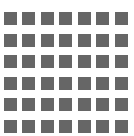
Towards Summarization for Social Media

Results of the TL;DR Challenge

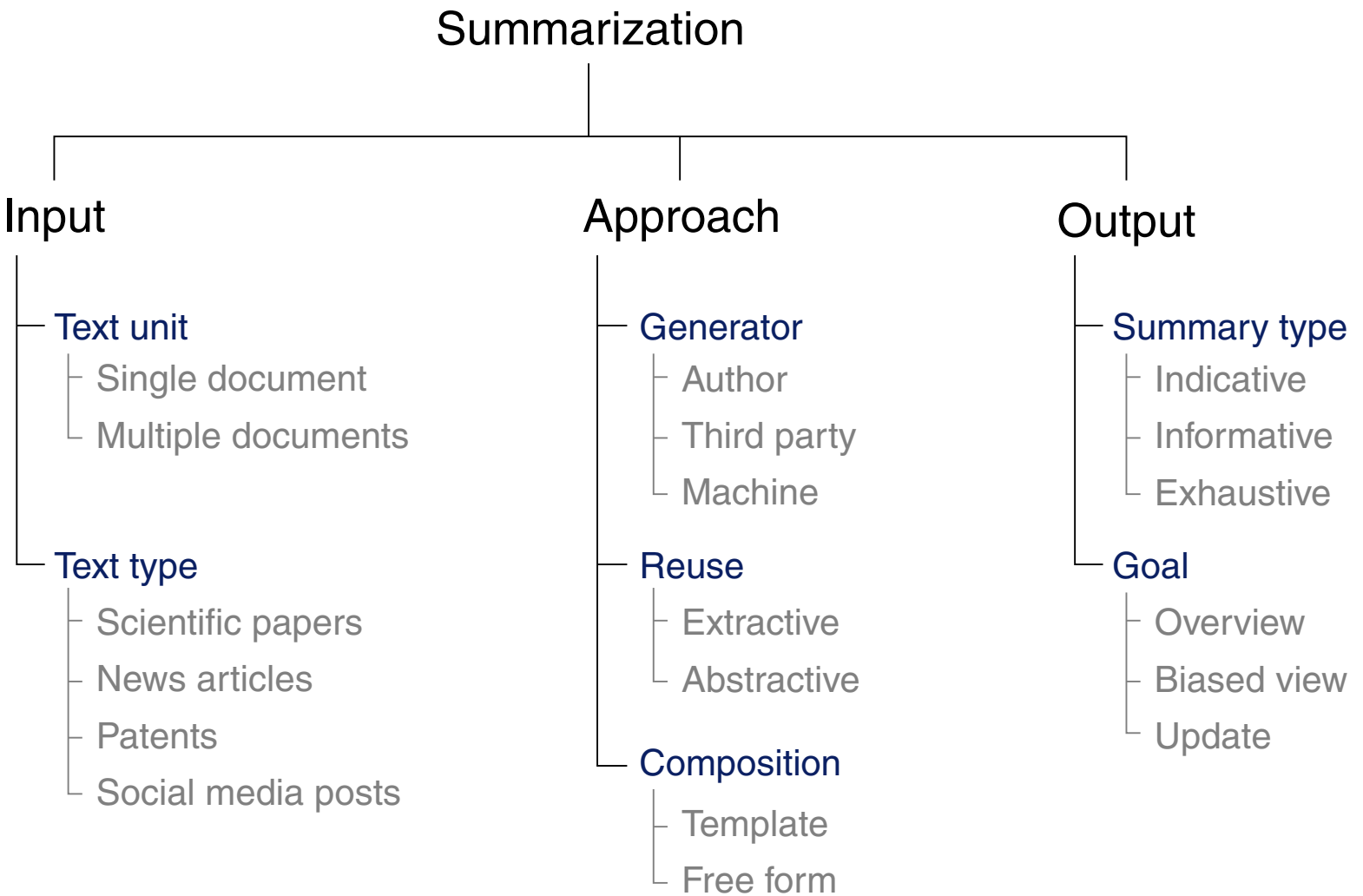
Shahbaz Syed Michael Völske Nedim Lipka

Benno Stein Hinrich Schütze Martin Potthast



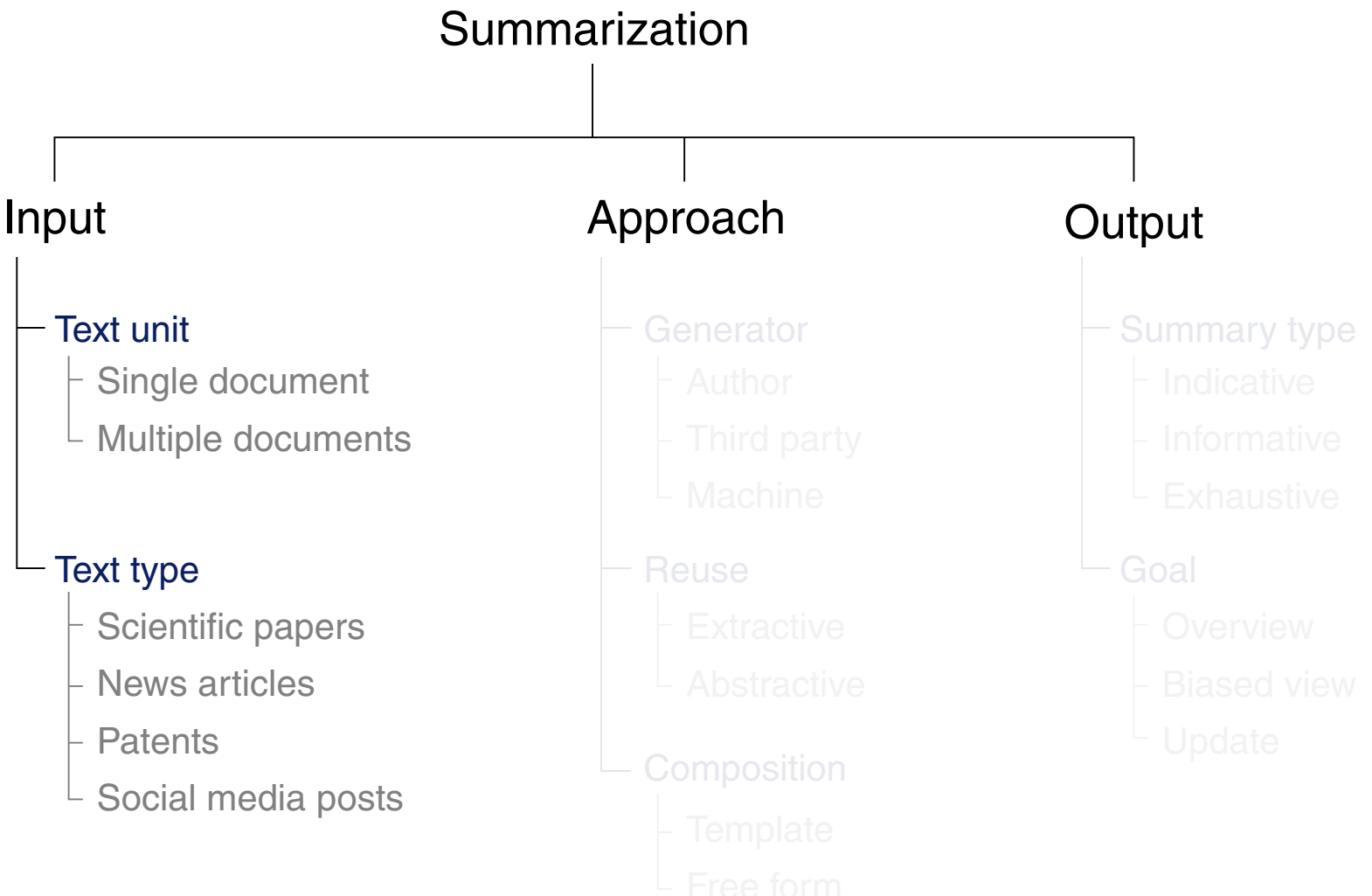
Introduction

Facets of Summarization



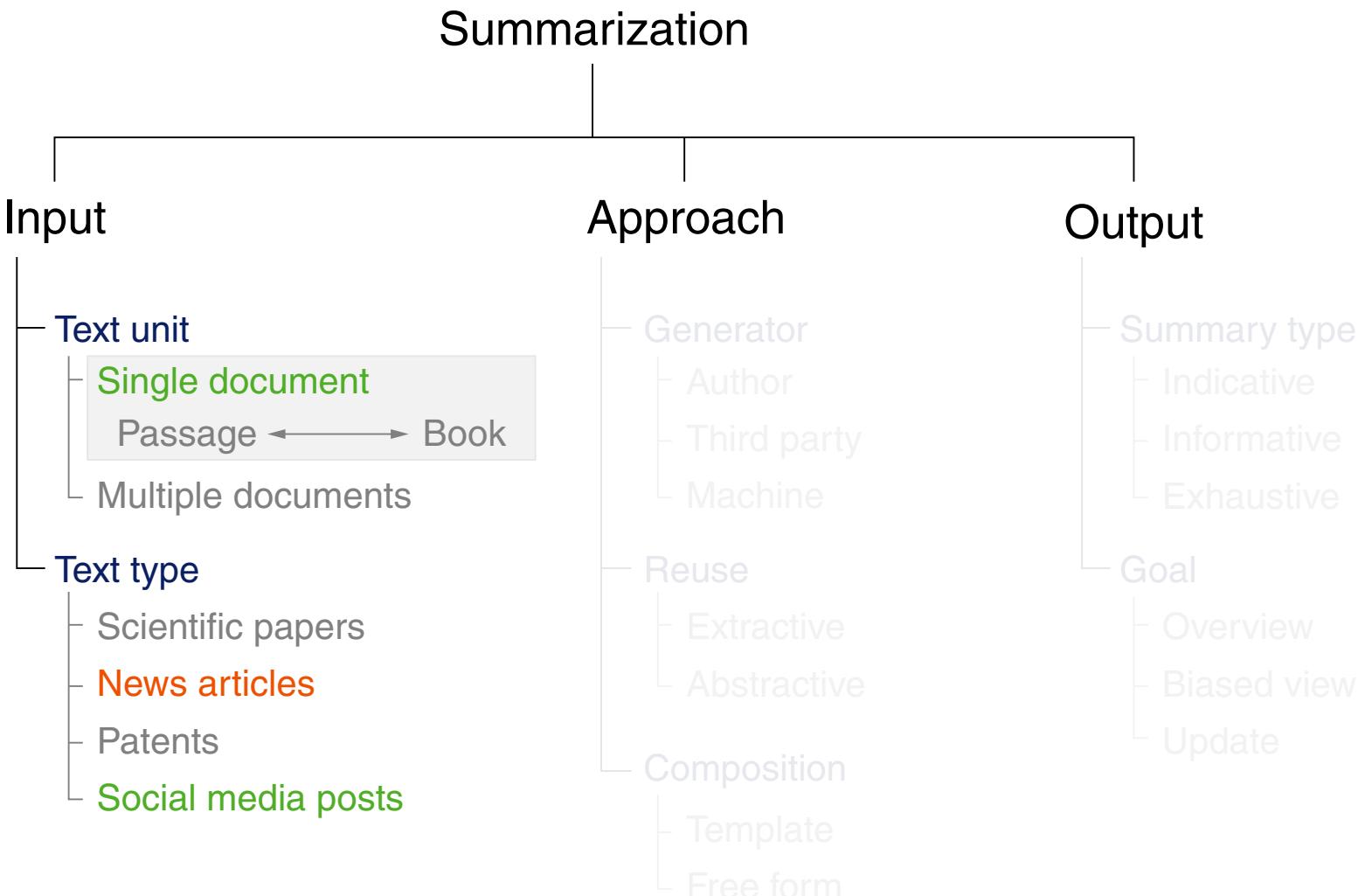
Introduction

Facets of Summarization



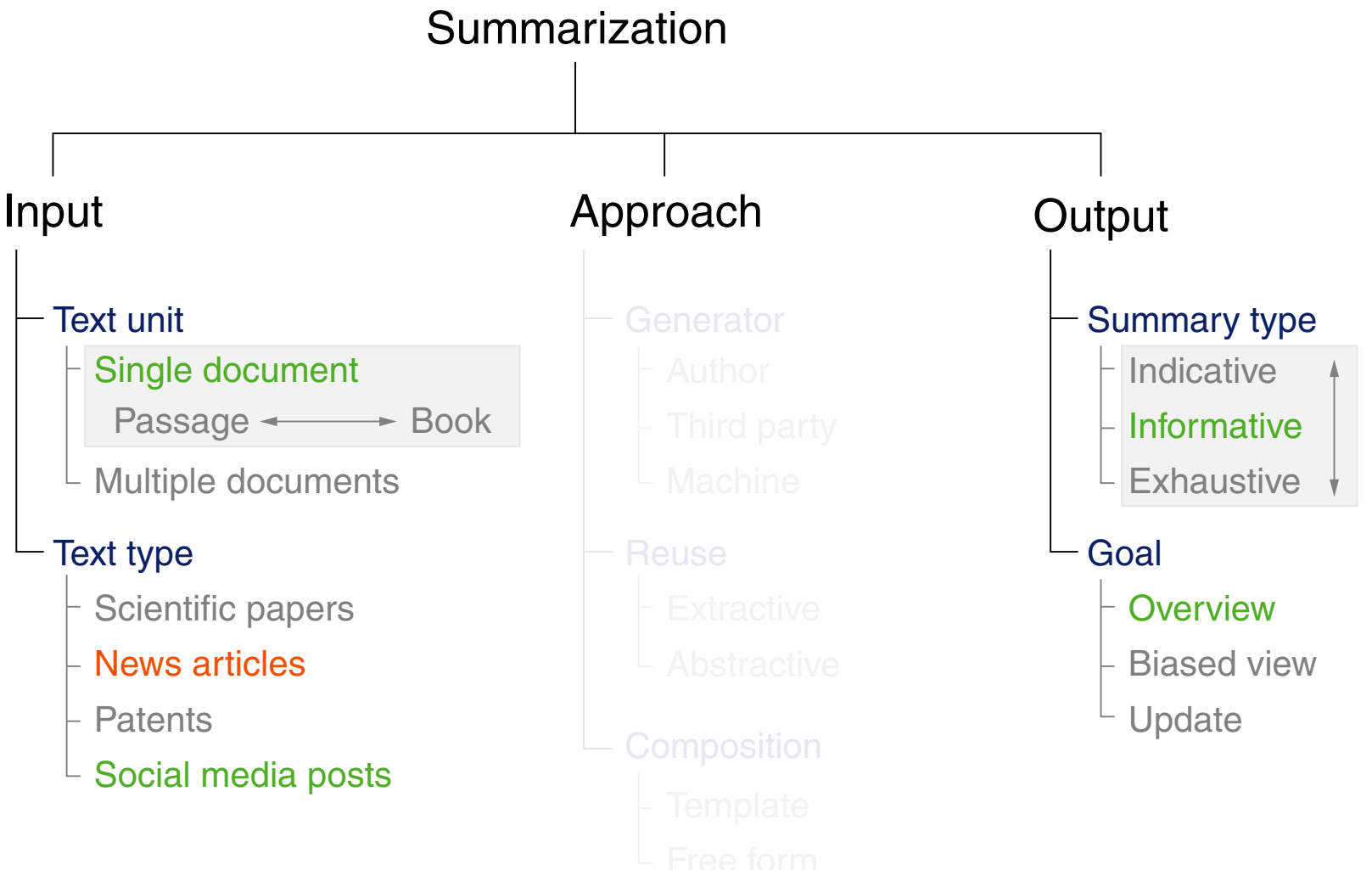
Introduction

Facets of Summarization



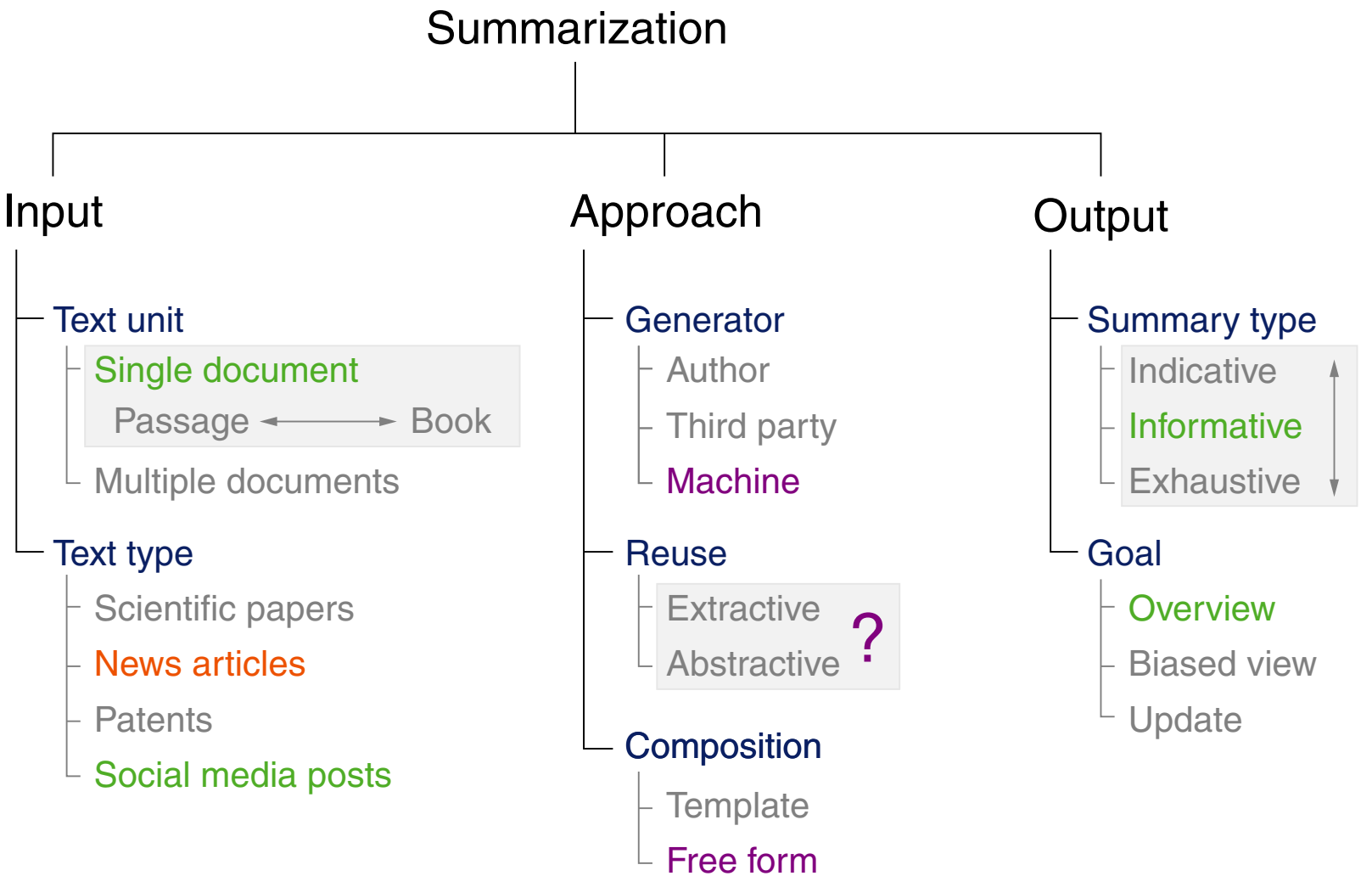
Introduction

Facets of Summarization



Introduction

Facets of Summarization



Introduction

Extractive vs. Abstractive Summarization

Extractive summaries reuse text from the original; abstractive summaries don’t.

Who cares?

Introduction

Extractive vs. Abstractive Summarization

Extractive summaries reuse text from the original; abstractive summaries don't.

Who cares?

Writer

- Writers reuse and / or abstract as needed to compose a "good" summary.

Introduction

Extractive vs. Abstractive Summarization

Extractive summaries reuse text from the original; abstractive summaries don't.

Who cares?

Writer

Reader

- Writers reuse and / or abstract as needed to compose a "good" summary.
- Readers just want "good" summaries.

Introduction

Extractive vs. Abstractive Summarization

Extractive summaries reuse text from the original; abstractive summaries don't.

Who cares?

Writer

Reader

Computer scientist

- Writers reuse and / or abstract as needed to compose a "good" summary.
- Readers just want "good" summaries.
- Computer scientists care about natural language understanding.

Introduction

Extractive vs. Abstractive Summarization

Extractive summaries reuse text from the original; abstractive summaries don't.

Background:

- For decades, summarization relied primarily on extractive techniques.
- Deep learning is the first true contender in abstractive summarization.

Introduction

Extractive vs. Abstractive Summarization

Extractive summaries reuse text from the original; abstractive summaries don't.

Background:

- For decades, summarization relied primarily on extractive techniques.
- Deep learning is the first true contender in abstractive summarization.

Obstacles to abstractive summarization:

- Widespread summary ground truth is extractive in nature.
- Training loss / evaluations rely on text overlap (ROUGE) with ground truth.
- Qualitative evaluations are non-standardized and often small-scale.

Introduction

Extractive vs. Abstractive Summarization

Extractive summaries reuse text from the original; abstractive summaries don't.

Background:

- For decades, summarization relied primarily on extractive techniques.
- Deep learning is the first true contender in abstractive summarization.

Obstacles to abstractive summarization:

- Widespread summary ground truth is extractive in nature.
- Training loss / evaluations rely on text overlap (ROUGE) with ground truth.
- Qualitative evaluations are non-standardized and often small-scale.

Our approach:

➜ TL;DRs as abstractive ground truth; crowdsourcing for qualitative evaluation.

Shared Task

Ground Truth: TL;DRs on Reddit

What are the most ridiculous things you believed to be true as a child?

My dad spent the better part of his childhood believing his early memories of his family having a pet monkey were just something his kid brain made up. When he asked his parents it turned out they actually did have one and that they gave it to a local zoo before his sister was born.

TL;DR
My dad believed his family didn't have a pet monkey

Shared Task

Ground Truth: TL;DRs on Reddit

- Author-supplied, abstractive summaries.
- Everyday topics, informal writing, slang, abbreviations.
- No tight control of summary types, noisy summaries.

- Mining Reddit yields 3 million content-summary pairs. ➜ Webis-TLDR-17

Shared Task

Submissions

- 16 registrations, 3 successful submissions, 5 models.
  Many non-academic registrants, most dropouts due to lack of time or resources.

  - Gehrmann et al.: transf-seq2seq and pseudo-self-attn
    "Generating Abstractive Summaries with Finetuned Language Models"
  - Choi et al.: unified-vae-pgn and unified-pgn
    "VAE-PGN based Abstractive Model in Multi-stage Architecture for Text Summarization"
  - Kalinowski: tldr-bottom-up
    (No system description submitted)

- Model training at home; model testing via TIRA (tira.io).
  Hidden test set for blind evaluation; archival of virtual machines for reproducibility

- Synopsis of generated summaries: tldr.webis.de > Summaries

Evaluation

Quantitative Analysis

Model	ROUGE			Novelty (n-grams)				Len.
	1	2	L	1	2	3	4	
pseudo-self-attn	18	4	13	1.49	7.21	9.54	9.98	12.1
transf-seq2seq	19	**5**	14	0.82	4.28	6.44	7.54	14.5
unified-pgn	19	4	**15**	0.80	5.15	8.67	11.42	33.5
unified-vae-pgn	19	4	**15**	0.86	5.04	8.90	11.92	32.8
tldr-bottom-up	**20**	4	**15**	1.90	5.29	8.32	10.73	37.3
ground truth	–	–	–	**9.48**	**21.94**	**24.86**	**25.20**	26.1

- Two length classes, under- and overshooting the average TL;DR length.
- ROUGE cannot properly measure model differences.
- Little novelty in generated summaries [See et al., 2017]. ➜ Extractive models

Evaluation

Quantitative Analysis

Model	ROUGE			Novelty (n-grams)				Len.
	1	2	L	1	2	3	4	
pseudo-self-attn	18	4	13	1.49	7.21	9.54	9.98	12.1
transf-seq2seq	19	**5**	14	0.82	4.28	6.44	7.54	14.5
unified-pgn	19	4	**15**	0.80	5.15	8.67	11.42	33.5
unified-vae-pgn	19	4	**15**	0.86	5.04	8.90	11.92	32.8
tldr-bottom-up	**20**	4	**15**	1.90	5.29	8.32	10.73	37.3
ground truth	–	–	–	**9.48**	**21.94**	**24.86**	**25.20**	26.1

- Two length classes, under- and overshooting the average TL;DR length.
- ROUGE cannot properly measure model differences.
- Little novelty in generated summaries [See et al., 2017]. ➜ Extractive models

Evaluation

Quantitative Analysis

Model	ROUGE			Novelty (n-grams)				Len.
	1	2	L	1	2	3	4	
pseudo-self-attn	18	4	13	1.49	7.21	9.54	9.98	12.1
transf-seq2seq	19	**5**	14	0.82	4.28	6.44	7.54	14.5
unified-pgn	19	4	**15**	0.80	5.15	8.67	11.42	33.5
unified-vae-pgn	19	4	**15**	0.86	5.04	8.90	11.92	32.8
tldr-bottom-up	**20**	4	**15**	1.90	5.29	8.32	10.73	37.3
ground truth	–	–	–	**9.48**	**21.94**	**24.86**	**25.20**	26.1

- Two length classes, under- and overshooting the average TL;DR length.
- ROUGE cannot properly measure model differences.
- Little novelty in generated summaries [See et al., 2017]. ➜ Extractive models

Evaluation

Quantitative Analysis

Model	ROUGE			Novelty (n-grams)				Len.
	1	2	L	1	2	3	4	
pseudo-self-attn	18	4	13	1.49	7.21	9.54	9.98	12.1
transf-seq2seq	19	**5**	14	0.82	4.28	6.44	7.54	14.5
unified-pgn	19	4	**15**	0.80	5.15	8.67	11.42	33.5
unified-vae-pgn	19	4	**15**	0.86	5.04	8.90	11.92	32.8
tldr-bottom-up	**20**	4	**15**	1.90	5.29	8.32	10.73	37.3
ground truth	–	–	–	**9.48**	**21.94**	**24.86**	**25.20**	26.1

- Two length classes, under- and overshooting the average TL;DR length.
- ROUGE cannot properly measure model differences.
- Little novelty in generated summaries [See et al., 2017]. ➜ Extractive models

Evaluation

Quantitative Analysis

Model	ROUGE			Novelty (n-grams)				Len.
	1	2	L	1	2	3	4	
pseudo-self-attn	18	4	13	1.49	7.21	9.54	9.98	12.1
transf-seq2seq	19	**5**	14	0.82	4.28	6.44	7.54	14.5
unified-pgn	19	4	**15**	0.80	5.15	8.67	11.42	33.5
unified-vae-pgn	19	4	**15**	0.86	5.04	8.90	11.92	32.8
tldr-bottom-up	**20**	4	**15**	1.90	5.29	8.32	10.73	37.3
ground truth	–	–	–	**9.48**	**21.94**	**24.86**	**25.20**	26.1

- Two length classes, under- and overshooting the average TL;DR length.
- ROUGE cannot properly measure model differences.
- Little novelty in generated summaries [See et al., 2017]. ➜ Extractive models
- Neither does our ground truth include *all possible* “good” summaries per post, nor always the best one. ➜ Qualitative evaluation

Evaluation

Qualitative Analysis: Preference scoring

- Given: Text + all summaries (random order)
  - Scoring of each summary on a 4-point Likert scale.
  - Written justification of each judgment.
  - Relative preference entails a mixture of sufficiency and text quality.

Evaluation

Qualitative Analysis: Preference scoring

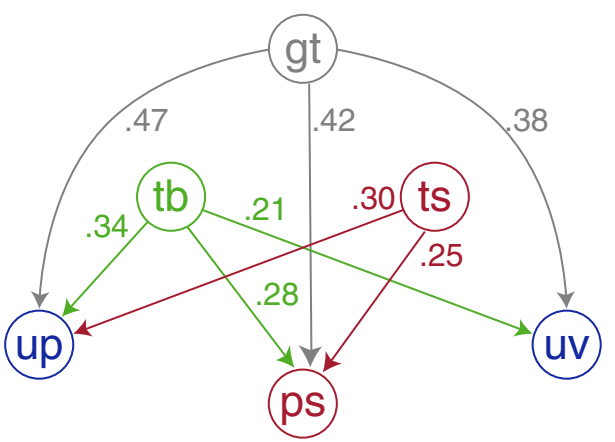
- Given: Text + all summaries (random order)
  - Scoring of each summary on a 4-point Likert scale.
  - Written justification of each judgment.
  - Relative preference entails a mixture of sufficiency and text quality.

Ranking:

1. ground truth (gt)
   tldr-bottom-up (tb)
2. transf-seq2seq (ts)
3. unified-vae-pgn (uv)
4. pseudo-self-attn (ps)
   unified-pgn (up)

Statistical analysis:

Arrows indicate significantly higher scores ($p < 0.001$) as per Mann-Whitney U with Bonferroni correction, labels indicate effect size.



Evaluation

Qualitative Analysis: Quality scoring

- Given: Text + one summary
  - Sufficiency: incomplete/unrelated – missing the main point – OK
  - Text quality: badly written – needs improvement – well-written
  - Tests workers' ability to understand and differentiate between dimensions.

Evaluation

Qualitative Analysis: Quality scoring

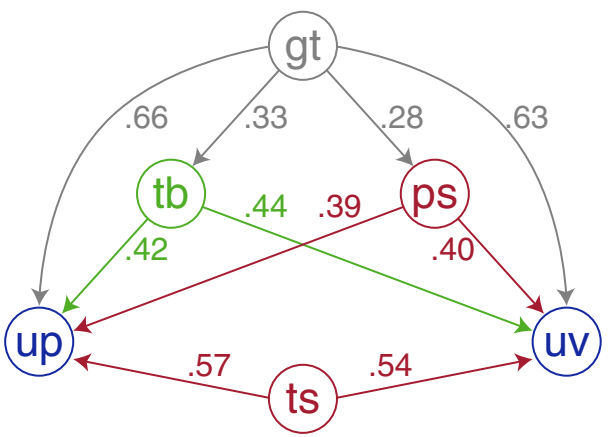
- Given: Text + one summary
  - Sufficiency: incomplete/unrelated – missing the main point – OK
  - Text quality: badly written – needs improvement – well-written
  - Tests workers' ability to understand and differentiate between dimensions.

Ranking:

1. ground truth (gt)
2. transf-seq2seq (ts)
   tldr-bottom-up (tb)
   pseudo-self-attn (ps)
3. unified-pgn (up)
   unified-vae-pgn (uv)

Statistical analysis:

Arrows indicate significantly higher scores ($p < 0.001$) as per Mann-Whitney U with Bonferroni correction, labels indicate effect size.



Evaluation

Error Analysis: Aligning justifications with the quality dimensions

Justification examples: (n=2100)

“basic theme is present but not much context”

“This goes directly opposite of what the text says”

“so close but one word being wrong ruins this summary”

“hits some key phrases but is not a summary”

“Only copies a few sentences.”

“decent but grammar and order of events needs fixed”

“it is choppy but it contains the basics”

“Starts to repeat words and more context is necessary”

“The first part of it is good but the end of it is bad grammar and makes no sense”

Evaluation

Error Analysis: Aligning justifications with the quality dimensions

Sufficiency:

- Missing context
  “basic theme is present but not much context”
- Wrong sentiment
  “This goes directly opposite of what the text says”
- Factually incorrect
  “so close but one word being wrong ruins this summary”
- Overly simplistic
  “hits some key phrases but is not a summary”

Text quality:

- Bad grammar
  “decent but grammar and order of events needs fixed”
- Incoherence
  “it is choppy but it contains the basics”
- Repetition
  “Starts to repeat words and more context is necessary”
- Bad continuity
  “The first part of it is good but the end of it is bad grammar and makes no sense”

Evaluation

Error Analysis: Aligning justifications with the quality dimensions

Sufficiency:

- Missing context
  "basic theme is present but not much context"
- Wrong sentiment
  "This goes directly opposite of what the text says"
- Factually incorrect
  "so close but one word being wrong ruins this summary"
- Overly simplistic
  "hits some key phrases but is not a summary"

Ranking		Neg.	Pos.
1	ground truth	122	178
2	tldr-bottom-up	163	137
	transf-seq2seq	172	128
3	unified-vae-pgn	200	100
4	pseudo-self-attn	217	83
5	unified-pgn	244	56

Text quality:

- Bad grammar
  "decent but grammar and order of events needs fixed"
- Incoherence
  "it is choppy but it contains the basics"
- Repetition
  "Starts to repeat words and more context is necessary"
- Bad continuity
  "The first part of it is good but the end of it is bad grammar and makes no sense"

Conclusions and Future Work

Take-away messages

- Generating abstractive summaries is still a work in progress.
- More sources of ground truth for abstractive summarization are needed.
- Quantitative evaluation still can't replace qualitative evaluation in this task.

Conclusions and Future Work

Take-away messages

- Generating abstractive summaries is still a work in progress.
- More sources of ground truth for abstractive summarization are needed.
- Quantitative evaluation still can't replace qualitative evaluation in this task.

TL;DR Challenge @ INLG 2020?

- Key goals: New summarization tasks and resources, better evaluation.
- Reducing noise from the existing TL;DR dataset.
- In progress: abstractive snippet generation and conclusion generation.
- Accessibility: software submission + GPUs for participants in need.

Conclusions and Future Work

Take-away messages

- Generating abstractive summaries is still a work in progress.
- More sources of ground truth for abstractive summarization are needed.
- Quantitative evaluation still can't replace qualitative evaluation in this task.

TL;DR Challenge @ INLG 2020?

- Key goals: New summarization tasks and resources, better evaluation.
- Reducing noise from the existing TL;DR dataset.
- In progress: abstractive snippet generation and conclusion generation.
- Accessibility: software submission + GPUs for participants in need.

Thank you!

Special thanks to our participants and the INLG committee!

tldr.webis.de

Appendix

Summarization Datasets

Overview

Corpus	Genre	Training pairs
Gigaword	News articles	4 million
Cornell Newsroom	News articles	1.3 million
NYT	News articles	655,000
CNN/Daily Mail	News articles	300,000
XSum (BBC)	News articles	226,000
arXiv	Scientific papers	215,000
PubMed	Scientific papers	133,000
TIPSTER	Magazine articles	33,000
DUC 2003	Newswire	624
DUC 2004	Newswire	500
Webis-TLDR-17	**Social Media**	**3 million**

Summarization Datasets

Webis-TLDR-17

Mining Reddit:

1. Filter posts from known bot accounts and repetitive / reused posts.

2. For any remaining post, check if contains the string **TL;DR**.
   TL;DR is written in many different forms: tl dr, tl;dr, tldr, tl:dr, tl/dr, tl; dr, tl,dr, tl, dr, tl-dr, tl'dr, ...

3. Skip posts with multiple occurrences of a TL;DR pattern.

4. Skip posts where the post text is shorter than its summary (e.g., edits).

Summarization Datasets

Example: Webis-TLDR-17

Webis-TLDR-17

Post

I'm so upset at myself. My boyfriend surprised me with an amazing, fancy dinner for our one year anniversary yesterday. I already wasn't feeling well when he told me we were going to dinner but when I saw what he planned I didn't have the heart to tell him I wasn't that hungry. In the end I pushed myself to eat the fixed menu he ordered for us and the bill was over 500, I couldn't handle it and after dessert I ended up going to the bathroom and throwing it all up.

I can't believe I wasted so much of his money and am so disappointed in myself for not speaking up and simply saying I didn't feel well. I feel like I've wasted the effort he put into planning this. I also feel like I missed out on some amazing food that we would usually never splurge for. He doesn't know I threw it up and I just told him I loved it because regardless of how I felt health wise I loved that he put in so much effort to make sure I felt special. But I can't stop stewing in my own feelings. Help.

TL;DR

my boyfriend is amazing and bought us an expensive anniversary dinner. Threw it all up, he doesn't know. Feel horrible guilt and FOMO

Summarization Datasets

Example: CNN/Daily Mail

CNN/DailyMail

Article

NASA will launch Space Shuttle Endeavour on February 7, which will be the first of five launches this year before the shuttle fleet is retired. Endeavour will blast off from the Kennedy Space Center in Florida on a 13-day mission to the international space station. The mission will include three spacewalks, NASA said. The shuttle will also deliver the final U.S. portion of the space station. This portion will provide more room for crew members. NASA plans to retire its space shuttles Discovery, Endeavour and Atlantis later this year. The space agency has been looking for places, such as museums, to house the shuttles after they are retired. Space Shuttle Discovery will be transferred to the Smithsonian National Air and Space Museum in Washington. The privilege of showing off a shuttle won't be cheap – about $29 million, NASA said.

Highlights

- This will be first of five launches this year before the shuttle fleet is retired
- NASA is scheduled to launch Space Shuttle Endeavour on February 7.
- Shuttle will deliver final U.S. portion of the international space station
- NASA has been looking for places to house the shuttles once they are retired