Summarization Evaluation Using Transformed Basic Elements

Stephen Tratz and Eduard Hovy

Information Sciences Institute
University of Southern California
History

• BLEU: ngrams for machine translation eval (Papineni et al., 2002)
• ROUGE: ngrams for text summarization eval (Lin and Hovy, 2003)
• Basic Elements (BE): short syntactic units for summarization eval (Hovy et al. 2006)
• ParaEval (Zhou et al. 2006)
• BEwT-E: Basic Elements with Transformations for Evaluation
ROUGE

• N-gram approach to summarization evaluation
  – Count ngram overlaps between peer summary and reference summaries
  – Various kinds of ngrams: unigrams, bigrams ... ‘skip’ ngrams

• Recall-oriented: measure percentage of reference text ngrams covered
  – In contrast, BLEU is precision oriented: measure percentage of peer text (translation) ngrams covered
  – Recall is appropriate for summarization
Problems with ROUGE

• Same information conveyed in many different ways
  – Information omitted, word order rearranged, names abbreviated, etc.

• N-gram matching restricted to surface form
  – “large green car” ! = “large car”
  – “large green car”! = “heavy emerald vehicle”
  – “USA” ! = “United States”, “America”
Basic Elements

• Uses syntax to capture long range dependencies, avoid the locality limitations of ngrams

• Original BE system uses syntactically-related word pairs

• New BE system's Basic Elements vary in length
  – Unigram BEs: nouns, verbs, and adjs
  – Bigram BEs: like original system
  – Trigram BEs: two head words plus prep
BEwT-E

• Overview:
  – Read, Parse, perform NER
  – Identify minimal syntactic units independently ([large car], [green car], etc.) — **Basic Elements** (BEs)
  – Apply **transformations** to each BE
  – Match against reference set
  – Compute recall as **evaluation** score
Pre-processing

1. Basic data cleanup (e.g. canonicalize quote characters)

2. Parsing
   - Charniak parser (Charniak and Johnson, 2005)
   - Using a non-Treebank-style parser would require modified rules to extract BEs from parse tree

3. Named Entity Recognition
   - LingPipe (Baldwin and Carpenter)
BE Extraction

• TregEx: Regular expressions over trees
  – (Levy and Andrew, 2006)
  – BE extraction TregEx rules built manually

John’s cat drank milk.

Charniak parse:
(S1 (S (NP (NP (NNP John) (POS ‘s)) (NN cat)) (VP (VBD drank) (NP (NN milk)))) (. .)))

Rule Name: Verb to NPHead
Tregex: VP [<# __=x & < (NP <# !POS=y)]
Tokens to Extract: xy
Extracted BEs: drankVBD+milkNN

Rule Name: Possessor of NPHead
Tregex: NP [< (NP <# (POS $- __=x)) & <# __=y]
Tokens to Extract: xy
Extracted BEs: John1Person+cat1NN
Transformations 1

• 15 transformations implemented:
  – Lemma-based matching
    • "running" vs "ran"
  – Synonyms
    • "jump" vs "leap"
  – Preposition generalization
    • "book on JFK" vs "book about JFK"
  – Abbreviations
    • "USDA" vs "US Department of Agriculture"
    • "mg" vs "milligram"
  – Add/Drop Periods
    • "U.S.A." vs "USA"
Transformations 2

- Hyper/Hyponyms
  - "news" vs "press"

- Name Shortening/Expanding
  - "Mr. Smith" vs "John" vs "John S. Smith"
  - "Google Inc." vs "Google"

- Pronouns
  - "he" vs "John", "they" vs "General Electric"

- "Pertainyms"
  - "biological" vs "biology", "Mongol" vs "Mongolia"

- Capitalized Membership Mero/Holonyms
  - "China" vs "Chinese"
Transformations 3

- Swap IS-A nouns
  • ”John, a writer ...,” vs ”a writer, John ...,”
- Prenominal Noun <-> Prepositional Phrase
  • ”refinery fire” <-> ”fire in refinery”
- ”Role”
  • ”Shakespeare authored” <-> ”author Shakespeare”
- Nominalization / Denominalization
  • ”gerbil hibernated” → ”hibernation of gerbil”
  • ”invasion of Iraq” → ”Iraq invasion”
- Adjective <-> Adverb
  • [”effective treatment”, ”effective at treating”] vs ”effectively treat”
Transformation pipeline

- Many paths through pipeline
- Different ordering of transformations may affect results
- Each transformed BE is passed to all remaining transformations; results gathered at end
Duplicates and Weighting

Include duplicates: Yes or No?

BE weights based upon number of references containing the BE
• All BEs worth 1
• Total number of references it occurs in
• $\sqrt{T}$ (Total number of references it occurs in)
Calculating scores

• As result of transformations, each BE may match multiple reference BEs

• Require that each BE may match at most one reference BE

• Search to find optimal matching

• Weighted assignment problem

\[
\begin{align*}
\text{maximize} & \quad \sum_{i=0}^{N} \sum_{j=0}^{M} C(i,j) W(j) x_{ij} \\
\text{subject to} & \\
\sum_{i=0}^{N} x_{ij} \in \{0,1\} & \text{forall } j \text{ where } 0 \leq j \leq M \\
\sum_{j=0}^{M} x_{ij} \in \{0,1\} & \text{forall } i \text{ where } 0 \leq i \leq N \\
\end{align*}
\]

\[x_{ij} \in \{0,1\}\]
Handling Multiple References

• Compare summary against each reference, take highest score
• In order to have fair comparison against reference document scores, jacknifing was used.
  – Create N subsets of N references, each missing 1 reference, and average multi-reference scores
## Results on TAC08 Part A

### vs Responsiveness

<table>
<thead>
<tr>
<th></th>
<th>Spearman</th>
<th></th>
<th></th>
<th>Pearson</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>Auto</td>
<td>Hu</td>
<td></td>
<td>All</td>
</tr>
<tr>
<td>BEwT-E</td>
<td>0.864</td>
<td>0.802</td>
<td>0.539</td>
<td>0.925</td>
<td>0.840</td>
<td>0.549</td>
</tr>
<tr>
<td>Original BE</td>
<td>0.873</td>
<td>0.815</td>
<td>0.467</td>
<td>0.887</td>
<td>0.817</td>
<td>0.595</td>
</tr>
<tr>
<td>ROUGE2</td>
<td>0.905</td>
<td>0.867</td>
<td>0.539</td>
<td>0.851</td>
<td>0.829</td>
<td>0.645</td>
</tr>
<tr>
<td>ROUGESU4</td>
<td>0.884</td>
<td>0.832</td>
<td>0.874</td>
<td>0.852</td>
<td>0.802</td>
<td>0.846</td>
</tr>
<tr>
<td>Mod Pyramid</td>
<td><strong>0.917</strong></td>
<td><strong>0.878</strong></td>
<td>0.611</td>
<td><strong>0.968</strong></td>
<td><strong>0.900</strong></td>
<td>0.509</td>
</tr>
</tbody>
</table>

### vs Modified Pyramid

<table>
<thead>
<tr>
<th></th>
<th>Spearman</th>
<th></th>
<th></th>
<th>Pearson</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>Auto</td>
<td>Hu</td>
<td></td>
<td>All</td>
</tr>
<tr>
<td>BEwT-E</td>
<td>0.955</td>
<td>0.935</td>
<td>0.833</td>
<td><strong>0.950</strong></td>
<td><strong>0.950</strong></td>
<td>0.665</td>
</tr>
<tr>
<td>Original BE</td>
<td>0.934</td>
<td>0.904</td>
<td>0.762</td>
<td>0.917</td>
<td>0.913</td>
<td>0.663</td>
</tr>
<tr>
<td>ROUGE2</td>
<td>0.936</td>
<td>0.907</td>
<td>0.857</td>
<td>0.869</td>
<td>0.907</td>
<td>0.544</td>
</tr>
<tr>
<td>ROUGESU4</td>
<td>0.919</td>
<td>0.883</td>
<td>0.857</td>
<td>0.871</td>
<td>0.886</td>
<td>0.543</td>
</tr>
<tr>
<td>Responsiveness</td>
<td><strong>0.917</strong></td>
<td><strong>0.878</strong></td>
<td>0.611</td>
<td><strong>0.968</strong></td>
<td><strong>0.900</strong></td>
<td>0.509</td>
</tr>
</tbody>
</table>

- Duplicates off, SQRT weights, all transforms except Hyper/Hyponyms
Results on TAC08 Part B

<table>
<thead>
<tr>
<th></th>
<th>Spearman</th>
<th></th>
<th></th>
<th></th>
<th>Pearson</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Auto</td>
<td>Hu</td>
<td></td>
<td>All</td>
<td>Auto</td>
<td>Hu</td>
<td></td>
</tr>
<tr>
<td>BEwT-E</td>
<td>0.926</td>
<td>0.891</td>
<td>0.802</td>
<td></td>
<td>0.925</td>
<td>0.924</td>
<td>0.642</td>
<td></td>
</tr>
<tr>
<td>Original BE</td>
<td>0.917</td>
<td>0.877</td>
<td>0.683</td>
<td></td>
<td>0.905</td>
<td>0.912</td>
<td>0.464</td>
<td></td>
</tr>
<tr>
<td>ROUGE2</td>
<td>0.920</td>
<td>0.882</td>
<td>0.587</td>
<td></td>
<td>0.882</td>
<td>0.909</td>
<td>0.579</td>
<td></td>
</tr>
<tr>
<td>ROUGESU4</td>
<td>0.927</td>
<td>0.893</td>
<td>0.898</td>
<td></td>
<td>0.835</td>
<td>0.901</td>
<td>0.796</td>
<td></td>
</tr>
<tr>
<td>Mod Pyramid</td>
<td><strong>0.948</strong></td>
<td><strong>0.925</strong></td>
<td>0.695</td>
<td></td>
<td><strong>0.980</strong></td>
<td><strong>0.949</strong></td>
<td>0.741</td>
<td></td>
</tr>
</tbody>
</table>

- Duplicates off, SQRT weights, all transforms except Hyper/Hyponyms
Effect of Transformations

<table>
<thead>
<tr>
<th>One Transform Off</th>
<th>All</th>
<th>Auto</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Hyper/Hyponyms</td>
<td>140</td>
<td>101</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>88</td>
<td>71</td>
<td>86</td>
</tr>
</tbody>
</table>

- Hyper/Hyponyms transformation generally has negative impact at the individual topic level
- Topics include DUC05 (50), DUC06 (50), DUC07 (45), TAC08A (48), TAC08B (48)
Effect of Transformations

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Auto</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>26</td>
<td>31</td>
<td>11</td>
</tr>
<tr>
<td>-</td>
<td>19</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>DUC07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>30</td>
<td>29</td>
<td>14</td>
</tr>
<tr>
<td>-</td>
<td>20</td>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>DUC06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>38</td>
<td>35</td>
<td>18</td>
</tr>
<tr>
<td>-</td>
<td>12</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>DUC05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>25</td>
<td>24</td>
<td>13</td>
</tr>
<tr>
<td>-</td>
<td>23</td>
<td>24</td>
<td>23</td>
</tr>
<tr>
<td>TAC08 Base</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>27</td>
<td>23</td>
<td>11</td>
</tr>
<tr>
<td>-</td>
<td>21</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>TAC08 Update</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>146</td>
<td>142</td>
<td>67</td>
</tr>
<tr>
<td>-</td>
<td>95</td>
<td>99</td>
<td>90</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of topics across DUC05-07, TAC08A, TAC08B whose summary-level Pearson correlation was affected (positively/negatively) when the remaining tranformations are enabled
Conclusions

• Observations:
  – BEwT-E tends to outperform old BE
  – Transformations help less than expected
  – Duplicate BEs usually hurt performance
  – SQRT weighting most consistent

• Improvements:
  – Parameter tuning to improve correlation
  – Coreference resolution
  – Additional transformation rules
Questions?

• Code will be made available soon via www.isi.edu