




ANLP: Survey Generation

Sally Ma
2020.9.29



Outline

- Introduction and Background
 - Multi-document Summarization and Survey Generation
 - Traditional Approaches and Recent Work
 - Datasets and Evaluation
- Papers
 - [Sauper and Barzilay. Automatically Generating Wikipedia Articles: A Structure-aware Approach. In ACL 2009](#)
 - [Banerjee and Mitra. Wikiwrite: Generating Wikipedia Articles Automatically. In IJCAI 2016](#)
 - [Deutsch and Roth. Summary Cloze: A New Task for Content Selection in Topic-Focused Summarization. In EMNLP 2019](#)
- Discussion Questions
- Reference

Multi-document Summarization (MDS)

- Goal: output summaries from document clusters on the same topic
 - \Leftrightarrow single-document summarization, where the input is a single document
- Types
 - Generic: makes no assumption about the domain; where the majority of the work is in
 - Domain-specific: uses domain-specific knowledge e.g. summarize biomedical documents
 - Query-based: generate summary that contains only information requested by the query about given inputs.
- Methods
 - Extractive: select sentences from inputs to form a summary; most approaches
 - Abstractive: paraphrase and rewrite; more coherent and avoids copyright violations, but much more difficult to achieve

Survey Generation

- A subset of MDS: it's query-based MDS, the query = a general topic
- Goal: automatically build informative surveys of a topic
- Motivation:
 - Scientific domain: rapid growth of publications => human-written surveys are limited in the coverage of topics and often become outdated quickly
 - Wikipedia domain: limited by the availability of knowledgeable authors and cannot keep pace with ever increasing demands of readers
 - => Survey generation helps us address information overload and find relevant information on a topic more efficiently

Traditional Approaches

- Abstractive
 - SUMMONS ([McKeown and Radev, 1995](#); [Radev and McKeown, 1998](#))
 - Pioneering work in MDS

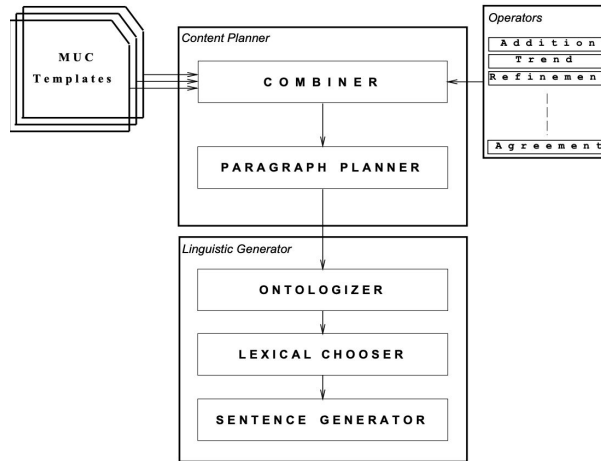


Figure 2: System Architecture.

Traditional Approaches

- Abstractive
 - SUMMONS ([McKeown and Radev, 1995](#); [Radev and McKeown, 1998](#))
 - Pioneering work in MDS
 - Opinosis ([Ganesan et al., 2010](#))
 - A graph-based approach to abstractive summarization of highly redundant opinions

Traditional Approaches

- Extractive
 - Maximum Marginal Relevance (MMR) ([Carbonell and Goldstein, 1998](#))
 - A ranking score used produce a ranked list of the candidate sentences based on the relevance to the query and redundancy to the selected sentences, which can be used to extract sentences.
 - The MMR score is calculated as follows:

$$\text{MMR} = \underset{D_i \in R \setminus S}{\operatorname{argmax}} [\lambda(\text{Sim}_1(D_i, Q) - (1 - \lambda)\max_{D_j \in S} \text{Sim}_2(D_i, D_j))]$$

Traditional Approaches

- Extractive
 - Maximum Marginal Relevance (MMR) ([Carbonell and Goldstein, 1998](#))
 - Graph-based e.g. Lexrank ([Erkan and Radev, 2004](#)); Textrank ([Mihalcea and Tarau, 2004](#))
 - Turn sentences into vector representations (word embeddings); calculate similarity matrix of the vectors and convert into graph (sentences as nodes and similarity scores as edges); ranks sentences based on centrality (each edge being a vote), and selects the top-ranked sentences as summaries

Recent Work

- Survey generation for scientific topics
 - [Qazvinian and Radev, 2008](#): C-Lexrank (a state-of-the-art system for survey generation)
 - Content model based on lexical network -- fully connected network of an article: node = each sentence in the citation summary of the article; weight of an edge = cosine similarity of pairs of sentences
 - Cluster the nodes s.t. intra-cluster similarity is max & inter-cluster similarity is min
 - => Communities of sentences discussing the same scientific contributions
 - Apply Lexrank in each cluster (largest to smallest) to find the most central sentences in each cluster for building the summary

Recent Work

- Survey generation for scientific topics
 - [Qazvinian and Radev, 2008](#): C-Lexrank (a state-of-the-art system for survey generation)
 - [Jha et al., 2015](#): HITSUM
 - New formulation of the lexical network to include additional lexical information:
 - $P(\text{citing})$ --the introductory sections of papers citing important papers on a topic (as in C-lexrank)
 - $P(\text{cited})$ -- sentences from the papers cited by these introductory sections
 - Form a directed edge from $s(i)$ in $P(\text{citing})$ and $s(j)$ in $P(\text{cited})$ if the tf-idf cosine similarity between the two $>$ threshold
 - Assigns hubs and authority scores to each node in a mutually reinforcing way
 - high authority scores = report important contributions; high hub scores = summarize important contributions
 - Select sentences w/ high hub scores in $P(\text{citing})$

Recent Work

- Survey generation for scientific topics
 - [Qazvinian and Radev, 2008](#): C-Lexrank (a state-of-the-art system for survey generation)
 - [Jha et al., 2015](#): HITSUM
 - [Jha et al., 2015](#). Surveyor
 - Focus on generating coherent surveys--those with well-defined and ordered subtopics
 - A content model, modeling subtopics in a scientific paper with Hidden Markov Model
 - A discourse model, modeling discourse-level dependencies for locally coherent summaries
 - Minimum Independent Discourse Contexts (MIDC) of a sentence $s(i)$: the minimum set of sentences preceding it such that $s(i)$ can be interpreted independently of the other sentences in its text segment
 - => 36% more coherent than C-Lexrank in human evaluation

Recent Work

- Single-document summarization (SDS) has been driven in recent years by:
 - The availability of large-scale datasets containing hundreds of thousands of document summary pairs
 - Sandhaus, 2008; Hermann et al., 2015; Grusky et al., 2018
 - Neural networks' ability to learn continuous representations without preprocessing tools / linguistic annotations, e.g. Seq2seq models that encode a source document and decode it into a summary.
 - See et al., 2017; Celikyilmaz et al., 2018; Paulus et al., 2018; Gehrmann et al., 2018

Recent Work

- MDS has been a lot less popular, as a result of bottlenecks e.g.
 - Data scarcity -- Human written summaries are sparse and expensive. High-quality multi-document summarization datasets (i.e., document clusters paired with multiple reference summaries written by humans) are small for training neural models:
 - the DUC 2004 (Paul and James, 2004) and TAC 2011 (Owczarzak and Dang, 2011) datasets have fewer than 100 document clusters
 - Difficulty applying end-to-end models, as the size and number of source documents can be very large
 - Given memory limitations of current hardware, it's infeasible to train a model that encodes all the documents into vectors and decode them to produce a summary

Large-scale Datasets

- To drive research efforts, large-scale datasets are introduced in most recent years:
- News domain: [Multi-News](#) ([Fabbri et al., 2019](#)) -- the first large-scale MDS news dataset
 - 56,216 articles-summary pairs
 - News articles and their summaries, professionally written by editors, from the site [newser.com](#)
 - Diverse: over 1,500 sites appear as source documents 5 times or greater \Leftrightarrow previous news datasets -- DUC and CNNDM both come from 2 sources; the Newsroom dataset (Grusky et al., 2018) comes from 38

Large-scale Datasets

- Wikipedia domain: [WikiSum](#) ([Liu et al., 2019](#)) -- a collection of over 1.5 millions Wikipedia pages (~300 GB)
 - Crawled Wikipedia articles and 78.9% of source reference documents (excluded invalid urls)
 - On average, a data instance has 525 paragraphs + a target summary with 139.4 tokens
 - A ranked version (for reduced size): top 40 source paragraphs (ranked by a learned ranker) for each data instance

Recent Work

- Neural methods for MDS
 - Exploit the graph structure among discourse relations in text clusters ([Yasunaga et al., 2017](#))
 - Use a Graph Convolutional Network (GCN) on the relation graphs + sentence embeddings from Recurrent Neural Networks (RNNs) as input node features.
 - GCN generates high-level hidden sentence features for salience estimation; use a greedy heuristic to extract salient sentences
 - Improvement from traditional graph-based extractive approaches and the vanilla sequence model (specifically GRU) with no graph

Recent Work

- Neural methods for MDS
 - Exploit the graph structure among discourse relations in text clusters ([Yasunaga et al., 2017](#))
 - Adapt models trained on SDS data to MDS
 - [Zhang et al., 2018](#): adds an additional document-level encoding, to adapt a hierarchical encoding framework trained on SDS data to MDS
 - [Lebanoff et al., 2018](#): introduces an external (MMR) module that does not require training on the MDS, to apply encoder-decoder models trained on SDS data to MDS
 - Common to truncate input articles to reduce MDS to SDS on longer documents
 - N tokens, S source documents; take the first N/S tokens from each source document, repeat iteratively if needed; concatenate the truncated documents into a single document

Evaluation

- Difficult because there doesn't exist an *ideal* summary given documents
- Human/manual evaluation
 - Rating on a scale (e.g. from 1 to 5) of grammaticality, non-redundancy, referential clarity, focus, structure, and coherence
 - Question-answering -- Compare the number of questions that can be answered from the surveys
 - Limitation: too expensive => want to find automatic evaluation metrics that have high correlation with human scores

Evaluation

- Automatic evaluation
 - [ROUGE \(Lin, 2004\)](#): Recall-Oriented Understudy for Gisting Evaluation
 - a set of metrics that have become the standards
 - ROUGE-N: n-gram recall

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)}$$

Evaluation

- Automatic evaluation
 - [ROUGE \(Lin, 2004\)](#): Recall-Oriented Understudy for Gisting Evaluation
 - a set of metrics that have become the standards, including
 - ROUGE-N: n-gram recall
 - ROUGE-L: based on longest common subsequence

$$\text{ROUGE-L}(s) = \frac{(1 + \beta^2)R_{\text{LCS}}P_{\text{LCS}}}{R_{\text{LCS}} + \beta^2 P_{\text{LCS}}}$$

$$R_{\text{LCS}}(s) = \frac{\sum_{i=1}^u \text{LCS}(r_i, s)}{\sum_{i=1}^u |r_i|}, \quad P_{\text{LCS}}(s) = \frac{\sum_{i=1}^u \text{LCS}(r_i, s)}{|s|}$$

Evaluation

- Automatic evaluation

- ROUGE (Lin, 2004): Recall-Oriented Understudy for Gisting Evaluation
 - a set of metrics that have become the standards
 - ROUGE-N: n-gram recall
 - ROUGE-L: based on longest common subsequence
 - Limitation: lexically based, so doesn't capture semantic similarity
- Alternative evaluation metrics
 - [Sun et al., 2019](#): ROUGE is highly sensitive to summary length => gives unfair advantage to methods that produce longer summaries => alternative metric that normalize ROUGE scores with those from a random system producing summaries of the same length
 - [Goodrich 2019](#): an alternative metric that correlates with human evaluation of factual accuracy better

Outline

- Introduction and Background
 - Multi-document Summarization and Survey Generation
 - Traditional Approaches and Recent Work
 - Datasets and Evaluation
- Papers
 - [Sauper and Barzilay. Automatically Generating Wikipedia Articles: A Structure-aware Approach. In ACL 2009](#)
 - [Banerjee and Mitra. Wikiwrite: Generating Wikipedia Articles Automatically. In IJCAI 2016](#)
 - [Deutsch and Roth. Summary Cloze: A New Task for Content Selection in Topic-Focused Summarization. In EMNLP 2019](#)
- Discussion Questions
- Reference

Paper #1

Automatically Generating Wikipedia Articles: A Structure-Aware Approach

Christina Sauper and Regina Barzilay

Computer Science and Artificial Intelligence Laboratory

Massachusetts Institute of Technology

{csauper, regina}@csail.mit.edu

Abstract

In this paper, we investigate an approach for creating a comprehensive textual overview of a subject composed of information drawn from the Internet. We use the high-level structure of human-authored texts to automatically induce a domain-specific template for the topic structure of a new overview. The algorithmic innova-

such as searching the Internet. Moreover, the challenge of maintaining output readability is magnified when creating a longer document that discusses multiple topics.

In our approach, we explore how the high-level structure of human-authored documents can be used to produce well-formed comprehensive overview articles. We select relevant material for an article using a domain-specific automatically generated content template. For example, a tem-

Motivation

- Challenge of maintaining output readability and structure is magnified while creating a long document that discusses multiple subtopics
 - Automatically inducing templates from documents helps with structure and coverage
- Want to optimize both local relevance of information for each subtopic and global coherence across the entire articles
 - Past work in MDS selects from documents of each subtopic individually; then combine the results while eliminating redundancy
 - Having a separated selection and combination process may not be optimal for generating comprehensive, multi-paragraph summary, where balance between coverage and redundancy is harder to achieve
 - => learn parameters for content selection jointly across all topics in the template (rather than using topic-specific parameters)
- => proposes a new system termed Perceptron-ILP

Method

- Preprocess
 - Template induction for a domain
 - Cluster all section headings for all documents
 - Label each cluster with the most common heading within the cluster
 - Select the largest k clusters to be topics forming the template
 - For each survey, retrieve from the internet a set of r excerpts for each topic from the template

Method

- Learn content selection criteria for all the topics simultaneously to maximize local fit and global coherence
 - Rank each excerpt based on how representative they are of each topic: map each excerpt to a score; rank from high to low
- $\phi(e_{jl})$ — feature vector for the l th candidate excerpt for topic t_j
- $\mathbf{w}_1 \dots \mathbf{w}_k$ — parameter vectors, one for each of the topics $t_1 \dots t_k$

$$score_j(e_{jl}) = \phi(e_{jl}) \cdot \mathbf{w}_j$$

Feature	Value
UNI_ $word_i$	count of word occurrences
POS_ $word_i$	first position of word in excerpt
BI_ $word_i$ _ $word_{i+1}$	count of bigram occurrences
SENT	count of all sentences
EXCL	count of exclamations
QUES	count of questions
WORD	count of all words
NAME	count of title mentions
DATE	count of dates
PROP	count of proper nouns
PRON	count of pronouns
NUM	count of numbers
FIRST_ $word_1$	1 [*]
FIRST_ $word_1$ _ $word_2$	1 [†]
SIMS	count of similar excerpts [‡]

Table 1: Features employed in the ranking model.

* Defined as the first unigram in the excerpt.

† Defined as the first bigram in the excerpt.

‡ Defined as excerpts with cosine similarity > 0.5

Method

- Learn content selection criteria for all the topics simultaneously to maximize local fit and global coherence
 - Rank each excerpt based on how representative they are of each topic
 - Exclusivity constraints: for each topic, choose the best excerpt only
 - Redundancy constraints: if excerpts have cosine sim > 0.5 , only one excerpt may be selected for the final doc
 - Use Integer Linear Programming (ILP) to select excerpts that minimize ranks globally
 - Learn the optimal k parameter vectors $w(1)...w(k)$ for each topic with a simple neural network: iterate over the train set until parameters converge / maximum number of iterations reached; in an iteration:
 - For each document, rank excerpts. Use ILP over the global list of ranked excerpts to select one excerpt for each topic. For each topic, if the selected excerpt isn't similar enough, penalize using standard perceptron update rule

Results

- Example output, where a template for articles about diseases are induced to have topics: diagnosis, causes, symptoms, an treatment

Diagnosis ...No laboratories offering molecular genetic testing for prenatal diagnosis of 3-M syndrome are listed in the GeneTests Laboratory Directory. However, prenatal testing may be available for families in which the disease-causing mutations have been identified in an affected family member in a research or clinical laboratory.

Causes Three M syndrome is thought to be inherited as an autosomal recessive genetic trait. Human traits, including the classic genetic diseases, are the product of the interaction of two genes, one received from the father and one from the mother. In recessive disorders, the condition does not occur unless an individual inherits the same defective gene for the same trait from each parent. ...

Symptoms ...Many of the symptoms and physical features associated with the disorder are apparent at birth (congenital). In some cases, individuals who carry a single copy of the disease gene (heterozygotes) may exhibit mild symptoms associated with Three M syndrome.

Treatment ...Genetic counseling will be of benefit for affected individuals and their families. Family members of affected individuals should also receive regular clinical evaluations to detect any symptoms and physical characteristics that may be potentially associated with Three M syndrome or heterozygosity for the disorder. Other treatment for Three M syndrome is symptomatic and supportive.

Figure 1: A fragment from the automatically created article for 3-M Syndrome.

Results

- Tested on two domains: American Film Actors and Diseases, both of which commonly used in prior work on summarization
- Use articles drawn from the corresponding categories in Wikipedia
 - 90/10 train/set; exclude Wikipedia sources during internet search phase of the method
- Oracle: selects excerpts with highest cosine sim to the target => upper bound
- Baselines
 - Search: search engine ranking for content selection; for each page, select the first k paragraphs
 - No Template: not inducing topics as constraints
 - Disjoint: learn weight parameters for each topic rather than globally

	Recall	Precision	F-score
Amer. Film Actors			
Search	0.09	0.37	0.13 *
No Template	0.33	0.50	0.39 *
Disjoint	0.45	0.32	0.36 *
Full Model	0.46	0.40	0.41
Oracle	0.48	0.64	0.54 *
Diseases			
Search	0.31	0.37	0.32 †
No Template	0.32	0.27	0.28 *
Disjoint	0.33	0.40	0.35 *
Full Model	0.36	0.39	0.37
Oracle	0.59	0.37	0.44 *

Table 3: Results of ROUGE-1 evaluation.

* Significant with respect to our full model for $p \leq 0.05$.

† Significant with respect to our full model for $p \leq 0.10$.

Assessment

- Contribution

- Proposed and demonstrated the benefit of inducing templates for structurally-aware content selection (compared to approaches that do not explicitly model topical structure), and of learning parameters for content selection jointly across all topics in the template (rather than using topic-specific parameters)

- Limitation

- Domain specific, and would need to learn parameters and templates for every Wikipedia category => not necessarily scalable
- There could be categories with higher structural variability, and their proposed method for inducing templates may not yield accurate results

Paper #2

WikiWrite: Generating Wikipedia Articles Automatically

Siddhartha Banerjee[†]

[†]The Pennsylvania State University
University Park, PA, USA
sbanerjee@ist.psu.edu

Prasenjit Mitra^{†*}

^{*}Qatar Computing Research Institute
HBKU, Doha, Qatar
pmitra@ist.psu.edu

Abstract

The growth of Wikipedia, limited by the availability of knowledgeable authors, cannot keep pace with the ever increasing requirements and demands of the readers. In this work, we propose WikiWrite, a system capable of generating content for new Wikipedia articles automatically. First, our technique obtains feature representations of entities on Wikipedia. We adapt an existing work on document embeddings to obtain vector representations of words and paragraphs. Using the repre-

Wikipedia categories³ are known. Furthermore, even if the category information is known, articles often belong to multiple categories that are often not equally important. Learning from several categories may result in a mish-mash of sections from different categories being chosen resulting in a less polished article. In some cases, categories do not provide much information, for example, *1959 births* or *Living people* in *Octavio Solis*⁴ article. The second issue, copyright violations, imply that content on the entity retrieved from the web cannot be directly copied into Wikipedia [Banerjee *et al.*, 2014]. To tackle the issue of copyright violation, we proposed an abstractive summarization system [Banerjee and Mitra, 2015c]

Motivation

- Past work all assume that Wikipedia categories are known, which may not always be true
 - even if they are known, survey articles often belong to multiple categories often not equally important => past work did not address how to generate coherent summaries from multiple categories
- Purely extractive methods used in the past on internet information are subject to copyright violations => use abstractive summarization
- Past work does not address lacking coherence while selecting and/or paraphrasing from multiple documents
- => a new end-to-end system to address all of these issues

Method

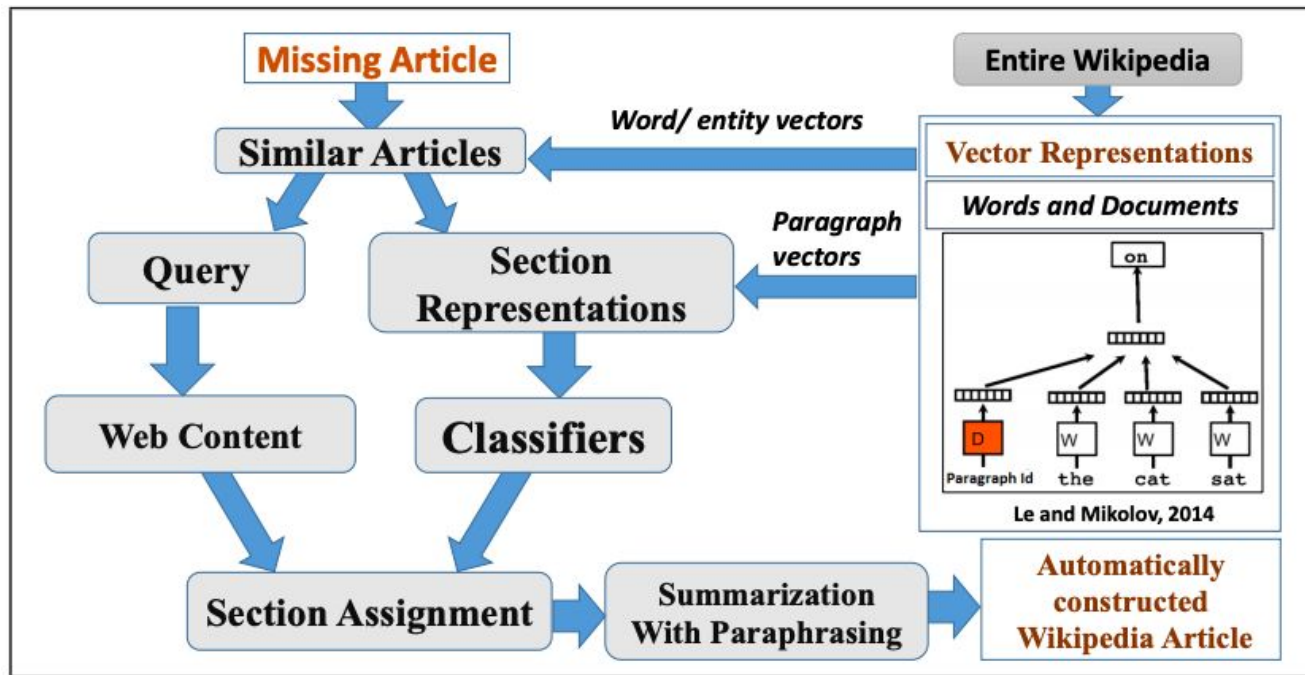


Figure 1: WikiWrite: Our Proposed Framework

Method

$$F = \sum_{i=1}^K w^{p_i} \cdot p_i + \lambda \sum_{a_{i,j} \in A} coh_{i,j} \cdot arc_{i,j} \quad (3)$$

$$w^{p_i} = I^{p_i} \cdot Sim^{intra}(p_i) \cdot LQ(p_i) \quad (4)$$

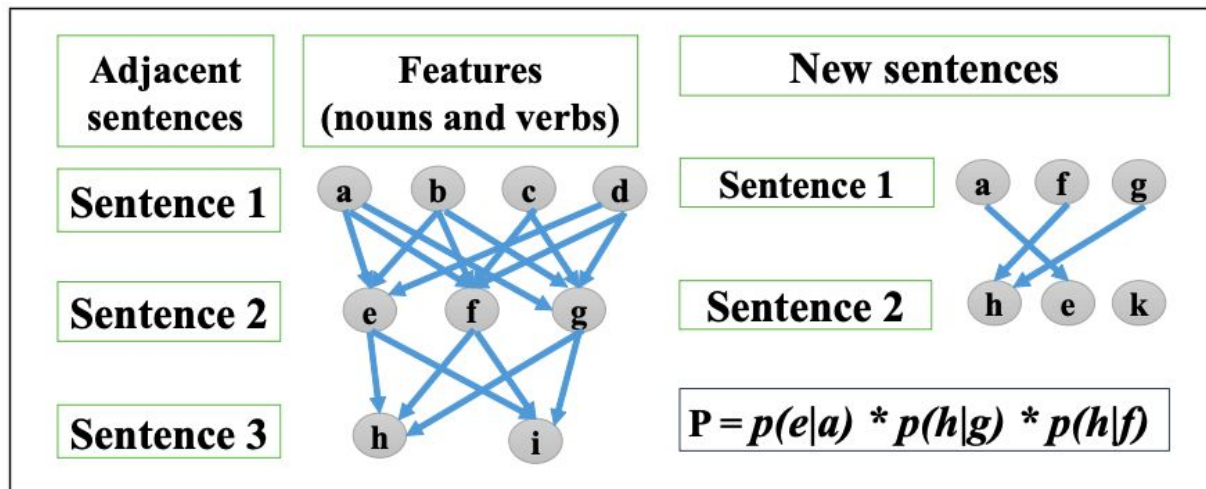


Figure 2: Local coherence estimation between sentences

Method

- "to considerable **economic benefits** for the" -- modification from **financial advantages** to **economics benefits**
- maximize this objective function:

$$R(t_1, \dots, t_{|T|}) = \sum_{i=1}^{|T|} Sim(seq_o, seq_i) \cdot LQ(seq_i) \cdot t_i \quad (8)$$

Results

- Table 1: WikiWrite able to assign content to sections more efficiently by learning from similar articles rather than restricting to only the most frequent sections in the categories; also assigns sections significantly faster
- Table 2: outperforms baselines (WikiKreator and Perceptron-ILP) according to both ROUGE scores.
 - WikiWrite(Ref) uses only the references listed in the Wikipedia article as an upper bound

Table 1: Section Classification Results

Technique	F1-score	Average Time
WikiWrite	0.622	~2 mins
WikiKreator	0.481	~10 mins

Table 2: Content Selection Results

Technique	ROUGE-1	ROUGE-2
WikiWrite	0.441	0.223
WikiWrite (Ref)	0.520	0.257
WikiKreator	0.371	0.183
Perceptron-ILP	0.342	0.169

Table 3: Statistics of Wikipedia stub content addition

Statistics	WikiKreator	WikiWrite
No. of stubs appended	40	40
Entire edit retained	15	32
Modification of content	5	5
Content Removed	20	3
Avg. change in size	287 bytes	424 bytes
Avg. no of edits	3.82	1.39

Assessment

- Contributions

- A new approach for Wikipedia article generation that
 - doesn't require information on Wikipedia categories
 - doesn't violate copyright (i.e. copy content from the web directly into the article) by paraphrasing (not using original words from source documents)
 - maintains coherence (optimizing the ordering of sentences) when choosing and paraphrasing from multiple documents

- Limitations

- Get sections by finding the most similar Wikipedia pages, and the query reformulation method relies on the most frequent nouns from the top 20 most similar Wikipedia pages. This may not be reliable for novel topics that do not have many existing Wikipedia pages that are similar

Paper #3

Summary Cloze: A New Task for Content Selection in Topic-Focused Summarization

Daniel Deutsch and Dan Roth

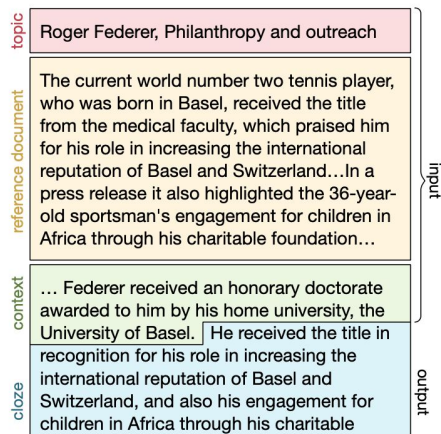
Department of Computer and Information Science

University of Pennsylvania

{ddeutsch, danroth}@seas.upenn.edu

Abstract

A key challenge in topic-focused summarization is determining what information should be included in the summary, a problem known as content selection. In this work, we propose a new method for studying content selection in topic-focused summarization called the *summary cloze* task. The goal of the summary cloze task is to generate the next sentence of a summary conditioned on the beginning of the summary, a topic, and a reference document(s). The main challenge is deciding what information in the references is relevant to the topic and partial summary and should be included in the summary. Although the

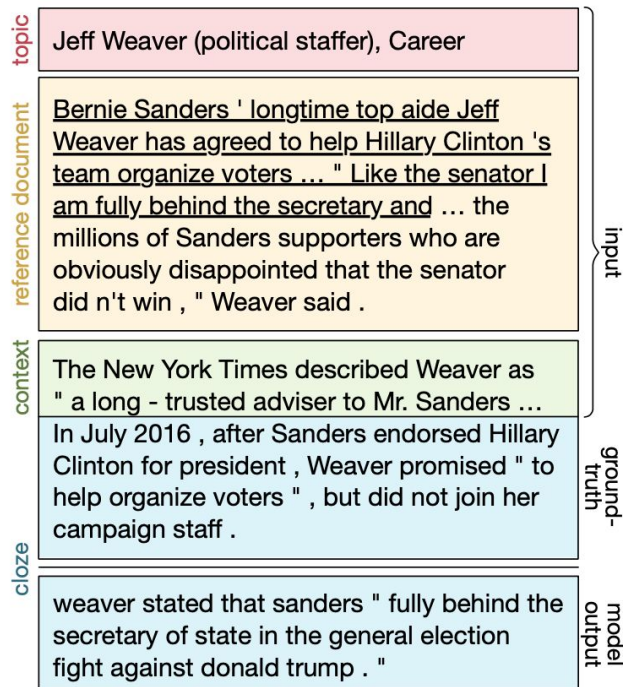


Motivation

- Narrow the scope of the problem makes it easier to collect a large-scale dataset tailored to the task => focus on the task of content selection (choosing what information to include in the summary), a key challenge in topic-focused summarization

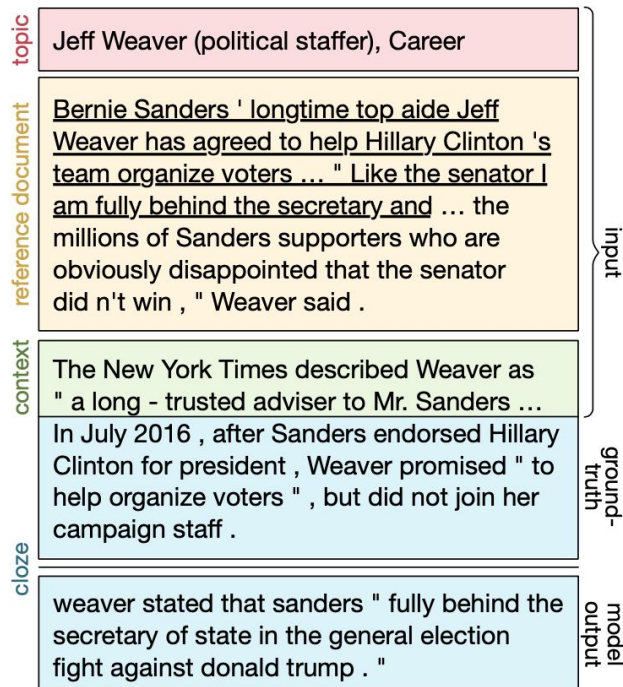
Method

- Formulate a new method for studying content selection: the summary cloze task
 - Goal: predict the next sentence of a summary (known as the cloze), conditioned on the beginning of the summary, a topic, and reference documents



Method

- Formulate a new method for studying content selection: the summary cloze task
 - Goal: predict the next sentence of a summary (known as the cloze), conditioned on the beginning of the summary, a topic, and reference documents
- Collected a new large-scale dataset from Wikipedia, called WikiCite
 - ~500k data instances
 - Each paragraph = a topic-focused summary of the references cited within the paragraph
 - Topic = article title and section headings
 - Context = previously generated summary

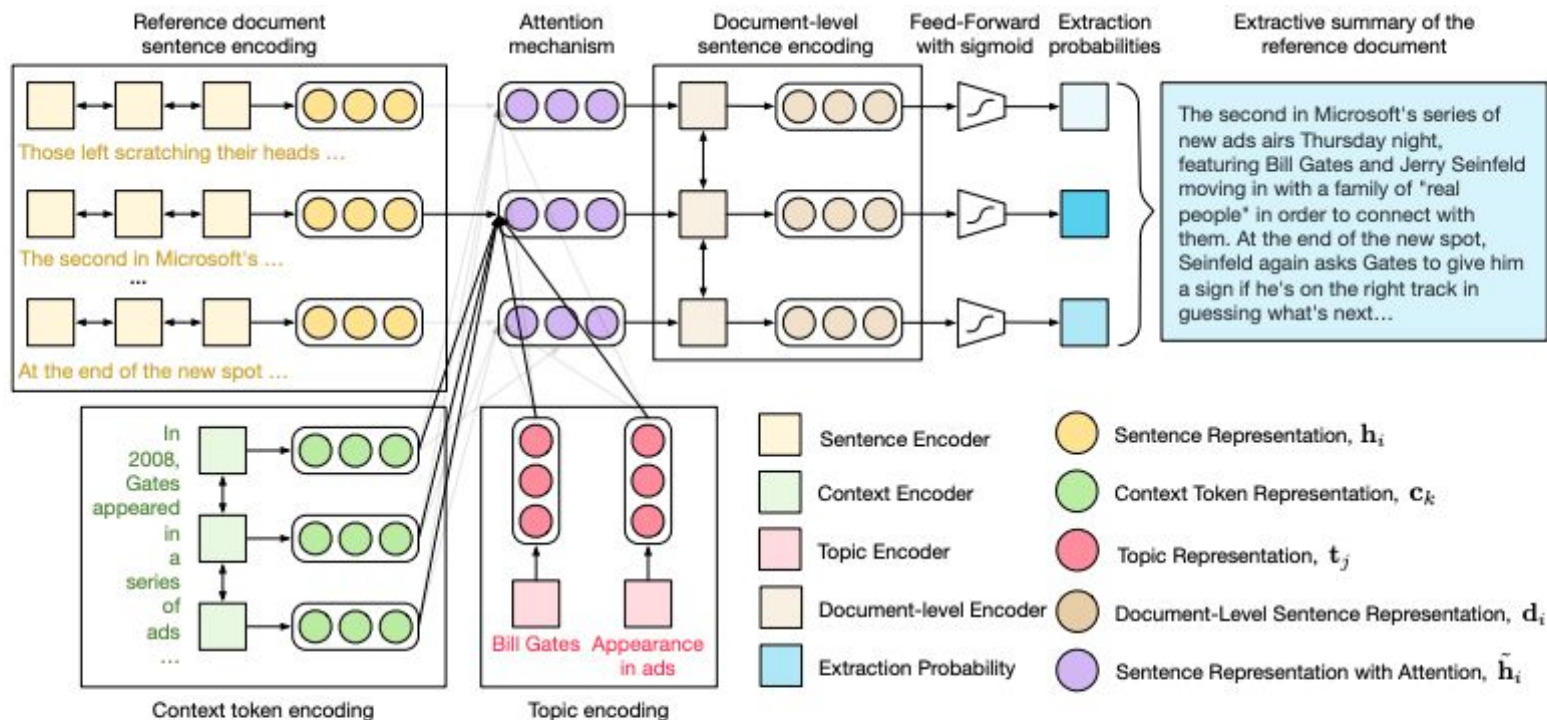


Method

- Approached this task with an extractive model and a two-step abstractive model
 - Extractive: combines representations of the topic and partial summary w/ representations of the document sentences through an attention mechanism to extract one reference sentence
 - Abstractive: reduces the length of the input data by extractively selecting a small num of sentences and abstractively summarizes them w/ a decoder that has an initial hidden state which depends on the partial summary

Section-body creation

- Extractive model architecture:



Method

- Abstractive step: replace the list of documents with the extractive model output, and use an extension of Pointer-Generator + Coverage network as the abstractive model for outputting a paraphrased summary
 - Pointer-Generator network: built on seq2seq model with attention; the reference document is encoded using an RNN, and a 2nd RNN produces the summary. Augmented with a copy mechanism: allows the attention distribution to influence the decoder's probability distribution over vocabulary => copy words from input more easily
 - Coverage mechanism: discourages attention weights from assigning high values to the same input tokens across decoding time steps repeatedly, by adding a new term to the loss function that penalizes the behavior => reduce redundancy in generated summary

Results

- Extractive model:
 - Oracle: selects the sentence from the reference set that maximizes the ROUGE score using ground truth
 - Providing context and topic shows improvement in performance; but topic may not be utilized properly since its impact is not statistically significant
 - Room for improvement

Model	R1	R2	RL
NO REFERENCE	14.47	1.43	11.41
NO CONTEXT	21.79	6.18	17.83

Table 3: The ROUGE F1 results of the baseline models that do not have access to the references or context.

Model	R1	R2	RL
ORACLE-1	44.32	26.17	37.62
LEAD-1	17.50	4.70	13.45
BM25	21.33	5.76	16.18
SUMFOCUS	17.78	4.61	13.86
CONTENTSELECTOR	25.13	9.48	19.65
-TOPIC	25.21	9.49	19.68
-CONTEXT	22.11	6.53	16.66
-TOPIC,-CONTEXT	21.29	6.01	16.05

Table 4: The ROUGE F1 scores for the extractive models, all of which are significantly lower than the oracle model, indicating there is room for improvement.

Results

- Abstractive model: ROUGE scores w/ 3 different extractive steps
 - Outperforms LEAD-200 (select first 200 tokens of reference); including context improves performance across all models=> Priming the decoder with context produces a better cloze
 - Sizable gap between the performance of the abstractive model when it uses the heuristic labels (modeling a perfect extractive step) vs. extractive model for preprocessing => improving the extractive model will provide large downstream abstractive improvement

Ext. Model	Abs.	R1	R2	RL	PPL
HEUR. LAB.	+C	34.30	16.34	28.73	18.08
	-C	33.62	15.80	27.96	18.71
LEAD-200	+C	23.56	7.15	19.3	39.73
	-C	21.79	6.18	17.83	40.90
CONTSEL.	+C	24.67	8.55	20.44	33.22
	-C	22.71	7.33	18.71	32.77

Table 6: The ROUGE F1 and perplexity results for abstractive models with and without the context (+/-C) with heuristic labels, lead, and CONTENTSELECTOR extractive preprocessing steps.

Results

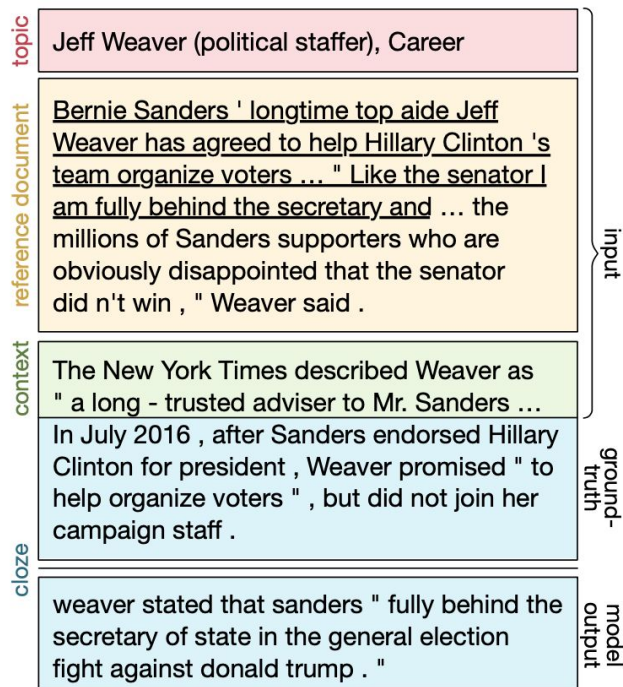


Figure 4: Example outputs from the abstractive model that uses the context. The model often copies sequences from the references which are sometimes correct (top) or incorrect but sensible (bottom), highlighting the difficulty of automatic evaluation. (Documents shortened for space. Sentences which are underlined were selected by the extraction step.)

Assessment

- Contribution:
 - Apply query-based summarization to specific sections of Wikipedia pages => a more self-contained version of Wikipedia-style topic summarization
 - A new large-scale dataset for future evaluation
- Limitation:
 - Extractive model has room for improvement
 - Topic does not seem to be used properly, despite topic attention in earlier layers

Discussion #1

- In what other ways / domains could survey generation be applied to?

Discussion #2

- How is evaluation of survey generation different from that of general text generation?

Discussion #3

- What are some limitations of current evaluation metrics for survey generation?

Discussion #4

- What new evaluation metrics would you propose?

Discussion #5

- What might be some ways to enhance abstractive methods?

Discussion #6

- What's your view on applying language model for survey generation?
(Benefit / potential problems)

Discussion #7

- What do you think are some unsolved challenges / future directions of survey generation?

References

- Rahul Jha, Catherine Finegan-Dollak, Ben King, Reed Coke, Dragomir Radev. Content Models for Survey Generation: A Factoid-Based Evaluation. In ACL 2015.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising Sequence-to-sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv preprint arXiv:1910.13461
- Xinyu Hua and Lu Wang. 2017. A pilot study of domain adaptation effect for neural abstractive summarization. In Proceedings of the Workshop on New Frontiers in Summarization, pages 100–106, Copenhagen, Denmark. Association for Computational Linguistics.
- Christina Sauper and Regina Barzilay. Automatically Generating Wikipedia Articles: A Structure-aware Approach. In ACL 2009
- Siddhartha Banerjee and Prasenjit Mitra. Wikiwrite: Generating Wikipedia Articles Automatically. In IJCAI 2016
- Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating Wikipedia by Summarizing Long Sequences. In ICLR 2018
- Daniel Deutsch and Dan Roth. Summary Cloze: A New Task for Content Selection in Topic-Focused Summarization. In EMNLP 2019
- Yang Liu and Mirella Lapata. Hierarchical Transformers for Multi-document Summarization. In ACL 2019
- Laura Perez-Beltrachini, Yang Liu, and Mirella Lapata. Generating Summaries with Topic Templates and Structured Convolutional Decoders. In ACL 2019
- Markus Zopf, Eneldo Loza Menćia, and Johannes Fűrnkranz. 2018. Which scores to predict in sentence regression for text summarization? In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1782–1791, New Orleans, Louisiana. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A Robustly Optimized Bert Pretraining Approach. arXiv preprint arXiv:1907.11692.
- Rahul Jha, Reed Coke, and Dragomir Radev. 2015. Surveyor: A System for Generating Coherent Survey Articles for Scientific Topics. In Twenty-Ninth AAAI conference on artificial intelligence.
- Yan Zhao, Mohammad Saleh, Peter J. Liu. SEAL: Segment-wise Extractive-Abstractive Long-form Text Summarization <https://arxiv.org/pdf/2006.10213.pdf>