Richer Annotation?

- Usually, an annotator indicates *what* the correct answer is.
- We propose the annotator *also* indicate *why*.

⇒ Each training example provides data about its class *and* why.
⇒ Richer annotation provides more data.

- Idea #1: richer annotation can aid ML.
- Idea #2: richer better use of our time than more.
Rationales in Text Categorization

The following segments were taken from movie reviews. Did the reviewer have a positive or negative opinion of the movie?

- Trust me,  American Pie.
- He continues to be of the most exciting artists on the big screen, performing his own stunts and
- ...and the romance was
- The movie is that even the most casual viewer may notice the
- ...and it even makes watching Eddie Murphy

---

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 rating is R for Reality. Tom Hanks is as Capt 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 come April, as with be in Tom's possession.
**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 "return to sender".

---

**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 Capt 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, come April, as with be in Tom’s possession.

---

How sure was the annotator that this is a positive review?

How sure was the annotator that this is a positive review?

If a rationale is masked out, the annotator would not be as sure that this is a positive review.

Intuition: a good model should also be less sure.
"Contrast" Examples

Intuition: a good model should be less sure of a positive classification on contrasts than on the original.

Our work: modified SVM that takes this intuition into account.

Standard SVM

Minimize:

$$\frac{1}{2}||w||^2$$

subject to:

$$w \cdot x_i \geq 1$$
**Incorporating Contrasts**

Minimize:

\[ \frac{1}{2} \|w\|^2 \]

subject to:

\[ w \cdot x_i \geq 1 \]

\[ w \cdot x_{ij} \geq 1 \]

**Slack Variables**

Minimize:

\[ \frac{1}{2} \|w\|^2 + C \sum \xi_i + C_{\text{contrast}} \sum \xi_{ij} \]

subject to:

\[ y_i (w \cdot x_i - w \cdot v_{ij}) \geq \mu(1 - \xi_j) \]

\[ y_i (w \cdot x_i - w \cdot v_{ij}) \geq 1 - \xi_j \]

**((Include Negative Examples))**

Minimize:

\[ \frac{1}{2} \|w\|^2 + C \sum \xi_i + C_{\text{contrast}} \sum \xi_{ij} \]

subject to:

\[ y_i (w \cdot x_i - w \cdot v_{ij}) \geq \mu(1 - \xi_j) \]

\[ y_i (w \cdot x_i - w \cdot v_{ij}) \geq 1 - \xi_j \]

**The Modified SVM**

Minimize:

\[ \frac{1}{2} \|w\|^2 + C \sum \xi_i + C_{\text{contrast}} \sum \xi_{ij} \]

subject to:

\[ y_i (w \cdot x_i - w \cdot v_{ij}) \geq \mu(1 - \xi_j) \]

\[ y_i (w \cdot x_i - w \cdot v_{ij}) \geq 1 - \xi_j \]
What this Means in Practice

Standard SVM cares about this margin

---

Recap

- Training examples: \((x_1, y_1), (x_2, y_2), \ldots\)
- \(y_i\) has \(n_i\) rationales: \(r_{i1}, r_{i2}, \ldots, r_{in}\)
- \(x_i\) gives \(n_i\) contrast examples: \(v_{i1}, v_{i2}, \ldots, v_{in}\)
  (obtain \(j^{th}\) contrast by masking out \(j^{th}\) rationale.)
- We extend the SVM to determine best hyperplane subject to:
  - Constraints for standard margin, and also
  - Constraints for original/contrast separating margin.

What this is **not**

- In tasks like digit recognition, one can "generate" more training data from the existing examples

Class-preserving transformations
What this is **not**

<table>
<thead>
<tr>
<th>Class-preserving</th>
<th>Not necessarily (contrast)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information from new examples similar to that from real examples</td>
<td>Information from contrast examples is of a different kind</td>
</tr>
<tr>
<td>Can get benefit by automatic preprocessing (rescale, deslant, etc)</td>
<td>Actually provides new information via human insight</td>
</tr>
</tbody>
</table>

The Dataset

- The movie review dataset (Pang & Lee)
  - 1000 positive reviews
  - 1000 negative reviews

- For each document, given the class annotation, we added the rationale annotation
  - Annotation process: in an HTML editor, rationale segments are boldfaced.

Annotation Time

- How big is the overhead for annotating rationales?
- Ought to establish that richer annotation is a good use of an annotator’s time.
  - vs. just annotating more documents
- One can imagine three annotation tasks:
  - **T1**: given document, annotate the class.
  - **T2**: given document and gold standard class, annotate the rationales.
  - **T3**: given document, annotate both the class and the rationales.
- 50 docs/task given to four annotators

Annotation Time

- **T1**: given document, annotate the class.
- **T2**: given document and gold standard class, annotate the rationales.
- **T3**: given document, annotate both the class and the rationales.
- We found that $\text{Time(T3)} \approx 2 \times \text{Time(T1)}$
- Even though on average 8.3 rationales/doc + class!
- Annotator already needs to find rationales to determine class. Extra work is only to make them explicit: $\text{Time(T3)} < \text{Time(T1)} + \text{Time(T2)}$ by about 20%
Annotation Time

- **T1**: given document, annotate the class.
- **T2**: given document and gold standard class, annotate the rationales.
- **T3**: given document, annotate both the class and the rationales.

- **Synergy**: $\text{Time(T3)} < \text{Time(T1)} + \text{Time(T2)}$

- Extra time reduced with better annotation setup (e.g., automatic boldfacing of highlighting, stylus, etc) or smart use of eye tracking.

- Note: the task of classifying full docs is a worst-case scenario for rationales.
  - Other tasks would have simpler/fewer rationales and more complex classes.

Feature Vector

<table>
<thead>
<tr>
<th>Word</th>
<th>Feature</th>
<th>val</th>
</tr>
</thead>
<tbody>
<tr>
<td>!</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>&quot;</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>a</td>
<td>333</td>
<td>0</td>
</tr>
<tr>
<td>a+</td>
<td>334</td>
<td>0</td>
</tr>
<tr>
<td>a-list</td>
<td>335</td>
<td>1</td>
</tr>
<tr>
<td>aaron</td>
<td>336</td>
<td>0</td>
</tr>
<tr>
<td>zoolander</td>
<td>17741</td>
<td>1</td>
</tr>
<tr>
<td>zorro</td>
<td>17742</td>
<td>0</td>
</tr>
<tr>
<td>zucker</td>
<td>17743</td>
<td>0</td>
</tr>
<tr>
<td>zwick</td>
<td>17744</td>
<td>0</td>
</tr>
</tbody>
</table>

Let's see some experimental results...
**Standard vs. Modified SVM**

- **Baseline**
- **Contrasts Introduced**
- **SVM**
- **Modified SVM**

- **Significantly different**
- **Removing rationales hurts performance**
- **Keeping only rationales hurts performance**
- **Rationales Only (as concatenations)**
Using Rationales from some (and not all) Documents

- We showed what happens if you use all the rationales in all the training documents.
- What if you use all the rationales from some training documents instead of all training documents?

Use Rationales from $R$ Documents

$T = 800$: Class annotation from 800 documents.
$R = 200$: Rationales from 200 documents only.

⇒ Pieces to solving classification puzzle cannot be found solely in the rationales
• Observation #1: much of the benefit can be obtained without annotating 100% of the documents
  – e.g. (0%, 50%, 100%) for $T=800$ and $T=1600$

• Observation #2: if you have a lot of training documents, adding more may not help much (curves flatten out). BUT adding more rationales provides a fresh benefit.
  ⇒ Benefit from $R$ even if $T$ "reaches its potential"
Simulating a “Lazy Annotator”

• In last few experiments, we kept all rationales from some training documents.
  – R=200 and T=800 means 600 documents contributed no contrast examples. Each of the 200 R documents contributes all its rationales.

• What if we keep some rationales from all documents?
  – Instead of using all the rationales in 200 documents, use the same number of rationales spread out over all 800 documents.


The two (T=800,R=200) points are comparable: same number of rationales. Difference is in distribution only.

Simulating a “Lazy Annotator”

• Experiment simulates a not-so-diligent annotator
  – This might be more common in reality.
  – Likely to pick ‘obvious’ rationales, yielding faster rationale annotation.
  – Also, obvious rationales may prove to be better.
    (Though experiment doesn’t test for that; rationales were picked at random)
Big Picture

- Idea #1: **richer annotation** can aid ML.
- Idea #2: **richer** better use of our time than **more**.
- Example of richer annotation: rationales.
- Developed and tested one method to use rationales (our extended SVM).
- Simulated degree of annotator laziness.

*Bonus*: annotator knows nothing about the ML method (or even feature set).

Future Work

- More datasets:
  - Different task may require different kind of rationales.
  - Might also require different annotation tool.
- More experiments:
  - Examination of annotation process.
  - Real experiments to see effect of a lazy annotator.
- More models:
  - Generative models: model annotation of rationales as a noisy process (annotators are not perfect).
  - Potentially other discriminative methods.

On The Internets

- The enriched dataset (and slides) here: http://cs.jhu.edu/~ozaidan/rationales

FYI – BibTeX entry:
@InProceedings{zaidan-eisner-piatko:2007:MainConf,
    author    = {Zaidan, Omar and Eisner, Jason and Piatko, Christine},
    title     = {Using "Annotator Rationales" to Improve Machine Learning for Text Categorization},
    booktitle = {Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference},
    month     = {April},
    year      = {2007},
    address   = {Rochester, New York},
    publisher = {Association for Computational Linguistics},
    pages     = {260--267},
    url       = {http://www.aclweb.org/anthology/N/N07/N07-0133} }