Basic Elements: A Framework for Automated Evaluation of Summary Content

Eduard Hovy,
Chin-Yew Lin,
Liang Zhou,
Junichi Fukumoto
USC/ISI
Goals

• **Automated evaluation** of summaries
  – and possibly, other texts (produced by algorithms) that can be compared to human reference texts, (incl. MT, NLG)

• Evaluation of **content only**: can focus on fluency, style, etc. in later work

• **Desiderata** for resulting automated system:
  – must reproduce rankings of human evaluators
  – must be reliable
  – must apply across domains
  – must port to other languages without much effort
Desiderata for SummEval metric

• Match **pieces of the summary against ideal summary/ies:**
  – Granularity: somewhere between unigrams and whole sentences
  – Units: EDUs (SEE; Lin 03), “nuggets” (Harman), “factoids” (Van Halteren and Teufel 03), SCUs (Nenkova et al. 04)…
  – **Question:** How to delimit the length? Which units?

• Match the **meanings** of the pieces:
  – **Questions:** How to obtain meaning? What paraphrases? What counts as a match? Are there partial matches?

• Compute a **composite score** out of lots of matches
  – **Questions:** How to score each unit? Are there partial scores? Are all units equally important? How to compose the scores?
Framework for SummEval

1. Obtain units ("breaker")
2. Match units against ideals ("matcher")
3. Assemble scores ("scorer")

Create ideal summaries

Create test summary
1. Breaking

- Simplest approach: sentences
  - E.g., SEE manual scoring, DUC 2000–03
  - **Problem**: sentence contains too many separate pieces of information; cannot match all in one

- Ngrams of various kinds (also skip-ngrams, etc.)
  - E.g., ROUGE
  - **Problem**: not all ngrams are equally important
    - **Problem**: no single best ngram length (multi-word units)

- Let each assessor choose own units
  - **Problem**: too much variation

- One or more Master Assessor(s) chooses units
  - E.g., Pyramid in DUC 2005

- Is there an automated way?
Automating BE unit breaking

- We propose using Basic Elements as units: minimal-length fragments of ‘sensible meaning’
- Automating this: parsers + ‘cutting rules’ that chop tree:
  - Charniak parser + CYL rules
  - Collins parser + LZ rules
  - Minipar + JF rules
  - Chunker including CYL rules
  - Microsoft’s Logical Form parser + LZ rules
- Result: BEs of variable length/scope:
- Working definition: Each constituent Head, and each relation (between Head and Modifier) in a dependency tree is a candidate BE. Only the most important content-bearing ones are actually used for SummEval:
  - Head nouns and verbs
  - Verb plus its arguments
  - Noun plus its adjective/nominal/PP modifiers
    - Examples: [verb-Subj-noun], [noun-Mod-adj], [noun], [verb]
BEs: Syntactic or semantic?

• Objection: these are syntactic definitions!
• BUT:
  – multi-word noun string is a single BE (“kitchen knife”)
  – Proper Name string is a single BE (“Bank of America”)
  – Each V and N is a BE: the smallest measurable units of meaning — if you don’t have these, how can you score for individual pieces of info?
  – Each head-rel-mod is a BE: it’s not enough to know that there was a parade and that New York is mentioned; you have to know that the parade was in New York
  – This goes up the parse tree: in “he said there was a parade in New York”, also the fact that the saying was about the parade is important
• So: while the definition is syntactic, the syntax-based rules delimit the semantic units we need
Example from MS: Parse and LF

Thanks to Lucy Vanderwende and colleagues, Microsoft
Ex BEs, merging multiple breakers

SUMMARY: D100.M.100.A.G.
New research studies are providing valuable insight into the probable causes of schizophrenia.

Tsub | study provide [MS_LF MINI]
Tobj | provide insight [MS_LF COLLINS]
Prep_into | insight into cause [MS_LF MINI]
Prep_of | cause of schizophrenia [MS_LF MINI]
Attrib jj | new study MS_LF MINI COLLINS CHUNK
Mod nn | research study [MS_LF MINI COLLINS CHUNK]
Attrib jj | valuable insight [MS_LF MINI COLLINS CHUNK]
jj | probable cause [MINI COLLINS CHUNK]
np | study [COLLINS CHUNK]
vp | provide [COLLINS CHUNK]
np | insight [COLLINS CHUNK]
np | cause [COLLINS CHUNK]
np | schizophrenia [COLLINS CHUNK]
Using BEs to match Pyramid SCUs (MINIPAR + Fukumoto cutting rules)

<table>
<thead>
<tr>
<th>Pyramid judgments</th>
<th>total overlap</th>
<th>BE</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.b2 D.b2 E.b2 F.b2 P.b2 Q.b2 R.b2 S.b2 U.b2 V.b2 df</td>
<td>&lt;&lt;BE element&gt;&gt;</td>
<td></td>
</tr>
<tr>
<td>1 0 1 1 1 0 0 0 1 0 5</td>
<td>defend &lt;- themselves (obj)</td>
<td></td>
</tr>
<tr>
<td>0 1 1 1 1 0 0 0 0 0 4</td>
<td>security &lt;- national (mod)</td>
<td></td>
</tr>
<tr>
<td>1 0 1 0 0 1 0 0 0 0 3</td>
<td>charge &lt;- subvert (of)</td>
<td></td>
</tr>
<tr>
<td>0 1 0 0 0 1 1 0 0 0 3</td>
<td>civil &lt;- and (punc)</td>
<td></td>
</tr>
<tr>
<td>0 1 0 0 0 1 1 0 0 0 3</td>
<td>civil &lt;- political rights (conj)</td>
<td></td>
</tr>
<tr>
<td>1 0 0 0 1 0 0 1 0 0 3</td>
<td>incite &lt;- subversion (obj)</td>
<td></td>
</tr>
<tr>
<td>0 0 1 0 0 0 1 1 0 0 3</td>
<td>president &lt;- jiang zemin (person)</td>
<td></td>
</tr>
<tr>
<td>0 0 0 1 0 0 0 0 1 1 3</td>
<td>release &lt;- china (subj)</td>
<td></td>
</tr>
<tr>
<td>1 0 0 0 1 0 0 0 0 0 2</td>
<td>action &lt;- its (gen)</td>
<td></td>
</tr>
<tr>
<td>0 0 0 1 0 0 0 0 0 1 2</td>
<td>ail &lt;- china (subj)</td>
<td></td>
</tr>
<tr>
<td>1 0 0 0 0 0 0 0 1 0 2</td>
<td>charge &lt;- serious (mod)</td>
<td></td>
</tr>
<tr>
<td>1 0 0 0 1 0 0 0 0 0 2</td>
<td>defend &lt;- action (obj)</td>
<td></td>
</tr>
<tr>
<td>1 0 0 0 1 0 0 0 0 0 2</td>
<td>defend &lt;- china (subj)</td>
<td></td>
</tr>
<tr>
<td>0 0 0 1 0 0 0 0 1 0 2</td>
<td>defend &lt;- dissident (subj)</td>
<td></td>
</tr>
<tr>
<td>1 0 0 1 0 0 0 0 0 0 2</td>
<td>democracy &lt;- multiparty (nn)</td>
<td></td>
</tr>
<tr>
<td>0 1 0 0 0 0 0 0 1 0 2</td>
<td>dissident &lt;- prominent (mod)</td>
<td></td>
</tr>
<tr>
<td>0 1 0 0 0 0 0 0 1 0 2</td>
<td>dissident &lt;- three (nn)</td>
<td></td>
</tr>
</tbody>
</table>
Using BEs to match Pyramid SCUs (Charniak + Lin cutting rules)

<table>
<thead>
<tr>
<th>Pos in text</th>
<th>Type of rel</th>
<th>Surface form</th>
<th>With semantic type for matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>* (1 10)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(103_CD</td>
<td>-</td>
</tr>
<tr>
<td>* (1 12 0)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(1988_CD</td>
<td>-</td>
</tr>
<tr>
<td>* (1 14)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(U.N._NNP</td>
<td>-</td>
</tr>
<tr>
<td>* (1 15 0)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(Security_NNP</td>
<td>-</td>
</tr>
<tr>
<td>* (1 16 0)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(Council_NNP</td>
<td>-</td>
</tr>
<tr>
<td>* (1 17 0)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(approves_VBZ</td>
<td>-</td>
</tr>
<tr>
<td>* (1 17 11)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(approves_VBZ</td>
<td>in_IN</td>
</tr>
<tr>
<td>* (1 17 12)</td>
<td>&lt;PP&gt;</td>
<td>(approves_VBZ</td>
<td>1988_CD</td>
</tr>
<tr>
<td>* (1 17 16)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(approves_VBZ</td>
<td>Council_NNP</td>
</tr>
<tr>
<td>* (1 17 18)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(approves_VBZ</td>
<td>plan_NN</td>
</tr>
<tr>
<td>* (1 17 2)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(approves_VBZ</td>
<td>decade_NN</td>
</tr>
<tr>
<td>* (1 17 24)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(approves_VBZ</td>
<td>to_TO</td>
</tr>
<tr>
<td>* (1 17 25)</td>
<td>&lt;TO&gt;</td>
<td>(approves_VBZ</td>
<td>try_VB</td>
</tr>
<tr>
<td>* (1 17 3)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(approves_VBZ</td>
<td>after_IN</td>
</tr>
<tr>
<td>* (1 17 5)</td>
<td>&lt;PP&gt;</td>
<td>(approves_VBZ</td>
<td>bombing_NN</td>
</tr>
<tr>
<td>* (1 17 9)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(approves_VBZ</td>
<td>Flight_NNP</td>
</tr>
<tr>
<td>* (1 18 0)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(plan_NN</td>
<td>-</td>
</tr>
<tr>
<td>* (1 18 19)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(plan_NN</td>
<td>proposed_VBN</td>
</tr>
<tr>
<td>* (1 19 0)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(proposed_VBN</td>
<td>-</td>
</tr>
<tr>
<td>* (1 19 20)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(proposed_VBN</td>
<td>by_IN</td>
</tr>
<tr>
<td>* (1 19 21)</td>
<td>&lt;PP&gt;</td>
<td>(proposed_VBN</td>
<td>U.S._NNP</td>
</tr>
<tr>
<td>* (1 2 0)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(decade_NN</td>
<td>-</td>
</tr>
<tr>
<td>* (1 2 1)</td>
<td>&lt;HEAD-MOD&gt;</td>
<td>(decade_NN</td>
<td>A_DT</td>
</tr>
</tbody>
</table>
2. Matching

- Input: ideal summary/ies units + test summary units
- Simplest approach: string match
  - **Problem 1:** cannot pool ideal units with same meaning: test summary may score twice by saying the same thing in different ways, matching different ideal units
  - **Problem 2:** cannot match ideal units when test summary uses alternative ways to say same thing
- **Solution 1:** Pool ideal units—a human groups together paraphrase-equal units into equivalence class (like BLEU)
- **Solution 2:** Humans judge semantic equivalence
  - **Problem:** expensive and difficult to decide
  - **Problem:** distributing meaning across multiple words
    - “a pair was arrested” “two men were arrested” “more than one person was arrested” — are these identical?
  - **Problem:** the longer the unit, the more bits require matching

Is there a way to automate this?
Using BEs to match Pyramid and DUC scores

• **Aim**: can we *exactly* reproduce Pyramid scoring, where each Pyramid fragment consists of a set of BEs?

• **Approach** tried: spectrum of matching tests, from exact to very general

• **Result**: cannot do automatically without smart matching function: refs too diversified

SCU1: the crime in question was the Lockerbie {Scotland} bombing
A1 [for the Lockerbie bombing]1
B1 [for blowing up]1 [over Lockerbie, Scotland]1
C1 [of bombing]1 [over Lockerbie, Scotland]1
D1 [was blown up over Lockerbie, Scotland.]1
P1 [the bombing of Pan Am Flight 103]1
Q1 [bombing over Lockerbie, Scotland.]1
R1 [for Lockerbie bombing]1
S2 [bombing of Pam Am flight 103 over Lockerbie.]1
U1 [linked to the Lockerbie bombing]1
V1 [in the Lockerbie bombing case.]1
# Merging BE to build SCUs

<table>
<thead>
<tr>
<th>BE</th>
<th>file_loc</th>
<th>doc_freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>lockerbie bombing</td>
<td>1988 bombing</td>
<td>lockerbie bombing</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>lockerbie bombing</td>
<td>1988 bombing</td>
<td>pan bombing</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>two libyans</td>
<td>two libyans</td>
<td>two suspects</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>pan jet</td>
<td>pan jet</td>
<td>lockerbie suspects</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>am jet</td>
<td>am jet</td>
<td>libyan suspects</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>december 1988</td>
<td>moammar gadhafi</td>
<td>moammar gadhafi</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>pan flight</td>
<td>u.n. council</td>
<td>united states</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>blowing up</td>
<td>blown up</td>
<td>==change</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>indicted in 1991</td>
<td>indicted in 1991</td>
<td>==charge, accuse</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>hand over</td>
<td>turning over</td>
<td>==transfer</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>were indicted</td>
<td>==be</td>
<td>charge, accuse</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>try in court</td>
<td>==act, move</td>
<td>in</td>
</tr>
</tbody>
</table>

---

**SENTENCE: Q1**

[BE_0 ] "agents"
[BE_0_0 ] "Two" BE_0
[BE_0_1 ] "Libyan" BE_0
[BE_0_2 ] "intelligence" BE_0
[BE_4 ] "States"
[BE_4_0 ] "United" BE_4
[BE_6 ] "Britain"
[BE_7 ] "bombing"
[BE_7_0 ] "1988" BE_7
[BE_7_1 ] "Pan" BE_7
[BE_7_2 ] "Am" BE_7
[BE_11 ] "Lockeberie"
[BE_12 ] "Scotland"
[BE_13 ] "implicated"
[BE_13_0 BE_4_1 ] BE_13 "by" BE_4 "and" BE_6
[BE_13_1 BE_7_3 ] BE_13 "in" BE_7
[BE_13_2 BE_11_0 ] BE_13 "over" BE_11 BE_12
[BE_17 ] "trial"
[BE_18 ] "Netherlands"
[BE_19 ] "location"
[BE_19_0 ] "neutral" BE_19
[BE_21 ] "Gadhafi"
[BE_21_0 ] "Libyan" BE_21
[BE_21_1 ] "leader" BE_21
[BE_21_2 ] "Col." BE_21
[BE_21_3 ] "Moammar" BE_21
[BE_26 ] "agreed"
[BE_26_0 ] BE_26 "upon"
[BE_26_1 BE_21_4 ] BE_26 "by" BE_21
[BE_29 ] "stand"
[BE_29_0 BE_17_0 ] BE_29 BE_17 "in" BE_18 BE_19 BE_26
Fragmented units and partial scores

• Why do we need small-grain units?

Reference unit:
[A B C D]

Doc 1:
[A D] x x x x x
x x x x [B C D]
x x x x

Doc 2:
 x x x x [B C D]
 x x x x

Partial score, or problems!

SEE (Lin 2001)
Issues in comparing BEs

• A central motivation for BEs is that each piece of semantic info can be counted (if important)
• To count once only, we need a smart BE matcher
• BEs’ small size makes (limited) paraphrase match feasible
• But it’s still not trivial:
  – Numbers: need to reason about sizes:
    • “almost $20 million” — 1 BE, or 2 [$20M + almost]?
    • If 2 BEs, then how to match this with “$19.9M”?
  – Names: need to handle pseudonyms and abbrevs:
    • USA = “United States” = “America” etc.
  – Reference: need to handle coref:
    • “Joe said” = “he said”
  – Metonymy: need to de-coerce:
    • “Washington announced” = “A spokesperson for the Gov’t said”
Semantic/paraphrase matching

What to do?

...this is an ideal research topic for the next few years:
- More specific than general entailment…
- Can start with simple term expansion…
- Can use syntactic transformations (Hermjakob et al. TREC-02)…
- Can try web-based reformulation validation…
- etc.
3. Scoring

• **Question 1**: How should each unit be scored? Is each unit equally important?

• **Approaches**:
  - Simplest: Each matched unit gets 1 point (like TREC relevance, simple ROUGE) — not ideal
  - Next: Each unit assigned an intrinsic ‘value’ depending on its *information content*: word entropy, (e.g., inverse term freq *itf* against regular English) — downgrades closed-class units
  - Next: each unit assigned score based on its *popularity* in the ideal summaries — proposed by Van Halteren and Teufel 03, used in Pyramid method

• **Question 2**: How should scores be combined?

• **Approaches**:
  - Simplest: just sum scores
  - Other models: weight scores by some policy (e.g, reflect coherence of sentence containing BE, etc.)
BE scoring

- Direct popularity score, as in pyramids
- BE scoring variations:
  - H — head-only match (BE-F does not have this)
  - HM — head and mod match (does not include head-only)
  - HMR — head, mod and relation match (relation can’t be NIL)
  - HM1 — H + HM (head and mod plus head only)
  - HMR1 — HM + HMR (mod cannot be NIL but relation can be)
  - HMR2 — H + HM + HMR (mod and relation can be NIL)

- Summary: BE is like ROUGE (skip bigrams), with some uninteresting bigrams removed, using popularity weighting
BE scores for DUC 05

• Recall differentiates well
BE correlations, DUC 2002

### DUC 2002 Single 100

<table>
<thead>
<tr>
<th></th>
<th>Original Pearson</th>
<th>Spearman</th>
<th>Stemmed Pearson</th>
<th>Spearman</th>
<th>Stopped and Stemmed Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>0.986</td>
<td>0.836</td>
<td>0.986</td>
<td>0.836</td>
<td>0.995</td>
<td>0.889</td>
</tr>
<tr>
<td>R2</td>
<td>0.988</td>
<td>0.957</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R3</td>
<td>0.997</td>
<td>0.981</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R4</td>
<td>0.996</td>
<td>0.990</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RL</td>
<td>0.989</td>
<td>0.849</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RS4</td>
<td>0.998</td>
<td>0.957</td>
<td>0.998</td>
<td>0.952</td>
<td>0.997</td>
<td>0.977</td>
</tr>
<tr>
<td>RSU4</td>
<td>0.996</td>
<td>0.900</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### DUC 2002 Single 100

<table>
<thead>
<tr>
<th></th>
<th>BE-L</th>
<th>BE-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.993</td>
<td></td>
</tr>
<tr>
<td>HM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMR</td>
<td>0.987</td>
<td>0.951</td>
</tr>
<tr>
<td>HMR1</td>
<td>0.995</td>
<td>0.954</td>
</tr>
<tr>
<td>HMR2</td>
<td>0.995</td>
<td>0.951</td>
</tr>
</tbody>
</table>

### DUC 2002 Multi 100

<table>
<thead>
<tr>
<th></th>
<th>Original Pearson</th>
<th>Spearman</th>
<th>Stemmed Pearson</th>
<th>Spearman</th>
<th>Stopped and Stemmed Pearson</th>
<th>Spearman</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>0.697</td>
<td>0.578</td>
<td>0.701</td>
<td>0.588</td>
<td>0.770</td>
<td>0.828</td>
</tr>
<tr>
<td>R2</td>
<td>0.896</td>
<td>0.842</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R3</td>
<td>0.931</td>
<td>0.867</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R4</td>
<td>0.910</td>
<td>0.782</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RL</td>
<td>0.679</td>
<td>0.648</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RS4</td>
<td>0.857</td>
<td>0.867</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSU4</td>
<td>0.808</td>
<td>0.600</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### DUC 2002 Multi 100

<table>
<thead>
<tr>
<th></th>
<th>BE-L</th>
<th>BE-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.876</td>
<td></td>
</tr>
<tr>
<td>HM</td>
<td>0.865</td>
<td>0.924</td>
</tr>
<tr>
<td>HMR</td>
<td>0.815</td>
<td>0.934</td>
</tr>
<tr>
<td>HMR1</td>
<td>0.880</td>
<td>0.924</td>
</tr>
<tr>
<td>HMR2</td>
<td>0.866</td>
<td>0.934</td>
</tr>
</tbody>
</table>

H => head only match (BE-F does not have this)
HM => head and mod match (does not include head-only)
HMR => head, mod and relation match (relation can't be NIL)
HMR1 => H + HM (head and mod plus head only)
HMR2 => HM + HMR (mod cannot be NIL but relation can be)
HMR2 => H + HM + HMR (mod and relation can be NIL)
BE correlations, DUC 2003

<table>
<thead>
<tr>
<th>DUC 2003</th>
<th>Original</th>
<th>Stemmed</th>
<th>Stopped and Stemmed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pearson</td>
<td>Spearman</td>
<td>Pearson</td>
</tr>
<tr>
<td>R1</td>
<td>0.961</td>
<td>0.965</td>
<td>0.956</td>
</tr>
<tr>
<td>R2</td>
<td>0.751</td>
<td>0.626</td>
<td>0.749</td>
</tr>
<tr>
<td>R3</td>
<td>0.712</td>
<td>0.591</td>
<td>0.700</td>
</tr>
<tr>
<td>R4</td>
<td>0.665</td>
<td>0.547</td>
<td>0.640</td>
</tr>
<tr>
<td>RL</td>
<td>0.974</td>
<td>0.969</td>
<td>0.968</td>
</tr>
<tr>
<td>RSU4</td>
<td>0.889</td>
<td>0.785</td>
<td>0.891</td>
</tr>
<tr>
<td></td>
<td>0.976</td>
<td>0.978</td>
<td>0.973</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DUC 2003</th>
<th>Pearson</th>
<th>Spearman</th>
<th>BE-L</th>
<th>BE-F</th>
<th>BE-L</th>
<th>BE-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>0.916</td>
<td>-</td>
<td>0.938</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HM</td>
<td>0.774</td>
<td>0.733</td>
<td>0.670</td>
<td>0.657</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMR</td>
<td>0.610</td>
<td>0.688</td>
<td>0.385</td>
<td>0.622</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMR1</td>
<td>0.968</td>
<td>0.733</td>
<td>0.956</td>
<td>0.657</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMR2</td>
<td>0.762</td>
<td>0.688</td>
<td>0.670</td>
<td>0.622</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DUC 2003</th>
<th>Pearson</th>
<th>Spearman</th>
<th>BE-L</th>
<th>BE-F</th>
<th>BE-L</th>
<th>BE-F</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>0.622</td>
<td>0.711</td>
<td>0.612</td>
<td>0.695</td>
<td>0.787</td>
<td>0.824</td>
</tr>
<tr>
<td>R2</td>
<td>0.803</td>
<td>0.678</td>
<td>0.800</td>
<td>0.686</td>
<td>0.901</td>
<td>0.876</td>
</tr>
<tr>
<td>R3</td>
<td>0.684</td>
<td>0.453</td>
<td>0.670</td>
<td>0.450</td>
<td>0.678</td>
<td>0.434</td>
</tr>
<tr>
<td>R4</td>
<td>0.488</td>
<td>0.326</td>
<td>0.488</td>
<td>0.336</td>
<td>0.501</td>
<td>0.344</td>
</tr>
<tr>
<td>RL</td>
<td>0.539</td>
<td>0.647</td>
<td>0.512</td>
<td>0.640</td>
<td>0.732</td>
<td>0.782</td>
</tr>
<tr>
<td>RSU4</td>
<td>0.744</td>
<td>0.692</td>
<td>0.757</td>
<td>0.707</td>
<td>0.889</td>
<td>0.879</td>
</tr>
<tr>
<td></td>
<td>0.723</td>
<td>0.687</td>
<td>0.727</td>
<td>0.707</td>
<td>0.867</td>
<td>0.883</td>
</tr>
</tbody>
</table>

H => head only match (BE-F does not have this)
HM => head and mod match (does not include head-only)
HMR => head, mod and relation match (relation can't be NIL)
HMR1 => H + HM (head and mod plus head only)
HMR2 => H + HM + HMR (mod cannot be NIL but relation can be)
HMR3 => H + HM + HMR (mod and relation can be NIL)
BE correlations 1, DUC 2005

- All comparisons over exactly the same 20 topics and 25 systems
- All 9 references (not just 7)
- Recall scores
- $S = $ Spearman
- $P = $ Pearson
BE correlations 2, DUC 2005

QuickTime™ and a TIFF (Uncompressed) decompressor are needed to see this picture.

- Comparisons over all DUC 05 topics
- Recall scores
- $S = $ Spearman
- $P = $ Pearson
## BE Framework

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SEE</td>
<td>sentences, auto</td>
<td>manual, partial ok</td>
<td>add partial points</td>
</tr>
<tr>
<td>ROUGE</td>
<td>auto ngrams, various kinds</td>
<td>string match, stemmed/not</td>
<td>single-point, also weighted</td>
</tr>
<tr>
<td>Van Halteren &amp; Teufel</td>
<td>factoids, manual</td>
<td>manual, assessors</td>
<td>popularity score</td>
</tr>
<tr>
<td>Pyramid</td>
<td>SCUs, manual</td>
<td>manual, community</td>
<td>popularity score</td>
</tr>
<tr>
<td>BE method</td>
<td>BEs, auto</td>
<td>string match</td>
<td>popularity</td>
</tr>
</tbody>
</table>
Conclusion 1

1. We propose a **general framework** in which various approaches can be embedded and compared
   - Framework provides ‘slots’ for:
     - Units of comparison (words, phrases, SCUs, BEs, etc.)
     - Relative strength/goodness of units
     - Methods of comparing units between summary and references
     - Methods of combining scores of individual units into an overall score
   - Anybody can insert their modules in the framework

2. We propose using **Basic Elements** as units: minimal-length fragments of ‘sensible meaning’
   - BEs of variable length: either a semantic ‘head’ or a head+relation+modifier
     - Head nouns and verbs
     - Verb plus its arguments
     - Noun plus its adjective/nominal/PP modifiers
Conclusion 2

• Please download the BE package and use it: http://www.isi.edu/~cyl/BE/

• Please build and insert your own modules!
  – Unit breakers
  – Matchers
  – Scorers
Thank you!
Automated Evaluation: The General Method

• Use $N$ human-created summaries as references
• For a given test summary, find its ‘average distance’ from the reference summaries — the closer, the higher it should score
• Distance measures:
  – Word overlap (test on word identity, root identity, word+synonyms, etc.)
  – Fragment correspondence (various kinds of fragments: SCUs, etc.)
• (NOTE: same general method as used in MT)
Questions and Problems

• The problem with words:
  – Single words are too indiscriminate: the summary may use ‘good’ words in the wrong contexts—should they be counted?
  – Ngrams are too fixed: the elements of pertinent information need different amounts of words—“Bank of America”=1 point
  – Not all words are equally important

• The problem with fragments:
  – It’s not clear how to define them
  – Some methods choose longest-common-substring fragments out of (some of) the references; but when more references are added, the fragment lengths may change—unstable
  – Fragments have to be built by hand—expensive and subjective

• Other questions:
  – Methods of comparing words/phrases when they’re not identical (“the Pope”, “John Paul II”, etc.)
  – Methods of combining overlap counts, scores—simple addition?
Proposed Framework: 4 Modules

1. How to create the units? Text ‘breaker’:
   - Input: running text
   - Output: units to be evaluated
   - Examples of units: words, word roots, SCUs, Basic Elements

2. What’s the score of each unit? Unit scorer:
   - Input: list(s) of units
   - Output: list of units, each unit with score
   - Examples of results: Pyramid, Madrid group combination list

3. When are two units the ‘same’? Unit matcher:
   - Input: 2 units (one from reference list, one from text)
   - Output: goodness-of-match score
   - Examples: word identity, root identity, paraphrase equivalence

4. What’s the overall score? Score adder function:
   - Input: list of units, each with individual score
   - Output: overall score for text
General Framework Procedure

• **Preparation phase (on references):** Using reference summaries:
  1. ‘Break’ text into individual units of content
  2. Rate quality/value of each unit
  3. Result: ranked/scored list of reference units

• **Evaluation phase (on test docs):** On system or human summary:
  1. ‘Break’ text to create its units of content
  2. Compare units against ranked/scored reference list to obtain individual unit scores
  3. Result: merge unit scores to compute overall score for the text
Various Parts Built So Far

• **Framework:**
  – Architecture: ISI is building
  – Module APIs: ISI has built

• **Modules:** Anyone can build their favorite module(s):
  – ISI is building one or more examples of each of the 4 modules
  – Columbia has built a Unit Scorer (the Pyramid)
  – Van Halteren-Teufel and Madrid have built Unit Scorers
  – ISI has built a word-level Breaker, Scorer, and Adder (unigram function inside ROUGE)

• **Evaluation** of modules:
  – Plug in a set of modules
  – Apply to standard set of texts for which human score ranking is known
  – Compare resulting ranking of texts against human ranking
  – …the better correspondence, the better the module(s)
Issue 1: Eval Gold Standard

• We need to choose the Truth:
  – We have various candidates for BEs and BE scoring methods, so we must compare them against some Truth
  – Which evaluation / ranking of texts will we use to determine what works best?
• Candidates:
  – Pyramid results (3 topics from DUC 03)
  – DUC 03, 04 rankings (NIST used SEE)
  – SEE results from DUC 01, 02
  – Results from Madrid
  – Results from Hans and Simone
  – ??
• Methodology: we need to decide on standard ranking comparison functions (Kendall, Krippendorff, etc.)
Issue 2: Size of Units

- **Words (unigram ROUGE):** Good as a starting point only, because:
  - not all words are equally important (closed-class)
  - word sequences form semantic units (‘Bank of America’)
- **SCUs (Pyramid):** Better, but not ideal because:
  - better: retain only sequences of words that are selected in multiple reference summaries (useful semantic units)
  - but: unit length varies according to the reference summs available, so units change when new ref summs are used
  - also: each unit gets same score, regardless of semantic content
  - also: SCUs are large; how to score partial matches?
- **Basic Elements (BEs):**
  - better: unchanging, minimal-length semantic units
  - also: potentially created automatically
  - problem: how are BEs defined?
  - working definition: Each relation (between Head and Modifier) in a dependency tree is a candidate BE. Only the most important content-bearing ones are actually used for SummEval
  - examples: [verb-Subj-noun], [noun-Mod-adj], [noun], [verb]
BEs vs. unigrams

• Unigram-matching assigns equal weight to each word, regardless of its importance
• BE match assigns weight only to important words (basic BEs) and to their relations (triple BEs)
  – Some words are double-counted (basic and in relation)
  – Some words are not counted (unimportant determiners, etc.)
• The challenge for BEs is to correlate better with human scores than unigram scores do
ISI Work on BEs: Approach

1. Parse or chunk the text (using one or more BE breakers)
   - Multiple BE creation engines deployed:
     • Parsers: Charniak (Brown), Collins (MIT), Contex (ISI), Minipar (Alberta)
     • Other systems: Lin chunker (ISI), Logical Forms parser (Microsoft)

2. Apply BE extraction rules to parse tree or chunks
   - Multiple extraction rulesets built:
     • Extraction rules: Fukumoto rules, Zhou rules, Lin rules
     • Results: Minipar+Fukumoto, Collins+Zhou, Lin-chunker, MS-LF, Charniak+Lin

3. Convert all results to standardized BE form and merge them
   - Done: results show that no single engine does it all

4. Obtain BEs also for reference texts (Pyramid and DUC 03)
   - Done for individual BE breakers but not yet multi-breaker version
   - Result: lists of BEs, ranked by reference popularity (Pyramid method)

5. Compare sets of BEs: find best breaker and rank BEs
   - Compare summary BE list to reference BE list and rank summaries
     • Comparison functions: equality and supertype-substitution equality
   - Goal: try to match Pyramid and DUC rankings for same texts