Recovering Implicit Thread Structure in Newsgroup Style Conversations

Yi-Chia Wang¹, Mahesh Joshi¹, William W. Cohen², Carolyn Rosé³

¹ Language Technologies Institute
² Machine Learning Department
³ Language Technologies Institute /Human-Computer Interaction Institute

School of Computer Science
Carnegie Mellon University
Taking Advantage of Thread Structure

- Newsgroup search (Xi et al., 2004)
  - Meta features extracted from the discussion threads
- Email summarization (Carenini et al., 2007)
  - Quotation graph to organize conversations
- Thread visualization (Trausan-Matu et al., 2007)
  - Assist analysts in identifying salient parts
- Text classification (Wang et al., 2007)
  - ...
Discussion Forums and Thread Structure

- Often thread structure is explicitly represented
- Sometimes thread structure is implicit
- Our task:
  - Recover implicit thread structure
  - Investigate the contribution of time and similarity

---

Interesting...although he fails to mention in the article if it is the 16 or 8gb models.

As I posted yesterday there is a 5-7 day shipping on both the 8 and 16gb iPhones at Apple.com

http://store.apple.com/1-600-MY-APPLE/iphone

Don’t know about NYC, but I grabbed a 16GB iPhone yesterday at an ATT store (Denver). They had plenty in stock.

SR MBP 2.4
Mac Mini Core Duo 1.6GHz/2GB
PB 1.6GHz (DDR)/1GB/96GB
Mac Pro 2.66 Quad/4GB/1TB and ATI x1900

Last edited by MacDuck : 03-25-2008 at 09:12 PM. Reason: add info

Originally Posted by MacDuck:
Don’t know about NYC, but I grabbed a 16GB iPhone yesterday at an ATT store. They had plenty in stock.

Which ATT store...I had to go to the Park Meadows one yesterday to get their last one (or so I thought). I still haven’t busted the shrink wrap on it yet if Apple announces a price drop or new model...and articles like this make me want to wait even longer...but I’m so tempted.

Yeah, but that doesn’t sound like the sort of problem that gets fixed by learning C++, which has a nasty habit of adding loads of complexity in all the wrong places. If you want tight, fast code and your project doesn’t need OOP, use C.
LegSim Corpus

- Dear colleague letters (group specific e-mail messages)
- Committee hearings (on-line discussion threads and recorded instant messaging)
- Floor debates (on-line discussion threads)
- On-line assignments (such as the ‘journal’ where the instructor asks students to connect experiential activities to specific materials from the course)
- Legislation (bills; committee reports on legislation, floor amendments)

Data provided by John Wilkerson
University of Washington
LegSim Corpus

- Civics course in Fall 2006
- 71 students
- 48 separate discussion streams (478 messages)
Goal of Thread Recovery Task

- Assume we have a conversation stream in which messages are ordered by time.
- The goal is to link the referred parent message(s) and the responding child message
  - Parent-child relationship
- Two assumptions
  - Each message can only link back to those messages posted before it
  - No meta information is available

**R**

A: We should raise the federal minimum wage to $7.25 by 2008.

**M1**

B: I agree. Raising the minimum wage can help low-income family

**M2**

C: No. It will simply add more money to the economy, and devalue the dollar.

**M3**

D: Yes. A higher minimum wage is necessary in showing worker empathy by the government.

**M4**

E: Poor people cannot benefit by it. Increasing the minimum wage will increase the cost of living.

**M5**

F: I disagree with D. We should allow states to decide the minimum wage for themselves.

**M6**

A: I don’t think raising the minimum wage will increase the living cost nor devalue the dollar. Businesses would only be able to raise prices so much as the market will bear it.
Gold Standard of Thread Structure

- Manually reconstruct the parent-child links
  - Based on semantic relationship and discourse markers
- Inter-agreement of our link annotation scheme
  - Two human annotators
  - Cohen’s kappa: 0.87

A: We should raise the federal minimum wage to $7.25 by 2008.

M1: B: I agree. Raising the minimum wage can help low-income family.

M2: C: No. It will simply add more money to the economy, and devalue the dollar.

M3: D: Yes. A higher minimum wage is necessary in showing worker empathy by the government.

M4: E: Poor people cannot benefit by it. Increasing the minimum wage will increase the cost of living.

M5: F: I disagree with D. We should allow states to decide the minimum wage for themselves.

M6: A: I don’t think raising the minimum wage will increase the living cost nor devalue the dollar. Businesses would only be able to raise prices so much as the market will bear it.
Two Baselines for Thread Recovery

**Temporal information**

- **LinkPrevious**
- Select the immediately previous message as the parent

**Content similarity**

- **SimilarityOnly**

\[ w_{ij} = \begin{cases} \frac{\vec{m}_i \cdot \vec{m}_j}{\|\vec{m}_i\| \cdot \|\vec{m}_j\|}, & \text{if } i > j \\ 0, & \text{otherwise} \end{cases} \]

\[
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0.15 & 0 & 0 & 0 & 0 & 0 \\
0.30 & 0.10 & 0 & 0 & 0 & 0 \\
0.25 & 0.20 & 0.15 & 0 & 0 & 0 \\
0.05 & 0.20 & 0.25 & 0.10 & 0 & 0 \\
0.10 & 0.35 & 0.05 & 0.30 & 0.05 & 0
\end{pmatrix}
\]
Two Baselines for Thread Recovery

**Temporal information**

- **LinkPrevious**
- Select the immediately previous message as the parent

**Content similarity**

- **SimilarityOnly**

\[
W_{ij} = \begin{cases} 
\frac{\langle \vec{m}_i, \vec{m}_j \rangle}{\| \vec{m}_i \| \cdot \| \vec{m}_j \|}, & \text{if } i > j \\
0, & \text{otherwise}
\end{cases}
\]

From

\[
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0.15 & 0 & 0 & 0 & 0 & 0 \\
0.30 & 0.10 & 0 & 0 & 0 & 0 \\
0.25 & 0.20 & 0.15 & 0 & 0 & 0 \\
0.05 & 0.20 & 0.25 & 0.10 & 0 & 0 \\
0.10 & 0.35 & 0.05 & 0.30 & 0.05 & 0
\end{pmatrix}
\]

To
Combinations of Time and Similarity

- **FixedWindow**
- Only messages within a certain window size are considered to be candidate parents

<table>
<thead>
<tr>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.35</td>
</tr>
<tr>
<td>0.05</td>
<td>0.20</td>
</tr>
<tr>
<td>0.25</td>
<td>0.15</td>
</tr>
<tr>
<td>0.30</td>
<td>0.10</td>
</tr>
<tr>
<td>0.15</td>
<td>0.00</td>
</tr>
<tr>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Combinations of Time and Similarity

- **FixedWindow**
- Only messages within a certain window size are considered to be candidate parents

\[
\begin{array}{cccccc}
\text{From} & 0 & 0.15 & 0.30 & 0 & 0.20 & 0.25 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.15 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.30 & 0.10 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.20 & 0.15 & 0 & 0 & 0 & 0 \\
0 & 0 & 0.25 & 0.10 & 0 & 0 & 0 \\
0 & 0 & 0 & 0.30 & 0.05 & 0 & 0 \\
\end{array}
\]
Combinations of Time and Similarity

- **FixedWindow**
- Only messages within a certain window size are considered to be candidate parents

- **TimeDistance**
- Penalize the similarity by the time distance between the two messages

From

| From | 0 | 0 | 0 | 0 | 0 | 0 | 0.15 | 0 | 0 | 0 | 0 | 0 | 0.30 | 0.10 | 0 | 0 | 0 | 0 | 0 | 0.20 | 0.15 | 0 | 0 | 0 | 0 | 0 | 0.25 | 0.10 | 0 | 0 | 0 | 0 | 0.30 | 0.05 | 0 | 0 | 0 |
To    | 0 | 0 | 0 | 0 | 0 | 0 | 0.15 | 0 | 0 | 0 | 0 | 0 | 0 | 0.30 | 0.10 | 0 | 0 | 0 | 0 | 0 | 0.20 | 0.15 | 0 | 0 | 0 | 0 | 0 | 0.25 | 0.10 | 0 | 0 | 0 | 0 | 0 | 0.30 | 0.05 | 0 | 0 | 0 |

Apr-1 ICWSM2008, Seattle 12
Combination of Time and Similarity

- **FixedWindow**
  - Only messages within a certain window size are considered to be candidate parents

- **TimeDistance**
  - Penalize the similarity by the time distance between the two messages
Evaluation Metrics

Precision \[= \frac{|real\_links \cap predicted\_links|}{predicted\_links}\]

Recall \[= \frac{|real\_links \cap predicted\_links|}{real\_links}\]

F-measure \[= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\]
Effect of Tuning Similarity or Time Thresholds

**Similarity Only**
(Highest F-measure = .28 while threshold = .165)

**Fixed Window**
(Highest F-measure = .39 while window size = 3)
Evaluation for Each Approach

- Leave-one-thread-out cross validation
- Significant difference between Baseline and Fixed Window approach

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>LinkPrevious</th>
<th>SimilarityOnly</th>
<th>FixedWin</th>
<th>TimeDistance</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>0.36</td>
<td>0.28</td>
<td>0.39</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Threads of Different Lengths

Thread length increases

- Performance decreases
- LinkPrevious is effective for short threads

In the most difficult two bins, FixedWindow is most effective

Approach within Bin

1. FixedWindow
2. LinkPrevious
3. SimilarityOnly

2-5
6-13
14-70
Different Average Message Lengths

- Average message length is computed by averaging the parent and child message length within the pair.

Performance increases

FixedWin is most useful for medium size messages

Message length increases

- 0-40
- 60
- 85
- 120
- >120
R
I propose to celebrate the success of the Idaho Potato and remind all Americans how great they are.

M1 I vote “yes.”

M2 I vote yes.

M3 yes

M4 I vote no on the premise of spending. I just don't know if this is a one time fixed cost or an annual amount...

M5 Yes

M6 yes

M7 The bill passes.

M8 I like this bill. But what is the 10k in appropriations going to be used for?

M9 Yeah, I don't think this bill really matters. So pass it, whatever. I don't think 10K will get them anywhere, and I doubt that appropriations will give them anything...but then again we do have an imaginary budget...
Contributions

• Compare several approaches for recovering thread structure
  • Contribution of time and similarity are investigated
  • Both time and similarity are important
    • FixedWindow is significantly better than SimilarityOnly
• Fine-grained analysis
  • FixedWindow is advantageous for recovering longer threads
  • FixedWindow is most useful for medium size messages
Current Work

• For the similarity comparisons, taking discourse focus (i.e. salience) into account to select the most relevant part of longer messages

• Taking discourse function (i.e. conversation act) of contributions into account
  • Proposal, Counter-proposal, Acceptance, Rejection …

• Investigating the thread recovery problem on other kinds of social media, such as web blogs
Question?

My Homepage: www.cs.cmu.edu/~yichiaw
My Email: yichiaw@cs.cmu.edu