Do Human Rationales Improve Machine Explanations?

The University of Texas at Austin
Department of Computer Science
College of Natural Sciences

Julia Strout, Ye Zhang, and Raymond J. Mooney
Department of Computer Science, The University of Texas at Austin

Introduction

Work on “learning with rationales” shows that humans providing explanations to a machine learning system can improve the system's predictive accuracy. However, this work has not been connected to work in “explainable AI” which concerns machines explaining their reasoning to humans.

In this work, we show that learning with rationales can also improve the quality of the machine's explanations as evaluated by human judges. Specifically, we present experiments showing that, for CNN-based text classification, explanations generated using “supervised attention” are judged superior to explanations generated using normal unsupervised attention.

Models and Data

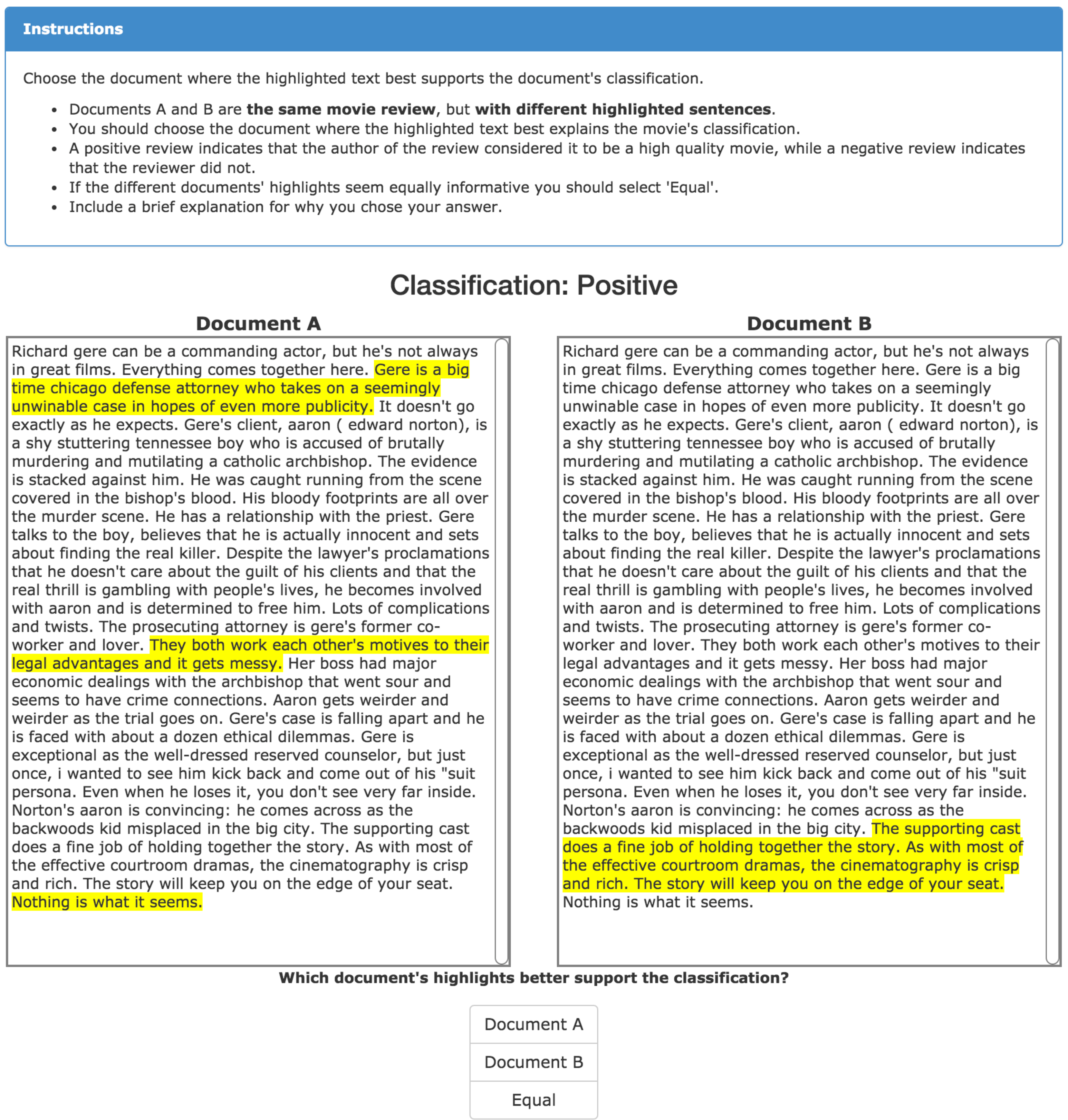
AT-CNN[1] is a CNN based text classification model with an attention mechanism that enforces a weighting over the sentences in the document. **RA-CNN**[1] is similar to AT-CNN but learns the weighting over the sentences from training labels called rationales, we call this “supervised attention.”

Classification Accuracy	
AT-CNN	RA-CNN
88.50%	90.00%

The dataset consists of 2000 movie reviews[2], 1800 used for training and 200 for test. Each document is either a positive or negative review.
Rationales: extra human annotations in the dataset that mark which sentences most support the classification.

Methods

Our Human Intelligence Task (HIT) shows a worker two copies of a test document along with the document's classification. Each copy of the document has a subset of sentences highlighted as explanations for the final classification. This subset is chosen as the 3 sentences with the largest weights from either AT-CNN's attention weights or RA-CNN's supervised weighting.



Instructions

Choose the document where the highlighted text best supports the document's classification.

- Documents A and B are **the same movie review**, but **with different highlighted sentences**.
- You should choose the document where the highlighted text best explains the movie's classification.
- A positive review indicates that the author of the review considered it to be a high quality movie, while a negative review indicates that the reviewer did not.
- If the different documents' highlights seem equally informative you should select 'Equal'.
- Include a brief explanation for why you chose your answer.

Classification: Positive

Document A

Richard gere can be a commanding actor, but he's not always in great films. Everything comes together here. Gere is a big time chicago defense attorney who takes on a seemingly unwinable case in hopes of even more publicity. It doesn't go exactly as he expects. Gere's client, aaron (edward norton), is a shy stuttering tennessee boy who is accused of brutally murdering and mutilating a catholic archbishop. The evidence is stacked against him. He was caught running from the scene covered in the bishop's blood. His bloody footprints are all over the murder scene. He has a relationship with the priest. Gere talks to the boy, believes that he is actually innocent and sets about finding the real killer. Despite the lawyer's proclamations that he doesn't care about the guilt of his clients and that the real thrill is gambling with people's lives, he becomes involved with aaron and is determined to free him. Lots of complications and twists. The prosecuting attorney is gere's former co-worker and lover. They both work each other's motives to their legal advantages and it gets messy. Her boss had major economic dealings with the archbishop that went sour and seems to have crime connections. Aaron gets weirder and weirder as the trial goes on. Gere's case is falling apart and he is faced with about a dozen ethical dilemmas. Gere is exceptional as the well-dressed reserved counselor, but just once, i wanted to see him kick back and come out of his "suit persona. Even when he loses it, you don't see very far inside. Norton's aaron is convincing: he comes across as the backwoods kid misplaced in the big city. The supporting cast does a fine job of holding together the story. As with most of the effective courtroom dramas, the cinematography is crisp and rich. The story will keep you on the edge of your seat. Nothing is what it seems.

Document B

Richard gere can be a commanding actor, but he's not always in great films. Everything comes together here. Gere is a big time chicago defense attorney who takes on a seemingly unwinable case in hopes of even more publicity. It doesn't go exactly as he expects. Gere's client, aaron (edward norton), is a shy stuttering tennessee boy who is accused of brutally murdering and mutilating a catholic archbishop. The evidence is stacked against him. He was caught running from the scene covered in the bishop's blood. His bloody footprints are all over the murder scene. He has a relationship with the priest. Gere talks to the boy, believes that he is actually innocent and sets about finding the real killer. Despite the lawyer's proclamations that he doesn't care about the guilt of his clients and that the real thrill is gambling with people's lives, he becomes involved with aaron and is determined to free him. Lots of complications and twists. The prosecuting attorney is gere's former co-worker and lover. They both work each other's motives to their legal advantages and it gets messy. Her boss had major economic dealings with the archbishop that went sour and seems to have crime connections. Aaron gets weirder and weirder as the trial goes on. Gere's case is falling apart and he is faced with about a dozen ethical dilemmas. Gere is exceptional as the well-dressed reserved counselor, but just once, i wanted to see him kick back and come out of his "suit persona. Even when he loses it, you don't see very far inside. Norton's aaron is convincing: he comes across as the backwoods kid misplaced in the big city. The supporting cast does a fine job of holding together the story. As with most of the effective courtroom dramas, the cinematography is crisp and rich. The story will keep you on the edge of your seat. Nothing is what it seems.

Which document's highlights better support the classification?

Quality Control

For our experiments on Amazon Mechanical Turk (AMT), we use gold standard questions to weed out poor workers, and require a group consensus on each test document.

Results

Percentage of examples where each model provided the best explanations		
RA-CNN	AT-CNN	Equal
43.47%	20.48%	36.14%

RA-CNN provides better explanations for the largest percentage of test documents (43.47%). The explanations are considered equal 36.14% of the time, and the remaining 20.48% of the documents were better explained by AT-CNN.

Comparing AT-CNN to Random Baseline		
AT-CNN	Random	Equal
57.23%	15.66%	27.12%

We ran a baseline test to ensure that AT-CNN explanations are reasonable and can at least beat a weak baseline. From the results above we can see that AT-CNN is beating the random baseline the majority of the time, demonstrating that attention, even without human supervision, can provide helpful explanations for a model's decision.

Percentage Of Test Documents where AT-CNN and RA-CNN share n explanation sentences			
0	1	2	3
33.5%	43.1%	22.2%	1.2%

Conclusion

Training with human rationales improves explanations for a model's classification decisions as evaluated by human judges. We show that while an unsupervised attention based model does provide some valuable explanations, as proven in the experiments comparing to a random baseline, a supervised attention model that trains on human rationales outperforms those results.

Acknowledgments

This research was supported by the DARPA XAI program through a grant from AFRL. The view sand conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government.

References

[1] Ye Zhang, Iain Marshall, and Byron C Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In Proceedings of the Conference on Empirical Methods in Natural Language Processing., volume 2016, page 795. NIH Public Access.

[2] Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using annotator rationales to improve machine learning for text categorization. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 260–267.