Automated Detection of Allusions in Toni Morrison's Novels

Felix Oke
Information Sciences Phd Student
University of Illinois at Urbana-Champaign

Introduction and Motivation

Allusion as a form of reference has been studied as an important interpretive device in literature. It's often challenging to systematically identify and analyze allusions in large scale due to its forms of representation without relying on computation methods (Bamman & Crane, 2011). While computational approaches have been successfully applied to classical texts (Coffee et al., 2013; Forstall et al., 2015), African American literature remains understudied in this domain. This project aims to develop text mining methods for automated detection of biblical allusions in Toni Morrison's two key novels. These two novels represent different periods and themes in Morrison's work: *Song of Solomon* (1977), with its explicit biblical names and Old Testament themes, and *Beloved* (1987), featuring more subtle New Testament allusions. Using a dataset from HathiTrust Digital Library, we propose to combine rule-based techniques with information retrieval approaches and transformer-based NLP models to identify biblical allusions with higher accuracy. With human annotation in the loop, we would assess model performance which will serve as a foundation resource in allusion detection in African American literature. This research contributes to studies on intertextuality more broadly (Kristeva, 1980) and to interpretation of literary texts enhanced through computations specifically (Manjavacas et al., 2019).

Questions:

1. What types and frequencies of biblical allusions appear in *Beloved* and *Song of Solomon*?
2. How can we systematically annotate biblical allusions to create reliable training data for classification models?
3. How do explicit vs. implicit allusions distribute across the selected texts?

Related Work

Computational allusion detection has evolved from simple string matching to sophisticated machine learning approaches (Manjavacas et al., 2019). Early work by Bamman and Crane (2011) established foundational principles for textual allusion logic, while Coffee et al. (2013) demonstrated machine learning applications to classical literature. Recent advances include enhanced n-gram matching (Forstall et al., 2015) and transformation-aware text reuse detection (Moritz et al., 2016). However, existing approaches primarily target classical texts with well-documented intertextual relationships. Modern literary works, particularly those by marginalized authors, present different challenges including copyright restrictions, varied allusive strategies, and limited computational resources. Our work extends these methodologies to contemporary African American literature while addressing practical constraints of working with copyrighted materials.

Methodology

Since allusion might be explicit or implicit, we developed a multi-phase methodology combining rule-based detection, named entity recognition, and information retrieval techniques, supported by a custom web-based annotation tool [morrison-biblical-annotation-tool.html] for ground truth creation.

Dataset:

We obtained Morrison's two novels from the collection of eight novels as pilot study through HathiTrust Digital Library's Extended Features API (Downie et al., 2014), which provides computational access while respecting copyright restrictions. The texts include (with publication year and HathiTrust volume ID): **The Bluest Eye** (1970) – *HathiTrust ID: uc1.32106018657251*, **Sula** (1973) – *HathiTrust ID: uc1.32106019072633*, **Song of Solomon** (1977) – *HathiTrust ID: mdp.39015032749130*, **Tar Baby** (1981) – *HathiTrust ID: uc1.32106005767956*, **Beloved** (1987) – *HathiTrust ID: mdp.49015003142743*, **Jazz** (1992) – *HathiTrust ID: ien.35556029664190*, **Paradise** (1998) – *HathiTrust ID: mdp.39015066087613*, **A Mercy** (2008) – *HathiTrust ID: mdp.39076002787351*. Morrison's works draw from the Bible either directly or indirectly with references to other features in the Bible. For instance, *Beloved* (1987) opens with an epigram from Romans 9:25. Another instance is *Song of Solomon* (1977) use of character names such as Hagar, Pilate and First Corinthians. For this project, biblical references will be drawn from public-domain Bible texts (KJV). This will be used as reference text or allusion source to identify biblical allusions in the datasets. Each verse or chapter of the Bible will serve as a candidate 'source' for allusions in the novels. The outcome of this task showing all detected biblical allusions will be regarded as part of the TM - Allusion dataset comprising all forms of allusive texts and their sources.

Our biblical reference corpus consists of the *King James Version Bible*, segmented into verses and indexed with book, chapter, and verse identifiers. Biblical allusions are broadly classified into main types which are explicit and implicit. But for the purpose of this study, we classified six types of biblical allusions (direct quote, paraphrase, thematic reference, character reference, structural echo, and no biblical reference) and six functional categories: characterization, thematic development, narrative structure, cultural commentary, ironic contrast, and spiritual dimension.

Data Preprocessing:

This project proposes the following text mining techniques on the novels such as data cleaning and tokenization, segmentation, normalization (lemmatization), and annotation. For data cleaning, we will ensure all non-textual elements like page numbers are removed from the novels and check if they are properly OCRed before performing a tokenization of the novels.. During the segmentation stage, we will divide the Bible into verses or passages and index them. Even though we will normalize (lowercasing) the data, we will maintain case and punctuation where necessary because capitalization for instance will help us identify proper nouns that are biblical names. One of the approaches we intend to use in this case is to keep two versions of the data. One data that is normalized and the other not normalized for different tasks. Lastly, we intend to undertake annotations by manually labelling instances of biblical allusion on small sample novels. The goal is to mark instances of biblical allusions and their corresponding verses or passages they refer to.

Computational Detection Pipeline

For this study, we propose the following computational pipeline in detecting biblical allusions. This is in three phases. The first phase is rule-based detection. We aim at exacting string matching for direct quotations, recognizing biblical names using enhanced spaCy NER with custom biblical dictionaries, matching patterns for common biblical phrases and citation formats, and using keyword-based detection for thematic concepts. The second phase deals with the information retrieval methods. By this, we intend to carry out an n-gram overlap analysis (3-5 grams) between novel segments and biblical verses, use TF-IDF similarity scoring for lexical overlap detection, fuzzy string matching for paraphrases and variations, and semantic similarity

using pre-trained word embeddings. Lastly in the third phase, we combine different methods (ensemble classification) of classification by using weighted combination of detection methods, confidence scoring based on multiple evidence sources, and threshold optimization using annotated ground truth.

Evaluation Plan

Evaluating an allusion detection system is complex and challenging at the same time. One of the reasons is the interpretative nature of what counts as allusive in literature and what does not. As a result, we will carry out the following metrics:

Error Analysis

We will conduct systematic error analysis by:

- False Positive Analysis: Examining why non-allusions were detected (common biblical vocabulary, coincidental matches)
- False Negative Analysis: Understanding missed allusions (implicit references, transformed language)
- Method Comparison: Analyzing which approaches perform best for different allusion types
- Iterative Improvement: Using error patterns to refine detection algorithms

Performance Metrics

- Precision, recall, F1-score for each method and allusion type
- Inter-annotator agreement scores
- Temporal analysis of detection accuracy across Morrison's career

Comparative analysis of baseline vs. advanced methods
As a result, we will calculate precision, recall, and F1-score on all allusive passages.

Bibliography

- Bamman, David, and Gregory Crane. "The Logic and Discovery of Textual Allusion." Proceedings of the 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, 2011, pp. 1-9.
- Coffee, Neil, et al. "Modelling the Interpretation of Literary Allusion with Machine Learning Techniques." Digital Scholarship in the Humanities, vol. 28, no. 4, 2013, pp. 692-712.
- Downie, J. Stephen, et al. "The HathiTrust Research Center: Using Large-Scale Analytics to Support Digital Humanities Scholarship." Digital Humanities 2014, 2014.
- Forstall, Christopher, et al. "Modeling the Scholars: Detecting Intertextuality through Enhanced Word-Level N-Gram Matching." Digital Scholarship in the Humanities, vol. 30, no. 4, 2015, pp. 503-515.
- Kristeva, Julia. Desire in Language: A Semiotic Approach to Literature and Art. Columbia University Press, 1980.
- Manjavacas, Enrique, Brian Long, and Mike Kestemont. On the Feasibility of Automated Detection of Allusive Text Reuse. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 104–114, Minneapolis, USA.

Catalog Record: Sula | HathiTrust Digital Library
https://catalog.hathitrust.org/Record/004737103
Catalog Record: Song of Solomon | HathiTrust Digital Library
https://catalog.hathitrust.org/Record/002473454
Catalog Record: Tar baby | HathiTrust Digital Library
https://catalog.hathitrust.org/Record/000169417
Catalog Record: Beloved : a novel | HathiTrust Digital Library
https://catalog.hathitrust.org/Record/000870183
Catalog Record: Jazz | HathiTrust Digital Library
https://catalog.hathitrust.org/Record/008325775
Catalog Record: Paradise | HathiTrust Digital Library
https://catalog.hathitrust.org/Record/003959081
Catalog Record: A mercy | HathiTrust Digital Library
https://catalog.hathitrust.org/Record/005898814