdel Test. Scene scores are counted, and if at least one scene is scored to pass the third level of the Bechdel criteria, we assign a passing score prediction.

Model	Loss	Accuracy	F1
Baseline Model	0.692	0.558	0.558
BC	0.681	0.488	0.602
LC	0.683	0.428	0.578
MRPC	0.681	0.530	0.578
CoRef	-	0.701	0.692

Table 1: Results

4 Results and Discussions

Table 1 shows the *Loss*, *Accuracy*, and *F1* scores for the baseline model and the three BERT pretrained models (bert-base-cased, BC; bert-large-cased, LC; and bert-base-cased-finetuned-mrpc, MRPC) as well as the *Accuracy* and *F1* scores for the Coreference Model application. The standard metric for comparing performance on large datasets is either the *accuracy* (%) as averaged over five runs or the average *micro-F1* score (Jaiswal et al., 2023) as long as the classes are well balanced. Of BERT models, the BC pre-trained model showed the best improvement with an increase in *F1* score of 0.044 above the baseline. The LC and MRPC models do show improvement from the baseline but not a significant amount.

The coreference approach demonstrates the largest improvement in performance compared to the baseline. By combining coreference resolution with scoring heuristics, this method effectively targets the more nuanced details of criteria 2 and 3 on a scene-by-scene basis. This enables a more precise examination of scene interactions and character relationships, while accommodating the document size of long screenplays.

5 Conclusion

Our results demonstrate that the application of modern NLP approaches is a viable option for automating the Bechdel Test. Combining coreference identification with a BERT model presents an exciting avenue for future research. However, we acknowledge key limitations in our dataset that influenced our approach and outcomes: first, the dataset was limited in the number of labeled data points, since Bechdel

scores were assigned to an entire script rather than individual scenes this allowed for less flexibility in model training; second, the individual scripts were large in size which required applying modern techniques like chunking and coreference resolution in novel ways. These challenges highlight areas of improvement for dataset design and methodological refinement.

Future tasks include expanding the dataset by scraping the internet for movie scripts and connecting associated Bechdel scores from bechdeltest.com and its extensive library. Additionally, acquiring more resources to allow the application of larger pretrained models, such as Longformer (Beltagy et al., 2020), could handle long documents more effectively. We also hypothesize that misclassifications in the coreference resolution model stemmed from instances where the applied heuristics inaccurately identified conversation spans. We experimented conversation span detection with various NLP approaches such as QA classification with T5 models, directed graph network predictions, and fine-tuning transformer models on the Cornell Movie Corpus. Although time and resource constraints limited our ability to fully explore these methods, they remain promising avenues for enhancing coreference resolution in future work.

Automating the Bechdel Test has significant implications for future research. A notable departure from the work of Agarwal et al. (2015) is that our approach requires no additional data points beyond the screenplays themselves, making them scalable and practical for broader applications. This work could enhance the accuracy and efficiency of human scorers, enriching valuable insights to the ongoing discussion of representation of women in media.

It is important to disclaim that the Bechdel Test itself measures the representation of women directly: not all movies that pass necessarily feature prominent female characters or female plotlines. The Bechdel Test should not be acknowledged as the standard of representation for the role of women, but instead a helpful indicator when evaluating singular movies.

Finally, as noted in Jentzsch et al. (2020), language models, specifically BERT, have been shown to produce biased output and responsible usage should be applied. Careful consideration should be taken when applying a machine learning model and interpreting the results.

While the automation of the Bechdel test would be a great achievement in the NLP space, it should not take away from the social responsibility that the output of this research requires. Not only should a conscious effort be made to produce real social changes, but it should be understood that this does not mean further research into empowering the female voice is not required.

6 References

Apoorv Agarwal, Jiehan Zheng, Shruti Kamath, Sriramkumar Balasubramanian, and Shirin Ann Dey. 2015. Key Female Characters in Film Have More to Talk About Besides Men: Automating the Bechdel Test. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 830–840, Denver, Colorado. Association for Computational Linguistics.

Scheiner-Fisher, C., & Russell, W. B. (2012). Using Historical Films to Promote Gender Equity in the History Curriculum. *The Social Studies*, *103*(6), 221–225. https://doi.org/10.1080/00377996.2011.616239

Garcia, D., Weber, I., & Garimella, V. (2014). Gender Asymmetries in Reality and Fiction: The Bechdel Test of Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 131-140. https://doi.org/10.1609/icwsm.v8i1.14522

Umairican Umi, 2014, The-Bechdel-Test, GitHub, Umairican/The-Bechdel-Test: This is an NLP project where I first attempt to see if I can create a binary classification model that can tell if a screenplay passes the Bechdel Test. Then I attempt to train a model to produce scenes that pass The Bechdel Test using LSTM Neural Networks

Hickey, Walt. 2014. The Dollar-And-Cents Case Against Hollywood's Exclusion of Women. FiveThirtyEight.

Appel, M., & Gnambs, T. 2023. Women in fiction: Bechdel-Wallace Test results for the highest-grossing movies of the last four decades. Psychology of Popular Media, 12(4), 499–504. https://doi.org/10.1037/ppm0000436

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *arXiv (Cornell University)*, *30*, 5998–6008. https://arxiv.org/pdf/1706.03762

Dobrovolskii, V. (2021). Word-Level Coreference resolution. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2021.emnlp-main.605yle "A, B, C".

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://arxiv.org/abs/1907.11692

Rattan, A. (2024, March 27). *Tackling the underrepresentation of women in media*. Harvard Business Review. https://hbr.org/2019/06/tackling-the-underrepresentation-of-women-in-media

Baruah, S., Narayanan, S., & University of Southern California. (2023). Character coreference resolution in movie screenplays. In Association for Computational Linguistics, *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 10300–10313). https://aclanthology.org/2023.findings-acl.654.pdf

Apoorv Agarwal, Sriramkumar Balasubramanian, Jiehan Zheng, and Sarthak Dash. 2014b. Parsing screenplays for extracting social networks from movies. EACL CLFL 2014, pages 50–58.

Beltagy, I., Peters, M. E., & Cohan, A. (2020, April 10). *Longformer: The Long-Document Transformer*. arXiv.org. https://arxiv.org/abs/2004.05150

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019, May 14). *How to Fine-Tune BERT for text classification?* arXiv.org. https://arxiv.org/abs/1905.05583

Jaiswal, A., & Milios, E. (2023, October 31). *Breaking the Token Barrier: Chunking and Convolution for Efficient Long Text Classification with BERT*. arXiv.org. https://arxiv.org/abs/2310.20558

Jentzsch, Sophie and Turan, Cigdem. 2022. Gender Bias in BERT - Measuring and Analysing Biases through Sentiment Rating in a Realistic Downstream Classification Task. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 184–199, Seattle, Washington. Association for Computational Linguistics.