Basic Elements:
A Framework for Automated Evaluation of Summary Content

Eduard Hovy,

Chin-Yew Lin,

Liang Zhou,

Junichi Fukumoto

USC/ISI

USC INFORMATION SCIENCES INSTITUTE

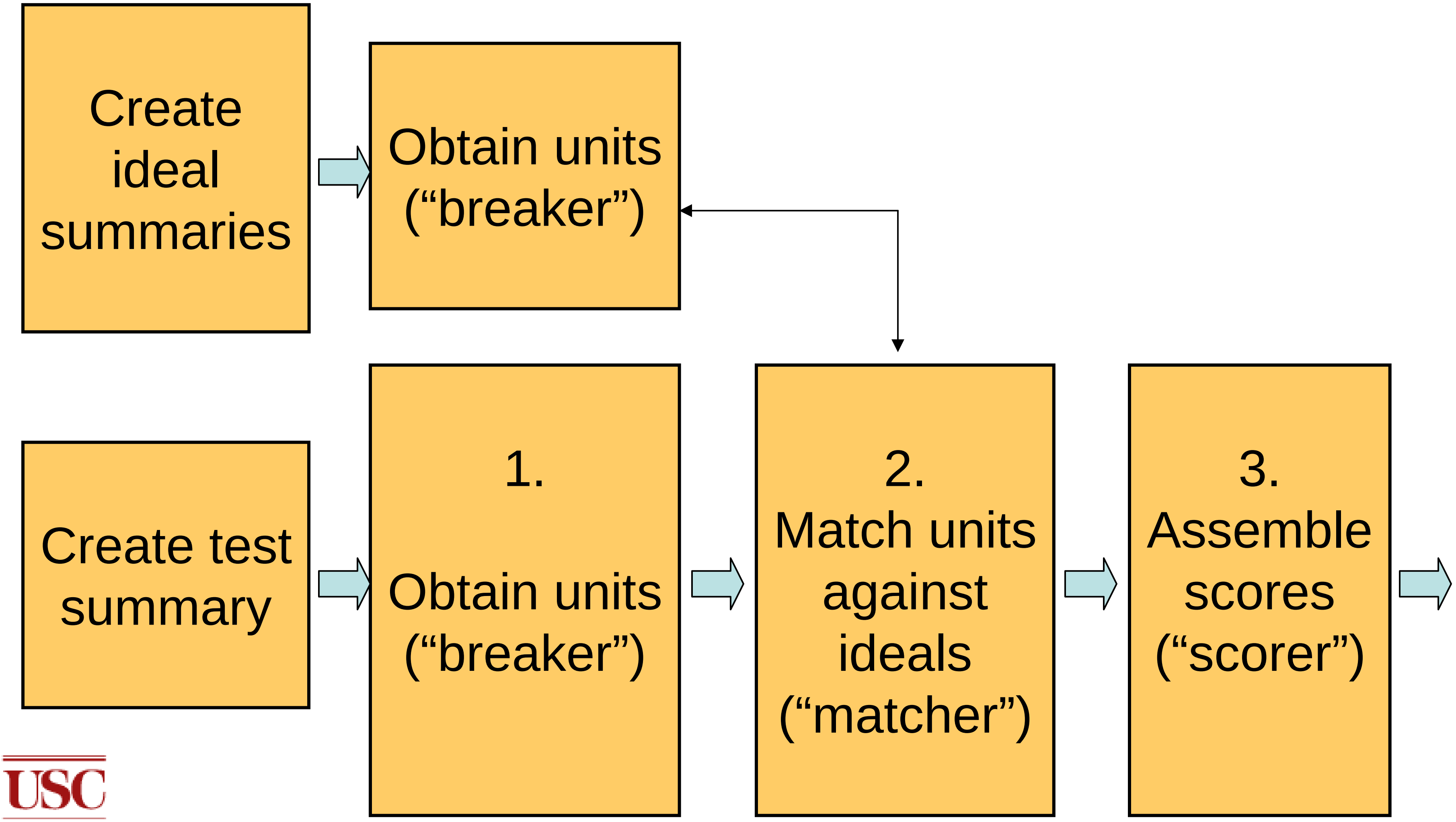
Goals

- **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

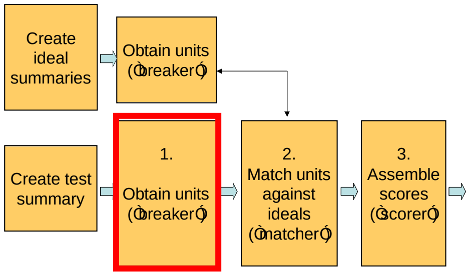
Desiderata for SummEval metric

- 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


Create ideal summaries
Obtain units (“breaker”)
Create test summary
1. Obtain units (“breaker”)
2. Match units against ideals (“matcher”)
3. Assemble scores (“scorer”)

1. Breaking



Create ideal summaries

Obtain units (Òbreaker Ó)

Create test summary

1.

Obtain units (Òbreaker Ó)

2. Match units against ideals (Òmatcher Ó)

3. Assemble scores (Òscorer Ó)

- 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

(thanks to Lucy Vanderwende et al., Microsoft)

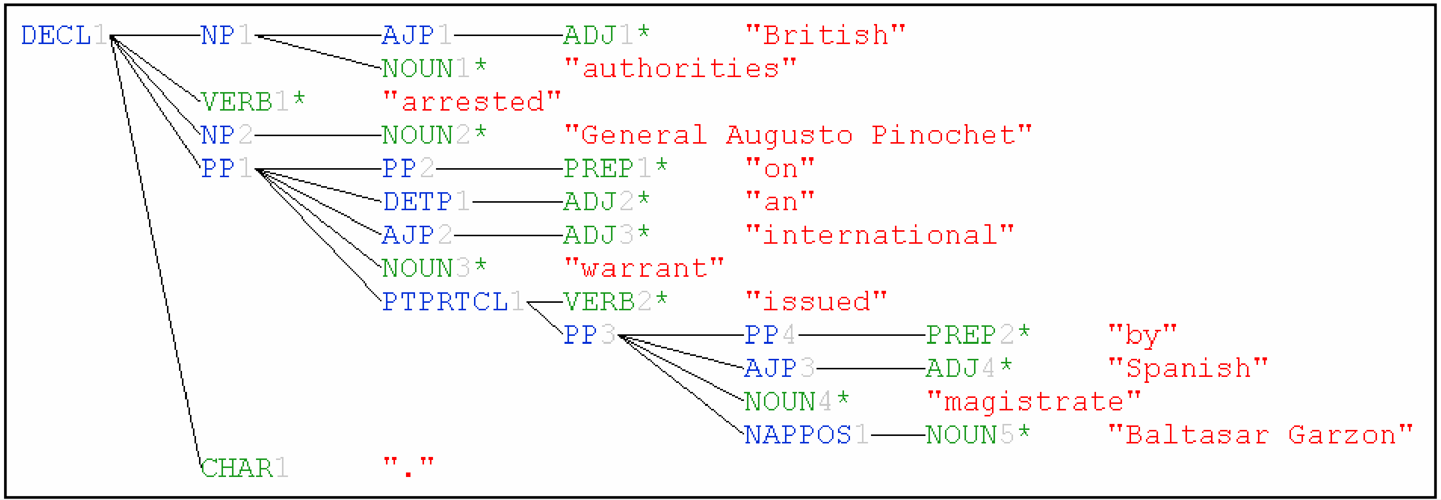
- Result: BEs of variable length/scope:
- Working definition: Each constituent Head, and each relation (between Head and Modifier) in a dependency tree is a candidate BE. Only the most important content-bearing ones are actually used for SummEval:
    - Head nouns and verbs
    - Verb plus its arguments
    - Noun plus its adjective/nominal/PP modifiers

  – Examples: [verb-Subj-noun], [noun-Mod-adj], [noun], [verb]

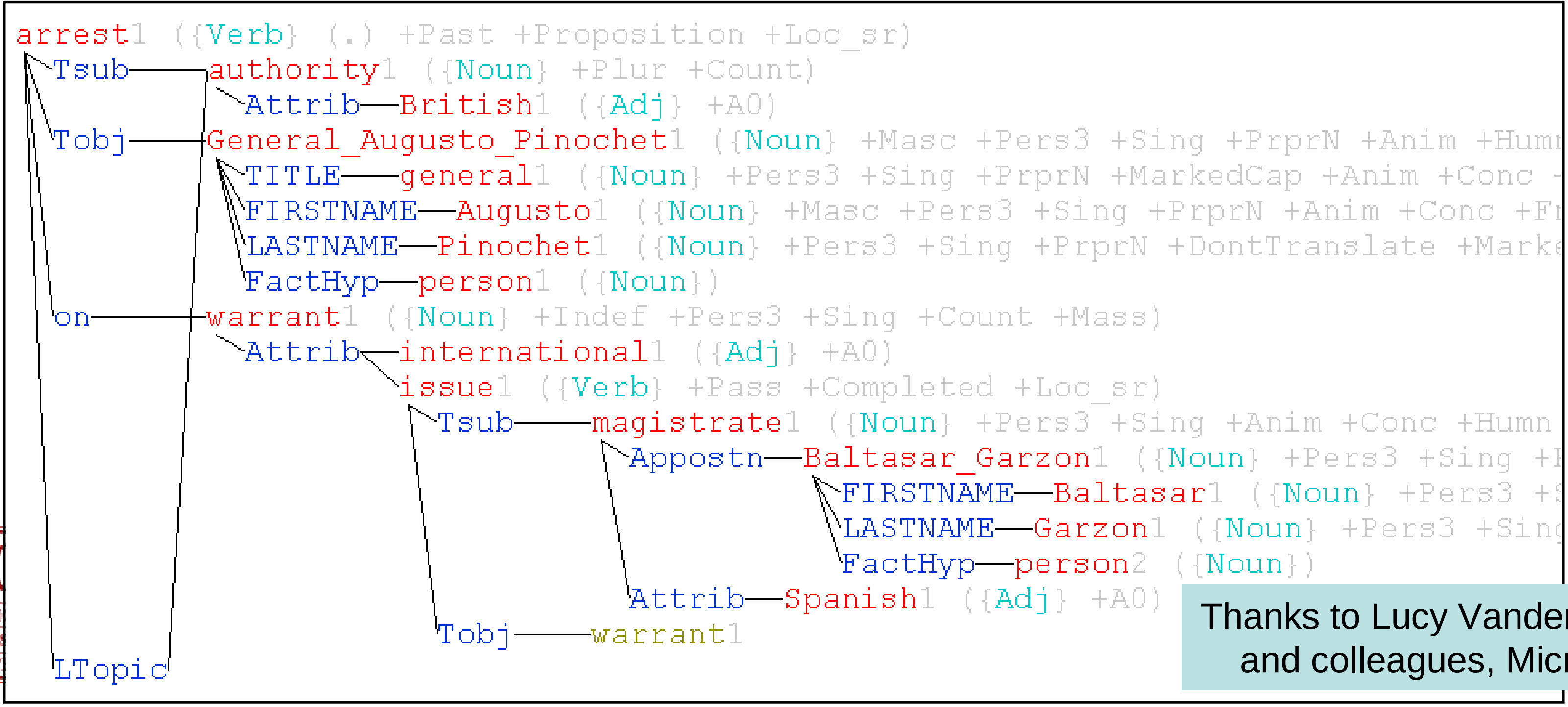
BEs: Syntactic or semantic?

- 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



```
DECL1──NP1──────AJP1─────ADJ1*     "British"
           └────NOUN1*    "authorities"
     ├─VERB1*    "arrested"
     ├─NP2──────NOUN2*    "General Augusto Pinochet"
     ├─PP1──────PP2──────PREP1*    "on"
     │     ├────DETP1────ADJ2*     "an"
     │     ├────AJP2─────ADJ3*     "international"
     │     ├────NOUN3*    "warrant"
     │     └────PTPRTCL1─VERB2*    "issued"
     │                 └─PP3──────PP4──────PREP2*    "by"
     │                      ├────AJP3─────ADJ4*     "Spanish"
     │                      ├────NOUN4*    "magistrate"
     │                      └────NAPPOS1──NOUN5*    "Baltasar Garzon"
     └─CHAR1     "."
```



```
arrest1 ({Verb} (.) +Past +Proposition +Loc_sr)
 ├Tsub─────authority1 ({Noun} +Plur +Count)
 │          └Attrib──British1 ({Adj} +A0)
 ├Tobj─────General_Augusto_Pinochet1 ({Noun} +Masc +Pers3 +Sing +PrprN +Anim +Humn
 │          ├TITLE─────general1 ({Noun} +Pers3 +Sing +PrprN +MarkedCap +Anim +Conc +
 │          ├FIRSTNAME─Augusto1 ({Noun} +Masc +Pers3 +Sing +PrprN +Anim +Conc +Fr
 │          ├LASTNAME──Pinochet1 ({Noun} +Pers3 +Sing +PrprN +DontTranslate +Marke
 │          └FactHyp───person1 ({Noun})
 ├on───────warrant1 ({Noun} +Indef +Pers3 +Sing +Count +Mass)
 │          └Attrib──international1 ({Adj} +A0)
 │                 └─issue1 ({Verb} +Pass +Completed +Loc_sr)
 │                    ├Tsub──────magistrate1 ({Noun} +Pers3 +Sing +Anim +Conc +Humn
 │                    │           ├Appostn──Baltasar_Garzon1 ({Noun} +Pers3 +Sing +F
 │                    │           │          ├FIRSTNAME──Baltasar1 ({Noun} +Pers3 +S
 │                    │           │          ├LASTNAME───Garzon1 ({Noun} +Pers3 +Sing
 │                    │           │          └FactHyp────person2 ({Noun})
 │                    │           └Attrib───Spanish1 ({Adj} +A0)
 │                    └Tobj──────warrant1
 └LTopic
```

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]

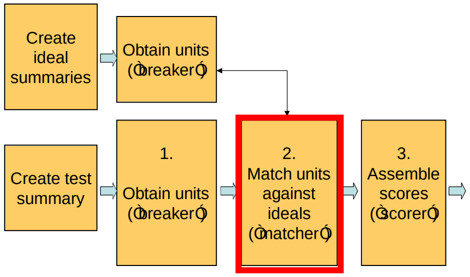
Using BEs to match Pyramid SCUs (MINIPAR + Fukumoto cutting rules)

Pyramid judgments										total overlap	BE
C.b2	D.b2	E.b2	F.b2	P.b2	Q.b2	R.b2	S.b2	U.b2	V.b2	df	<<BE element>>
1	0	1	1	1	0	0	0	1	0	5	defend <- themselves (obj)
0	1	1	1	1	0	0	0	0	0	4	security <- national (mod)
1	0	1	0	0	1	0	0	0	0	3	charge <- subvert (of)
0	1	0	0	0	1	1	0	0	0	3	civil <- and (punc)
0	1	0	0	0	1	1	0	0	0	3	civil <- political rights (conj)
1	0	0	0	1	0	0	1	0	0	3	incite <- subversion (obj)
0	0	1	0	0	0	1	1	0	0	3	president <- jiang zemin (person)
0	0	0	1	0	0	0	0	1	1	3	release <- china (subj)
1	0	0	0	1	0	0	0	0	0	2	action <- its (gen)
0	0	0	1	0	0	0	0	0	1	2	ail <- china (subj)
1	0	0	0	0	0	0	0	1	0	2	charge <- serious (mod)
1	0	0	0	1	0	0	0	0	0	2	defend <- action (obj)
1	0	0	0	1	0	0	0	0	0	2	defend <- china (subj)
0	0	0	1	0	0	0	0	1	0	2	defend <- dissident (subj)
1	0	0	1	0	0	0	0	0	0	2	democracy <- multiparty (nn)
0	1	0	0	0	0	0	0	1	0	2	dissident <- prominent (mod)
0	1	0	0	0	0	0	0	1	0	2	dissident <- three (nn)

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 0)</td><td>&lt;HEAD-MOD&gt;</td><td>(103_CD|-|-)</td><td>&lt;103:CARDINAL|-:NA&gt;</td></tr>
<tr><td>* (1 11 12)</td><td>&lt;HEAD-MOD&gt;</td><td>(in_IN|1988_CD|R)</td><td>&lt;in:NA|1988:DATE&gt;</td></tr>
<tr><td>* (1 12 0)</td><td>&lt;HEAD-MOD&gt;</td><td>(1988_CD|-|-)</td><td>&lt;1988:DATE|-:NA&gt;</td></tr>
<tr><td>* (1 14 0)</td><td>&lt;HEAD-MOD&gt;</td><td>(U.N._NNP|-|-)</td><td>&lt;U.N. Security Council:ORGANIZATION|-:NA&gt;</td></tr>
<tr><td>* (1 15 0)</td><td>&lt;HEAD-MOD&gt;</td><td>(Security_NNP|-|-)</td><td>&lt;U.N. Security Council:ORGANIZATION|-:NA&gt;</td></tr>
<tr><td>* (1 16 0)</td><td>&lt;HEAD-MOD&gt;</td><td>(Council_NNP|-|-)</td><td>&lt;U.N. Security Council:ORGANIZATION|-:NA&gt;</td></tr>
<tr><td>* (1 16 14)</td><td>&lt;HEAD-MOD&gt;</td><td>(Council_NNP|U.N._NNP|L)</td><td>&lt;U.N. Security Council:ORGANIZATION|U.N. Security Council:ORG&gt;</td></tr>
<tr><td>* (1 16 15)</td><td>&lt;HEAD-MOD&gt;</td><td>(Council_NNP|Security_NNP|L)</td><td>&lt;U.N. Security Council:ORGANIZATION|U.N. Security Council:ORG&gt;</td></tr>
<tr><td>* (1 17 0)</td><td>&lt;HEAD-MOD&gt;</td><td>(approves_VBZ|-|-)</td><td>&lt;approves:NA|-:NA&gt;</td></tr>
<tr><td>* (1 17 11)</td><td>&lt;HEAD-MOD&gt;</td><td>(approves_VBZ|in_IN|L)</td><td>&lt;approves:NA|in:NA&gt;</td></tr>
<tr><td>* (1 17 12)</td><td>&lt;PP&gt;</td><td>(approves_VBZ|1988_CD|in_DATE)</td><td></td></tr>
<tr><td>* (1 17 16)</td><td>&lt;HEAD-MOD&gt;</td><td>(approves_VBZ|Council_NNP|L)</td><td>&lt;approves:NA|U.N. Security Council:ORGA&gt;</td></tr>
<tr><td>* (1 17 18)</td><td>&lt;HEAD-MOD&gt;</td><td>(approves_VBZ|plan_NN|R)</td><td>&lt;approves:NA|plan:NA&gt;</td></tr>
<tr><td>* (1 17 2)</td><td>&lt;HEAD-MOD&gt;</td><td>(approves_VBZ|decade_NN|L)</td><td>&lt;approves:NA|A decade:DATE&gt;</td></tr>
<tr><td>* (1 17 24)</td><td>&lt;HEAD-MOD&gt;</td><td>(approves_VBZ|to_TO|R)</td><td>&lt;approves:NA|to:NA&gt;</td></tr>
<tr><td>* (1 17 25)</td><td>&lt;TO&gt;</td><td>(approves_VBZ|try_VB|to_NA)</td><td></td></tr>
<tr><td>* (1 17 3)</td><td>&lt;HEAD-MOD&gt;</td><td>(approves_VBZ|after_IN|L)</td><td>&lt;approves:NA|after:NA&gt;</td></tr>
<tr><td>* (1 17 5)</td><td>&lt;PP&gt;</td><td>(approves_VBZ|bombing_NN|after_NA)</td><td></td></tr>
<tr><td>* (1 17 9)</td><td>&lt;HEAD-MOD&gt;</td><td>(approves_VBZ|Flight_NNP|L)</td><td>&lt;approves:NA|Flight:NA&gt;</td></tr>
<tr><td>* (1 18 0)</td><td>&lt;HEAD-MOD&gt;</td><td>(plan_NN|-|-)</td><td>&lt;plan:NA|-:NA&gt;</td></tr>
<tr><td>* (1 18 19)</td><td>&lt;HEAD-MOD&gt;</td><td>(plan_NN|proposed_VBN|R)</td><td>&lt;plan:NA|proposed:NA&gt;</td></tr>
<tr><td>* (1 19 0)</td><td>&lt;HEAD-MOD&gt;</td><td>(proposed_VBN|-|-)</td><td>&lt;proposed:NA|-:NA&gt;</td></tr>
<tr><td>* (1 19 20)</td><td>&lt;HEAD-MOD&gt;</td><td>(proposed_VBN|by_IN|R)</td><td>&lt;proposed:NA|by:NA&gt;</td></tr>
<tr><td>* (1 19 21)</td><td>&lt;PP&gt;</td><td>(proposed_VBN|U.S._NNP|by_GPE)</td><td></td></tr>
<tr><td>* (1 2 0)</td><td>&lt;HEAD-MOD&gt;</td><td>(decade_NN|-|-)</td><td>&lt;A decade:DATE|-:NA&gt;</td></tr>
<tr><td>* (1 2 1)</td><td>&lt;HEAD-MOD&gt;</td><td>(decade_NN|A_DT|L)</td><td>&lt;A decade:DATE|A decade:DATE&gt;</td></tr>
</tbody>
</table>

2. Matching



Create ideal summaries

Obtain units (ÒbreakerÓ)

Create test summary

1. Obtain units (ÒbreakerÓ)

2. Match units against ideals (ÒmatcherÓ)

3. Assemble scores (ÒscorerÓ)

- 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?

USC INFORMATION SCIENCES INSTITUTE

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

Level of specificity	Feb 05 tests
WordNet replacement, top-level	91%
WordNet replacement, mid-level	⇧
Paraphrase	?%
Related-word expansion	⇩
Synonyms	
Derivational alternatives	
Root identity	
Word identity	40–50%

➡ **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

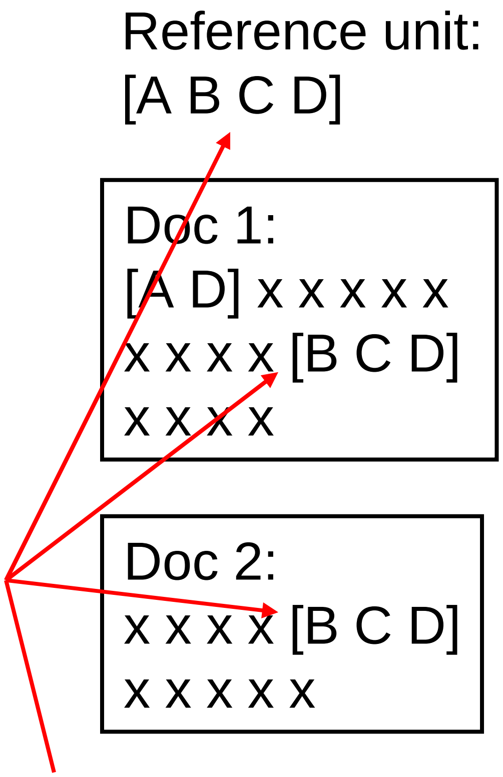
BE	file_loc	doc_freq
lockerbie bombing|1988 bombing|lockerbie bombing|lockerbie case |am bombing|bombing case|blowing over lockerbie scotland in 1988 .|wanted in bombing of flight over lockerbie|bombing suspects|linked to bombing|turning suspects in case|implicated over lockerbie scotland|blown over lockerbie scotland|implicated in bombing|indicted for bombing|wanted for bombing in 1988 which killed|==NE|event|	R1 S2 U1 V1 Q1 B1 P1 D1 A1	0.9
lockerbie bombing|1988 bombing|pan bombing|lockerbie bombing|1988 bombing|lockerbie bombing|lockerbie case |am bombing|bombing case|blowing over lockerbie scotland in 1988 .|wanted in bombing of flight over lockerbie|bombing suspects|linked to bombing|turning suspects in case|implicated over lockerbie scotland|blown over lockerbie scotland|implicated in bombing|indicted for bombing|wanted for bombing in 1988 which killed|==NE|act, human action, human activity|	A1 Q1 R1 S2 U1 V1 B1 P1 D1	0.9
two libyans|two libyans|two suspects|two agents|two suspects|two suspects|two libyans|two suspects|hand suspects wanted|libyan agents|try suspects in netherlands|bombing suspects|turning suspects in case|jumbo jet|intelligence agents|==two|entity|	A1 B1 P1 Q1 R1 S2 U1 V1	0.8
pan jet|pan jet|lockerbie suspects|hand suspects wanted|blowing jet|blowing over lockerbie scotland in 1988 .|wanted in bombing of flight over lockerbie|try suspects in netherlands|bombing suspects|turning suspects in case|implicated over lockerbie scotland|jumbo jet|intelligence agents|blown over lockerbie scotland|==NE|entity|	B1 D1 P1 R1 S2 V1 Q1	0.7
am jet|am jet|libyan suspects|hand suspects wanted|blowing jet|try suspects in netherlands|bombing suspects|turning suspects in case |jumbo jet|intelligence agents|==entity|entity|	B1 D1 P1 R1 V1 Q1	0.6
december 1988|moammar gadhafi|moammar gadhafi|libyan gadhafi |libyan gadhafi|blowing over lockerbie scotland in 1988 .|agreed by gadhafi|leader gadhafi|col. gadhafi |leader gadhafi|wanted for bombing in 1988 which killed|==NE|	D1 Q1 V1 B1 R1	0.5
pan flight|u.n. council|united states|pam flight|wanted in bombing of flight over lockerbie|flight 103|flight 103|am flight|==NE|group, grouping|	P1 Q1 S2	0.3
blowing up|blown up|==change	B1 D1	0.2
indicted in 1991|indicted in 1991|==charge, accuse|in|NE|	A1 B1	0.2
hand over|turning over|==transfer	R1 V1	0.2
were indicted|==be|charge, accuse|	B1	0.1
try in court|==act, move|in|group, grouping|	P1	0.1

----------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] "Lockerbie"
[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?

SEE (Lin 2001)

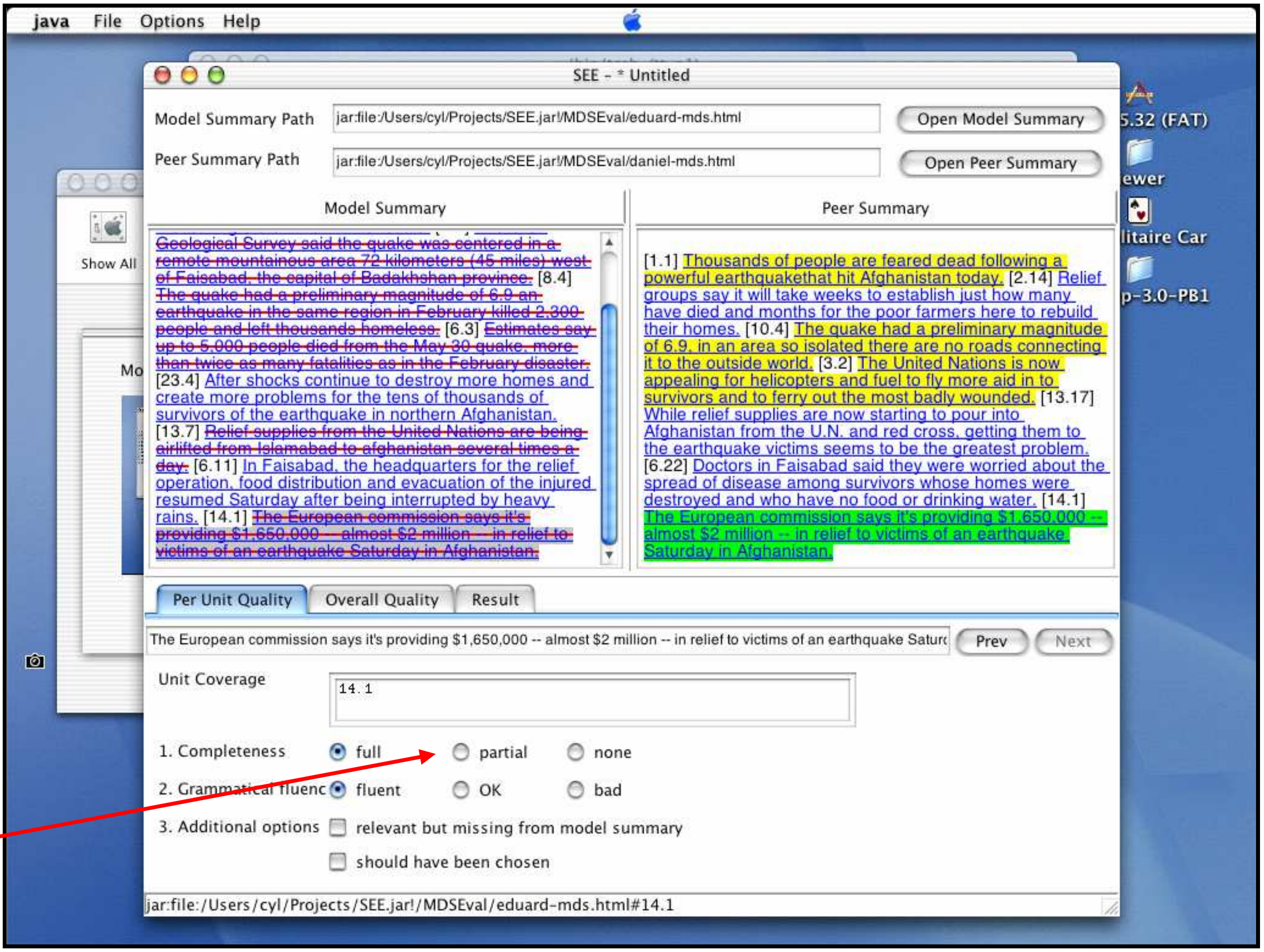


Reference unit:
[A B C D]

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 x

Partial score,
or problems!



Model Summary Path: jar:file:/Users/cyl/Projects/SEE.jar!/MDSEval/eduard-mds.html

Peer Summary Path: jar:file:/Users/cyl/Projects/SEE.jar!/MDSEval/daniel-mds.html

Model Summary

~~Geological Survey said the quake was centered in a remote mountainous area 72 kilometers (45 miles) west of Faisabad, the capital of Badakhshan province.~~ [8.4] ~~The quake had a preliminary magnitude of 6.9 an earthquake in the same region in February killed 2,300 people and left thousands homeless.~~ [6.3] ~~Estimates say up to 5,000 people died from the May 30 quake, more than twice as many fatalities as in the February disaster.~~ [23.4] After shocks continue to destroy more homes and create more problems for the tens of thousands of survivors of the earthquake in northern Afghanistan. [13.7] ~~Relief supplies from the United Nations are being airlifted from Islamabad to afghanistan several times a day.~~ [6.11] In Faisabad, the headquarters for the relief operation, food distribution and evacuation of the injured resumed Saturday after being interrupted by heavy rains. [14.1] ~~The European commission says it's providing $1,650,000 -- almost $2 million -- in relief to victims of an earthquake Saturday in Afghanistan.~~

Peer Summary

[1.1] Thousands of people are feared dead following a powerful earthquakethat hit Afghanistan today. [2.14] Relief groups say it will take weeks to establish just how many have died and months for the poor farmers here to rebuild their homes. [10.4] The quake had a preliminary magnitude of 6.9, in an area so isolated there are no roads connecting it to the outside world. [3.2] The United Nations is now appealing for helicopters and fuel to fly more aid in to survivors and to ferry out the most badly wounded. [13.17] While relief supplies are now starting to pour into Afghanistan from the U.N. and red cross, getting them to the earthquake victims seems to be the greatest problem. [6.22] Doctors in Faisabad said they were worried about the spread of disease among survivors whose homes were destroyed and who have no food or drinking water. [14.1] The European commission says it's providing $1,650,000 -- almost $2 million -- in relief to victims of an earthquake Saturday in Afghanistan.

The European commission says it's providing $1,650,000 -- almost $2 million -- in relief to victims of an earthquake Saturc

Unit Coverage: 14.1

1. Completeness ● full ○ partial ○ none

2. Grammatical fluenc ● fluent ○ OK ○ bad

3. Additional options ☐ relevant but missing from model summary

☐ should have been chosen

jar:file:/Users/cyl/Projects/SEE.jar!/MDSEval/eduard-mds.html#14.1

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"

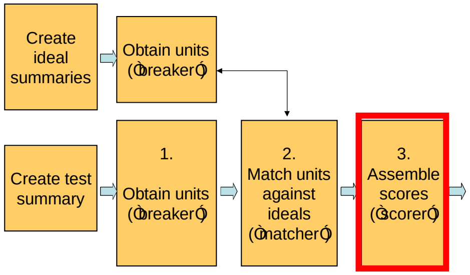
Semantic/paraphrase matching

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



Create ideal summaries

Obtain units (ÒbreakerÓ)

Create test summary

1. Obtain units (ÒbreakerÓ)

2. Match units against ideals (ÒmatcherÓ)

3. Assemble scores (ÒscorerÓ)

- **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.)

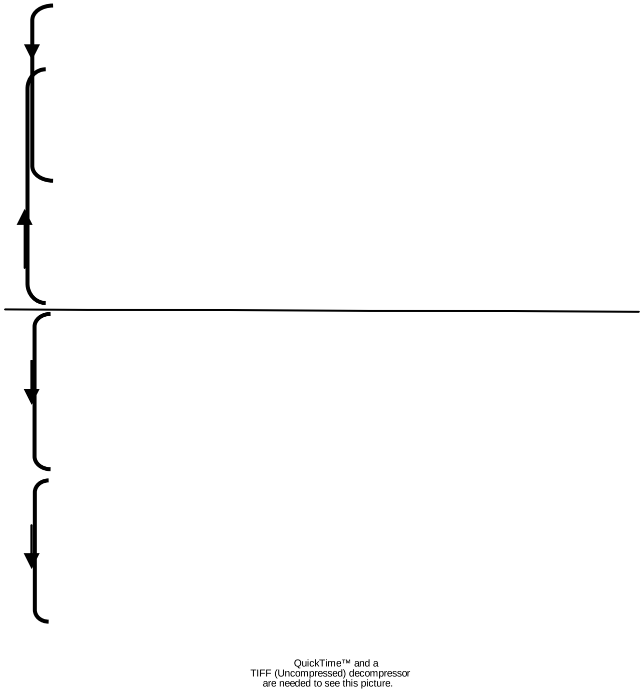
BE scoring

- 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

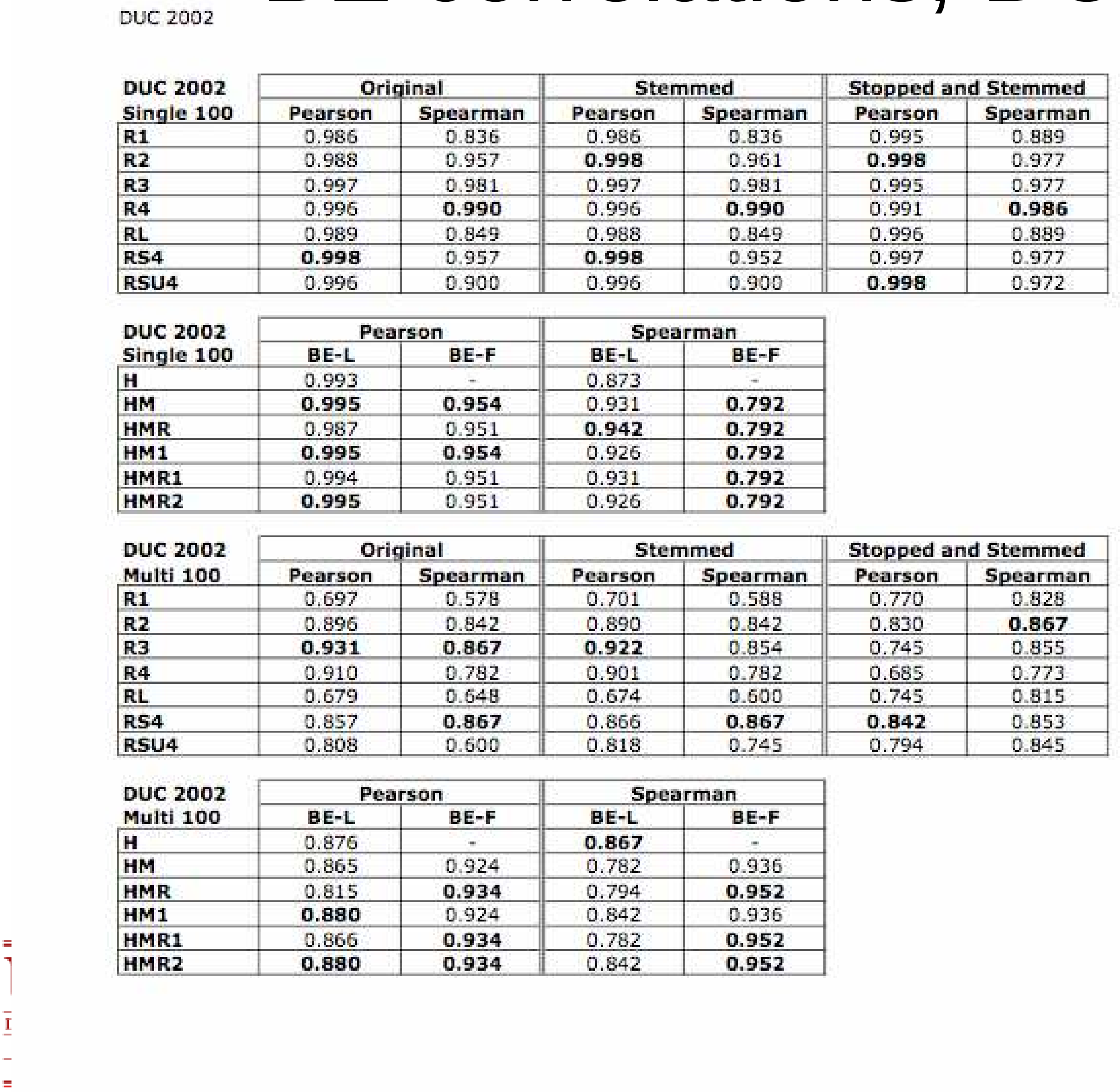
BE scores for DUC 05

- Recall differentiates well


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

USC
INFORMATION
SCIENCES
INSTITUTE

BE correlations, DUC 2002



DUC 2002

DUC 2002	Original		Stemmed		Stopped and Stemmed	
Single 100	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
R1	0.986	0.836	0.986	0.836	0.995	0.889
R2	0.988	0.957	**0.998**	0.961	**0.998**	0.977
R3	0.997	0.981	0.997	0.981	0.995	0.977
R4	0.996	**0.990**	0.996	**0.990**	0.991	**0.986**
RL	0.989	0.849	0.988	0.849	0.996	0.889
RS4	**0.998**	0.957	**0.998**	0.952	0.997	0.977
RSU4	0.996	0.900	0.996	0.900	**0.998**	0.972

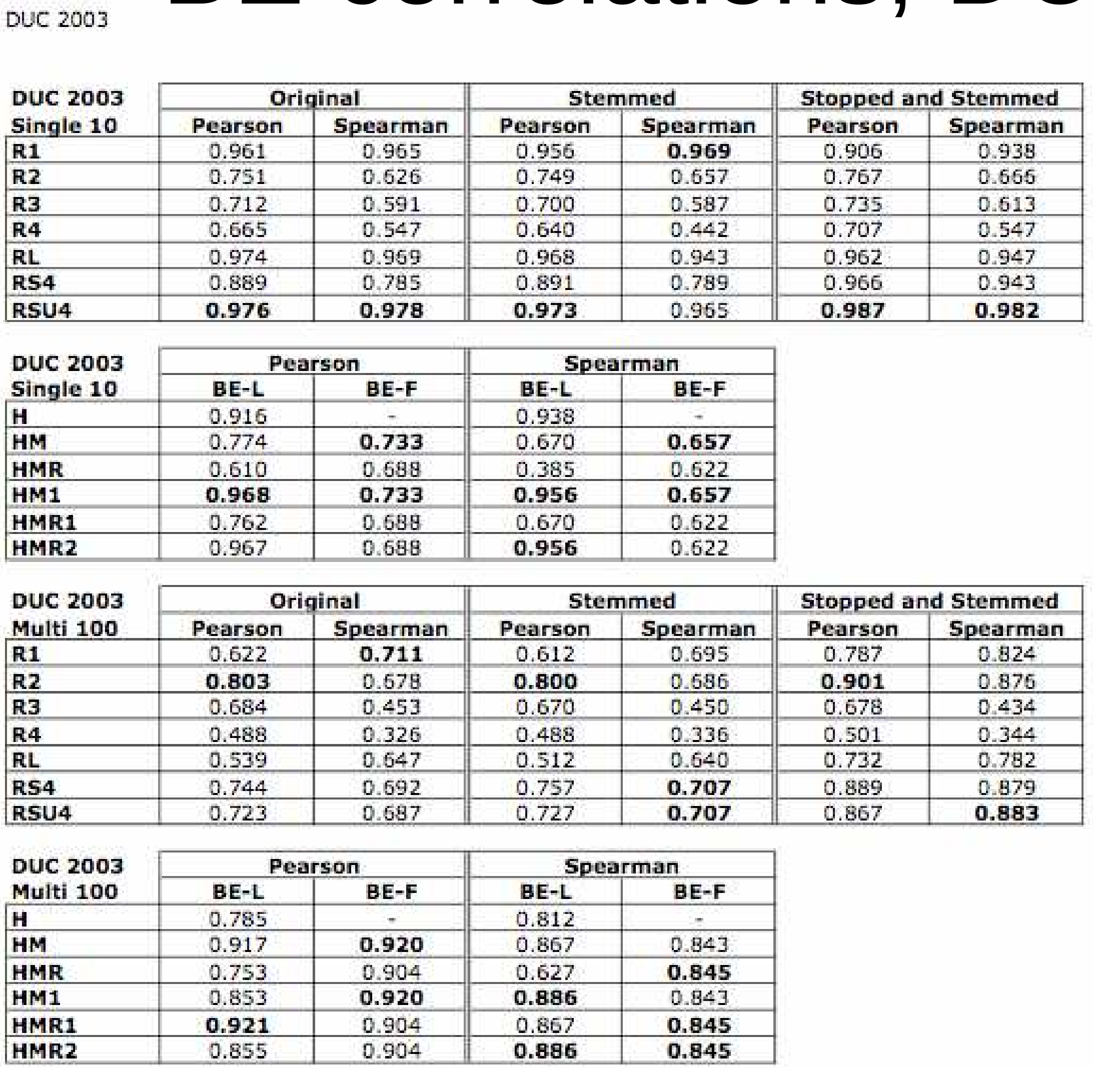
DUC 2002	Pearson		Spearman	
Single 100	BE-L	BE-F	BE-L	BE-F
H	0.993	-	0.873	-
HM	**0.995**	**0.954**	0.931	**0.792**
HMR	0.987	0.951	**0.942**	**0.792**
HM1	**0.995**	**0.954**	0.926	**0.792**
HMR1	0.994	0.951	0.931	**0.792**
HMR2	**0.995**	0.951	0.926	**0.792**

DUC 2002	Original		Stemmed		Stopped and Stemmed	
Multi 100	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
R1	0.697	0.578	0.701	0.588	0.770	0.828
R2	0.896	0.842	0.890	0.842	0.830	**0.867**
R3	**0.931**	**0.867**	**0.922**	0.854	0.745	0.855
R4	0.910	0.782	0.901	0.782	0.685	0.773
RL	0.679	0.648	0.674	0.600	0.745	0.815
RS4	0.857	**0.867**	0.866	**0.867**	**0.842**	0.853
RSU4	0.808	0.600	0.818	0.745	0.794	0.845

DUC 2002	Pearson		Spearman	
Multi 100	BE-L	BE-F	BE-L	BE-F
H	0.876	-	**0.867**	-
HM	0.865	0.924	0.782	0.936
HMR	0.815	**0.934**	0.794	**0.952**
HM1	**0.880**	0.924	0.842	0.936
HMR1	0.866	**0.934**	0.782	**0.952**
HMR2	**0.880**	**0.934**	0.842	**0.952**

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)

BE correlations, DUC 2003



DUC 2003

DUC 2003	Original		Stemmed		Stopped and Stemmed	
Single 10	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
R1	0.961	0.965	0.956	**0.969**	0.906	0.938
R2	0.751	0.626	0.749	0.657	0.767	0.666
R3	0.712	0.591	0.700	0.587	0.735	0.613
R4	0.665	0.547	0.640	0.442	0.707	0.547
RL	0.974	0.969	0.968	0.943	0.962	0.947
RS4	0.889	0.785	0.891	0.789	0.966	0.943
RSU4	**0.976**	**0.978**	**0.973**	0.965	**0.987**	**0.982**

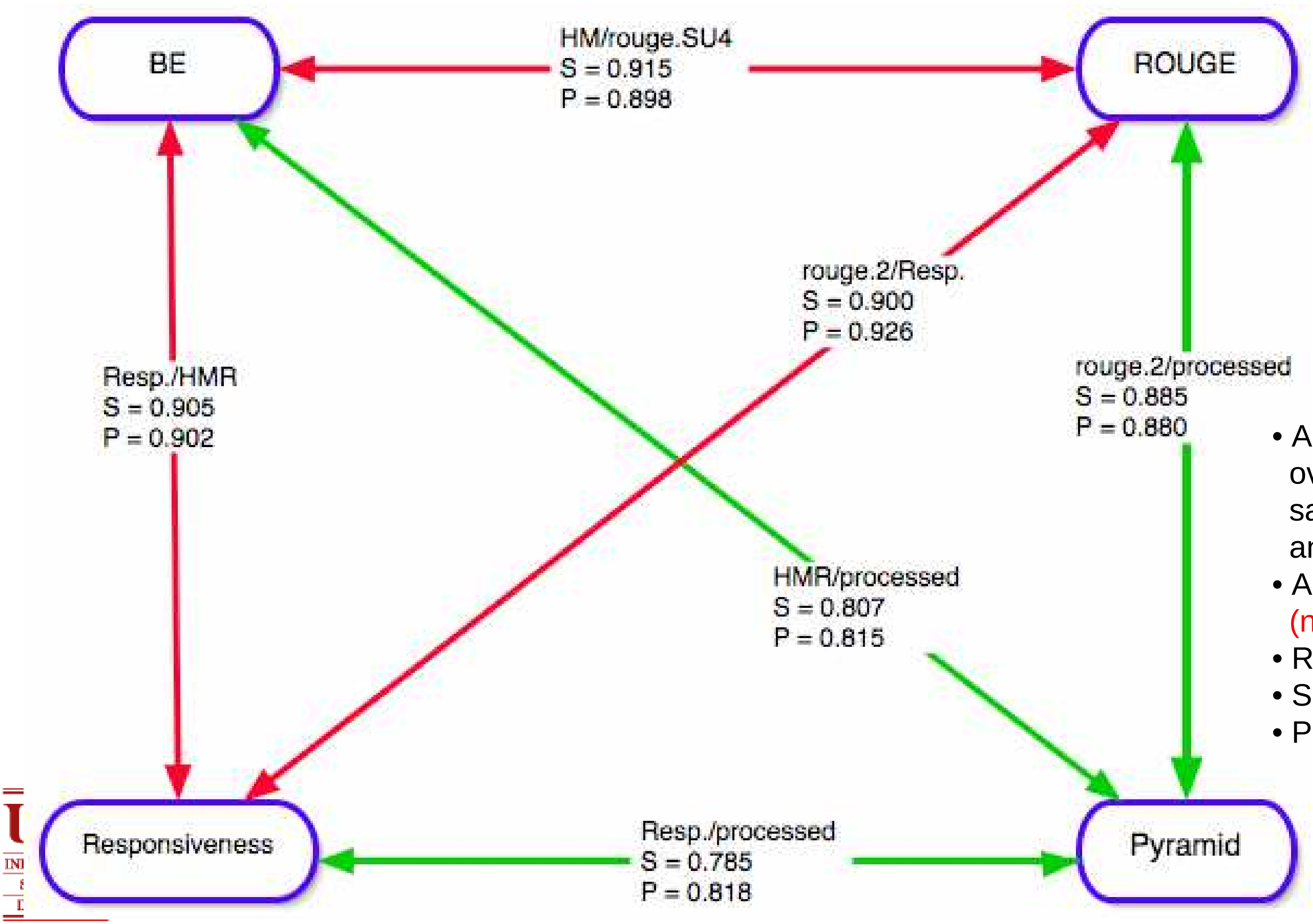
DUC 2003	Pearson		Spearman	
Single 10	BE-L	BE-F	BE-L	BE-F
H	0.916	-	0.938	-
HM	0.774	**0.733**	0.670	**0.657**
HMR	0.610	0.688	0.385	0.622
HM1	**0.968**	**0.733**	**0.956**	**0.657**
HMR1	0.762	0.688	0.670	0.622
HMR2	0.967	0.688	**0.956**	0.622

DUC 2003	Original		Stemmed		Stopped and Stemmed	
Multi 100	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
R1	0.622	**0.711**	0.612	0.695	0.787	0.824
R2	**0.803**	0.678	**0.800**	0.686	**0.901**	0.876
R3	0.684	0.453	0.670	0.450	0.678	0.434
R4	0.488	0.326	0.488	0.336	0.501	0.344
RL	0.539	0.647	0.512	0.640	0.732	0.782
RS4	0.744	0.692	0.757	**0.707**	0.889	0.879
RSU4	0.723	0.687	0.727	**0.707**	0.867	**0.883**

DUC 2003	Pearson		Spearman	
Multi 100	BE-L	BE-F	BE-L	BE-F
H	0.785	-	0.812	-
HM	0.917	**0.920**	0.867	0.843
HMR	0.753	0.904	0.627	**0.845**
HM1	0.853	**0.920**	**0.886**	0.843
HMR1	**0.921**	0.904	0.867	**0.845**
HMR2	0.855	0.904	**0.886**	**0.845**

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)

BE correlations 1, DUC 2005



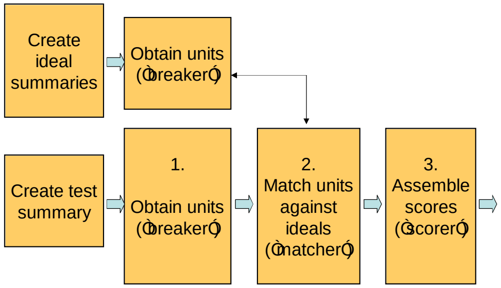
- All comparisons over exactly the same 20 topics and 25 systems
- All 9 references (not just 7)
- Recall scores
- S = Spearman
- P = Pearson

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



Create ideal summaries

Obtain units (ÒbreakerÓ)

Create test summary

1. Obtain units (ÒbreakerÓ)

2. Match units against ideals (ÒmatcherÓ)

3. Assemble scores (ÒscorerÓ)

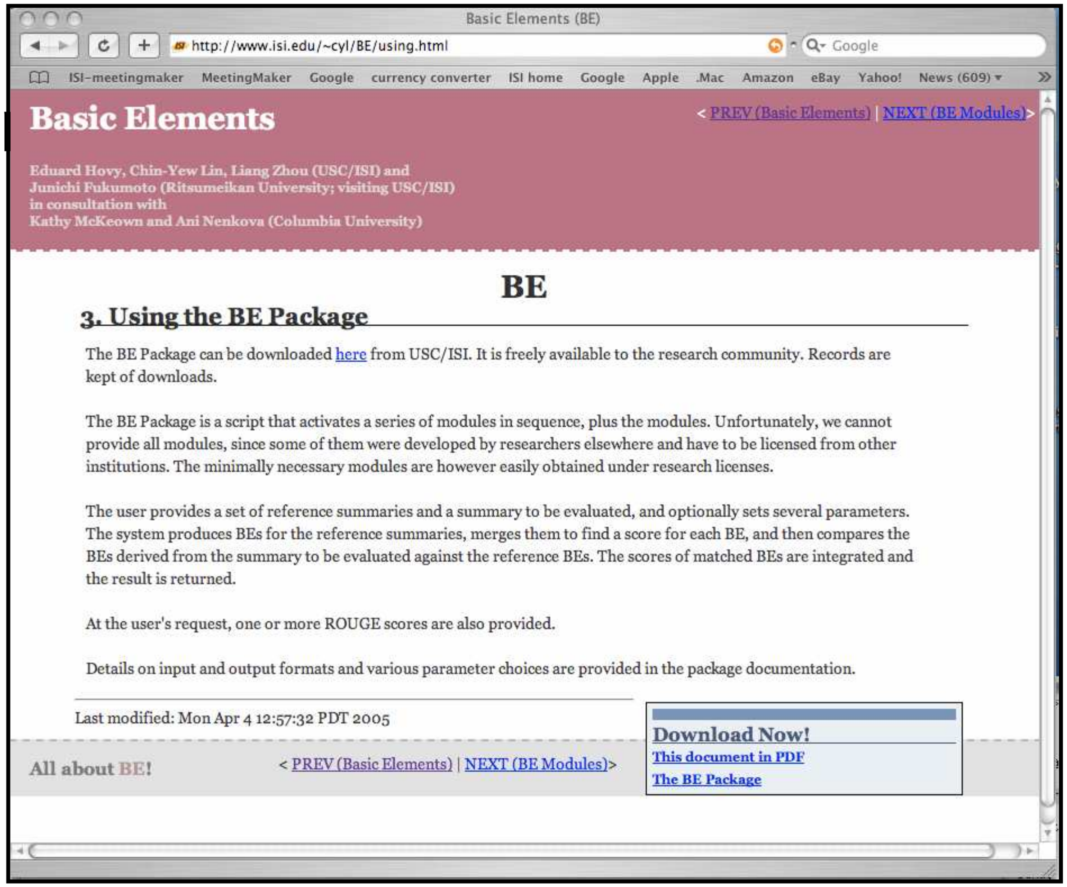
Method	1. Units	2. Matching	3. Scoring
SEE	sentences, auto	manual, partial ok	add partial points
ROUGE	auto ngrams, various kinds	string match, stemmed/not	single-point, also weighted
Van Halteren & Teufel	factoids, manual	manual, assessors	popularity score
Pyramid	SCUs, manual	manual, community	popularity score
BE method	BEs, auto	string match	popularity

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 ow modules!
  - Unit breakers
  - Matchers
  - Scorers



Basic Elements

< PREV (Basic Elements) | NEXT (BE Modules)>

Eduard Hovy, Chin-Yew Lin, Liang Zhou (USC/ISI) and
Junichi Fukumoto (Ritsumeikan University; visiting USC/ISI)
in consultation with
Kathy McKeown and Ani Nenkova (Columbia University)

BE

3. Using the BE Package

The BE Package can be downloaded here from USC/ISI. It is freely available to the research community. Records are kept of downloads.

The BE Package is a script that activates a series of modules in sequence, plus the modules. Unfortunately, we cannot provide all modules, since some of them were developed by researchers elsewhere and have to be licensed from other institutions. The minimally necessary modules are however easily obtained under research licenses.

The user provides a set of reference summaries and a summary to be evaluated, and optionally sets several parameters. The system produces BEs for the reference summaries, merges them to find a score for each BE, and then compares the BEs derived from the summary to be evaluated against the reference BEs. The scores of matched BEs are integrated and the result is returned.

At the user's request, one or more ROUGE scores are also provided.

Details on input and output formats and various parameter choices are provided in the package documentation.

Last modified: Mon Apr 4 12:57:32 PDT 2005

Download Now!
This document in PDF
The BE Package

All about BE!

< PREV (Basic Elements) | NEXT (BE Modules)>

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