**Modeling Annotators:**
A Generative Approach to Learning from Annotator Rationales

Omar F. Zaidan  
Jason Eisner

The Center for Language and Speech Processing  
Johns Hopkins University

EMNLP 2008 – Honolulu, HI  
Saturday October 25th, 2008

Annotations are Underutilized?

This is a disaster flick which is a disaster alright. Directed by Tony Scott (Top Gun), it's the story of an asteroid the size of Texas caught on a collision course with Earth. After a great opening, in which an American spaceship plus NYC are completely destroyed by a comet shower, NASA detects a asteroid and go into a frenzy. They hire the world's best oil driller (Bruce Willis), and send him and his crew up into space to fix our global problem.

The action scenes are over the top and too ludicrous for words. So much so, I had to sigh and hit my head with my notebook a couple of times. Also, to see a wonderful actor like Billy Bob Thornton in a film like this is a waste of his talents. The only real reason for making this film was to somehow out-perform Deep Impact. Bottom line is, Armageddon is a failure.

**Annotations are Underutilized!**

This film, directed by Tony Scott (Top Gun), is the story of an asteroid the size of Texas caught on a collision course with Earth. After a great opening, in which an American spaceship plus NYC are completely destroyed by a comet shower, NASA detects an asteroid and go into a frenzy. They hire the world's best oil driller (Bruce Willis), and send him and his crew up into space to fix our global problem.

The action scenes are over the top and too ludicrous for words. So much so, I had to sigh and hit my head with my notebook a couple of times. Also, to see a wonderful actor like Billy Bob Thornton in a film like this is a waste of his talents. The only real reason for making this film was to somehow out-perform Deep Impact. Bottom line is, Armageddon is a failure.

Annotations are Underutilized...

A lot of thought goes into annotation, but little of that is captured.

Annotations are Underutilized!

This film, directed by Tony Scott (Top Gun), is the story of an asteroid the size of Texas caught on a collision course with Earth. After a great opening, in which an American spaceship plus NYC are completely destroyed by a comet shower, NASA detects an asteroid and go into a frenzy. They hire the world's best oil driller (Bruce Willis), and send him and his crew up into space to fix our global problem.

The action scenes are over the top and too ludicrous for words. So much so, I had to sigh and hit my head with my notebook a couple of times. Also, to see a wonderful actor like Billy Bob Thornton in a film like this is a waste of his talents. The only real reason for making this film was to somehow out-perform Deep Impact. Bottom line is, Armageddon is a failure.

Annotations are Underutilized!!!

This film, directed by Tony Scott (Top Gun), is the story of an asteroid the size of Texas caught on a collision course with Earth. After a great opening, in which an American spaceship plus NYC are completely destroyed by a comet shower, NASA detects an asteroid and go into a frenzy. They hire the world's best oil driller (Bruce Willis), and send him and his crew up into space to fix our global problem.

The action scenes are over the top and too ludicrous for words. So much so, I had to sigh and hit my head with my notebook a couple of times. Also, to see a wonderful actor like Billy Bob Thornton in a film like this is a waste of his talents. The only real reason for making this film was to somehow out-perform Deep Impact. Bottom line is, Armageddon is a failure.
Annotators are Underutilized!

“Hey annotator, is this review positive or negative?”

Annotator says:

(\(y = -1\))

Why would Rationales be Useful?

\(\text{Knowledge}\)

Annotator

Class labels

Rationales

Related Work

- “Annotators are underutilized. Let’s ask them to do more than just annotate class.”
- Let’s ask annotator to identify relevant features.
- But:
  - Features could be hard to describe.
  - Features could be hard for annotators to understand.
  - We might want different features in the future.
Non-annotated documents

Class and also “rationales”

Annotated documents

<\( x, y, r \)>

Zaidan & Eisner – Modeling Annotators

Saving Private Ryan

War became a reality to me after seeing Saving Private Ryan. Steve Spielberg with his latest production. Keep the kids home as the R rating is for Reality. Tom Hanks is Captain John Miller, set out in France during WW II to rescue and return home a soldier, Private Ryan (Matt Damon) who lost three brothers in the war. Spielberg takes us inside the heads of these individuals as they face death during the horrific battle scenes. Private Ryan is not for everyone, but for a movie like this to be made. The movie reminds us of the sacrifices made by our fighting men and women. For this I thank them and for Steve Spielberg And I’m sure the Academy will not forget Tom Hanks come April, as with be in Tom’s possession.

Zaidan & Eisner – Modeling Annotators

The Postman

Question: after the disaster that was Waterworld, were the execs who gave Costner the money to make another movie thinking??

In this 3 hour advertisement for his new hair weave, Costner plays a nameless drifter who dons a long dead postal employee’s uniform and gradually turns a nuked-out USA into an idealized hippy-dippy society. (The main accomplishment of this brave new world is in re-inventing polyester.) When he’s not pointing the camera directly at himself, director Costner does have a nice visual sense, but by the time the second hour rolled around, Mark this one

Class and also “rationales”

OK … now what??
Linear Classifier

- In our experiments, we use a linear classifier.
- Classifier is represented by a weight vector $\vec{\theta}$.
- Rationales will play a role when learning $\vec{\theta}$.
- Improvements **solely due to a better-learned $\vec{\theta}$** (vs. no rationales)
- At test time:
  - No change to decision rule.
  - No new features.
  - No need for rationales.

$\vec{\theta}$

Trees Lounge is the directorial debut from one of my favorite actors, Steve Buscemi. He gave memorable performances in in The Soup, Fargo, and Reservoir Dogs. Now he tries his hand at writing, directing and acting all in the same flick. The movie starts out awfully slow with Tommy (Buscemi) hanging around a local bar the “trees lounge” and him pestering his brother. It’s obvious he a loser. But as he says ”it’s better I’m a loser and know I am, then being a loser and not thinking I am.” Well put. The story starts to take off when his uncle dies, and Tommy, not having a job, decides to drive an ice cream truck. Well, the movie starts to pick up with him finding a love interest in a 17 year old girl named Debbie (Chloe Sevigny) and... I liked this movie alot even though it did not reach my expectation. After you've seen him in Fargo and Reservoir Dogs, you know he is capable of a better performance. I think his brother, Michael, did an excellent job for his debut performance. Mr. Buscemi is off to a good career as a director!
This $\theta$ was learned in a standard way:

Choose $\tilde{\theta} = \arg \max_\theta \prod_{i=1}^n p(y_i | x_i, \theta)$

choose $\tilde{\theta}$ that models class labels well

(of training data)

PS: $p(y | x, \tilde{\theta}) \propto \exp(y \cdot (\tilde{\theta} \cdot f(x)))$ i.e. log-linear model

From a review for “Prince of Egypt” ($y = +1$):

51 weeks into ’98 , a champ has emerged .

the prince of egypt succeeds where other movies failed .

• We encode rationales as a tag sequence.

• The model $p(r | y, x, \tilde{\theta})$ should predict tag sequence with some help from $\tilde{\theta}$. Let’s describe it…

This $\tilde{\theta}$ was learned in a novel way (our algorithm!):

Choose $\tilde{\theta} = \arg \max_\theta \prod_{i=1}^n p(y_i | x_i, \theta) \cdot p(r_i | y_i, x_i, \theta)$

choose $\tilde{\theta}$ that models class labels & rationales well

(of training data)

PS: $p(y | x, \tilde{\theta}) \propto \exp(y \cdot (\tilde{\theta} \cdot f(x)))$  $p(r | y, x, \tilde{\theta}) = ???$
Designing the Model $\mathbb{P}(r \mid y, x, \overrightarrow{\theta})$

- Mission: model $\{I,O\}$ tag sequence with $\overrightarrow{\theta}$'s help.

$$r_1, r_2, r_3, \ldots, r_M \quad \text{(Tags)}$$

$$x_1, x_2, x_3, \ldots, x_M \quad \text{(Words)}$$

Hmm... $\overrightarrow{\theta}$ has no role here.

- CRF: Used in other labeling tasks too (POS, NER).

Designing the Model $\mathbb{P}(r \mid y, x, \overrightarrow{\theta})$

- Mission: model $\{I,O\}$ tag sequence with $\overrightarrow{\theta}$'s help.

$$\theta_{x_1}, \theta_{x_2}, \theta_{x_3}, \ldots, \theta_{x_M} \quad \text{(Words)}$$

- Makes sense: if for some word $w$, $(y \cdot \theta_w) >> 0$, then annotator is likely to mark it $I$.

- How likely? We learn a parameter $\phi_{rel}$ from that annotator’s rationale data.

- $\phi_{rel}$ is learned jointly with $\overrightarrow{\theta}$.

Designing the Model $\mathbb{P}(r \mid y, x, \overrightarrow{\theta})$

- Mission: model $\{I,O\}$ tag sequence with $\overrightarrow{\theta}$'s help.

$$\theta_{x_1}, \theta_{x_2}, \theta_{x_3}, \ldots, \theta_{x_M} \quad \text{(Words)}$$

- Problem with this model: it is led to believe that any word $w$ marked with $I$ has high $(y \cdot \theta_w)$.

- BUT: temptation to start shouting slogans, ...

- So, learn 4 more weights: $\phi_{1-I}$ $\phi_{O-O}$ $\phi_{I-O}$ $\phi_{O-I}$

- Better! Now: “$w$ is marked with $I$ either because of high $(y \cdot \theta_w)$ or being around others marked with $I$.”

- CRF: Used in other labeling tasks too (POS, NER).
Recap

- Linear classifier: Choose \( y^* = \begin{cases} +1 & \text{if } \theta \cdot f(x) > 0 \\ -1 & \text{if } \theta \cdot f(x) < 0 \end{cases} \)

- Our work: choose \( \theta \) that models all of annotator’s data well: both class labels and rationales.

Choose \( \theta = \arg \max_{\theta} \prod_{i=1}^n \left[ p(y_i | x_i, \theta) \prod_{j=1}^m p_z(r_j | y_i, x_i, \theta) \right] \)

\( p(y | x, \theta) \): Standard log-linear model

\( p_z(r | y, x, \theta) \): CRF predicting I/O tag sequence

- Remember, at test time:
  - No change to decision rule.
  - No new features.
  - No need for rationales.

Experiments


- 2000 document, each annotated with class label and enriched with rationales (Zaidan et al., 2007).

- We train a classifier jointly maximizing likelihood of class and rationale data.

- Compared it to 2 baseline models that account only for class data: log-linear model and SVM.
  - …and to “SVM contrast” method of Zaidan et al. (2007).


I like ‘em curves, but…

- Q: “I see that including rationales helps with performance. But they do take extra time to collect. Is it worth the extra time?”

- A: “Yes!”

- Question addressed in Zaidan et al. (2007).
  - Collecting rationales doesn’t take too long.
  - You don’t even need so many to get much of benefit.

- You should collect some in your next annotation project!
Fancy $\phi$-Features

- Remember this?

$$
\begin{array}{cccc}
\theta_1 & \theta_2 & \ldots & \theta_M \\
(x_1, y_1) & (x_2, y_2) & \ldots & (x_M, y_M) \\
\end{array}
$$

- Actually, we used an expanded set that includes about 100 “conditional” transition features, e.g.:
  - How often do you see I→O around punctuation?
  - How often do you see I→O around syntactic boundaries?
- Extra features model rationales better, but don’t help learn $\theta$ any better than basic feature set.

$\phi$ Captures Annotator’s Style

- So, learned $\phi$ captures style of a specific annotator.
- Empirical evidence?
  - An annotator’s own $\phi$ predicts their rationales best.
- “So, perhaps you got lucky with A0 (whose data you used in experiments). Would this work with others?”

<table>
<thead>
<tr>
<th></th>
<th>A0</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-linear baseline</td>
<td>71.0</td>
<td>73.0</td>
<td>71.0</td>
<td>70.0</td>
</tr>
<tr>
<td>Our method</td>
<td><strong>76.0</strong></td>
<td><strong>76.0</strong></td>
<td><strong>77.0</strong></td>
<td><strong>74.0</strong></td>
</tr>
<tr>
<td>SVM baseline</td>
<td>72.0</td>
<td>72.0</td>
<td>72.0</td>
<td>70.0</td>
</tr>
<tr>
<td>“SVM Contrast” method</td>
<td>75.0</td>
<td>73.0</td>
<td><strong>74.0</strong></td>
<td><strong>72.0</strong></td>
</tr>
</tbody>
</table>

Conclusions

- A model that accounts for class and rationale annotations outperforms two strong baselines.
- Generative approach generalizes to other classifiers (just incorporate $p_\phi(r \mid y, x, \theta)$ in objective).
- …and to other domains:
  - E.g. vision: just model how relevant portions of image tend to trigger rationales (“channel model”) and what size/shape rationales tend to be (“language model”).
- Doing an annotation project? Collect rationales!
  - Even small number could help.
  - “You may get more bang for the buck!”

![Salesman](1-800-for-RATS)
And finally…

• Jason and I thank Christine Piatko (co-author on Zaidan et al., 2007) for helpful discussions and feedback on paper and presentation.

• Thank you!

• A woman in peril. A confrontation. An explosion. The end. (Metro)

• As directed by Joel Schumacher, B&R is (Batman & Robin)

• (The Postman)

On The Internets

• Slides and more:
  
  http://cs.jhu.edu/~ozaidan/rationales

FYI–BibTeX entry:

@inproceedings{zaidan-eisner:2008:EMNLP,
  author = {Zaidan, Omar and Eisner, Jason},
  title = {Modeling Annotators: A Generative Approach to Learning from Annotator Rationales},
  booktitle = {Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)},
  month = {October},
  year = {2008},
  address = {Honolulu, Hawaii},
  publisher = {Association for Computational Linguistics},
  pages = {31--40},
  url = {http://www.aclweb.org/anthology/D08-1004}
}