Human versus Machine Attention in Document Classification

Nikolaos Pappas
Andrei Popescu-Belis

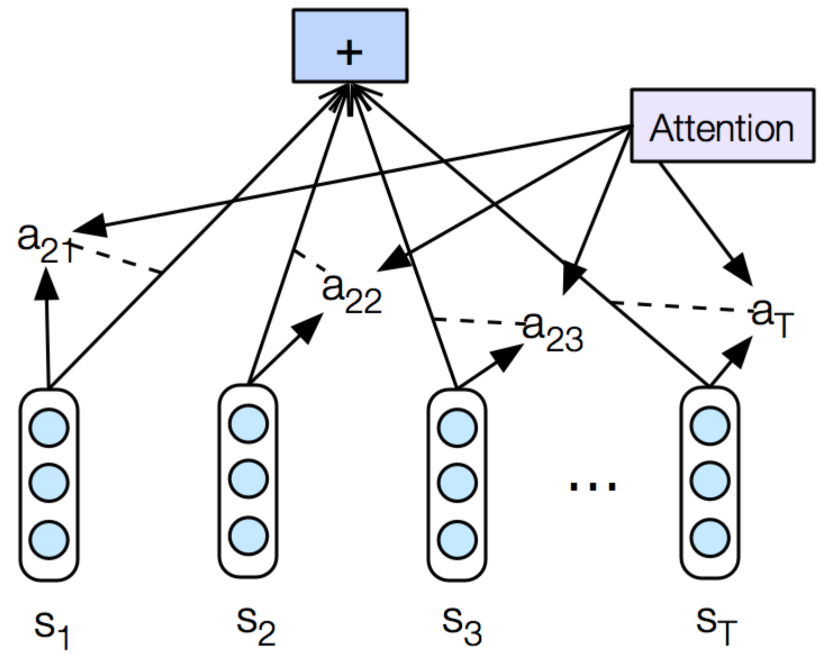
Idiap Research Institute Martigny, Switzerland

SocialNLP Workshop at EMNLP
November 1, 2016

Attention mechanism

"A mechanism which learns to focus on relevant parts of the input or intermediate states for a given task."

- Machine translation
  - Translate sequences of words
- Question answering
  - Collect relevant facts and answer comprehension questions
- Document classification
  - Predict one or more categories



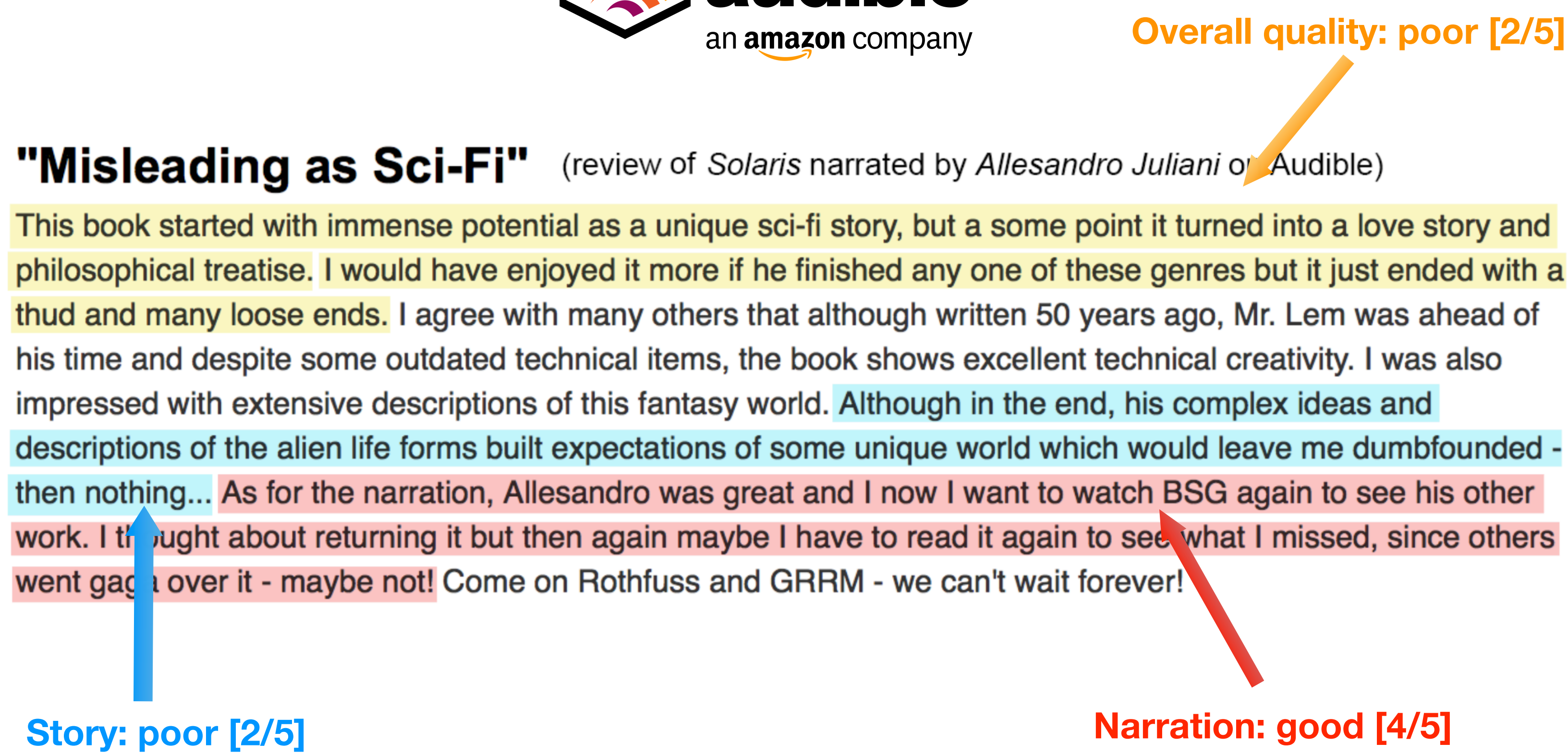
Contributions of this study

→ Captured human attention when classifying documents

→ Used this data to evaluate a document attention model (Pappas and Popescu-Belis, 2014)

Case study: Predicting aspect ratings of reviews

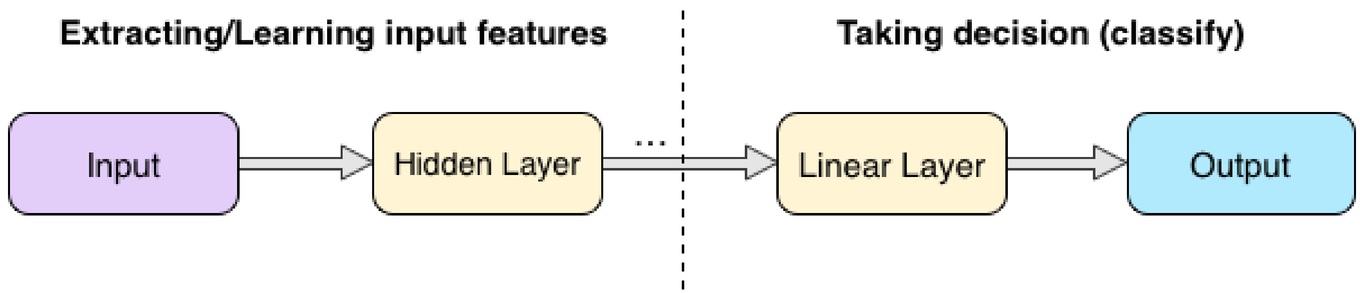
Given $\mathcal{D} = \{(x_i, y_i), \mid i = 1 \dots m\}$, find $\Phi_k : \mathcal{X} \rightarrow \mathcal{Y}_k$, where $x_i \in \mathbb{R}^d$ is a review and $y_i \in \mathbb{R}^k$ are the k target aspect ratings



- Such “weak” labels are abundant online (social sites)

Typical methods

- BOW, n-grams, topic models (Pang and Lee, 2005), (Titov and McDonald, 2008), (Zhu et al., 2012)
- Autoencoders, CNN, RNN (Maas et al., 2011), (Mikolov et al., 2013), (Mesnil et al., 2014), (Tang et al., 2015)
- Training on segmented text or with structured learning to capture label relations (McAuley et al., 2012)



→ Treat the text globally and ignore the "weak" nature of labels
→ Make simplistic assumptions when aggregating or pooling features

Attention-based methods

Use attention mechanism in one or more layers of the document modeling hierarchy (Pappas and Popescu-Belis, 2014), (Yang et al., 2016)

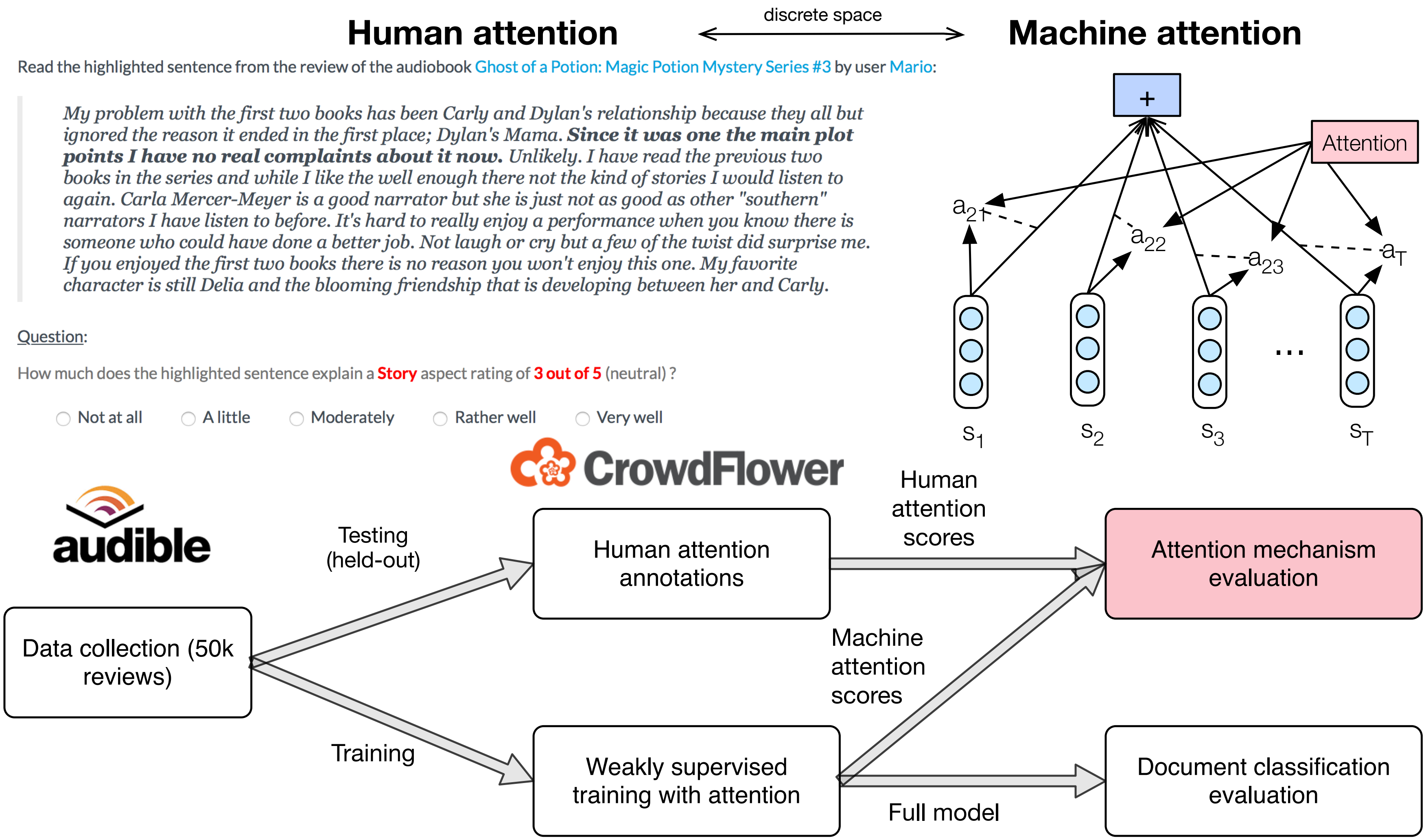
- Model the weak relation of categories to documents
- Provide a smarter way for aggregating or pooling features
- Perform better than typical methods without attention

Study	Datasets	Metric	Averaging	Attention
PPB14	5	MSE (μ)	4.34	**3.89**
Yang16	6	Acc (μ)	65.35	**66.41**

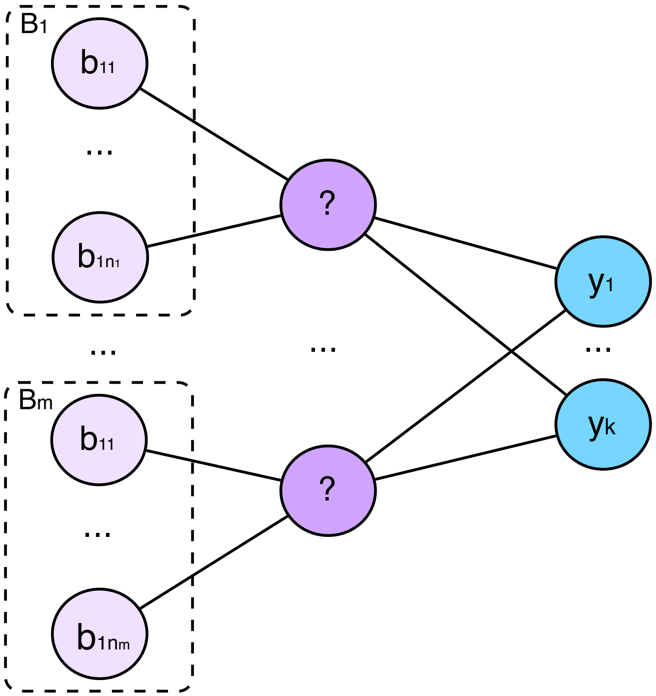
Limitations

- Evaluation makes use of extrinsic tasks only
- Visual analysis of attention is helpful but not grounded
- Lack of evidence of the quality of the learned structure

Overview of our proposal



Multiple-instance regression



Given $\mathcal{D} = \{(b_{ij}, y_i) \mid j = 1 \ldots n_i\}^m$,
find $\Phi_k : \mathcal{B} \to \mathcal{X} \to \mathcal{Y}_k$

- The bag B_i is a review represented by n_i instances b_{ij}, its sentences
- The labels $y_i \in \mathbb{R}^k$ are the aspect ratings of the review
- The exemplar (representation) $x_i \in \mathbb{R}^d$ of B_i is initially unknown

Advantages

$\rightarrow$ Supports several input assumptions (average, max, prime, instance)
$\rightarrow$ Better suited for weak (bag-level) labels, interpretable and flexible
$\rightarrow$ Subsumes traditional supervised learning methods

Instance relevance assumption

The method proposed in Pappas and Popescu-Belis (2014) models instance weights and target labels at the same time

$$x_i = \sum_{j=1}^{n_i} \psi_{ij} b_{ij},\ \psi_{ij} \geq 0 \ \text{ and } \sum_{j=1}^{n_i} \psi_{ij} = 1 \tag{1}$$

- Target labels model: $\hat{y}_i = f(\Phi, \Psi) = \Phi^T (B_i \psi_i)\ s.t.\ (1)$
- Instance weights model: $\hat{\psi}_i = g(O) = O^T B_i$
- Loss based on regularized least squares solved with Alternating Projections [2014] or Stochastic Gradient Descent [this study]

$\rightarrow$ JAIR paper underway

Learning parameters jointly with SGD

$$\sigma(B_i, O) = P(\psi = y_i | x) = \frac{e^{(O^T B_i)}}{\sum_{k=1}^{n_i} e^{(O^T B_{ik})}}$$

$$O, \Phi = \underset{O,\Phi}{arg\ min} \sum_{i=1}^{m} (y_i - \Phi^T (B_i \cdot \sigma(B_i, O)))^2 + \Omega(\Phi, O)$$

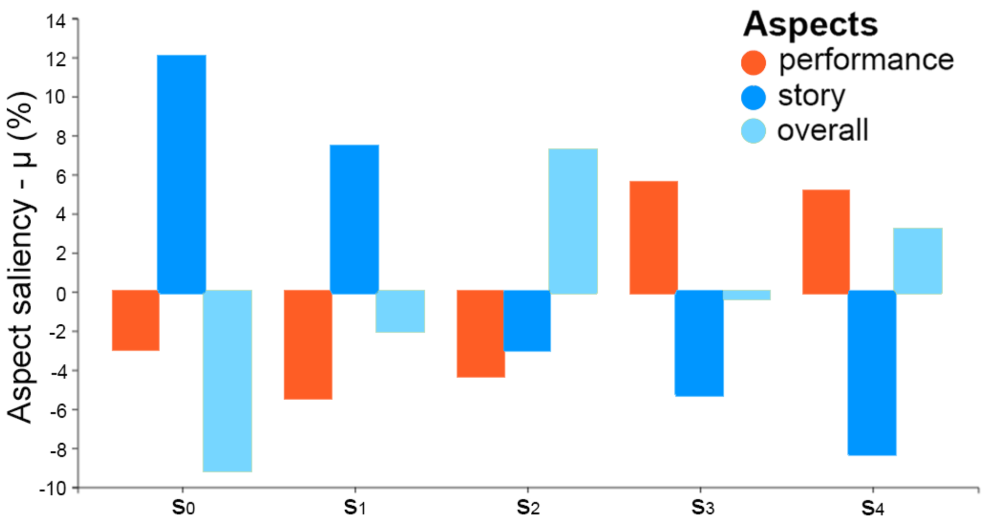
- Preserves constraints of instance relevance assumption
- Achieves similar performance to alternating projections
- Makes the learning procedure more scalable

Shared material

$\rightarrow$ Code: wmil, wmil-sgd

`https://github.com/nik0spapp/`

Estimated relevance of sentences



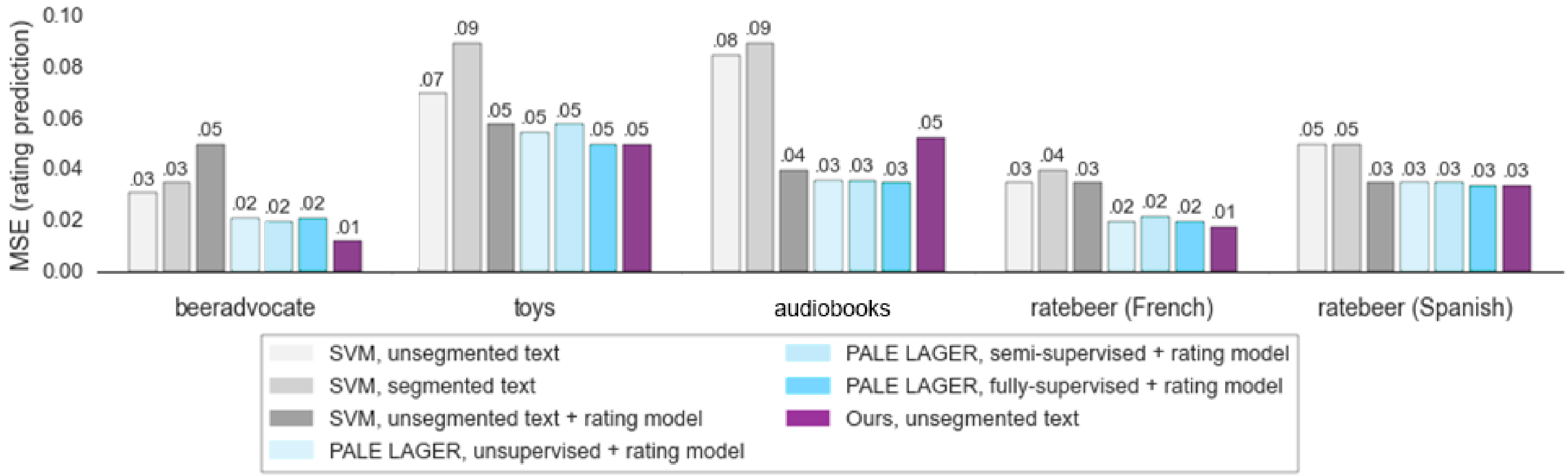
ID	Aspect	Sentence
s_0	story	This book was nearly as good as the first one in the series.
s_1	story	It seemed the ending was at least an hour or more too late.
s_2	story	When I thought it should be over, I checked how many minutes I had left and knew I was not even close.
s_3	perform.	I liked the narration, I thought he did a good job.
s_4	overall	Still a 4 star rating: good story, good characters, and looking forward to the third in the series.

MIR weights for a positive audiobook review (4 out of 5).

- Captures how relevant is a sentence to the aspect rating
- This is different from topicality, i.e. being “about” an aspect
  - zero relevance for a factual sentence about an aspect
  - high relevance sentences are more likely to discuss topic

- Introduction
  - Background and motivation
  - Related work
  - Overview
- System: A Model of Document Attention
  - Weakly-supervised learning
  - Structural assumptions
  - Instance relevance mechanism
- **Evaluation**
  - **Aspect-based rating prediction**
  - **Comparing mechanism to humans**
- Conclusion

Results: Document-level aspect rating prediction



- MIR document attention model achieves lower error than
  - methods trained with segmented text (SVM, PALE LAGER by McAuley et al. 2012)
  - structured learning (Structured SVM, PALE LAGER)

→ How can we evaluate the sentence relevance intrinsically?

Crowdsourcing task

Goal: capture human attention to sentences when attributing categories (aspect ratings) to documents (audiobook reviews)

- How much does each sentence explain the given aspect rating?

Data: reviews from Audible

- 4,986 micro-tasks = 1,662 sentences (100 reviews) $\times$ 3 aspects
- obtained 20k annotations ($\geq$4 annotators per micro-task)
- 0.60 agreement score by Crowdflower

Shared material

$\rightarrow$ HATDOC dataset

`https://www.idiap.ch/paper/hatdoc`

Crowdsourcing task: Screenshot

Read the highlighted sentence from the review of the audiobook Ghost of a Potion: Magic Potion Mystery Series #3 by user Mario:

> *My problem with the first two books has been Carly and Dylan's relationship because they all but ignored the reason it ended in the first place; Dylan's Mama.* ***Since it was one the main plot points I have no real complaints about it now.*** *Unlikely. I have read the previous two books in the series and while I like the well enough there not the kind of stories I would listen to again. Carla Mercer-Meyer is a good narrator but she is just not as good as other "southern" narrators I have listen to before. It's hard to really enjoy a performance when you know there is someone who could have done a better job. Not laugh or cry but a few of the twist did surprise me. If you enjoyed the first two books there is no reason you won't enjoy this one. My favorite character is still Delia and the blooming friendship that is developing between her and Carly.*

Question:

How much does the highlighted sentence explain a **Story** aspect rating of **3 out of 5** (neutral) ?

○ Not at all ○ A little ○ Moderately ○ Rather well ○ Very well

Example: Positive review

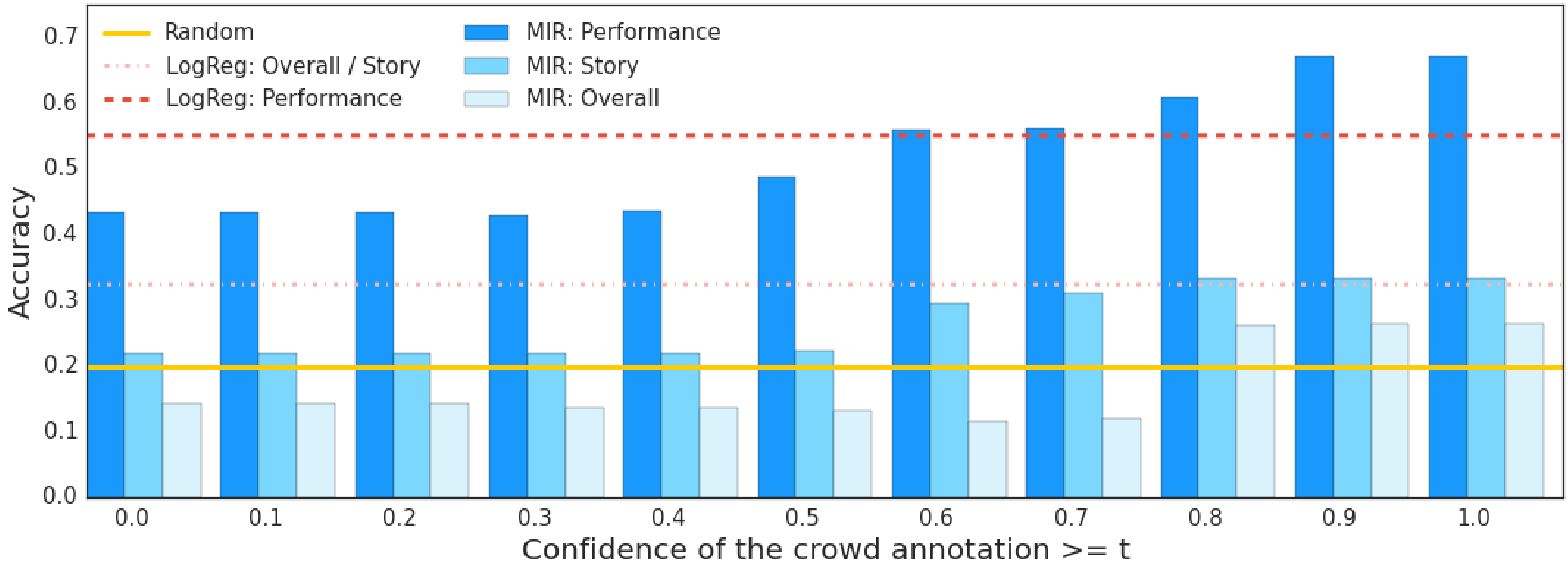
Ove. (5/5)	Perf. (5/5)	Story (5/5)	Document (id=969066)
0.45	0.56	0.18	Narrated by one of my favorite narrators, Scott Brick, I found this offering by Harlan Coben to be one of their best - for them both.
0.18	0.22	0.36	I found it very difficult to "put this down".
0.36	0.22	0.45	It is one of those no-brainer 5 star thillers!

Visualization of attention labels (normalized per aspect).

→ More examples online:

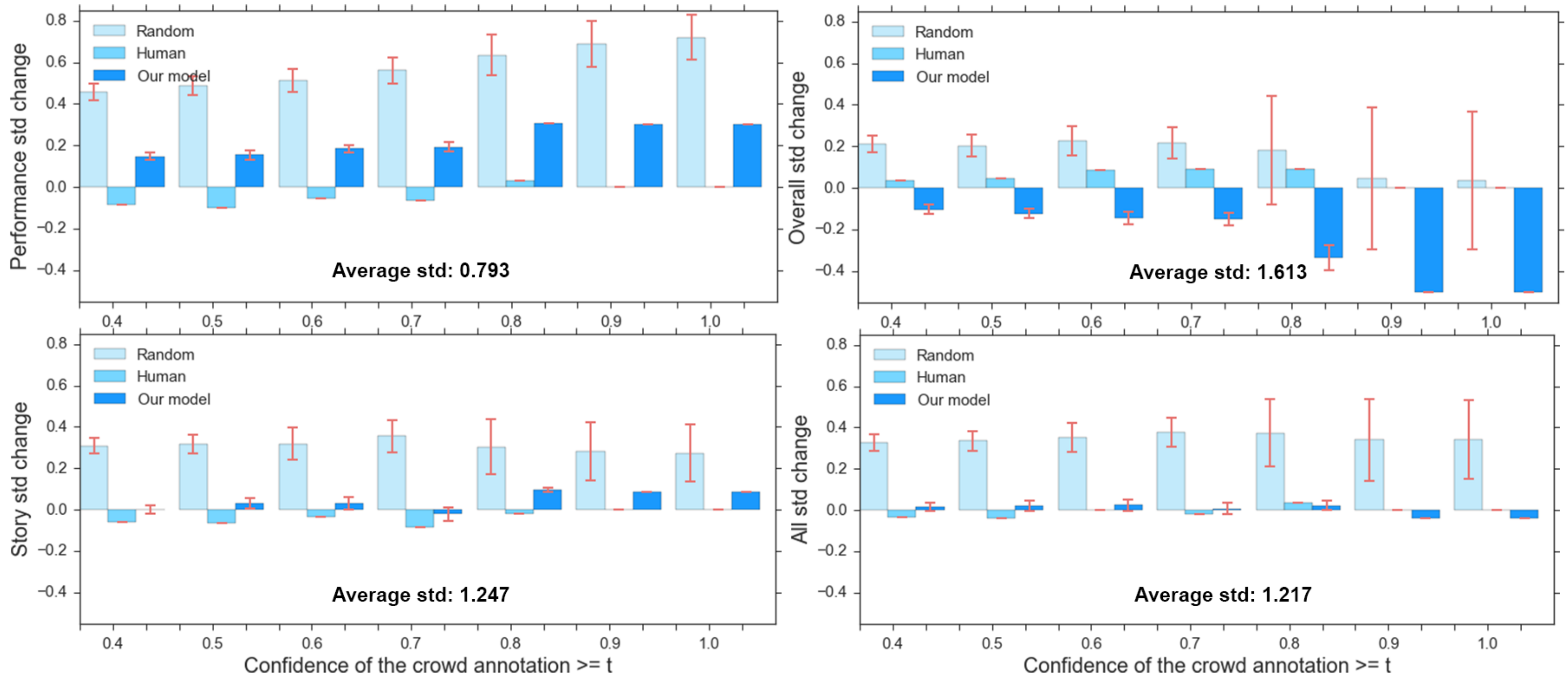
http://www.idiap.ch/paper/hatdoc/explore.html

Results: Human attention prediction (exact match)



- MIR outperforms Random for all three aspects (confidence ≥ 0.8)
- High MIR accuracy on the *Performance* aspect (least ambiguous)
- MIR compares favorably to fully-supervised LogReg (oracle)

Reliability analysis: STD change with label replacements (x100)



- MIR consistently outperforms Random for all aspects and levels
- MIR is comparable to qualified humans for *Story* and better than qualified humans for *Overall*

Conclusion

- New intrinsic benchmark for attention mechanisms
- Document attention models capture meaningful structure
    - Positive correlation of MIR accuracy with human confidence
    - Comparable results to qualified humans for two of the aspects
- Intuitive way to summarize the sentiment towards each aspect

Extensions:

- Refine evaluation and compare to other attention-based models
- Apply to other labels (e.g. topics) and linguistic levels (e.g. words)

Thank you!

Acknowledgments:

EU projects n. 287,872 and 688,139

References I

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 142–150, Portland, OR, USA, 2011.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *Proceedings of the 12th IEEE International Conference on Data Mining*, ICDM '12, pages 1020–1025, Brussels, Belgium, 2012. doi: 10.1109/ICDM.2012.110.

Grégoire Mesnil, Tomas Mikolov, Marc'Aurelio Ranzato, and Yoshua Bengio. Ensemble of generative and discriminative techniques for sentiment analysis of movie reviews. *CoRR*, abs/1412.5335, 2014. URL http://arxiv.org/abs/1412.5335.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Christopher Burges, Lon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 115–124, Ann Arbor, Michigan, 2005. doi: 10.3115/1219840.1219855.

Nikolaos Pappas and Andrei Popescu-Belis. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 455–466, Doha, Qatar, 2014.

Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 111–120, Beijing, China, 2008. doi: 10.1145/1367497.1367513.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. Hierarchical attention networks for document classification. In *HLT-NAACL*, 2016.

Jingbo Zhu, Chunliang Zhang, and Matthew Y. Ma. Multi-aspect rating inference with aspect-based segmentation. *IEEE Trans. on Affective Computing*, 3(4):469–481, 2012. ISSN 1949-3045. doi: 10.1109/T-AFFC.2012.18.