CATS: CREATIVE AUTOMATIC TITLES FROM SYNOPSIS

Jonas Molz (i6142067), Albert Negura (i6145864)

1 INTRODUCTION

The aim for this project is to create a system that will generate adequate titles for movies given a synopsis. We conjecture that given additional knowledge about the genre of a movie would increase performance of such a system, since the naming conventions between genres may differ strongly (i.e. it might be that the title of romantic comedies likely contain the word "love" or that a movie in the genre of superhero-fiction is more likely to be named after its protagonist). However, movies might belong to multiple genres.

Our approach with CATS (Creative Automatic Title from Synopsis generator) is hence three-fold: (1) Implement a system which can classify a movie to a particular set of genres, (2) Implement a system which extracts the most important keywords for a particular body of text and (3) Generate a title based on the selected tags and genres, further using the body of text to enhance the results.

2 LITERATURE REVIEW

A substantial amount of research went into automatic text and caption generation. In particular, to news [1], blogs or article titles [2] and, more recently, headline generation using RNNs [3]. Headline generation is an interesting topic, as its uses are not immediately obvious. Besides helping authors deal with their writer's block, these can also be used for spam filtering [4] or to deal with partisanism in news articles, movies or books [5]. Furthermore, headline generation is a method to explore automatic summarization of text, which can be used to extract relevant information from speech or text while simultaneously discarding irrelevant information.

Most methods dealing with headline generation utilize information retrieval and text mining to extract keywords which are then used to generate a headline [2]. The two methods we are using are Lopyrev's method [6] and Kar's method [7]. The latter is used to extract a specific set of tags or keywords, while the former is used to generate headlines based on the introductory paragraph of news articles.

3 DATA

Two datasets were found which contain useful data, both available through Kaggle. One contains an MPST dataset with 14828 movies,

their titles, synopses and custom tags , while the other contains approximately 35000 movie titles, their (partially filled) genres and their summaries as taken from Wikipedia [8]. Doing some cleaning on the Wikipedia dataset (removing duplicates and movies with unknown genres) results in about 28803 datapoints. The MPST dataset synopses have 1200 words on average, while the Wikipedia summaries have about 500 words on average. The Wikipedia summaries also contain names very commonly as seen in figure 4. Finally, there is a large disparity between the frequency of articles in the Wikipedia summaries and other words. Most articles have total frequencies in the thousands, while most non-articles and non-adverbs (including names) appear at most a few hundred times.

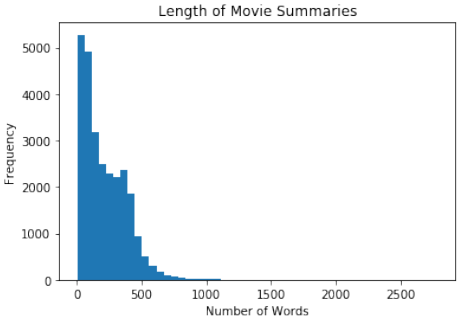

Figure 1: The distribution of the length of the summaries in the Wikipedia dataset.

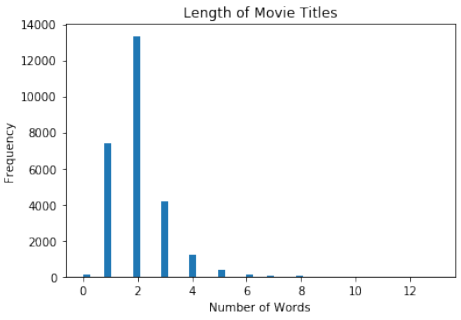

Figure 2: The distribution of the length of movie titles in the Wikipedia dataset.

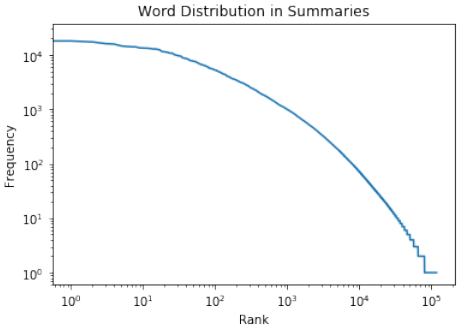

Figure 3: The distribution of frequencies of all words in the Wikipedia database, ranked inversely based on their frequency. There are a few words which appear a few hundred times more often than most words in the corresponding summaries.

Figure 4: Wordcloud showing the variety of words in the summary of a popular movie included in the Wikipedia dataset. Notice that most prominent words in these summaries are (usually) names.

4 Model

CATS is a multi-step system which is intended to generate creative titles or headlines given a synopsis. First, the system utilizes a Mutinomial Naive Bayes model trained on the cleaned data to produce a series of tags and genres corresponding to a specific synopsis. The model can then be used to predict the rel-

evant genres and tags for new synopses. Two different Multinomial Naive Bayesian models were trained on each of the datasets, one for the tags and one for the genres, and function together to produce tag/genre combinations.

The tag generation can also be done with a more advanced system described by Kar et al., which uses affective features (folksonomy) in the body of the text in order to produce genre-specific tags for the movie. The model used is available via their paper, but is very resource intensive to train.

Title generation is done using the method described by Lopyrev. The method involves the use of a Recurrent Neural Network architecture to generate news article headlines from news text bodies. It functions by first creating a distributed representation of the text via an encoder and subsequently generating the title using a decoder.

As shown in figure 5, the goal of this system is to utilize Lopyrev's network to produce the titles after it is fed with a series of tags (keywords) and genres which are produced by the text itself. A system utilizing only Lopyerv's model is also implemented in order to test the performance of the proposed model.

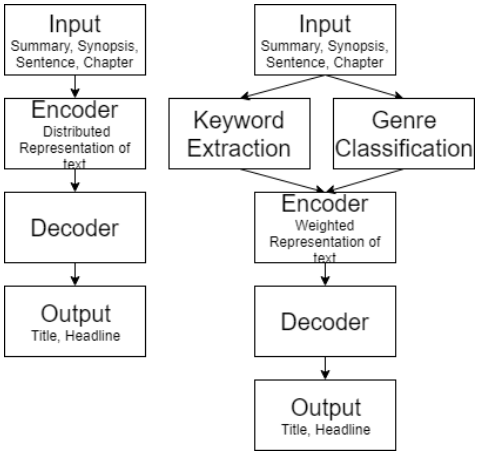

Figure 5: A flowchart depicting the general idea of the model used by Lopyrev (left) and of our proposed model (right).

5 RESULTS

The genre classification achieves a high performance in genres inversely proportional to their frequency in the dataset. There is a high class imbalance between the top 10 genres and the remaining 445 genres in the dataset - the genres were selected such that they appear at least 2 times in the whole dataset. Due

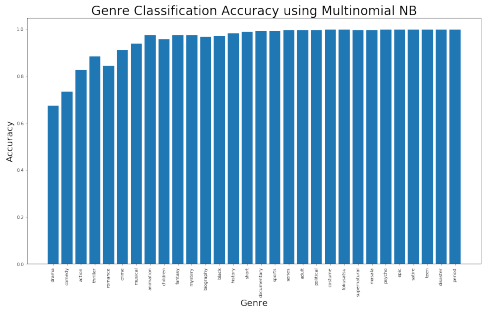

Figure 6: The classification accuracy for the Multinomial Naive Bayes genre classification for movies in the Wikipedia dataset.

to the large complexity of the task, even the simpler (Lopyrev) model could not be trained effectively on the machines used. As such, the model was trained for five iterations of 3 epochs each and a batch size of 32, all parameter values halved compared to Lopyrev's network. The network was trained and implemented using Keras [9]. The loss function can be seen in Figure 6. Note the validation score plateau, which could be due to the low

number of epochs dedicated to the training of the model. Furthermore, the dataset contains a variety of movies in different language (but their synopses are still in English). As such, the network learns to produce names in different languages when it's given the synopsis in English. Other than going through the entire dataset one by one and deleting all foreign entries, not much could have been done to avoid this.

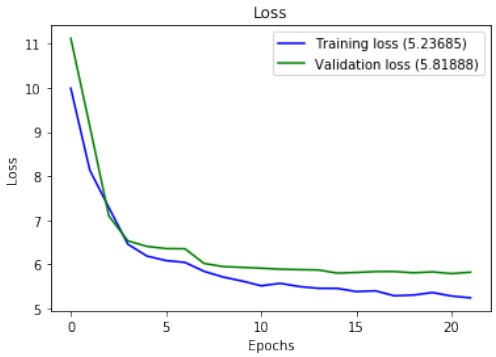

Figure 7: Training and validation loss for Lopyrev's model.

News headline generation is significantly easier than movie title generation. First of all, real news headlines utilize information (or words) directly available within the story, while movie titles can be obscure and have nothing to do with the plot. The network is set to output any titles it is confident of, but fails to produce results in many cases. These results can be seen in table 1.

Actual Title	Title 1	Title 2	Synopsis
Little Shop of Horrors	Farb	Stoolie	A three-girl "Greek chorus" Crystal, Ronnette, and Chiffon introduce the movie, warning the audience that some horror is coming their way...
Tarantula	Cross Man		A severely deformed man stumbles through the Arizona desert, falls and dies. Dr. Matt Hastings, a doctor in the nearby small...
Endless Love	Dangerous Man	Underworld	In suburban Chicago, teenagers Jade Butterfield and David Axelrod fall in love after they are introduced by Jade\'s brother Keith. Jade\'s family is known...
Nowhere to Run	Keeps Zoo Enchanted	Walk Two	In rural Texas, 1960 an age of good times and innocence, when growing up was supposed to be easy six high school seniors know the terrible secret...
The Things	Fly Planet	Girls Road	In Antarctica, in 1982, a Norwegian helicopter pursues a sled dog to an American research station . The Americans witness the Norwegian pilot accidentally blow up...

Table 1: A list of movies and the alternative titles generated. Note that all titles (including the actual titles) have very little in common with the synopsis.

6 Conclusion

Since the networks both required much more powerful computers than the ones available (Lopyrev's network failed to be trained on a personal machine and using Google Colab, crashing after the training started). As a result, the CATS system could not be fully implemented. The produced network can produce movies, sometimes with words reminiscent of the plot of the movie, but generally with very little direct correlation (no words from the text of the synopsis were reused).

In the future, besides the full implementation of the desired system, a possible improvement would be to utilize semantic and syntactic rules from the English language to further improve on the meaning of the title. Furthermore, additional datasets without foreign movies could also be used to help produce a more robust network.

REFERENCES

[1] Rong Jin and Alexander G Hauptmann. "Automatic title generation for spoken broadcast news". In: *Proceedings of the first international conference on Human language technology research*. Association for Computational Linguistics. 2001, pp. 1–3.

[2] Songhua Xu, Shaohui Yang, and Francis Lau. "Keyword extraction and headline generation using novel word features". In: *Twenty-Fourth AAAI Conference on Artificial Intelligence*. 2010.

[3] Yuko Hayashi and Hidekazu Yanagimoto. "Headline generation with recurrent neural network". In: *New Trends in E-service and Smart Computing*. Springer, 2018, pp. 81–96.

[4] Kai Shu et al. "Deep headline generation for clickbait detection". In: *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2018, pp. 467–476.

[5] Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. "From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles". In: *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. 2017, pp. 84–89.

[6] Konstantin Lopyrev. "Generating News Headlines with Recurrent Neural Networks". In: *CoRR* abs/1512.01712 (2015). arXiv: 1512.01712. URL: http://arxiv.org/abs/1512.01712.

[7] Sudipta Kar et al. "MPST: A Corpus of Movie Plot Synopses with Tags". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari (Conference chair) et al. Miyazaki, Japan: European Language Resources Association (ELRA), Dec. 7. ISBN: 979-10-95546-00-9.

[8] *Wikipedia Movie Plots*. https://www.kaggle.com/jrobischon/wikipedia-movie-plots. Accessed: 2019-05-13.

[9] François Chollet et al. *Keras*. https://keras.io. 2015.