Nutshell

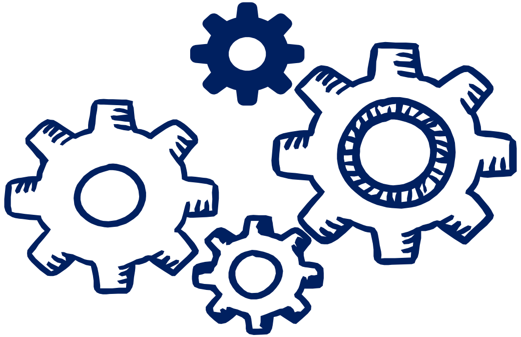
A Topic-Focused Text Summarizer

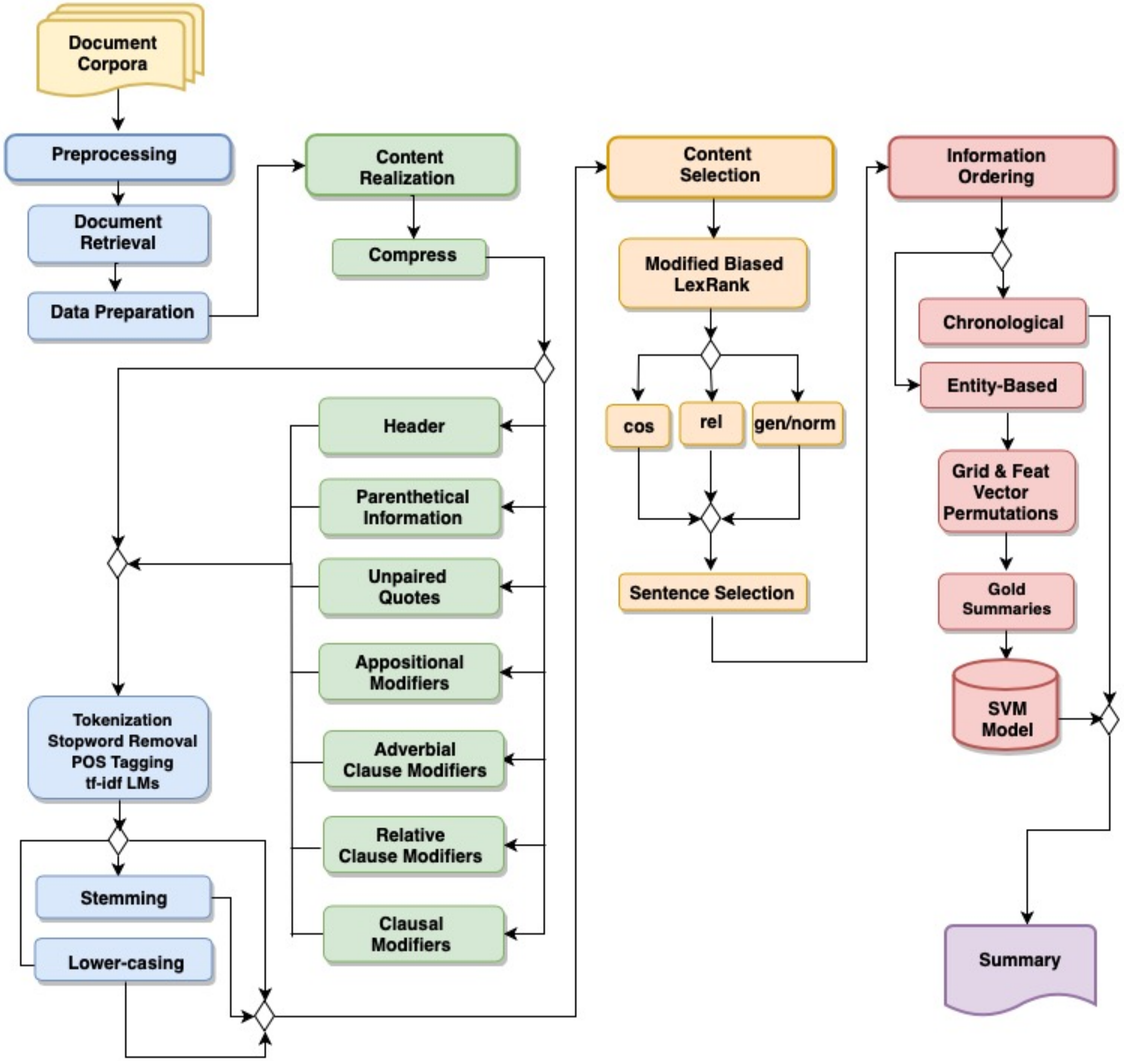
Amina Venton ◦ Shannon Ladymon ◦ Ben Longwill ◦ Hayley Lepp

Overview

- System Architecture
- D4 Configuration
- Improvements
- Analysis
- Results
- Future Work

System Architecture





Document Corpora

Preprocessing

Document Retrieval

Data Preparation

Tokenization
Stopword Removal
POS Tagging
tf-idf LMs

Stemming

Lower-casing

Content Realization

Compress

Header

Parenthetical Information

Unpaired Quotes

Appositional Modifiers

Adverbial Clause Modifiers

Relative Clause Modifiers

Clausal Modifiers

Content Selection

Modified Biased LexRank

cos

rel

gen/norm

Sentence Selection

Information Ordering

Chronological

Entity-Based

Grid & Feat Vector Permutations

Gold Summaries

SVM Model

Summary

D4 Configuration

Content Selection

Biased LexRank (Otterbacher et al., 2005) configuration

- PageRank style approach to sentence salience using a graph of inter-sentential similarity with a bias for sentence-topic similarity.

$$\mathrm{BLR}(s|q) = d\frac{\mathrm{rel}(s|q)}{\sum_{z\in C}\mathrm{rel}(z|q)} + (1-d)\sum_{v\in C}\frac{\mathrm{cos\_sim}(s,v)}{\sum_{z\in C}\mathrm{cos\_sim}(z,v)}$$

$$\mathrm{rel}(s|q) = \sum_{w\in q} log(\mathrm{tf}_{w,s}+1)\times log(\mathrm{tf}_{w,q}+1)\times \mathrm{idf}_w$$

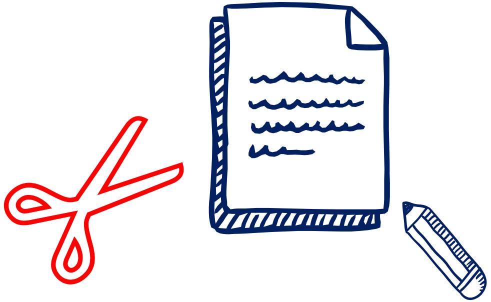
$$\mathrm{cos\_sim}(x,y) = \frac{\sum_{w\in x,y}(\text{tf-idf}_{w,x})\times(\text{tf-idf}_{w,y})}{\sqrt{\sum_{x_i\in x}(\text{tf-idf}_{x_i,x})^2}\times\sqrt{\sum_{y_i\in y}(\text{tf-idf}_{y_i,y})^2}}$$

Information Ordering

Entity-Based Ordering (Barzilay and Lapata, 2008)

Simple Approach:

- Entity Grids
  - Coreference Exact Match
  - Entities either present (X) or absent (–)
  - Transitions of length 2
- Training and Testing
  - Gold-standard summaries
  - SVM (Joachims 2006 SVM^{rank})

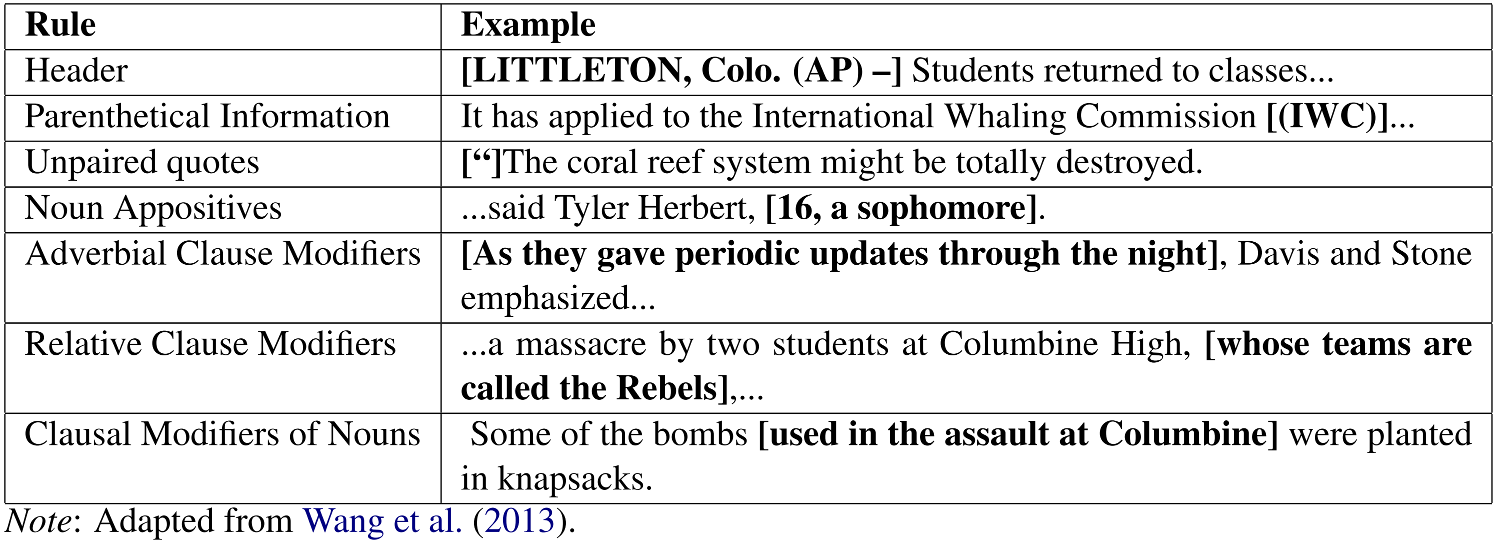


Improvements

Content Realization

Rule-Based Compression

Bold elements in brackets removed



Rule	Example
Header	**[LITTLETON, Colo. (AP) –]** Students returned to classes...
Parenthetical Information	It has applied to the International Whaling Commission **[(IWC)]**...
Unpaired quotes	**[“]**The coral reef system might be totally destroyed.
Noun Appositives	...said Tyler Herbert, **[16, a sophomore]**.
Adverbial Clause Modifiers	**[As they gave periodic updates through the night]**, Davis and Stone emphasized...
Relative Clause Modifiers	...a massacre by two students at Columbine High, **[whose teams are called the Rebels]**,...
Clausal Modifiers of Nouns	Some of the bombs **[used in the assault at Columbine]** were planted in knapsacks.

Note: Adapted from Wang et al. (2013).

Tools & Algorithm

Punctuation

- ○ Regex pattern-matching
- ○ Remove the item
- ○ Generate a new sentence

Grammatical Constituents

- ○ spaCy (Explosion AI, 2019) dependency parse
- ○ Navigate the parse tree
- ○ Remove the node
- ○ Generate a new sentence

Analysis

Samples

Content Selection Ranking

Sentence Compression

Before:

Seven near-simultaneous bomb blasts tore through crowded markets in the Indian tourist city of Jaipur Tuesday, killing at least 80 people and wounding 200 in what police said was a terror attack. Seven bombs ripped through the city of Jaipur, leaving at least 60 people dead. ***It’s a terror attack.*** *In the first terrorist attack in many months, seven bombs went off within minutes of one another on Tuesday evening in the crowded lanes of one of India’s main tourist hubs, the historic city of Jaipur, with reports of deaths* ***[ranging from 50 upward]****, with roughly 150 injured, officials said.*

After:

Seven near-simultaneous bomb blasts tore through crowded markets in the Indian tourist city of Jaipur Tuesday, killing at least 80 people and wounding 200 in what police said was a terror attack. Seven bombs ripped through the city of Jaipur, leaving at least 60 people dead. ***Pakistan denies any role in the bombings.*** *In the first terrorist attack in many months, seven bombs went off within minutes of one another on Tuesday evening in the crowded lanes of one of India’s main tourist hubs, the historic city of Jaipur, with reports of deaths, with roughly 150 injured, officials said.*

A Good Summary

Sentence Cleaning

Before:

PORT MORESBY, Papua New Guinea (AP) _ A tsunami spawned by a 7.0 magnitude earthquake crashed into Papua New Guinea’s north coast, crushing villages and leaving hundreds missing, officials said Sunday. Authorities at Aitape in the West Sepik province, on Papua New Guinea’s northwest coast, said the tsunami that hit the coast west of the village on Friday night had wiped out three villages and had almost completely destroyed another. “They’re dead . CANBERRA, July 18 (Xinhua) – Australia will provide transport for relief supplies and a mobile hospital to Papua New Guinea (PNG) following Friday’s tsunami tragedy.

After:

Authorities at Aitape in the West Sepik province, on Papua New Guinea’s northwest coast, said the tsunami that hit the coast west of the village on Friday night had wiped out three villages and had almost completely destroyed another. A tsunami spawned by a 7.0 magnitude earthquake crashed into Papua New Guinea’s north coast, crushing villages and leaving hundreds missing, officials said Sunday. The death toll in Papua New Guinea’s tsunami disaster has climbed to 599 and is expected to rise, a PNG disaster control officer said Sunday. They’re dead . It’s complete devastation. The beaches are clean, Cassey said.

A Good Summary

Sentence Compression

Bold elements in brackets removed

*Representatives, including the commercial fishing industry, government regulators and environmental groups, plan **[to discuss ways to strengthen information sharing and cooperation among regional organizations]** to better manage tuna stock and adopt an action plan, the Japanese Fisheries Agency said in a statement.*

Readability Errors

Entity Mention Violations

- Pronouns, definite articles, and last names used in first mention of an entity

Drought changed all ***that*** *...*
They *ascribed ...*
The letter was from Sudan's
foreign minister, ***Erwa*** *said.*

- Full name of an entity or explanation used in subsequent mentions

Debra Lafave, the former Tampa middle school teacher accused of having sex with a 14-year-old male student ...Debra LaFave had sex with a 14-year-old student

Readability Errors

Clausal Level Violations

- Inconsistent formatting

  Cox News Service RICHMOND, Va. ?-

- Discourse relation errors

  ***But** in Sichuan and in Shaanxi province, …*

  ***By comparison**, Floyd's hurricane-force wind …*

- Redundancy

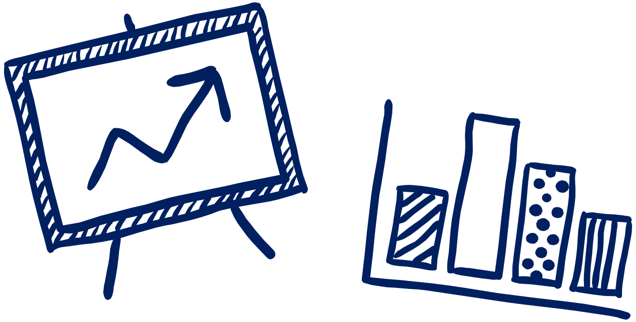
  on Friday night … on Friday night

- Grammatical violations leads to nonsense sentences

  They, he said.

  And it has China, again.

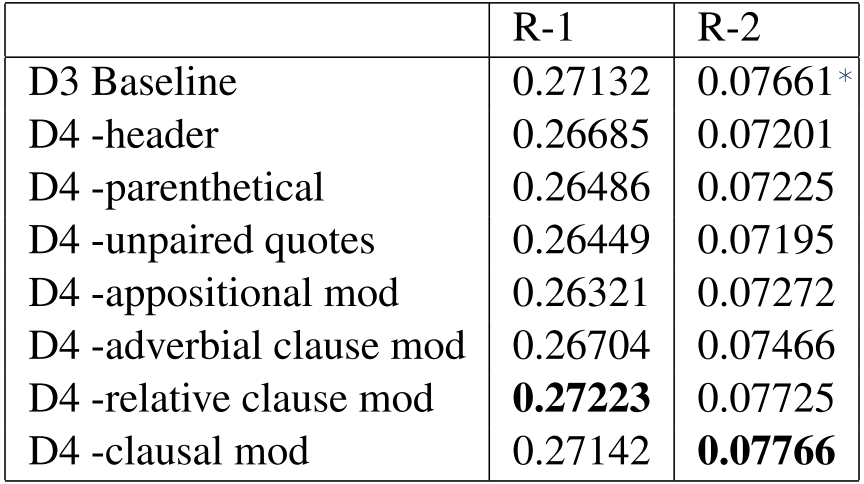
  Chen also stressed that.



Results

Evaluation

Devtest

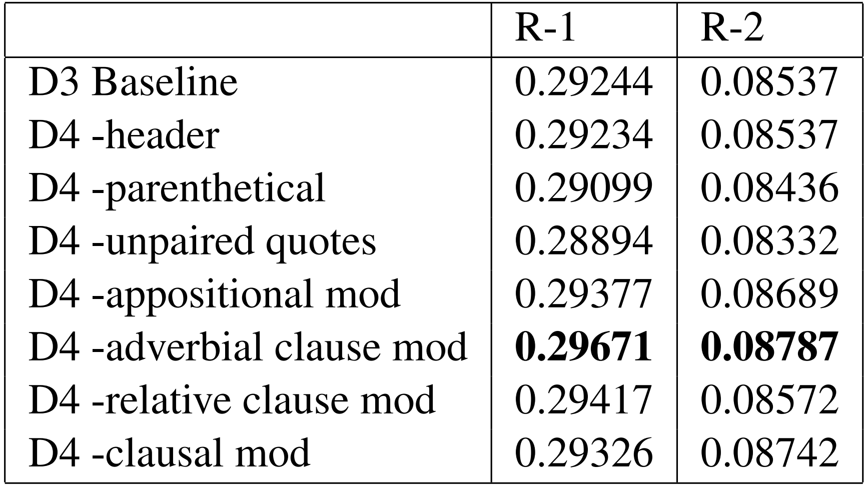


	R-1	R-2
D3 Baseline	0.27132	0.07661*
D4 -header	0.26685	0.07201
D4 -parenthetical	0.26486	0.07225
D4 -unpaired quotes	0.26449	0.07195
D4 -appositional mod	0.26321	0.07272
D4 -adverbial clause mod	0.26704	0.07466
D4 -relative clause mod	**0.27223**	0.07725
D4 -clausal mod	0.27142	**0.07766**

*R-2 is 0.07661 instead of 0.08017 due to some changes in the preprocessing of data

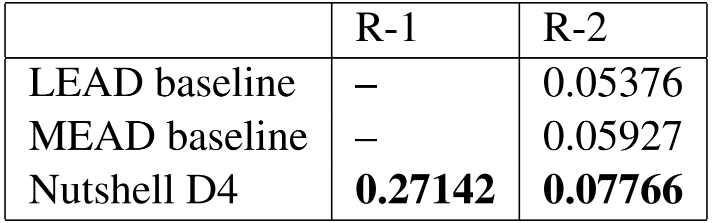
Evaluation

Evaltest

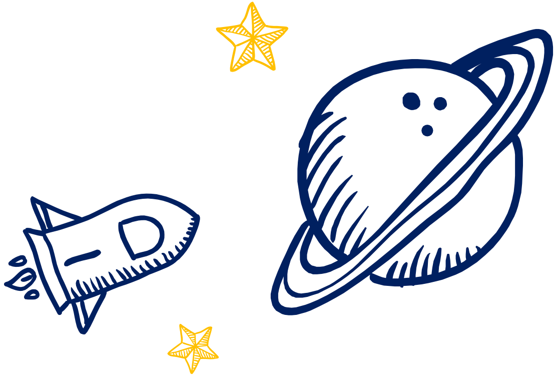


	R-1	R-2
D3 Baseline	0.29244	0.08537
D4 -header	0.29234	0.08537
D4 -parenthetical	0.29099	0.08436
D4 -unpaired quotes	0.28894	0.08332
D4 -appositional mod	0.29377	0.08689
D4 -adverbial clause mod	**0.29671**	**0.08787**
D4 -relative clause mod	0.29417	0.08572
D4 -clausal mod	0.29326	0.08742

LEAD and MEAD baselines TAC 2010



	R-1	R-2
LEAD baseline	–	0.05376
MEAD baseline	–	0.05927
Nutshell D4	**0.27142**	**0.07766**



Future Work

Module Improvements

- More sophisticated Entity Grids for Information Ordering
- Improvement of Content Realization sub-components
    - Continue to remove metadata and punctuation-based patterns
    - Coreference resolution and the use of named entity recognition to shorten entity references
    - Integration of machine learning into the selection of which parameters correlate most highly with ROUGE scores

Related Readings

Explosion AI. 2019. encorewebmd: English multi-task cnn trained on ontonotes.

Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research, 17:35– 55.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. Computational Linguistics, 34(1):1–34.

Jahna Otterbacher, Gunes Erkan, and Dragomir Radev. 2005. Using random walks for question-focused sentence retrieval. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 915-922, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. Information Processing and Management, 43(6):1606–1618.

Lu Wang, Hema Raghavan, Vittorio Castelli, Radu Florian, and Claire Cardie. 2013. A sentence compression based framework to query-focused multi- document summarization. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1384–1394. Association for Computational Linguistics.

David Zajic, Bonnie J. Dorr, Jimmy Lin, and Richard Schwartz. 2007. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. Information Processing and Management, 43(6):1549–1570.

Thanks!

Any questions?