Predicting News Headline Popularity with Syntactic and Semantic Knowledge Using Multi-task Learning

Hardt, Daniel; Hovy, Dirk; Lamprinidis, Sotiris

Predicting News Headline Popularity with Syntactic and Semantic Knowledge Using Multi-Task Learning

Daniel Hardt
Copenhagen Business School
dh.msc@cbs.dk

Dirk Hovy
Bocconi University
dirk.hovy@unibocconi.it

Sotiris Lamprinidis
Copenhagen University
sot.lampr@gmail.com

Abstract

Newspapers need to attract readers with headlines, anticipating their readers' preferences. These preferences rely on topical, structural, and lexical factors. We model each of these factors in a multi-task GRU network to predict headline popularity. We find that pre-trained word embeddings provide significant improvements over untrained embeddings, as do the combination of two auxiliary tasks, news-section prediction and part-of-speech tagging. However, we also find that performance is very similar to that of a simple Logistic Regression model over character n-grams. Feature analysis reveals structural patterns of headline popularity, including the use of forward-looking deictic expressions and second person pronouns.

1 Introduction

The data generated from online news consumption constitutes a rich resource, which allows us to explore the relation between news content and user opinions and behaviors. In order to stay in business, newspapers need to pay attention to this information. For example, what headlines do users click on, and why? With the volume of news being consumed online today, there is great interest in addressing this problem algorithmically. We collaborate with a large Danish newspaper, who gave us access to several years' worth of headlines, and the number of clicks generated by readers.

We aggregate the viewing logs to classify headlines as popular or unpopular, and build models to predict those classifications. We use an expanded version of the dataset investigated by Hardt and Rambow (2017). That work found that bag-of-word models based on headlines did indeed have predictive value concerning viewing behavior, although models based on the article body were more accurate. As Hardt and Rambow noted, this is somewhat paradoxical: how can a model based on the article text be better at predicting clicks? After all, the choice to click on an article must be based on the headline alone – the article is only seen *after* the clicking decision is made. Hardt and Rambow speculate that "it is possible that the headline on its own gives readers a lot of semantic information which we are not capturing with our features, but which the whole article does provide. So human readers can "imagine" the article before they read it and implicitly base their behavior on their expectation." (Hardt and Rambow, 2017)

In other words, readers are able to *anticipate* the contents of an article in advance from a headline, because of the linguistic and world knowledge that they bring to bear when assessing the headline. If we can incorporate this "future" knowledge into a prediction model, we are likely to improve performance.

We test this hypothesis by defining ways to model aspects of the lexical, structural, and topical knowledge of human news readers:

- **Lexical – Word Embeddings:** we provide our models with pretrained word embeddings from large datasets. This models aspects of the rich lexical information and association that human readers bring to bear in reading a headline.

- **Structural – POS Tagging:** part of speech information is a basic component of structural linguistic knowledge, reflected in the structure of common headline templates such as "Can X do Y?" or "You will not believe what happened when X".

- **Topical – Section Prediction:** Each article is labeled with a section (sports, politics, etc). We include a task which predicts the section of a headline. This models the ability of a news reader to understand the most salient

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 659–664
Brussels, Belgium, October 31 - November 4, 2018.

and interesting topical material in a headline text.

We use a multi-task learning (MTL) setup (Caruana, 1993), which provides a natural framework to test the above hypotheses: one of the first uses of MTL was to include the outcome of future diagnostic tests into a prediction task (Caruana et al., 1996).

We explore the effect of pretrained word embeddings, and the effects of auxiliary tasks involving POS tagging and section prediction. We find that the combination of all of these factors results in substantial improvements over the baseline and the previous work, which used a single-task system. We also build logistic regression models, both for word and character n-grams. The word-based models have the advantage that the predictiveness of individual words can be examined.

While the word n-gram models have performance comparable to the baseline neural net, the character n-gram model has higher performance, competing with the best MTL result. This finding is in line with the results from Zhang et al. (2015).

Our results indicate that MTL can indeed provide the tools to implement prediction processes that involve expectations about the future. Given the successful integration of two auxiliary tasks, we see this as a promising starting point for future research. However, the performance parity with the character model underscores the fact that simple model architectures still have a place. Our findings, in line with other current work (Benton et al., 2017), shine light on the question of auxiliary task selection and their interaction, and highlight that MTL results should be rigorously tested.

A good predictive model is a powerful diagnostic tool for editors, allowing them to select proposed headlines. However, journalism is a creative production process, so detection is only part of the application. We also want to be able to give strategic advice to headline *writers*. To this end, we report an analysis of common n-gram features in the word-based logistic regression model, that provide some insights into successful headline patterns.

Contributions We explore an MTL architecture with two auxiliary tasks for headline popularity prediction. We show how aspects of lexical, structural, and topical knowledge are all relevant for headline popularity. The positive results reported here provide a fruitful basis for further development of MTL models for news data. We also analyze lexical features that are predictive of headline popularity.

Figure 1: Example of Jyllands-Posten headline as seen by audience

2 Data

News Data The present work is based on a significantly expanded and cleaned version of the dataset used by Hardt and Rambow (2017). This dataset includes Jyllands-Posten articles and logs. Jyllands-Posten is a major Danish newspaper (and became known to an international audience over the cartoon controversy). The data covers a period from July 2015 through July 2017. We removed any articles from before July 2015, when the viewing logs began, since these older articles have unreliable numbers of clicks. The resulting dataset consists of 82,532 articles and a total of 281,005,390 user views. We furthermore extracted the news section each article belongs to (sports, politics, etc.) from the URL.

We bin the articles by numbers of clicks into 2 bins, thus defining a classification task: is the article in the top 50% of clicks or not? The data is divided into 80% training, and 10% each development and test data.

Figure 1 shows the top headline on the Jyllands-Posten web site for August 27, 2018. Our data does not include information such as the position of a headline on the page, and possible associated graphical material.

Additional Data In addition to the news data from JP, we obtained a corpus of 100 million words of Danish text from the Society for Danish Language and Literature, or DSL (Jørg Asmussen, 2018). This corpus was collected from diverse

sources over a period from 1990 to 2010. The corpus has been automatically annotated for part of speech and lemmatization, and we use this for our POS tagging task. We also downloaded the Danish Wikipedia, which consists of approximately 49 million words of Danish text. We use these corpora in conjunction with the JP article texts to induce pre-trained Danish word-embeddings.

Data Statement A. CURATION RATIONALE The dataset is collected by Jyllands-Posten as part of a general strategy to understand user behavior and preferences with respect to the news content on the site.

B. LANGUAGE VARIETY The data is Danish (da-DK).

C. SPEAKER DEMOGRAPHIC The text is produced by professional journalists.

D. ANNOTATOR DEMOGRAPHIC There is no manual annotation of the text.

E. SPEECH SITUATION The texts were produced from July 2015 until July 2017; the intended audience is Danish news consumers.

F. TEXT CHARACTERISTICS The text is standard, mainstream Danish journalism.

3 Models

Our task is to predict which articles get the most user clicks, based on the headline alone. We report results using logistic regression and a neural network, using MTL.

Logistic Regression We define the following features for logistic regression models:

1. *n-chars*: sequences of n characters, with n ranging from 2 to 6 in all experiments.

2. *word unigrams*: *tfidf* scores for all word unigrams

3. *word bigrams*: *tfidf* scores for all word bigrams

GRU Neural Network While the task is classification, which could be done with a feed-forward model, we want a sequential architecture, so that we can incorporate POS tagging as an auxiliary task, adding POS output at each time step.

Based on good results in recent work (Lee and Dernoncourt, 2016), (Liu et al., 2016), we choose a Recurrent Neural Network architecture and after a series of experiments on the training and validation set, we obtained the best results using GRU (Gated Recurrent Unit) units.

Each layer k consists of two sets of units, labeled fw and bw that process the sequence forwards and backwards respectively, so that information from the whole sequence is available on every timestep t. The two directions' activations are concatenated and fed to a fully-connected softmax (for multi-class classification) or sigmoid layer (for binary classification) to get the output probability y_t^k of the task associated with layer k. So that higher level tasks can benefit, we embed the output probabilities using the fully connected label embedding LE layer, a technique used on similar scenarios (Rønning et al., 2018). The embedded label gets concatenated with the GRU output to get the activation a_t^k that gets fed in the next layer, or the final fully connected prediction layer, as presented in figure 2.

In the sequential auxiliary task, i.e. POS tagging, this is done for every timestep, while for the classification tasks the prediction is made on the final timestep.

For regularization, we apply dropout on every layer of our network.

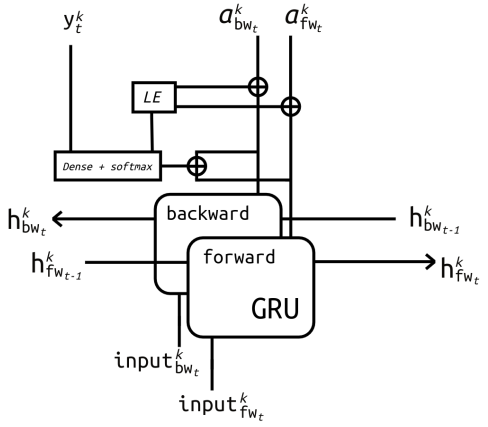

Figure 2: Representation of a single timestep t for a pair of forward-backward units on layer k where h_{t-1}^k is the previous hidden state.

Auxiliary Tasks In our setup, we use two auxiliary tasks:

1. *POS tagging*: we include POS tagging using the DSL dataset on the first recurrent layer of the GRU.

2. *Section prediction*: we include classification into one of the 227 sections of the Jyllands-

Posten website. The output for this task is based on the penultimate recurrent layer.

Hyper-parameters and Training We perform a grid search to find the best hyper-parameters for a single-task model (i.e., without any auxiliary tasks) and then keep those settings for all our experiments. We settle on a model with hidden size $H = 112$ and $N_k = 3$ layers, respectively. The dropout probability $p = 0.3$ gave best results for both models.

We train the model for 10 epochs using Adam optimizer with the default parameters, clipping the gradient updates so that their norm is not higher than 5. We train the different tasks sequentially for each epoch, with the lower level (POS tagging) first and the popularity prediction last. Additionally, we decay the learning rate by a factor of 0.9 after each epoch. While this is not common with adaptive methods such as Adam, it performed better. We stop training if the accuracy on the development set stops improving.

4 Results

Tables 1 and 2 report accuracy for logistic regression and neural classifiers. We also give the best score from Hardt and Rambow (2017) for comparison purposes (note, though, that the data sets are not identical and can therefore not be directly compared). We observe a substantial improvement over the baseline GRU when incorporating the pre-trained embeddings and both auxiliary tasks. It seems that pretrained embeddings and MTL act at least partly as regularizers, as these models trained for more epochs without overfitting. Interestingly, we observe a similar improvement over the word-based logistic regression models with a character n-gram model.

5 Analysis and Discussion

Our main focus in this paper is on MTL as a framework to explore the lexical, structural and topical knowledge involved in users' selection of headlines. However, recognizing a popular headline and giving advice on how to write one are not the same: we want to provide editors and journalists with insights as to what constructions are likely to attract more eyeballs.

One way to explore this is to examine individual words and their contribution to predictiveness. Table 3 displays the top 20 unigrams based on their coefficients in the logistic regression model. For each unigram we provide a translation (if needed) and a comment. We classify several unigrams as Deictic-reference. This follows Blom and Hansen (2015), who suggest that headline "clickbait" often relies on forward-looking expressions, such as "This", as in, e.g., "This is how you should eat an avocado". Here, "this" is a referring expression, but the reader understands that the antecedent will be found in the article body. Several of these top unigrams are names that are of specific topical interest in areas such as sports and politics. Others mention topics of more general interest (Researchers, dead, found). The second person pronoun is also on the list – in general, it was found that second person pronouns are far more predictive of popularity than first or third person pronouns. Finally, several unigrams identify sections of the newspaper of particular interest (car, weather, analysis, and satire).

6 Related Work

Prediction of news headline popularity is an increasingly important problem, as news consumption has moved online. The insights and models described here might well be applicable to related problems of interest: for example, Balakrishnan and Parekh (2014) and Jaidka et al. (2018) study the problem of predicting clicks on email subject lines.

Subramanian et al. (2018) show that a regression-based multitask approach can increase performance for the classification prediction of popularity. Their work looks at the popularity of online petitions, but the methodology applies to our subject as well, and ties in with the approaches taken in this project.

Benton et al. (2017) caution that in order to evaluate MTL results properly, we need to take the number of parameters into account. Our results to some extent support this finding, by showing that a simpler linear model can fare equally well on the task.

The choice of auxiliary tasks greatly influences the performance of MTL architectures, prompting several recent investigations into the selection process (Alonso and Plank, 2017; Bingel and Søgaard, 2017). However, it is still unclear whether these tasks serve as mere regularizers, or whether they can also impart some additional information.

model	input	accuracy
Hardt and Rambow (2017)	word unigrams	61.2
logistic regression	word unigrams	65.6
logistic regression	word bigrams	65.7
logistic regression	character 2-6grams	**67.4**

Table 1: Accuracy results for various Logistic Regression models

model	input	auxiliary tasks	epoch	accuracy
GRU 3 layers w/ 112 hidden	random embeddings	—	3	65.2
GRU 3 layers w/ 112 hidden	pretrained embeddings	—	5	66.8
GRU 3 layers w/ 112 hidden	pretrained embeddings	POS	5	65.7
GRU 3 layers w/ 112 hidden	pretrained embeddings	section	4	66.8
GRU 3 layers w/ 112 hidden	pretrained embeddings	POS+section	7	**67.4**

Table 2: Accuracy results for various GRU model implementations

Unigram	Translation	Comment
Magnussen		Name (Sports)
Trump		Name (Politics)
AGF		Name (Sports)
Test		
Her	Here	Deictic-reference
død	dead	topical
Wozniac		Name (Tech)
Trumps		Name (Politics)
Forskere	Researchers	topical
fundet	found	topical
du	you	pronoun
AGF-træner	AGF coach	Name (sports)
Se	Watch	Deictic-reference
Kevin		Name (Sports)
Islamisk	Islamic	Name (Politics)
Analyse	Analysis	Section
Sådan	This	Deictic-reference
Satire	Satire	Section
bil	car	Section
DMI	Weather	Section

Table 3: Top twenty Unigrams (Logistic Regression)

7 Conclusion

We presented an exploratory approach to predicting newspaper article popularity from headlines alone. Using pre-trained embeddings and a MTL setup, we are able to incorporate rich structural and semantic knowledge into the task and substantially improve performance. While the results are encouraging and allow the exploration of further auxiliary tasks (for example article word prediction), we find that a simple character-based n-gram model performs competitively. These findings highlight two aspects: 1) For any application of MTL, this is a strong case for comparing the results to non-deep models. While it is comparatively easy to show an improvement over the basic STL model, there might be other simple models that are competitive. 2) The selection of auxiliary tasks greatly influences the performance, even beyond simple regularization, and in a non-linear way. It does, however, provide us with a tool to test human intuitions about task interactions and the importance of certain problem aspects.

Acknowledgments

Thanks to A. Michele Colombo for help with the data and experiments. We also thank Jyllands-Posten for giving us access to the data, and to DSL for data for embeddings and POS annotations.

References

Héctor Martínez Alonso and Barbara Plank. 2017. When is multitask learning effective? semantic sequence prediction under varying data conditions. In *15th Conference of the European Chapter of the Association for Computational Linguistics*.

Raju Balakrishnan and Rajesh Parekh. 2014. Learning to predict subject-line opens for large-scale email marketing. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 579–584. IEEE.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask Learning for Mental Health Conditions with Limited Social Media Data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 152–162.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 164–169.

Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87 – 100.

Rich Caruana. 1993. Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 41–48.

Rich Caruana, Shumeet Baluja, Tom Mitchell, et al. 1996. Using the future to "sort out" the present: Rankprop and multitask learning for medical risk evaluation. *Advances in neural information processing systems*, pages 959–965.

Daniel Hardt and Owen Rambow. 2017. Predicting user views in online news. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 7–12.

Kokil Jaidka, Tanya Goyal, and Niyati Chhaya. 2018. Predicting Email and Article Clickthroughs with Domain-adaptive Language Models. In *Proceedings of the 10th ACM Conference on Web Science*, pages 177–184. ACM.

Jørg Asmussen. 2018. Society for Danish Language and Literature. http://dsl.dk. [Online; accessed 28-April-2018].

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

Ola Rønning, Daniel Hardt, and Anders Søgaard. 2018. Sluice Resolution without Hand-crafted Features over Brittle Syntax Trees. In *NAACL*.

Shivashankar Subramanian, Timothy Baldwin, and Trevor Cohn. 2018. Content-based Popularity Prediction of Online Petitions Using a Deep Regression Model. *Transactions of the Association for Computational Linguistics*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.