Towards Robust Abstractive Multi-Document Summarization: A Caseframe Analysis of Centrality and Domain

Jackie CK Cheung and Gerald Penn

{jcheung,gpenn}@cs.toronto.edu

Aug 6, 2013

UNIVERSITY OF TORONTO

Centrality and Extraction

- **Centrality**—a summary should contain the parts of the source text that are most representative of it
- Explicitly modelled as summarization objective
  - e.g., MMR (Carbonell and Goldstein, 1998)

    objective = centrality term + non-redundancy term
- Refined by more sophisticated methods
  - e.g. Term weighting (Lin and Hovy, 2000)
  - Core component of most successful current methods (Conroy et al., 2006)

Limits of Extraction

- Compression ratio
  - Text simplification e.g., (Knight and Marcu, 2000)
  - Sentence fusion e.g., (Barzilay and McKeown, 2005)
- Coherence
  - Avoid dangling referents
  - Text structuring e.g., (Christensen et al., 2013)
- **Aggregation** and **information synthesis**
  - Key part of potential utility of automatic summaries
  - Limited work outside of specific genres and domains

Message of This Paper

- Extractive centrality-based summarization systems currently dominate summarization shared tasks
- Advance towards robust abstraction **not** by better optimizing centrality-based measures
- Require return to more domain knowledge

- Studies on TAC Guided Summarization data
  - Compare characteristics of model summaries vs. state-of-the-art summarizers

Previous Studies on Summarization

- Best possible extractive system using word-overlap measures such as ROUGE

  (Lin and Hovy, 2003; Conroy et al., 2006)

  - Best possible extractive summary as good as humans
  - ROUGE not designed for this purpose
- Human-created extractive summaries

  (Genest et al., 2009)

  - Score in between current automatic systems and abstracts on responsiveness, linguistic quality, and Pyramid

More Related Studies

- **Cut-and-paste** operations

  (Jing and McKeown, 2000)

  - 19% of analyzed sentences cannot be explained by these processes
- (Saggion and Lapalme, 2002)
  - Definition and analysis of transformations necessary to convert source text to summary text
- (Copeck and Szpakowicz, 2004)
  - 55% of vocabulary items found in model summaries occur in source text

Novelty of Our Studies

- Analysis of impact of domain knowledge for multi-document summarization
  - Made possible by use of recent guided summarization data
- **Developmental** approach, not **evaluative**
  - Distinguish model and peer summaries in a useful way
  - Guide development of future systems
- Analysis at a shallow semantic level (**caseframes**)
  - In contrast to previous use of word overlap or syntactic measures

Overview of Studies

- **Study 1: How to measure aggregation?**
  - Quantitative measure of sentence aggregation
- **Study 2: How do humans aggregate information?**
  - Not just by centrality—automatic systems are already more “central” than human summarizers with respect to source text
- **Study 3: How to generate human-like summaries?**
  - Domain knowledge as a source of information for abstractive summarization systems

Unit of Analysis: Caseframes

- *(gov, role)* pairs extracted from dependency parse
  - *gov*: a proposition-bearing unit (verb, event noun, nominal or adjectival predicate)
  - *role*: semantic role derived from grammatical role
  - e.g. *(kill, dobj), (hurt, nsubj), (murder, prep_of)*
- Approximation of semantic role structure
  - Distinct from case frames in Case Grammar
- Can be automatically extracted
- Well-suited to characterize a domain
  - Abstracts away syntactic alternations, entity realizations, etc.

Example

- Cluster: Unabomber trial

  Theodore Kaczynski faces a federal indictment for 4 mail bomb attacks attributed to the Unabomber in which two people were killed. If found guilty, he faces a death penalty. He has pleaded innocent to all charges. District Judge Garland Burrel Jr. presides

  - DEFENDANT *(face, nsubj), (plead, nsubj)*
  - CHARGES *(face, dobj)*
  - REASON *(indictment, prep_for)*
  - SENTENCE *(face, dobj)*
  - PLEAD *(plead, dobj)*
  - JUDGE *(preside, nsubj)*

Data Set

- TAC 2010 Guided Summarization
  - 920 documents
  - 46 topic clusters in 5 domains
  - Templates provided to provide guidance to systems
- Initial vs. update summarization task
- Summarizers:
  - 8 human **model** summary writers (alphabetic: A – H)
  - 43 **peer** summarization systems (1 – 43)
    - Removed two systems that did not generate summaries for most topic clusters

Peer Comparison Conditions

- **Peer average**
  - Average of 41 peer summarizers
- **Peer 16**
  - Best in responsiveness in initial task
  - Best in ROUGE-2, responsiveness, Pyramid in update task
- **Peer 22**
  - Best in ROUGE-2, Pyramid in initial task
- **Peer 1**
  - NIST's leading baseline from most recent document
  - Best in linguistic quality in both tasks

Study 1: Sentence Aggregation

- Quantitative measure of degree of aggregation
- **Average sentence cover size**
  - Minimum # sentences from the source text needed to cover all of the caseframes found in a summary sentence (for those that can be found in the source text)
  - Take average of this over all summary sentences
  - Pure extraction = 1.0

e.g. Summary sentence: {1,2,3,4,5}

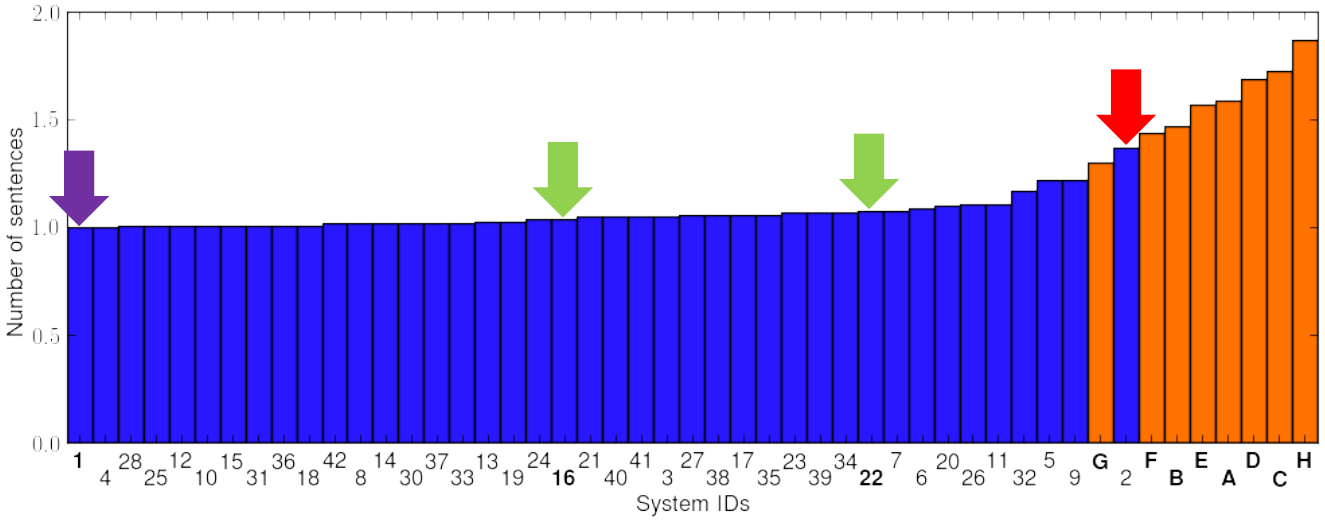
Source text: {{1, 3, 4}, {2,5,6}, {1,4,7}}

Cover size = 2: {{1,3,4}, {2,5,6}}

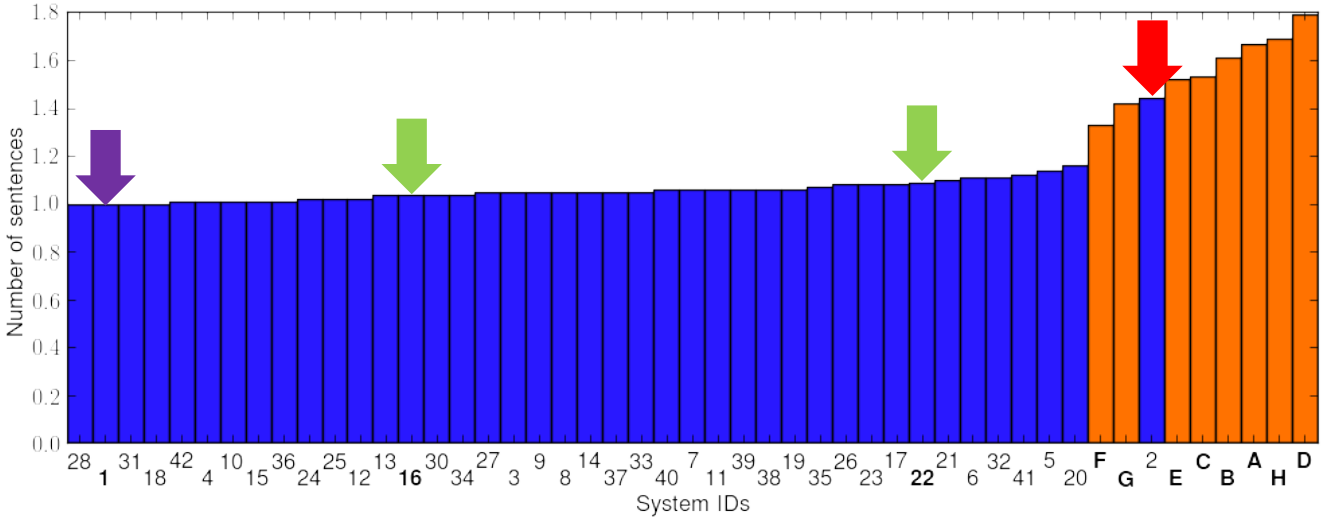
  - Solved optimally by ILOG CPLEX

Study 1: Sentence Aggregation

- Initial



- Update



Study 1: Sentence Aggregation

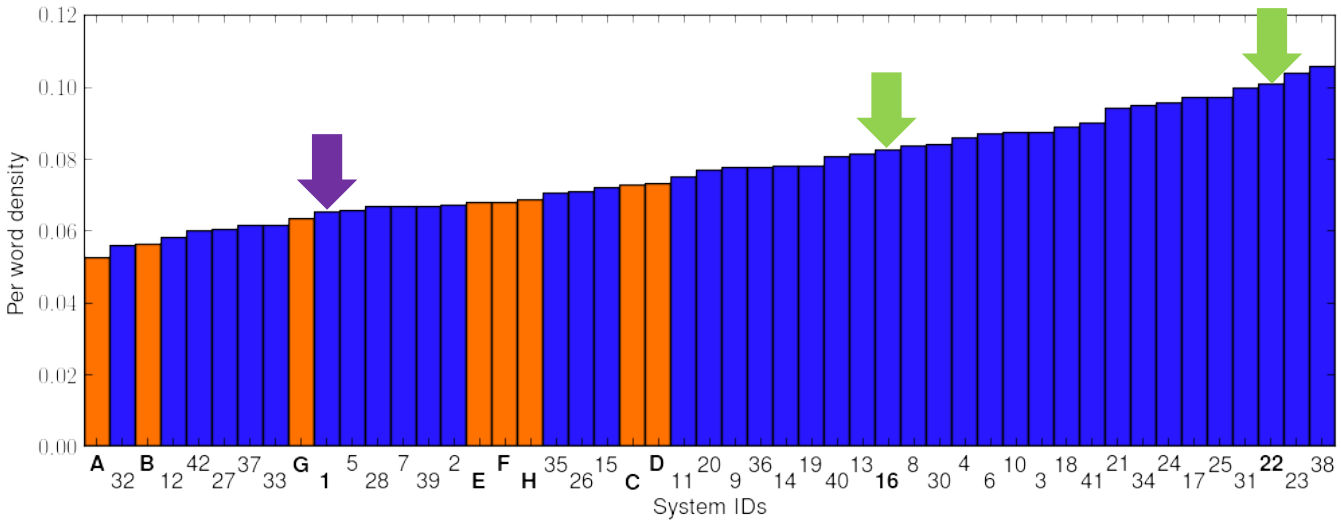
Condition	Initial	Update
Model average	1.58	1.57
Peer average	1.06	1.06
Peer 1	1.00	1.00
Peer 16	1.04	1.04
Peer 22	1.08	1.09

Study 2: Signature Caseframes

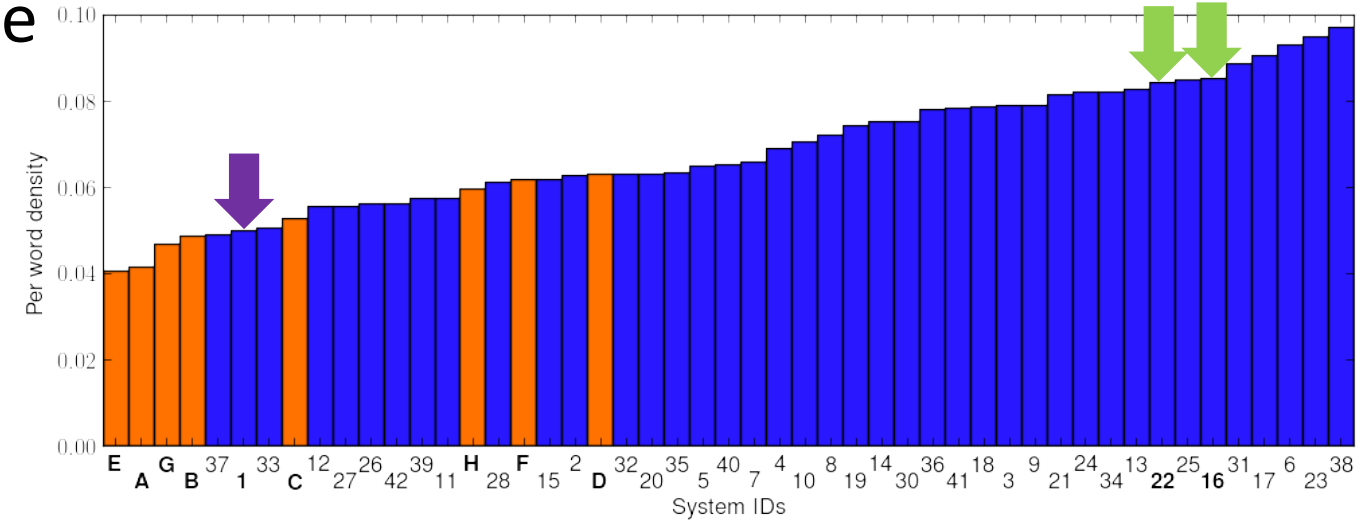
- How do humans aggregate information?
  - Option 1: better compaction, but still based on centrality
  - Option 2: novel sentences that synthesize information
- **Signature caseframes** are computed by method of Lin and Hovy, (2000), extended to caseframes
  - They appear in source text more often than expected by chance, compared to a background corpus
  - Log-likelihood ratio test based on binomial distribution
- Measure **signature caseframe density**
  - # signature caseframes in summaries / # words in summaries

Signature Caseframe Density

- Initial



- Update



Signature Caseframe Density

Condition	Initial	Update
Model average	0.065	0.052
Peer average	0.080*	0.072*
Peer 1	0.066	0.050
Peer 16	0.083*	0.085*
Peer 22	0.101*	0.084*

- Automatic systems are already more “central” than peer systems!

Accounting for Paraphrasing

- Results hold even after merging distributionally similar caseframes by agglomerative clustering

Condition	Initial	Update
Model average	0.062	0.047
Peer average	0.071*	0.063*
Peer 1	0.060	0.044
Peer 16	0.072*	0.077*
Peer 22	0.084*	0.075*

Threshold = 0.8

Consequences

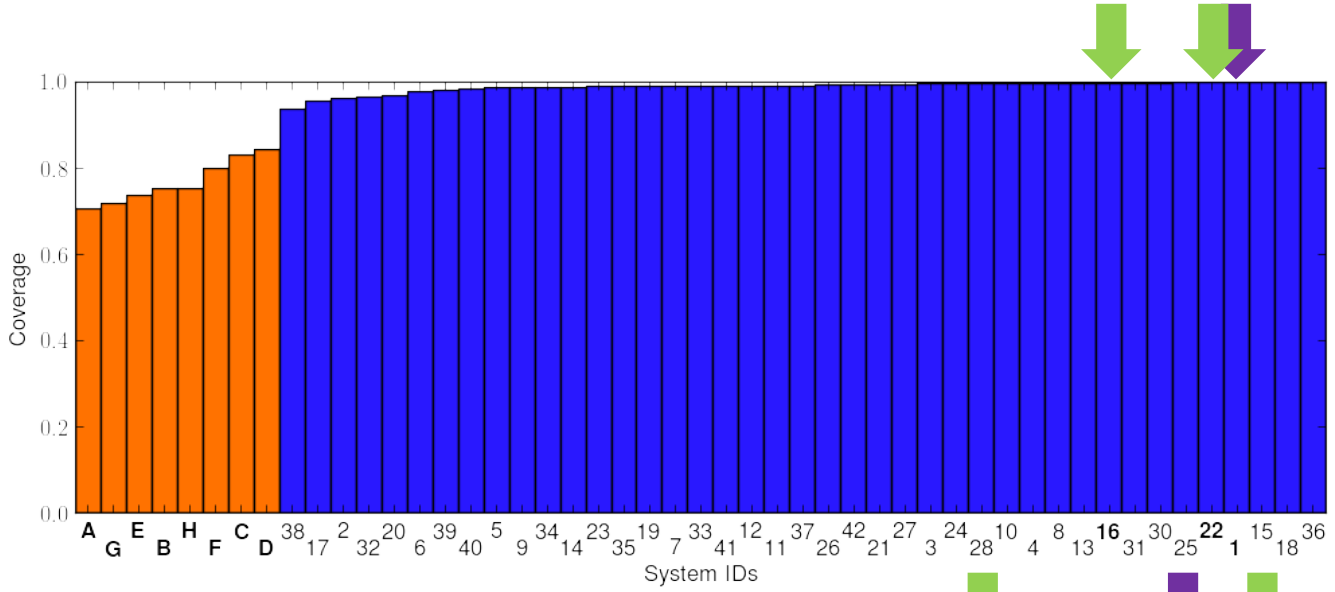
- How do humans aggregate information?
  - Option 1: better compaction, but still based on centrality
  - **Option 2: novel sentences that synthesize information**
- Better optimizing centrality-based measures unlikely to result in paradigm advancement
- Sentence simplification and fusion only part of the answer

Study 3: Summary Reconstruction

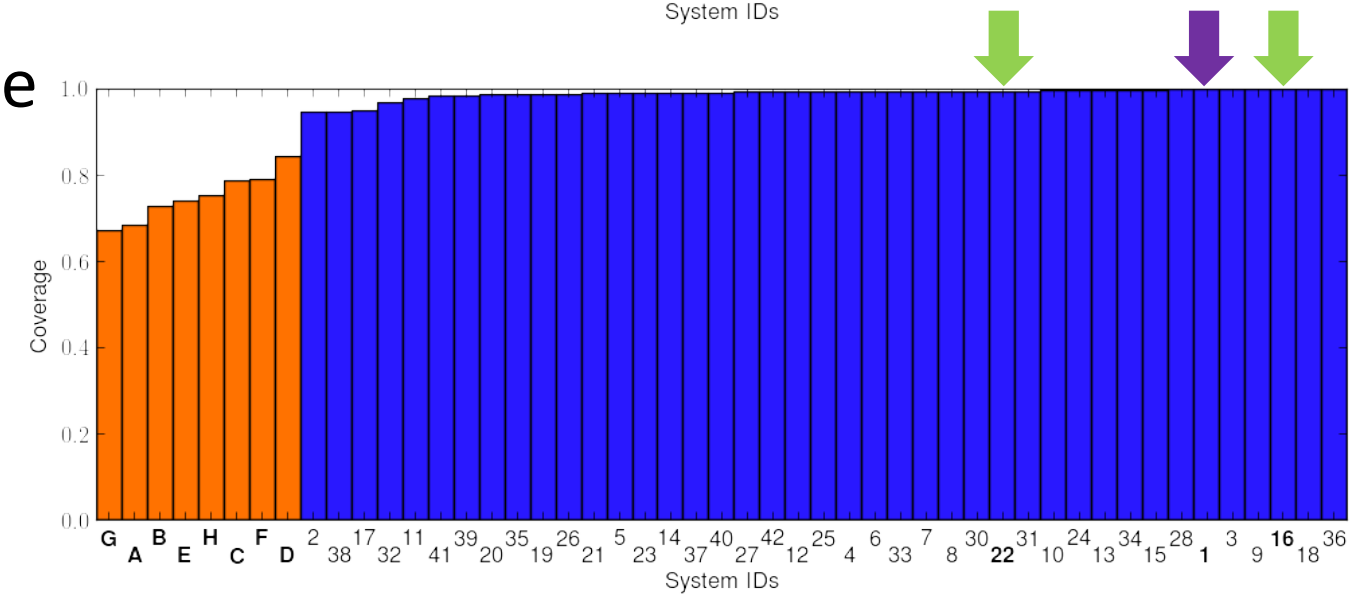
- How might model summaries be generated automatically at all?
  - Want hypothesis space that includes model summaries
- **Caseframe coverage**
  - Proportion of caseframes in a summary that is contained by some reference set
- What is the reference set?
  - Source text alone
  - Source text plus articles from the same domain
- Extends Copeck and Szpakowicz's (2004) analyses

Reconstruction from Source Text

- Initial



- Update



Reconstruction from Source Text

Condition	Initial	Update
Model average	0.77	0.75
Peer average	0.99	0.99
Peer 1	1.00	1.00
Peer 16	1.00	1.00
Peer 22	1.00	1.00

Adding In-domain Articles

- Include all articles from the same domain in reference set
- Baseline: same # of articles from another domain

Reference Set	Initial	Update
Source text	0.77	0.75
+out-of-domain	0.91	0.91
+in-domain	0.98	0.97

Conclusions

- Series of studies on guided summarization data by caseframes
- Can distinguish model vs. state-of-the-art peer summarizers by information content
- Human-written model summaries:
  - contain more aggregation
  - rely less on centrality, even after accounting for paraphrasing
  - cannot be reconstructed from source text alone

Using Domain Knowledge

- Aggregate statistics like Lin and Hovy, (2000) have been successful
  - Identify salient or topical features

- **Future work: more direct use of domain knowledge**
  - Mining in-domain documents for caseframes
  - Learning structured representation of a domain to learn typical slots and events. e.g., (Cheung and Penn, 2013)