NLP for Short Fiction

Pooja Hiranandani
ph1130@nyu.edu

Project description

For my project, I have built a system that attempts to identify some elements of a typical story, namely:

Point of view
The term point of view refers to who is telling a story, or who is narrating it. The narration of a story or novel can be told in three main ways: first person, second person, and third person.

Protagonist
A protagonist is the principal character of a literary work.

Theme
Theme is defined as a main idea or an underlying meaning of a literary work, which may be stated directly or indirectly.

Style
Style in literature is the literary element that describes the ways that the author uses words — the author's word choice, sentence structure, figurative language, and sentence arrangement all work together to establish mood, images, and meaning in the text.

Methods used to achieve this include:

- Tagging the story text with parts of speech and NER tags, finding patterns within it and using those patterns to build rule based classifiers
- Vectorising the stories and training a classifier with the data

Motivation for the project

While I have come across several studies that have extracted information from longer fiction such as novels and folk tales, I have not come across a similar effort for short fiction. Since fiction is an area of interest of mine, I decided to see if I could use the techniques learnt in class in its service.

Description of dataset

The dataset is comprised of 118 New Yorker short fiction stories found at: https://www.newyorker.com/magazine/fiction. I created a script that scraped the New Yorker website and saved relevant information to a csv file. The fields of the dataset are:
Title of story (scraped)

Story text (scraped)
Author (scraped)
Issue date (scraped)
URL (scraped)
Tag1 (scraped)
Tag2 (scraped)
Tag3 (scraped)
Tag4 (scraped)
Tag5 (scraped)
Tag6 (scraped)
Point of View (human annotated)
Protagonist (human annotated)
Theme (human annotated)

Basis for choosing the stories

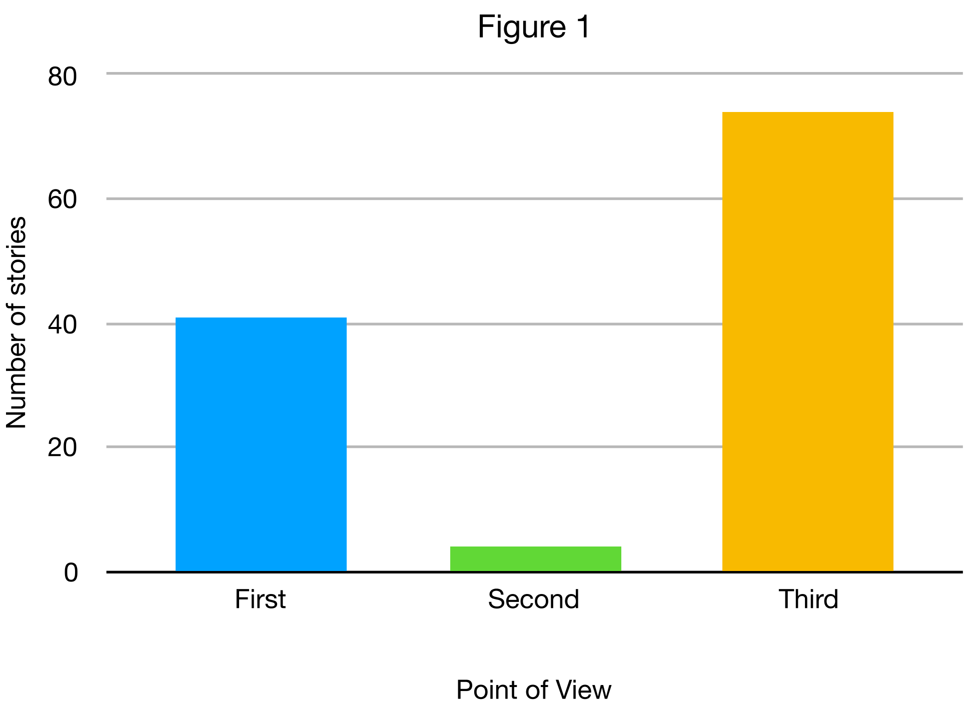
Each story in the New Yorker short fiction stories section is tagged with 1-6 one-word or two-word tags that denote the theme of the story. I analysed a large sample of the stories and chose four broad themes that had adequate representation in the repertoire. The stories were chosen from most recent onwards and the goal was to get a near-equal distribution of the four themes. In cases where a story was tagged with more than one of the four central themes, I read the story and categorised it as per my subjective judgement of where it most belonged.

60% of the dataset was my training set of which 15% served as a dev set and was subsequently merged into the training set while 40% served as my test set.

Methods Used, Results and Analysis

POINT OF VIEW IDENTIFICATION

In the dataset, the distribution of the point of view is as below:



In order to identify the point of view of a story, I ran the story text through the NLTK parts of speech tagger and kept a count of the various personal pronoun tags that are followed by a verb tag. I included the verb tag because I only wanted to keep track of pronouns that are the subject of a sentence. Based on the personal pronoun that had the highest count, I labelled the stories with a point of view. For instance, if the count of the pronoun ‘I’ followed by a verb tag is greater than any other personal pronoun, the point of view must be of first person. This simple technique worked surprisingly well as the results below indicate. Results are for both training and test sets as no model was trained on the training data, only patterns sought and rules made based on these patterns.

Table 1

	Precision	**Recall**	**F1**
Training set (72)	96.61	97.89	97.23
Test set (46)	96.67	94.12	95.02
Whole set (118)	96.54	96.29	96.38

Other techniques tried that detracted from the accuracy:
Excluding personal pronouns found inside dialogue.

Improvements to be made:
A way to identify multiple points of view in the same story.

PROTAGONIST IDENTIFICATION

In order to identify the protagonist of a story, I ran the story text through the Stanford NER parser and collected a list of names identified as ‘Person’. Using this list of names, I went through the text again and kept a count of the number of the times the name, when followed by a verb tag, appears in the story. After this list is gathered, I took the first two names with the best counts and if they were very close in number, I checked which of the two names appeared both in the first and the last paragraph of the story and chose the name that met this condition. If both appeared, I chose the name with the greatest count. If the difference between the counts of the first two names was large, I just took the name of the person with the largest number of mentions in the text. The above process was only done for stories predicted as having a third person point of view. For first and second person point of view stories, I classified the protagonist as unnamed and reader respectively since the probability of that being the case is quite high, as I learnt from the training set. Results are for both training and test sets as no model was trained on the training data, only patterns sought and rules made based on these patterns.

Table 2

	Precision	**Recall**	**F1**
Training set (72)	66.99	68.06	67.51
Test set (46)	84.55	80.43	82.39
Whole set (118)	74.44	72.88	73.52

Other techniques tried that detracted from the accuracy:

Checking if the gender of assumed protagonist matches the gender of the third person subject pronoun with the maximum count

Maximum entropy model to classify each name as protagonist or not protagonist based on features such as count of mentions, point of view, personal pronoun counts, etc.

Improvements to be made:

- First person point of view protagonist identification when name of narrator is mentioned in the story.
- Protagonist identification when point of view is first person but the protagonist is someone other than the narrator.
- Correct third person point of view protagonist identification when protagonist is unnamed but other characters are named.
- Incorporating gender of the assumed protagonist into the identification exercise.
- Building a model entropy model when more data is available.
- A way to identify more than one protagonist

THEME IDENTIFICATION

After analysing a sample set of New Yorker short fiction stories, I identified four broad themes that had sufficient coverage in the set. The four themes are Adultery, Children, Death and Immigrants. Each theme is represented by ~30 (+/-3) stories. In order to identify the theme of the stories, I converted each story in the dataset into trigram, bigram and unigram features using Tf-IDF. I then trained a Multilayer Perceptron Classifier to assign labels. The results, while better than a random assignment of labels, are not entirely satisfactory. Results are only for test set as the model was trained on the training set and achieved a near perfect accuracy score on the training set.

Table 3

	Precision	**Recall**	**F1**
Test set (46)	57.35	54.75	54.27

Possible reasons for poor performance:

- Paucity of data might be the biggest reason. Gathering more data should lead to an improvement in performance.
- Overlap between themes. For instance, many stories of adultery also include children, same too for stories of immigrants. I confirmed this suspicion by studying the confidence level probabilities assigned by the MLP classifier for each theme for each story. In the case of a wrong label assignment, the gold label usually has the second highest confidence level indicating that the model has learnt something about what themes are represented in the text except that there might be another theme contained in the story that is overpowering the classifier. Again, gathering more data will likely alleviate this issue.

Other techniques tried that did not improve performance:

Count based Bag of Words

Doc2Vec

Convolutional Neural Networks

SVM Classifier
Multinomial Naive Bayes Classifier
Multilabel classification

STYLE ANALYSIS

I did a simplistic analysis of the writing style of the stories. Style here is denoted by four factors:

- Difficulty of the text indicated by the number of large words in the text
- Descriptiveness of the text indicated by the number of adjectives in the text
- Extent of dialogue in the text indicated by the number of double quotation marks in the text
- Wordiness of the text indicated by the number of words in the text

All the stories are ranked using percentile calculations on a scale of 0-9 with 0 being the lowest rating on a factor and 9 being highest.

This is an exercise in description so there is no performance to measure.

Some interesting insights from the analysis:

- The most wordy stories are on the theme of immigrants while the least wordy stories are on the theme of death.
- Stories about immigrants are among the highest in difficulty level.
- Difficulty of a text and its descriptiveness are highly positively correlated.
- First person point of view stories are generally longer and more descriptive than third person point of view stories.
- Third person point of view stories generally have more dialogue in them than first person stories.

Details of files included in the submission

Data files:
stories-death.txt - scraped data for the theme ‘death’, tab separated
stories-adultery.txt - scraped data for the theme ‘adultery’, tab separated
stories-children.txt - scraped data for the theme ‘children’, tab separated
stories-immigrants.txt - scraped data for the theme ‘immigrants’, tab separated
training-set.npy - training set dataset
test-set.npy - test set dataset

Program files:
ny.py - Scraping script
storyteller.py - Program to read in the data files, divide data into training and test dictionaries and save them as .npy files
storyteller1.py - Program to identify voice, protagonist, theme and writing style
Stanford NER tagger

Other files:
whole_dataset.csv - A more reader friendly version of the dataset
dataset_style.csv - Style rankings for all stories
project_description_analysis.pdf - This file

Details of implementation

Requirements:
python 3.6.3
numpy 1.14.2
tabulate 0.8.2
pandas 0.20.3
nltk 3.2.4
scikit-learn 0.19.1
beautifulsoup4 4.6.0
urllib3 1.22
requests 2.18.4

All data and program files should be in the same folder except the Stanford NER tagger which should be in a folder called ‘apps’.

To run the program please type into a terminal:

```
python ./storyteller1.py
```

The expected output is:
F1, precision and recall scores for training, test and whole sets where applicable
Style rankings of every story

References

https://thewritepractice.com/elements-of-fiction/

http://examples.yourdictionary.com/examples-of-point-of-view.html

https://literarydevices.net/theme/

http://www.readwritethink.org/files/resources/lesson_images/lesson209/definition_style.pdf

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

Fernandez et al. 2015. Extracting Social Network from Literature to Predict Antagonist and Protagonist, retrieved from https://nlp.stanford.edu/courses/cs224n/2015/reports/14.pdf

Groza, Adrian, and Lidia Corde. “Information Retrieval in Falktales Using Natural Language Processing.” 2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), 2015, retrieved from https://arxiv.org/pdf/1511.03012.pdf.

Luyckx, K., Daelemans, W., Vanhoutte, E.: Stylogenetics: Clustering based stylistic analysis of literary corpora. In: Workshop Toward Computational Models of Literary Analysis (2006), retrieved from https://www.clips.uantwerpen.be/~kim/Papers/LDV06.pdf

Vala, Hardik & Jurgens, David & Piper, Andrew & Ruths, Derek. (2015). Mr. Bennet, his coachman, and the Archbishop walk into a bar but only one of them gets recognized: On The Difficulty of Detecting Characters in Literary Texts. 10.18653/v1/D15-1088, retrieved from aclweb.org/anthology/D/D15/D15-1088.pdf

Reagan, et al. “The Emotional Arcs of Stories Are Dominated by Six Basic Shapes.” [1606.07772] The Emotional Arcs of Stories Are Dominated by Six Basic Shapes, 26 Sept. 2016, retrieved from arxiv.org/abs/1606.07772.