Machine Learning with Annotator Rationales to Reduce Annotation Cost

Omar F. Zaidan Jason Eisner Christine D. Piatko

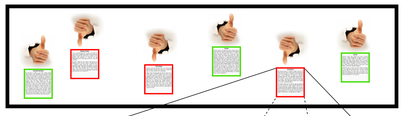
JOHNS HOPKINS UNIVERSITY

Motivation

Task: classification of movie reviews into positive and negative reviews. Typically, an annotator is given a set of unannotated documents, and annotates the correct class for each one (⇨).

This classification task is actually hard for a machine learner. Place yourself in its position, and imagine the task in Arabic instead (⇩). If you are not fluent in Arabic, class data alone might not be very helpful. This is truly the situation from a computer's point of view!

But what if, in addition to class, you were also told which segments of the text actually support that class (⇘)? That should make it easier for you to learn the true model. Presumably, the same is true for a machine learner as well.



Armageddon

This disaster flick 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 said 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.

يوم الحساب

هذا الفيلم الكارثي هو بحق فيلم كارثي. الفيلم، من إخراج توني سكوت (مخرج فيلم توب جن)، يروي قصة كويكب بحجم ولاية تكساس في طريقه للاصطدام بالكرة الأرضية. بعد مشهد افتتاحي رائع، تدمر فيه سفينة فضاء أمريكية، إضافة إلى مدينة نيويورك، تكتشف ناسا الكويكب أنف الذكر ويجن جنونها. فيستعينون بأفضل منقب عن النفط في العالم (بروس ويليس)، ويرسلونه وطاقمه إلى الفضاء لإصلاح مشكلتنا العالمية هذه.

مشاهد الإثارة والأكشن مبالغ بها ويصعب وصف سخفها بأي كلمات، إلى حد وجدت نفسي فيه أتنهد، بل وضربت جبهتي بمدونتي بضع مرات. و أيضا فإن رؤية ممثل مثل بيلي بوب ثورنتون في فيلم كهذا هو بحق مضيعة لمواهبه. السبب الوحيد لصنع هذا الفيلم هو محاولة التفوق على فيلم ديب امباكت بطريقة ما. خلاصة الأمر، يوم الحساب فيلم فاشل.

يوم الحساب

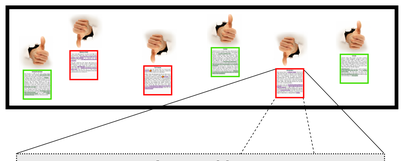
هذا الفيلم الكارثي هو بحق فيلم كارثي. الفيلم، من إخراج توني سكوت (مخرج فيلم توب جن)، يروي قصة كويكب بحجم ولاية تكساس في طريقه للاصطدام بالكرة الأرضية. بعد مشهد افتتاحي رائع، تدمر فيه سفينة فضاء أمريكية، إضافة إلى مدينة نيويورك، تكتشف ناسا الكويكب أنف الذكر ويجن جنونها. فيستعينون بأفضل منقب عن النفط في العالم (بروس ويليس)، ويرسلونه وطاقمه إلى الفضاء لإصلاح مشكلتنا العالمية هذه.

مشاهد الإثارة والأكشن مبالغ بها ويصعب وصف سخفها بأي كلمات، إلى حد وجدت نفسي فيه أتنهد، بل وضربت جبهتي بمدونتي بضع مرات. و أيضا فإن رؤية ممثل مثل بيلي بوب ثورنتون في فيلم كهذا هو بحق مضيعة لمواهبه. السبب الوحيد لصنع هذا الفيلم هو محاولة التفوق على فيلم ديب امباكت بطريقة ما. خلاصة الأمر، يوم الحساب فيلم فاشل.

Proposal: we wish to better utilize annotators by having them tell us more about their classification process. We propose annotators indicate not only *what* the correct answers are, but also provide hints about *why*.

We propose that they should **highlight relevant portions of the example, such as substrings** (⇨), that help to justify their annotations. We call such hints ***rationales***.

We have collected rationales for the movie review dataset of [PL04], and have developed two methods that use the rationales during training. One is a discriminative method [ZEP07] and one is generative [ZE08]. Both methods yield significant accuracy improvements…



Armageddon

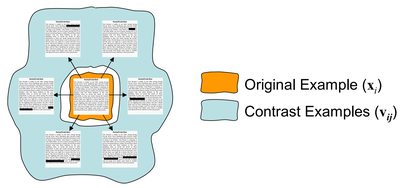
This disaster flick is a disaster alright. Directed by Tony Scott (Top Gun), it's the story of an asteroid the size of Texas caught on a collision course with Earth. After a great opening, in which an American spaceship, plus NYC, are completely destroyed by a comet shower, NASA detects said asteroid 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.

Approaches

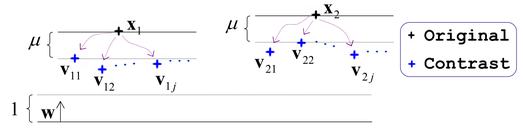
Discriminative Approach

From the rationale annotations on a positive example $\mathbf{x}_i$, we construct several "not-quite-as-positive" contrast examples $\mathbf{v}_{ij}$. Each contrast $\mathbf{v}_{ij}$ is obtained by starting with the original and "masking out" a rationale substring:



The intuition: a correct model should be less sure of a positive classification on the contrast example $\mathbf{v}_{ij}$ than on the original example $\mathbf{x}_i$, because $\mathbf{v}_{ij}$ lacks evidence the annotator found significant.

We express our intuition as additional constraints on an SVM-like model: we want (for each j) to have $\mathbf{w} \cdot \mathbf{x}_i - \mathbf{w} \cdot \mathbf{v}_{ij} \geq \mu$, where $\mu \geq 0$ controls the size of a margin between original and contrast examples:



Minimize:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i + C_{contrast}\sum_{i,j} \xi_{ij}$$

Subject to:

$$y_i(\mathbf{w}\cdot\mathbf{x}_i) \geq 1-\xi_i$$

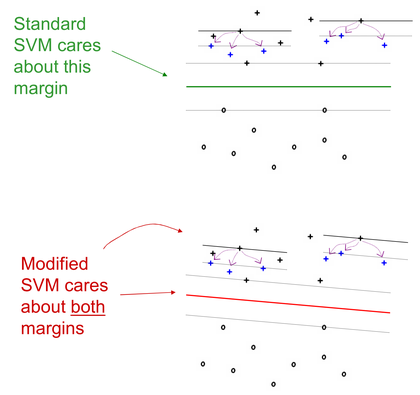
Subject to:

$$y_i(\mathbf{w}\cdot\mathbf{x}_i) \geq \mu(1-\xi_{ij})$$

$$y_i\left(\mathbf{w}\cdot\frac{\mathbf{x}_i-\mathbf{v}_{ij}}{\mu}\right) \geq 1-\xi_{ij}$$

$$y_i(\mathbf{w}\cdot\mathbf{x}_{ij}) \geq 1-\xi_{ij}$$

What this means in practice



Generative Approach

We typically choose parameters that explain class labels y of training data. With rationales, we propose that parameters be chosen to explain rationale data r in addition to class labels y. For instance, a conditional log-linear model:

$$p(y \mid x, \vec{\theta}) \propto \exp\left(\vec{\theta}\cdot\vec{f}(x,y)\right)$$

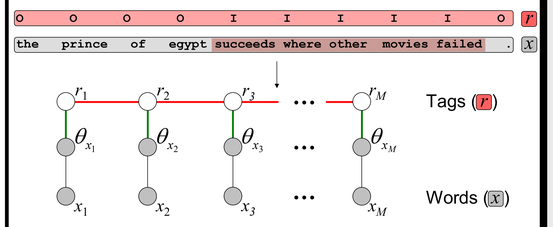
would have a parameter vector $\vec{\theta}$ chosen so that:

$$\vec{\theta} = \arg\max_{\vec{\theta},\vec{\phi}} \prod_{i=1}^{n} p(y_i \mid x_i, \vec{\theta}) \cdot p_{\vec{\phi}}(r_i \mid y_i, x_i, \vec{\theta})$$

Learned jointly

i.e. try to model class labels & rationales well

The second factor corresponds to a model for r. Its parameters, $\vec{\phi}$, capture how the true $\vec{\theta}$ influence the annotator. To do this, we encode the rationales as a tag sequence and model it using a CRF:



$$\text{CRF: } p_{\vec{\phi}}(r \mid x, y, \vec{\theta}) \propto \exp\left(\vec{\phi}\cdot\vec{g}(\vec{r},\vec{x},y,\vec{\theta})\right)$$

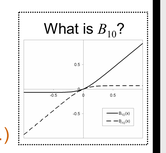
The first-order "emission" features of $\vec{g}(\cdot)$ relate the tag r_m to (x_m, y, θ_m), whereas the second-order "transition" features of $\vec{g}(\cdot)$ relate the tag r_m to r_{m-1}.

$$g_{\text{rel}}(\vec{r},\vec{x},y,\vec{\theta}) = \sum_{m=1}^{M} I(r_m = \texttt{I}) \cdot B_{10}(y\cdot\theta_{x_m})$$

$$g_{\texttt{O-I}}(\vec{r},\vec{x},y,\vec{\theta}) = \sum_{m=1}^{M} I(r_{m-1} = \texttt{O} \text{ and } r_m = \texttt{I})$$

(Plus several other g features…)

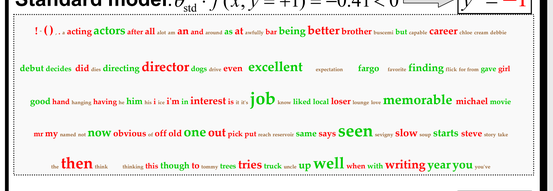
What is B_{10}?



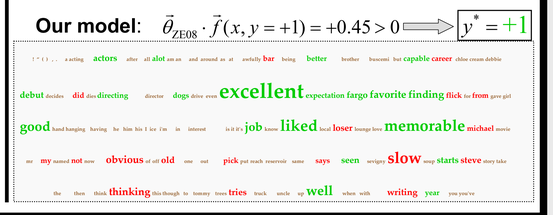
What this means in practice

Classifying `pos_987.txt` (correct class: +1):

Standard model: $\vec{\theta}_{\text{std}} \cdot \vec{f}(x, y=+1) = -0.41 < 0 \Rightarrow y^* = -1$

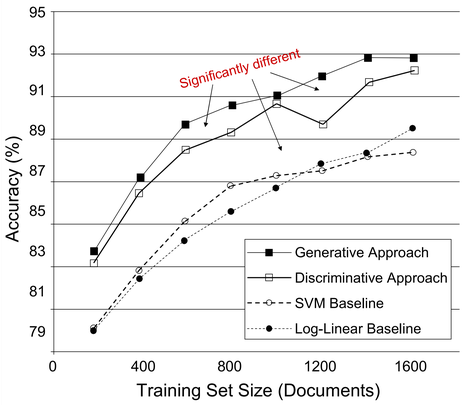


Our model: $\vec{\theta}_{\text{ZE08}} \cdot \vec{f}(x, y=+1) = +0.45 > 0 \Rightarrow y^* = +1$



Results

Both methods give significant improvements in accuracy over two strong baselines:



Above curves were generated using rationales from a single annotator, **A0**. What about other annotators? We solicited rationales from several other annotators (on 100 documents) and saw similar improvements in accuracy:

	A0	A3	A4	A5
Log-linear baseline	71.0	*73.0*	71.0	70.0
Generative Method	**76.0**	**76.0**	**77.0**	**74.0**
SVM baseline	72.0	*72.0*	72.0	70.0
Discriminative Method	*75.0*	*73.0*	*74.0*	*72.0*

Take-home Message:

- Annotators are underutilized! Richer annotations, such as rationales, can aid machine learning.
- Existing machine learning methods can be modified to exploit rationales.
- Remember, at test time:
  - No change to decision rule.
  - No new features.
  - No need for rationales.

  Improvements due solely to better-learned $\mathbf{w}/\vec{\theta}$
- Doing an annotation project? Collect rationales! Even a small number could help.

References:

[PL04] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proc. ACL, pages 271–278, 2004.

[ZEP07] O. Zaidan, J. Eisner, and C. Piatko. Using "annotator rationales" to improve machine learning for text categorization. Proc. NAACL HLT, pages 260–267, 2007.

[ZE08] O. Zaidan and J. Eisner. Modeling annotators: A generative approach to learning from annotator rationales. Proc. EMNLP, pages 31–40, 2008.