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

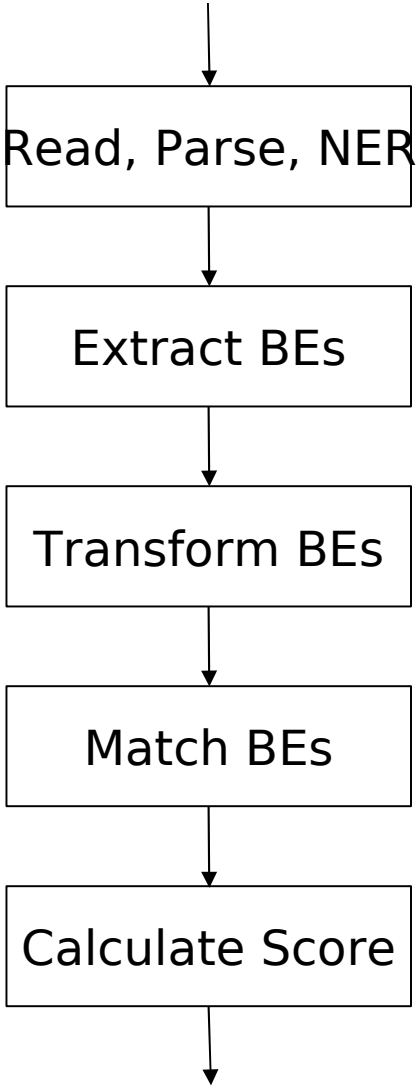
- Same information conveyed in many different ways
  - Information omitted, word order rearranged, names abbreviated, etc.
- N-gram matching restricted to surface form
  - “large green car” != “large car”
  - “large green car” != “heavy emerald vehicle”
  - “USA” != “United States”, “America”

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: drank|VBD+milk|NN

Rule Name: Possessor of NPHead
Tregex: NP [< (NP <# (POS $- __=x)) & <# __=y]
Tokens to Extract: xy
Extracted BEs: John|Person+cat|NN

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

USC INFORMATION SCIENCES INSTITUTE

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(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

$$maximize \sum_{i=0}^{N} \sum_{j=0}^{M} C(i,j) W(j) x_{ij}$$

$$subject\ \text{to}$$

$$\sum_{i=0}^{N} x_{ij} \in \{0,1\}\ forall j where\ 0 \le j \le M$$

$$\sum_{j=0}^{M} x_{ij} \in \{0,1\}\ forall i where\ 0 \le i \le N$$

$$x_{ij} \in \{0,1\}$$

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

	Spearman			Pearson		
	All	Auto	Hu	All	Auto	Hu
BEwT-E	**0.864**	**0.802**	0.539	**0.925**	**0.840**	0.549
Original BE	**0.873**	**0.815**	0.467	**0.887**	**0.817**	0.595
ROUGE2	**0.905**	**0.867**	0.539	**0.851**	**0.829**	*0.645*
ROUGESU4	**0.884**	**0.832**	**0.874**	**0.852**	**0.802**	**0.846**
Mod Pyramid	**0.917**	**0.878**	0.611	**0.968**	**0.900**	0.509

vs Modified Pyramid

	Spearman			Pearson		
	All	Auto	Hu	All	Auto	Hu
BEwT-E	**0.955**	**0.935**	**0.833**	**0.950**	**0.950**	*0.665*
Original BE	**0.934**	**0.904**	*0.762*	**0.917**	**0.913**	*0.663*
ROUGE2	**0.936**	**0.907**	**0.857**	**0.869**	**0.907**	0.544
ROUGESU4	**0.919**	**0.883**	**0.857**	**0.871**	**0.886**	0.543
Responsiveness	**0.917**	**0.878**	0.611	**0.968**	**0.900**	0.509

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

USC INFORMATION SCIENCES INSTITUTE

Results on TAC08 Part B

vs Responsiveness

	Spearman			Pearson		
	All	Auto	Hu	All	Auto	Hu
BEwT-E	**0.926**	**0.891**	*0.802*	**0.925**	**0.924**	*0.642*
Original BE	**0.917**	**0.877**	*0.683*	**0.905**	**0.912**	0.464
ROUGE2	**0.920**	**0.882**	0.587	**0.882**	**0.909**	0.579
ROUGESU4	**0.927**	**0.893**	**0.898**	**0.835**	**0.901**	**0.796**
Mod Pyramid	**0.948**	**0.925**	*0.695*	**0.980**	**0.949**	*0.741*

vs Modified Pyramid

	Spearman			Pearson		
	All	Auto	Hu	All	Auto	Hu
BEwT-E	**0.969**	**0.955**	0.595	**0.941**	**0.954**	0.474
Original BE	**0.957**	**0.938**	0.190	**0.915**	**0.943**	0.054
ROUGE2	**0.959**	**0.942**	-0.024	**0.896**	**0.942**	-0.014
ROUGESU4	**0.952**	**0.931**	0.357	**0.859**	**0.925**	0.333
Responsiveness	**0.948**	**0.925**	*0.695*	**0.980**	**0.949**	*0.741*

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

USC INFORMATION SCIENCES INSTITUTE

Effect of Transformations

One Transform Off	All		Auto		Human	
	+	-	+	-	+	-
...						
Hyper/Hyponyms	**140**	101	**153**	88	71	86
...						

- 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

	All		Auto		Human	
	+	-	+	-	+	-
DUC07	26	19	31	14	11	17
DUC06	30	20	29	21	14	16
DUC05	38	12	35	15	18	19
TAC08 Base	25	23	24	24	13	23
TAC08 Update	27	21	23	25	11	15
Total	**146**	95	**142**	99	67	90

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