Overview of TAC 2009 Summarization Track

Hoa Trang Dang, Karolina Owczarzak

National Institute of Standard and Technology

TAC 2009 Summarization Track

- Update Summarization task
  - multidocument summarization
    - initial summary (10 documents)
    - update summary (10 documents)

- Automatically Evaluating Summaries of Peers (AESOP) task
  - automatic metrics for evaluation of summary quality
  - model summaries available
  - source documents available

Update Summarization Task

- Topic-guided summarization of multiple documents
  - □ initial summary:

    A 100-word summary of a set of 10 documents concerned with a single topic.
  - □ update summary:

    A 100-word summary of a set of further 10 documents for the same topic, with the assumption that the content of the first 10 documents is already known to the reader.

ID: D0919

Topic: Marriage of Camilla Parker Bowles to Prince Charles

Narrative: Report on the marriage of Camilla Parker Bowles to Prince Charles. Include engagement activities, planning for the wedding, and reaction to the engagement. Do not include Camilla's activities prior to her engagement.

Update Summarization Task

- 8 NIST assessors
- 44 topics
- 20 documents selected for each topic
  - AQUAINT-2 collection:
    - 2.5 GB of text (about 907K documents)
    - October 2004 - March 2006
    - Agence France Presse, Central News Agency (Taiwan), Xinhua News Agency, Los Angeles Times-Washington Post News Service, New York Times, the Associated Press
- 20 documents divided in half:
  - Set A (first 10 documents) – source text for initial summary
  - Set B (second 10 documents) – source text for update summary
- 4 model summaries created for each subtopic (A & B)

Update Summarization Task

- Participants:
  - 27 teams
  - 52 runs (up to two per team)
- Baselines:
  - Baseline 1 (ID = 1): all the leading sentences (up to 100 words) in the most recent document.
  - Baseline 2 (ID = 2): a copy of one of the model summaries for the docset, but with the sentences randomly ordered.
  - Baseline 3 (ID = 3): a manual extractive summary provided by the University of Montreal.
- All runs evaluated manually:
  - Overall Responsiveness
  - Overall Readability
  - Pyramid

Update Summarization Task - Evaluation

- Overall Responsiveness (1 – 10)

  How well does the summary respond to the information need contained in the topic statement? How good is its linguistic quality?

- Overall Readability (1 – 10)
  - How fluent and readable is the summary?
    - grammaticality, non-redundancy, referential clarity, focus, structure, coherence

Very Poor Poor Barely Acceptable Good Very Good
1……….2……….3……….4……….5……….6……….7………..8……….9……….10

- System score = mean score of all its summaries
- System ranking
  - ANOVA
  - multiple comparison (Tukey’s honestly significant difference criterion)

Update Summarization Task - Evaluation

- Pyramid (Passonneau et al., 2005)

1. Extract all “information nuggets”, a.k.a. Summary Content Units (SCUs), from model summaries

D0919

SCU: The British Prime Minister was pleased about the engagement

contr1: The British Prime minister...pleased over the engagement

contr2: PM Tony Blair...approved the marriage

contr3: Prime Minister Tony Blair...gave their approval

contr4: British Prime Minister supported the announcement

2. Each SCU’s weight = number of model summaries that contain it

Update Summarization Task - Evaluation

- Pyramid (Passonneau et al., 2005)

3. Check how many SCUs are present in the candidate summary

$$\text{score} = \frac{\text{total weight of all SCUs present in the candidate}}{\text{total SCU weight possible for average-length summary}}$$

Evaluation - Responsiveness

ID	RESPONSIVENESS	
C	**9.3182**	**A**
F	**9.2727**	**A**
G	**9.0455**	**A**
D	**8.8636**	**A**
B	**8.7273**	**A**
H	**8.6818**	**A**
A	**8.4545**	**A**
E	**8.2727**	**A**
2	***6.3636***	***B***
3	***6.3409***	***B***
ICSI_UTD2	5.1591	B C
THUSUM1	4.9545	B C D
UWB.JRC.UT1	4.9545	B C D
RaliLatl1	4.9091	B C D
Siel_091	4.8636	C D
ICSI_UTD1	4.8409	C D
UWB.JRC.UT2	4.7955	C D
Siel_092	4.7273	C D
CLASSY1	4.6818	C D
ICTCAS2	4.5682	C D

Initial summaries

models

reordered model

HexTac

first 100 wrds

ID	RESPONSIVENESS	
C	**9.1364**	**A**
H	**8.6818**	**A**
G	**8.6818**	**A**
F	**8.5909**	**A**
A	**8.3636**	**A**
B	**8.3182**	**A**
E	**8.2273**	**A**
D	**8.0455**	**A B**
2	***6.1818***	***B C***
3	***6.1136***	***C D***
THUSUM1	5.0227	C D E
ICSI_UTD1	4.75	C D E F
uOttawa1	4.6591	C D E F G
Siel_091	4.6136	D E F G H
Siel_092	4.5682	D E F G H
ICSI_UTD2	4.5682	D E F G H
RaliLatl1	4.3409	E F G H I
ICTCAS1	4.3409	E F G H I
1	***4.3182***	***E F G H I***
UWB.JRC.UT1	4.3182	E F G H I

Update summaries

Evaluation - Readability

ID	READABILITY	
F	**9.2727**	**A**
G	**9.1364**	**A**
C	**9.1364**	**A**
B	**9.1364**	**A**
H	**8.8636**	**A**
D	**8.6818**	**A**
A	**8.6364**	**A**
E	**8.4545**	**A B**
3	***7.4773***	***A B C***
1	***6.7045***	***B C D***
UWB.JRC.UT1	5.9318	C D E
UWB.JRC.UT2	5.7727	D E F
THUSUM1	5.6818	D E F G
ICSI_UTD2	5.6364	D E F G H
RaliLatl1	5.6364	D E F G H
VensesTeam1	5.5909	D E F G H
TRI1	5.5455	D E F G H
2	***5.4773***	***D E F G H***
RaliLatl2	5.4091	D E F G H I
uOttawa1	5.3864	D E F G H I

Initial summaries

models

HexTac

first 100 wrds

reordered model

ID	READABILITY	
C	**9.3636**	**A**
H	**9.0909**	**A B**
F	**8.8182**	**A B**
E	**8.8182**	**A B**
G	**8.7273**	**A B**
A	**8.7273**	**A B**
B	**8.5455**	**A B**
D	**8.3636**	**A B C**
3	***7.25***	***B C D***
1	***6.4545***	***C D E***
THUSUM1	5.8864	D E F
2	***5.8864***	***D E F***
TRI1	5.8636	D E F
uOttawa1	5.7955	D E F G
RaliLatl2	5.6364	E F G
ICSI_UTD1	5.5227	E F G H
ISCI_UTD2	5.5	E F G H I
RaliLatl1	5.4773	E F G H I
AIATe1	5.4545	E F G H I J
abawakid2	5.4091	E F G H I J

Update summaries

Evaluation - Pyramid

ID	PYRAMID	
F	**0.77382**	**A**
C	**0.71991**	**A**
G	**0.70677**	**A**
A	**0.68486**	**A**
D	**0.65677**	**A**
E	**0.65595**	**A**
H	**0.65005**	**A**
2	***0.63518***	***A***
B	**0.6165**	**A**
ICSI_UTD2	0.37666	B
ICSI_UTD1	0.36777	B C
3	***0.35232***	***B C D***
WHU2	0.3333	B C D E
ICTCAS2	0.32645	B C D E
ICL_SUM1	0.32573	B C D E
EMLR2	0.31493	B C D E
ICTCAS1	0.31464	B C D E
WHU1	0.31357	B C D E
THUSUM1	0.31077	B C D E F
TRI1	0.31009	B C D E F

Initial summaries

models

reordered model

HexTac

ID	PYRAMID	
2	***0.67748***	***A***
F	**0.66745**	**A**
B	**0.66345**	**A**
G	**0.65764**	**A**
C	**0.64018**	**A B**
H	**0.61573**	**A B**
D	**0.56623**	**A B**
E	**0.55995**	**A B**
A	**0.48086**	**B**
3	***0.32391***	***C***
Siel_091	0.30309	C D
ICSI_UTD1	0.29889	C D E
Siel_092	0.29461	C D E F
THUSUM1	0.29207	C D E F G
ICTCAS2	0.28668	C D E F G H
ICSI_UTD2	0.286	C D E F G H I
ICTCAS1	0.28539	C D E F G H I
UWB.JRC.UT1	0.26259	C D E F G H I J
ICL_SUM1	0.25384	C D E F G H I J K
LIPN1	0.25336	C D E F G H I J K

Update summaries

Evaluation: Average scores

Responsiveness

	initial	update	
models	8.830	8.506	
automatic	4.149	3.866	*

Readability

	initial	update
models	8.915	8.807
automatic	4.859	4.838

Number of SCUs

	initial	update	
models	10.966	7.796	*
automatic	4.452	3.034	*

Pyramid

	initial	update	
models	0.683	0.606	*
automatic	0.260	0.209	*

AESOP Task

- Purpose: To emulate Pyramid and/or Responsiveness

- Test data:
  - 55 candidate summarizers
  - 8 human summarizers
  - 44 topics (A & B): summaries, source documents, topic statements, model summaries

- Participants:
  - 12 teams
  - 35 metrics (up to 4 per team)

- Baselines:
  - ROUGE-SU4: matching bigrams with skip distance up to 4 words, stemmed (Lin, 2004)
  - BE-HM: head-modifier pairs, stemmed (Hovy et al., 2005)

AESOP Task

- Use of resources
  - model summaries: 30 metrics
  - source documents: 10 metrics
  - topic statements: 3 metrics

- Conditions:
  - AllPeers: models + automatic summaries
    - Can automatic metrics distinguish between human and automatic summaries?
  - NoModels: only automatic summaries, model summaries as reference
    - Can automatic metrics accurately evaluate the quality of automatic summaries?

AESOP Task - Evaluation

- Correlations (Pearson, Spearman, Kendall) with:
  - Overall Responsiveness
  - Pyramid

- Discriminative power

AESOP Metric

C4	5.44	A
C17	5.2	A
C35	4.75	A B
C12	4.06	B C
C6	3.14	C
C3	2.37	C

Responsiveness

C4	9.60	A
C32	9.56	A
C6	8.62	A
C1	7.89	B C
C3	7.12	B C
C17	6.55	B C

AESOP Task - Evaluation

- Correlations with:
  - □ Overall Responsiveness
  - □ Pyramid

- Discriminative power

AESOP Metric

C4	5.44	A		
C17	5.2	A		
C35	4.75	A	B	
C12	4.06		B	C
C6	3.14			C
C3	2.37			C

C4 > C3

AGREEMENT

Responsiveness

C4	9.60	A		
C32	9.56	A		
C6	8.62	A		
C1	7.89		B	C
C3	7.12		B	C
C17	6.55		B	C

AESOP Task - Evaluation

- Correlations with:
  - Overall Responsiveness
  - Pyramid

- Discriminative power

AESOP Metric

C4	5.44	A
C17	5.2	A
C35	4.75	A B
C12	4.06	B C
C6	3.14	C
C3	2.37	C

C4 = C17

C4 > C17

DISAGREEMENT

Responsiveness

C4	9.60	A
C32	9.56	A
C6	8.62	A
C1	7.89	B C
C3	7.12	B C
C17	6.55	B C

AESOP Task - Evaluation

- Correlations with:
  - Overall Responsiveness
  - Pyramid

- Discriminative power

AESOP Metric

C4	5.44	A
C17	5.2	A
C35	4.75	A B
C12	4.06	B C
C6	3.14	C
C3	2.37	C

C17 > C6

C6 > C17

CONTRADICTION

Responsiveness

C4	9.60	A
C32	9.56	A
C6	8.62	A
C1	7.89	B C
C3	7.12	B C
C17	6.55	B C

Evaluation – Correlations with Pyramid

Pearson's *r*

ID	B2 incl	ID	B2 exl
CLASSY4	0.978	CLASSY2	0.972
PolyuU4	0.967	CLASSY4	0.967
UWB.JRC.UT2	0.967	ISI2	0.967
PolyU2	0.965	ISI1	0.967
IITKharagpur2	0.963	PRaSa4	0.960
PolyU3	0.962	CLASSY1	0.959
DemokritosGR1	0.954	DemokritosGR1	0.958
UWB.JRC.UT1	0.952	UWB.JRC.UT1	0.958
univille1	0.952	TRI1	0.954
UWB.JRC.UT4	0.951	UWB.JRC.UT2	0.952
ROUGE-SU4	*0.921*	*ROUGE-SU4*	*0.950*
BE-HM	*0.857*	*BE-HM*	*0.949*

Initial summaries

ID	B2 incl	ID	B2 exl
DemokritosGR1	0.970	ISI2	0.969
CLASSY4	0.970	ISI1	0.969
PolyU3	0.968	PRaSa1	0.961
PolyU4	0.962	*BE-HM*	*0.956*
UWB.JRC.UT1	0.962	TRI1	0.951
IITKharagpur2	0.957	CLASSY1	0.947
TRI1	0.946	DemokritosGR1	0.944
UWB.JRC.UT2	0.946	CLASSY2	0.941
PolyU2	0.944	IIITKharagpur2	0.940
univille1	0.944	CLASSY4	0.938
ROUGE-SU4	*0.940*	PolyU3	0.937
BE-HM	*0.924*	*ROUGE-SU4*	*0.904*

Update summaries

Evaluation – Correlations with Responsiveness

Pearson’s *r*

ID	B2 incl	ID	B2 exl
CLASSY4	0.872	CLASSY1	0.888
PolyU4	0.856	ISI2	0.880
UWB.JRC.UT4	0.854	ISI1	0.876
PolyU2	0.853	TRI1	0.875
UWB.JRC.UT2	0.851	CLASSY3	0.875
IITKharagpur2	0.851	DemokritosGR1	0.872
PolyU3	0.846	CLASSY2	0.872
univille1	0.839	CLASSY4	0.871
DemokritosGR1	0.829	PRaSa1	0.863
IITKharagpur3	0.827	DemokritosGR2	0.863
ROUGE-SU4	*0.767*	*BE-HM*	*0.849*
BE-HM	*0.692*	*ROUGE-SU4*	*0.839*

Initial summaries

ID	B2 incl	ID	B2 exl
IITKharagpur2	0.833	ISI2	0.871
PolyU4	0.825	ISI1	0.859
PolyU2	0.821	PRaSa1	0.855
CLASSY4	0.814	CLASSY1	0.847
PolyU3	0.814	DemokritosGR2	0.847
UWB.JRC.UT2	0.801	*BE-HM*	*0.846*
UWB.JRC.UT4	0.798	CLASSY3	0.845
DemokritosGR1	0.796	TRI1	0.842
univille1	0.792	DemokritosGR1	0.828
UWB.JRC.UT1	0.768	TRI3	0.827
ROUGE-SU4	*0.729*	CLASSY4	0.818
BE-HM	*0.694*	*ROUGE-SU4*	*0.784*

Update summaries

Evaluation – Discriminative power

AESOP metrics vs Pyramid on initial summaries

Differences between models and automatic summaries

METRIC	DIFFERENCE (max 432)	NO DIFFERENCE (max 8)	CONTRADICTIONS
DemokritosGR1	432	8	0
DemokritosGR2	432	8	0
PRaSa2	432	8	0
TRI1	432	8	0
univille1	432	8	0
MIRACL1	432	7	0
MIRACL2	432	7	0
ROUGE-SU4	*227*	*0*	*0*
BE-HM	*97*	*0*	*0*

Evaluation – Discriminative power

AESOP metrics vs Responsiveness on initial summaries

Differences between models and automatic summaries

METRIC	DIFFERENCE (max 440)	NO DIFFERENCE (max 0)	CONTRADICTIONS
DemokritosGR1	433	0	0
DemokritosGR2	432	0	0
MIRACL2	432	0	0
TRI1	432	0	0
univille1	432	0	0
PRaSa2	432	0	0
MIRACL1	432	0	1
ROUGE-SU4	*227*	*0*	*8*
BE-HM	*97*	*0*	*8*

Evaluation – Discriminative power

AESOP metrics vs Pyramid on initial summaries

Differences between automatic summaries

METRIC	DIFFERENCE (max 371)	NO DIFFERENCE (max 1114)	CONTRADICTIONS
UWB.JRC.UT1	364	1016	0
UWB.JRC.UT2	362	1026	0
DemokritosGR1	361	1025	0
CLASSY4	359	1055	0
CLASSY2	356	1058	0
PolyU2	353	1019	0
PolyU3	352	1015	0
ROUGE-SU4	*351*	*1042*	*0*
BE-HM	*271*	*1097*	*0*

Conclusions

- Update Summarization task
  - Quality gap between human models and automatic summaries (Responsiveness, Readability, Pyramid)
  - Quality gap between initial and update summaries for automatic systems (Responsiveness, Pyramid)

- Automatically Evaluating Summaries of Peers (AESOP) task
  - Several submissions achieve high correlations with manual metrics
  - High agreement with manual discriminative power
  - Outperforming both baselines

TAC 2010

You’re invited!