Wikipedia Play Pages Generation

Motivation

- Currently, users interested in plays have to search and collect information from multiple sites. To the best of our knowledge, there has been no effort to create a single information source in the domain of plays.
- While a few famous plays have existing Wikipedia articles (only 10% in our dataset), we believe that creating entries for other plays will help improve the completeness of Wikipedia with respect to plays.
- By putting relevant information in a concise form with pointers to the most important sources of information for further in-detail reading, we will help improve the availability of information about plays.

Technical Challenges

- Create a focused crawler that can perform focused crawling of playscripts with reasonable accuracy.
- Extracting relevant information from the playscripts – Author, play title, characters and publishing information.
- Retrieving other relevant information on the play plot, critical reviews, audience reception, etc. from various web sources.
- Creating Wikipedia articles automatically by collecting and selecting relevant information and summarizing them to create a human-readable Wikipedia entry of high quality. Unlike Sauper and Barzilay’s Wikipedia generation task (Sauper et al. 2009), we use a standard web template because we generate articles on a predefined set of topics on plays.

Contribution

- We are the first to address the problem of playscript classification and summarization. This problem will be made publicly available to the research community.
- Classifiers trained with our set of structural features outperform the classifiers trained on “bag of words”.
- We help improve coverage of play articles in Wikipedia. We automatically extract information related to plays and generate Wikipedia pages and provide references as pointers for detailed study.

A Sample Playscript

**THE GAME**

A morality play is one act

by Louise Dayton

The following morality play is a representation of The Presentation Play, New York City. By 1914, New York had a large Jewish community that was shaped by its unique cultural history.

CHARACTERS

LIFE

DEATH

YOUTH

DEATH

LIFE

(Counting abstractly) Fifty thousand, fifty-eleven, sixty-nine, sixty-nine, sixty-nine at the next speech.)

LIFE: Come, come, Life. Forget your troubles. It’s no fun playing with a dull partner. I had hoped for a good game tonight: although there’re little in it for me just a couple of dutch.

Dataset

- Training set: DS1 - 358 playscripts documents (positive class) and 265 other documents (negative class)
- Test set: DS2 - 40 playscripts (positive class) and 32 other documents (negative class)

Wikipedia Play Pages Generation

**Feature Design - Playscript Classification**

- **Feature Name**: Gamefreq, Authorfreq
- **Feature Description**: Frequency of keyword categories
- **Gamefreq**: Number of lines with capitalized words
- **Authorfreq**: Ratio of count of lines in a capital to count of all lines
- **Drugfreq**: Ratio of number of drugs to the number of lines
- **Bicycles**: Number of bicycles

**Wikipedia Page Generation**

- **Author and Text**
- **Tags**
- **Page Generation**
- **Trumpet**
- **Date**

Experiments and Results

- Classification: Structural Features vs “Bag of Words” Features – 10 fold Cross Validation (CV) experiments on five different classifiers using Weka.

Conclusions and Future Work

- Two out of ten play articles submitted have been accepted into Wikipedia mainspace.
- In future, improving article sections by using better abstractive summarization and natural language generation techniques.
- We are also planning to use speech to text recognition systems to generate playscripts from play performances of plays for which playscripts are not freely available online.