Plots of Children and Machines

Harry Halpin

September 26th, 2003
The Project

Overall Aims:

• **StoryStation**: Activating *Pinky the Plot Analyzer* for the story rewriting task.

• **Story rewriting task**: Comparing two stories for similarity. One story is the exemplar, the other story is a story rewritten from the exemplar.

• Not automated grading per se, but using automated plot evaluation to enable agent to help student.

• Agent suggests concrete changes or suggests student ask teacher for help.
Contributions of MSc. Project

- Implemented plot evaluation component, compared three different algorithms and three different machine-learners for computing event similarity.

- Not yet integrated into StoryStation...but will be!

Theoretical Hypothesis

Combine Formal Symbolic with Statistical Approach for better results.
Formal Symbolic Approach

**Formal:** to make explicit.

**Symbolic:** to use some mark to denote a discrete referent.

**In Project:** *Event Calculus, Plot-Comparison Algorithm*

Loss of Information in Event Calculus: Can’t discriminate between causality and temporality!

- *Dave fell, after* Chris pushed him.
- *Dave fell, then* Chris pushed him.
Statistical Approach

**Statistical:** Uses probabilities to describe either a class or the underlying data (i.e. words) themselves.

**Not explicit:** The properties are in the numbers!

The LSA “bag of words” approach lost information:

**Example:** “Nils leaves. City disappears.” = “Leaves City. Nils disappears.”
Uses of Statistics in Project

- Statistical **features** of the story: *LSA Scores*

- Statistical **machine-learners**: *Naive Bayes, MaxEnt, K-Nearest Neighbors*

**Example**: *LSA Scores* $\Rightarrow$ *K-Nearest Neighbors* $\Rightarrow$ *Grade*

Machine-learner always statistical. Features either statistical (*LSA*) or symbolic (*event calculus*).
Symbolic Story Representation

There are different symbolic frameworks for different purposes.

**Aim:** Automatic translation of story into knowledge representation.

Used **event calculus** because it captured relevant aspects of plot.

**Such as:** *Temporal flow, characters, events*

**Example:** \( \text{stay}(t=3)(\text{Nils, Sweden, geese}) \) where \( t \) is “time” in narrative.
XML-based Pipeline

Automatically abstracted from the data via a pipeline that uses mostly off-the-shelf tools.

Pipeline:
1. Word and Sentence Tokenization
2. POS Tagging and Lemmatization
3. Pronominal Resolution
4. Chunking to get tuples
5. Transformation of tuples into Event Calculus

The tools constrain the event-calculus.

Example: “Nils stays in Sweden and he is always playing with geese up the mountain.” ⇒ stay(Nils, Sweden) and play(geese, mountain)
Results

Grade matches with human rater using 10-fold cross validation:

- **Baseline**: 20.38%
- **K-Nearest Neighbors** *(LSA Alone)*: 44.66%
- **Naive Bayes** *(LSA with Event Calculus)*: 54.36%

Three-fold increase in accuracy. Actually more complex...
Sample Results

This confusion matrix shows the distribution of classes output by one particular machine-learner, the automatic, fast, fairly accurate Naive Bayes.

<table>
<thead>
<tr>
<th>Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Excellent)</td>
<td>0</td>
<td>17</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2 (Good)</td>
<td>1</td>
<td>29</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3 (Fair)</td>
<td>0</td>
<td>13</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>4 (Poor)</td>
<td>0</td>
<td>8</td>
<td>3</td>
<td>22</td>
</tr>
</tbody>
</table>

*Automatic Naive Bayes: Confusion Matrix*

Collapsing all categories except “good” and “poor”, but successful at classifying those. This level is useful for teachers, and approaches human rater agreement.
Explanation

Problems: Many fairly subjective classes (grades), small dataset, hard to tell what the cause of effects are precisely.

Machine-learning Explanation: Using symbolic knowledge representations shrinks the feature space that has to be searched, bringing prominent features to light.

Symbolic Explanation: The statistical encoding of subtle human judgements allowed the significance of particular parts to be evaluated with regards to real-world data - fighting the Frame Problem.

Conclusion: We need both! Multiple levels of abstraction should be combined. It’s hard to tell what level of abstraction is proper for a dataset, but I believe this method solves the story rewriting task in a broad sense. Still room for improvement in knowledge representation and statistical techniques.