Aggregation of Multiple Judgments for Evaluating Ordered Lists

ECIR’10 Mar 31, 2010

Hyun Duk Kim (hkim277@illinois.edu)
ChengXiang Zhai (czhai@cs.uiuc.edu)
Jiawei Han (hanj@cs.uiuc.edu)

Department of Computer Science
University of Illinois, Urbana-Champaign
How to evaluate the performance of a new system?

• Compare with human judgment

Input → New System → Human assessor → System Output → Gold Standard → More similar → Better system
Ordered List

• Important IR problem
  – Search engine ranking
  – Sentence ordering of summarization

New System

\[ O_{SYS} : F \ C \ E \ D \]

Human assessor

\[ O_{H} : F \ C \ D \ E \]

\[ s(O_{SYS}; O_{H}) \]

More similar
Better system
For difficult / subjective task

• More human judgments $\rightarrow$ reliable evaluation

Disagreement among human judgments

System

$O_{SYS}$ : C D E F

$O_{H1}$ : C D F E

$O_{H2}$ : E C D F

$O_{H3}$ : F C D E

Need to look into sub-structure

How to handle for evaluation?
Existing Solution 1

• 1 target vs. 1 gold standard
  Multiple evaluation averaged

Coarse Aggregation
Cannot handle human disagreement
Existing Solution 1

• If there is noise?
  – Biased / Lazy assessors

\[ s(O_{SYS}; O_{H1}) + s(O_{SYS}; O_{H2}) + s(O_{SYS}; O_{H3}) \]

\[
\frac{s(O_{SYS}; O_{H1}) + s(O_{SYS}; O_{H2}) + s(O_{SYS}; O_{H3})}{3}
\]

Noise \(\rightarrow\) Low quality
Existing Solution 2

• Pyramid method [Nenkova et al, 07]
  – Automatic summarization evaluation
  – Each unit has weights based on voting

\[\text{Input} \rightarrow \{A, C\} \quad \text{System 1 is better} \]
\[\{B, C\} \]
\[\{A, B, C\} \quad \text{Because ‘A’ received more votes} \]
\[\{A, B\} \]
\[\{A, C\} \]
Existing Solution 2

- Pyramid method [Nenkova et al, 07]
  - Automatic summarization evaluation
  - Each unit has weights based on voting

Not applicable for ordered list

- How to find agreement?
- How to weight?

Because ‘A’ received more votes
Judgments Aggregation for Ordered Lists

1. Weighted correlation aggregation (WCA)
   - Noise handling

2. Rank-based aggregation (RBA)
   - Noise handling + Look into sub structure

3. Frequent sequential pattern-based aggregation (FreSPA)
   - Noise handling + More Look into sub structure
Rest of the Talk

1. Methods
   - Weighted correlation aggregation (WCA)
   - Rank-based aggregation (RBA)
   - Frequent sequential pattern-based aggregation (FreSPA)

2. Experiments

3. Related work & Conclusion
Methods
Basic Measure

• 1:1 order comparison measures
  – Kendall’s tau [Lapata`06]
    \[
    \tau = 1 - \frac{2S(\pi, \sigma)}{N(N - 1) / 2}, \quad S(\pi, \sigma) : \# \text{ of disordant pairs between } \pi \text{ and } \sigma
    \]
  – Spearman’s rank correlation coefficient
    \[
    \text{Spearman} = 1 - \frac{6\sum_{i=1}^{N}(\pi(i) - \sigma(i))^2}{N(N^2 - 1)}, \quad \pi(i), \sigma(i) : \text{Rank of } i \text{ in } \pi \text{ and } \sigma
    \]
Basic Measure

• Average Correlation (AC)

\[ S_{AC}(O_S; O_{H1}, ..., O_{Hn}) = \frac{1}{n} \sum_{i=1}^{n} C(O_S, O_{Hi}) \]

\[ C(O_1, O_2) : \text{Correlation between two orderings} \]

\[ \rightarrow \text{AC- } \tau : \text{Kendall’s } \tau \]

\[ \text{AC-Sp: Spearman’s rank correlation} \]
Proposed Methods

1. Weighted correlation aggregation (WCA)

2. Rank-based aggregation (RBA)

3. Frequent sequential pattern-based aggregation (FreSPA)
1. Weighted correlation aggregation (WCA)

• More similar to other judges $\Rightarrow$ Higher weights

$$O_S \rightarrow O_{H1} \rightarrow O_{H2} \rightarrow O_{H3}$$

$$C(O_S, O_{H1}) \times w_1$$

$$C(O_S, O_{H2}) \times w_2$$

$$C(O_S, O_{H3}) \times w_3$$

$w_1, w_2, w_3$

Weighted Average Score
1. Weighted correlation aggregation (WCA)

\[ S_{WCA}(O_S; O_{H1}, \ldots, O_{Hn}) = \frac{\sum_{i=1}^{n} w_i C(O_S, O_{Hi})}{\sum_{i=1}^{n} w_i} \]

\[ w_i = \frac{1}{n} \sum_{j=1, j \neq i}^{n} C(O_{Hi}, O_{Hj}) \]

\[ C(O_1, O_2) : \text{Correlation between two orderings} \]
\[ \rightarrow \text{WCA-} \tau : \text{Kendall’s } \tau \]
\[ \text{WCA-Sp: Spearman’s rank correlation} \]

• WCA captures overall agreement.
Can we look into sub-structure?
2. Rank-based aggregation (RBA)

• Aggregate all ranks $\rightarrow$ Combined list

$OS \rightarrow \sum_{j=1}^{n} \text{Rank}_j(x_i)$

$OH1 \rightarrow O_{C}$

$OH2 \rightarrow O_{H3}$

$OC \rightarrow C(OS, OC)$

Kendall’s $\tau$ or Spearman

Combined Ranking Score of $x_i = \sum_{j=1}^{n} \text{Rank}_j(x_i)$

$\text{Rank}_j(x_i)$ is the rank of $x_i$ in $O_{Hj}$
2. Rank-based aggregation (RBA)

- Rank in combined list = sum of ranks in judgments
  [Similar to Borda Count]

<table>
<thead>
<tr>
<th>Rank</th>
<th>$O_{H1}$</th>
<th>$O_{H2}$</th>
<th>$O_{H3}$</th>
<th>$O_C$</th>
<th>Rank Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C</td>
<td>C</td>
<td>E</td>
<td>C</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>D</td>
<td>C</td>
<td>D</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>E</td>
<td>F</td>
<td>D</td>
<td>E</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>F</td>
<td>E</td>
<td>F</td>
<td>F</td>
<td>11</td>
</tr>
</tbody>
</table>

- Compensate small errors
  Can we capture at various levels of granularity?
3. Frequent sequential pattern-based aggregation (FreSPA)

- Evaluate with partial order represented by frequent sequential patterns

\[ O_S \rightarrow CD, EF, CDF \]
\[ O_{H1} \rightarrow CD, DF \]
\[ O_{H2} \rightarrow \ldots \]
\[ O_{H3} \rightarrow CDE, CDF \]

**Agreed Partial Orders**

**Weighted Scoring With P**
3. FreSPA

- Step1: Find frequent sequential pattern (sup ≥ minSup)

Frequent Sequential Pattern Mining

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>C D</td>
<td>3</td>
</tr>
<tr>
<td>D F</td>
<td>3</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>C D E</td>
<td>2</td>
</tr>
<tr>
<td>C D F</td>
<td>3</td>
</tr>
<tr>
<td>E C D F</td>
<td>1</td>
</tr>
</tbody>
</table>
FreSPA

• Step 2: Calculate score for target list with patterns mined.

```
<table>
<thead>
<tr>
<th>Pattern</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>C D</td>
<td>3</td>
</tr>
<tr>
<td>D F</td>
<td>3</td>
</tr>
<tr>
<td>C D E</td>
<td>2</td>
</tr>
<tr>
<td>C D F</td>
<td>3</td>
</tr>
</tbody>
</table>
```

System 1 is better
3. FreSPA

- Step2: Calculate score for target list with patterns mined.

1. CD : 3
2. DF : 3
3. CDE : 2
4. CDF : 3

- More frequent pattern:
- Longer pattern:
- More pattern:
3. FreSPA

• Step 2: Calculate score for target list with patterns mined.

\[
S_{FreSPA}(O; O_1, \ldots, O_n) = \sum_{pat_i \in O} \frac{\sum_{pat_i \in P} (1 + wLen^*(seqLen_i - 1)) \times (1 + wSup^*(sup_i - 1))}{\sum_{pat_i \in P} (1 + wLen^*(seqLen_i - 1)) \times (1 + wSup^*(sup_i - 1))}
\]

*wLen*: weight on length  \hspace{1cm}  *seqLen_i*: length  

*wSup*: weight on support  \hspace{1cm}  *sup_i*: support
Experiments
Experiment

• Data set
  – Sentence order set [Barzilay et al.‘01‘02]
  – 10 sets of sentence sequence
    • Average 8.8 sentence, 10.4 human judgment per set

• Baseline
  – AC-τ, AC-Sp

• Proposed Method
  – WCA-τ, WCA-Sp
  – RBA-τ, RBA-Sp
  – FreSPA
Metric

• Evaluation Discriminativeness (ED)
  – Score for the good order – score for the bad order
  – High ED → Better evaluation method
  – Use one of the human judgment as a good order, the reverse of it as a bad order

\[
ED = \frac{1}{n} \sum_{i=1}^{n} \left( S(O_i; O_1, \ldots, O_{i-1}, O_{i+1}, \ldots, O_n) - S(O_{iR}; O_1, \ldots, O_{i-1}, O_{i+1}, \ldots, O_n) \right)
\]

• Robustness to noisy judgments
  – Add random artificial accessors and check ED
  – Less affected by noise → More reliable evaluation method
Basic Comparison

<table>
<thead>
<tr>
<th></th>
<th>AC-τ</th>
<th>AC-Sp</th>
<th>WCA-τ</th>
<th>WCA-Sp</th>
<th>RBA-τ</th>
<th>RBA-Sp</th>
<th>FreSPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED</td>
<td>0.454</td>
<td>0.542</td>
<td>0.459</td>
<td>0.549</td>
<td>0.564</td>
<td>0.676</td>
<td>0.722</td>
</tr>
</tbody>
</table>

No noise,
FreSPA: wLen=1.0, wSup=1.0, minSup=0.75, minLen=2, maxLen=k

FreSPA > RBA > WCA > Baselines → Aggregating at finer granularity is more effective
Comparison on Noisy Data

Robustness to noise:
FreSPA > RBA > WCA > Baselines
Parameter Setting of FreSPA

- **ED with different wLen**

  ![Graph showing ED with different wLen](image)

  - All patterns
  - Equally important, stable performance

  - In noisy situation, longer pattern more trustable
Parameter Setting of FreSPA

- ED with different wSup

Can entirely trust supports computed, Stable performance

Cannot entirely trust supports computed
Related Work & Conclusion
Related Work

• Data fusion in IR
  – CombSUM/CombMNZ [Fox and Shaw’93], ProbFuse [Lillis et al.’06],
    Generative model [Efron’09], Unsupervised learning [Klementiev’07]

• Aggregation of judgments and votes
  – Borda count, Condorcet count
    [Lang’07, Dietrich and List’05, Hartmann and Sprenger’08, Drissi and Truchon’02]

• Pyramid method for summary content evaluation
  [Nenkova et al.’07]

• Evaluating sentence ordering
  – Qualitative by human [Okazaki et al.’04, Bollegala et al.’06],
    Kendall’s τ, Spearman’s rank correlation coefficient

• Frequent sequential pattern mining
  – GSP [Srikant and Agrawal’96], SPADE [Zaki’01], PrefixSpan [Pei et al.’01],
    CloSpan [Yan et al.’03]

• Subjectivity in human evaluators’ annotation
  [Reidsma et al.’08, Wilson’08, Beigman et al.’08, Passonneau et al.’08, Wiebe et al.’99]
Conclusion

• Three aggregation evaluation methods
  1. Weighted correlation aggregation (WCA)
  2. Rank-based aggregation (RBA)
  3. Frequent sequential pattern-based aggregation (FreSPA)

• Effectively aggregate multiple human annotation
  – High discriminativeness
  – Robust to noise

• Additional benefit of FreSPA
  – Parameter can be optimized flexibly
Future work

• More principled pattern weighting based on probabilistic significance

• Possibility of extension to non-ordering task evaluation
  – Ex. Summary evaluation using frequent patterns

• Adopt other rank fusion algorithms
  – Instead of RBA
Thank you

University of Illinois at Urbana-Champaign
Hyun Duk Kim (hkim277@illinois.edu)
ChengXiang Zhai (czhai@cs.uiuc.edu)
Jiawei Han (hanj@cs.uiuc.edu)
References

References


References

References

Backup Slides
Sentence ordering

• Last step of automatic summarization
• Find coherent order for reader
Problem Definition
Problem Definition

- **k items to be ordered**: $X = \{x_1, \ldots, x_k\}$
- **Ordered list to be evaluated**: $O = (y_1, \ldots, y_k)$
- Orders by $n$ human assessors

- **Goal**: find a score of $O$, $s(O; O_1, \ldots, O_n)$
- based on $O_1, \ldots, O_n$
2. Rank-based aggregation (RBA)

Combined Ranking Score of \( x_i = \sum_{j=1}^{n} \text{Rank}_j(x_i) \)

\( \text{Rank}_j(x_i) \) is the rank of \( x_i \) in \( O_j \)

- Score is correlation between a target and the combined orderings
  - \( \text{RBA-} \tau \) : Kendall’s \( \tau \)
  - \( \text{RBA-Sp} \) : Spearman’s rank correlation

- Compensate small errors
  Can we capture at various levels of granularity?
3. FreSPA

• Step 2: Calculate score for target list with patterns mined.
  – List having more patterns has higher score
  – Longer pattern more important
    • Longer sequential agreement: Difficult to happen
  – Higher support pattern more important
    • High Support → More agreement
Robustness to noise:
FreSPA > RBA > WCA > Baselines
## Basic comparison

<table>
<thead>
<tr>
<th>Noise ratio</th>
<th>AC-τ</th>
<th>AC-Sp</th>
<th>WCA-τ</th>
<th>WCA-Sp</th>
<th>RBA-τ</th>
<th>RBA-Sp</th>
<th>FreSPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.454</td>
<td>0.542</td>
<td>0.459</td>
<td>0.549</td>
<td>0.564</td>
<td>0.676</td>
<td>0.722</td>
</tr>
<tr>
<td>0.25</td>
<td>0.300</td>
<td>0.354</td>
<td>0.357</td>
<td>0.435</td>
<td>0.460</td>
<td>0.564</td>
<td>0.656</td>
</tr>
<tr>
<td>0.5</td>
<td>0.220</td>
<td>0.264</td>
<td>0.236</td>
<td>0.292</td>
<td>0.380</td>
<td>0.468</td>
<td>0.551</td>
</tr>
<tr>
<td>0.75</td>
<td>0.167</td>
<td>0.202</td>
<td>0.173</td>
<td>0.215</td>
<td>0.358</td>
<td>0.433</td>
<td>0.510</td>
</tr>
<tr>
<td>1</td>
<td>0.113</td>
<td>0.136</td>
<td>0.133</td>
<td>0.159</td>
<td>0.307</td>
<td>0.379</td>
<td>0.395</td>
</tr>
<tr>
<td>Max-Min</td>
<td>0.341</td>
<td>0.406</td>
<td>0.326</td>
<td>0.389</td>
<td>0.257</td>
<td>0.297</td>
<td>0.327</td>
</tr>
<tr>
<td>% Degradation</td>
<td>75.1</td>
<td>74.9</td>
<td>71.0</td>
<td>70.9</td>
<td>45.6</td>
<td>43.9</td>
<td>45.3</td>
</tr>
</tbody>
</table>
### ED with different wLen and wSup

<table>
<thead>
<tr>
<th>wLen</th>
<th>ED (original)</th>
<th>ED (noisy)</th>
<th>wSup</th>
<th>ED (original)</th>
<th>ED (noisy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.7301</td>
<td>0.5255</td>
<td>0</td>
<td>0.7106</td>
<td>0.5704</td>
</tr>
<tr>
<td>0.25</td>
<td>0.7261</td>
<td>0.5517</td>
<td>0.25</td>
<td>0.7186</td>
<td>0.5145</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7239</td>
<td>0.4736</td>
<td>0.5</td>
<td>0.7203</td>
<td>0.5504</td>
</tr>
<tr>
<td>0.75</td>
<td>0.7225</td>
<td>0.5327</td>
<td>0.75</td>
<td>0.7211</td>
<td>0.5141</td>
</tr>
<tr>
<td>1</td>
<td>0.7215</td>
<td>0.5482</td>
<td>1</td>
<td>0.7215</td>
<td><strong>0.6019</strong></td>
</tr>
<tr>
<td>5</td>
<td>0.7177</td>
<td>0.5642</td>
<td>5</td>
<td>0.7227</td>
<td>0.5694</td>
</tr>
<tr>
<td>10</td>
<td>0.7169</td>
<td>0.5828</td>
<td>10</td>
<td>0.7229</td>
<td>0.5616</td>
</tr>
<tr>
<td>20</td>
<td>0.7165</td>
<td>0.5874</td>
<td>20</td>
<td>0.7230</td>
<td>0.5275</td>
</tr>
<tr>
<td>50</td>
<td>0.7163</td>
<td><strong>0.6120</strong></td>
<td>50</td>
<td>0.7230</td>
<td>0.5361</td>
</tr>
<tr>
<td>100</td>
<td>0.7162</td>
<td>0.5403</td>
<td>100</td>
<td><strong>0.7230</strong></td>
<td>0.5130</td>
</tr>
</tbody>
</table>
### Parameter Setting of FreSPA

- **ED with different wLen**

<table>
<thead>
<tr>
<th>wLen</th>
<th>ED (original)</th>
<th>ED (noisy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.7301</td>
<td>0.5255</td>
</tr>
<tr>
<td>0.25</td>
<td>0.7261</td>
<td>0.5517</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7239</td>
<td>0.4736</td>
</tr>
<tr>
<td>0.75</td>
<td>0.7225</td>
<td>0.5327</td>
</tr>
<tr>
<td>1</td>
<td>0.7215</td>
<td>0.5482</td>
</tr>
<tr>
<td>5</td>
<td>0.7177</td>
<td>0.5642</td>
</tr>
<tr>
<td>10</td>
<td>0.7169</td>
<td>0.5828</td>
</tr>
<tr>
<td>20</td>
<td>0.7165</td>
<td>0.5874</td>
</tr>
<tr>
<td>50</td>
<td>0.7163</td>
<td><strong>0.6120</strong></td>
</tr>
<tr>
<td>100</td>
<td>0.7162</td>
<td>0.5403</td>
</tr>
</tbody>
</table>

- All patterns equally important
- In noisy situation, longer pattern more trustable
Parameter Setting of FreSPA

• **ED with different wSup**

<table>
<thead>
<tr>
<th>wSup</th>
<th>ED (original)</th>
<th>ED (noisy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.7106</td>
<td>0.5704</td>
</tr>
<tr>
<td>0.25</td>
<td>0.7186</td>
<td>0.5145</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7203</td>
<td>0.5504</td>
</tr>
<tr>
<td>0.75</td>
<td>0.7211</td>
<td>0.5141</td>
</tr>
<tr>
<td>1</td>
<td>0.7215</td>
<td>0.6019</td>
</tr>
<tr>
<td>5</td>
<td>0.7227</td>
<td>0.5694</td>
</tr>
<tr>
<td>10</td>
<td>0.7229</td>
<td>0.5616</td>
</tr>
<tr>
<td>20</td>
<td>0.7230</td>
<td>0.5275</td>
</tr>
<tr>
<td>50</td>
<td>0.7230</td>
<td>0.5361</td>
</tr>
<tr>
<td>100</td>
<td>0.7230</td>
<td>0.5130</td>
</tr>
</tbody>
</table>

Cannot entirely trust supports computed

Can entirely trust supports computed
## ED with different maxLen and minLen ratios

<table>
<thead>
<tr>
<th>maxLen/k (minLen=2)</th>
<th>ED (original)</th>
<th>ED (noisy)</th>
<th>minLen/k (maxLen=k)</th>
<th>ED (original)</th>
<th>ED (noisy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.7493</td>
<td>0.5265</td>
<td>0</td>
<td>0.7215</td>
<td>0.5841</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5818</td>
<td>0.4362</td>
<td>0.25</td>
<td>0.7215</td>
<td>0.5361</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7267</td>
<td>0.5331</td>
<td>0.5</td>
<td>0.4137</td>
<td>0.0707</td>
</tr>
<tr>
<td>0.75</td>
<td>0.7215</td>
<td>0.5320</td>
<td>0.75</td>
<td>0.0824</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>0.7215</td>
<td>0.5303</td>
<td>1</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
ED with different minimum support

<table>
<thead>
<tr>
<th>minSup</th>
<th>ED (original)</th>
<th>ED (noisy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.1741</td>
<td>0.0815</td>
</tr>
<tr>
<td>0.25</td>
<td>0.3678</td>
<td>0.2639</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5328</td>
<td>0.4040</td>
</tr>
<tr>
<td>0.75</td>
<td>0.7215</td>
<td>0.5855</td>
</tr>
<tr>
<td>1</td>
<td>0.8985</td>
<td>0.0230</td>
</tr>
</tbody>
</table>
## Parameter Setting of FreSPA

### • ED with different maxLen and minLen ratios

<table>
<thead>
<tr>
<th>maxLen/k (minLen=2)</th>
<th>ED (original)</th>
<th>ED (noisy)</th>
<th>minLen/k (maxLen=k)</th>
<th>ED (original)</th>
<th>ED (noisy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.7493</td>
<td>0.5265</td>
<td>0</td>
<td>0.7215</td>
<td>0.5841</td>
</tr>
<tr>
<td>0.25</td>
<td>0.5818</td>
<td>0.4362</td>
<td>0.25</td>
<td>0.7215</td>
<td>0.5361</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7267</td>
<td>0.5331</td>
<td>0.5</td>
<td>0.4137</td>
<td>0.0707</td>
</tr>
<tr>
<td>0.75</td>
<td>0.7215</td>
<td>0.5320</td>
<td>0.75</td>
<td>0.0824</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>0.7215</td>
<td>0.5303</td>
<td>1</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Generally using all the patterns desirable
maxLen=k, minLen=2
Parameter Setting of FreSPA

• **ED with different minSup**

![Graph showing the relationship between ED and minSup for original and noisy data.](image)

Better in higher minSup

Cannot be minSup=1 for noisy data