Text summarization: fundamentals

Prof. Luca Cagliero
Dipartimento di Automatica e Informatica
Politecnico di Torino

Lecture goal

- The data deluge
- Data summarization
- Problem statement
- The Rouge score

The data deluge

- Online news
  - Newspapers and TV provide content on the web

LA STAMPA

medialab —— ARCHIVIO STORICO —— RICERCA

The Mercury News

The Newspaper of Silicon Valley

- User Generated Content (Web & Mobile)
  - E.g., Facebook, Instagram, Yelp, TripAdvisor, Twitter, YouTube

The data deluge

- E-learning platforms
  - Course materials & notes, homeworks & reviews

coursera

- Log files
  - Web server log files, machine syslog files

Data summarization

- Summary

  “using few words to give the most important information about something“

- Data summarization pinpoints relevant information from large data collections
  - large body of diverse techniques
  - appropriate for different data types

Textual data summarization

- News
  - Extract summaries of news articles ranging over the same topic and acquired from different sources

- E-commerce data
  - Summarize the comments of the customers on the purchased products/services

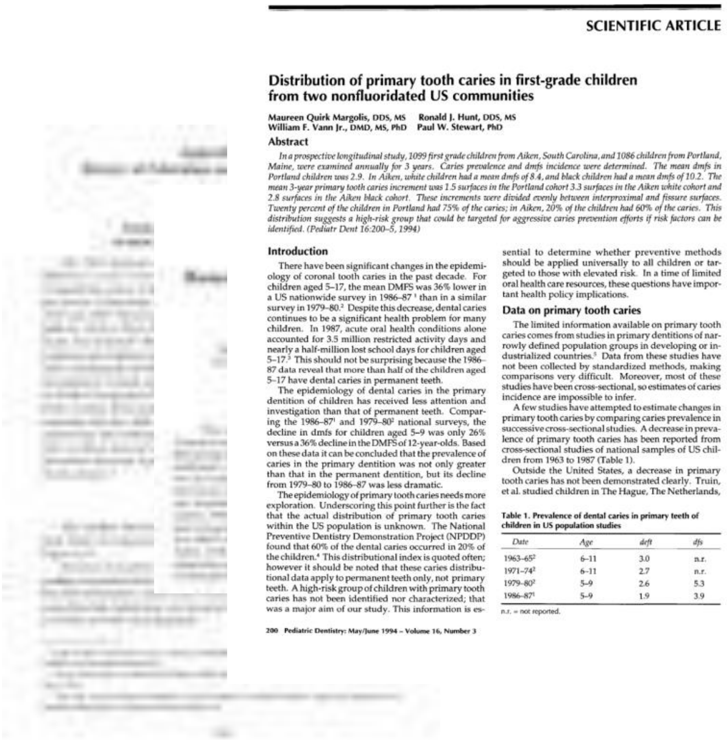
Textual data summarization

- E-learning platforms
  - Provide summaries of learning notes/teaching material to students in support of individual/collective learning activities

- Social networks
  - Summarize the opinions expressed in online communities through the analysis of blogs and/or posts
  - Topic, trend analysis

Why summarizing?

- Generate a big picture of the analyzed documents



SCIENTIFIC ARTICLE

Distribution of primary tooth caries in first-grade children from two nonfluoridated US communities

Maureen Quirk Margolis, DDS, MS Ronald J. Hunt, DDS, MS
William F. Vann Jr., DMD, MS, PhD Paul W. Stewart, PhD

Abstract

In a prospective longitudinal study, 1099 first grade children from Aiken, South Carolina, and 1086 children from Portland, Maine, were examined annually for 3 years. Caries prevalence and dmfs incidence were determined. The mean dmfs in Portland children was 2.9. In Aiken, white children had a mean dmfs of 8.4, and black children had a mean dmfs of 10.2. The mean 3-year primary tooth caries increment was 1.5 surfaces in the Portland cohort 3.3 surfaces in the Aiken white cohort and 2.8 surfaces in the Aiken black cohort. These increments were divided evenly between interproximal and fissure surfaces. Twenty percent of the children in Portland had 75% of the caries; in Aiken, 20% of the children had 60% of the caries. This distribution suggests a high-risk group that could be targeted for aggressive caries prevention efforts if risk factors can be identified. (Pediatr Dent 16:200–5, 1994)

Introduction

There have been significant changes in the epidemiology of coronal tooth caries in the past decade. For children aged 5–17, the mean DMFS was 36% lower in a US nationwide survey in 1986–87 [1] than in a similar survey in 1979–80.[2] Despite this decrease, dental caries continues to be a significant health problem for many children. In 1987, acute oral health conditions alone accounted for 3.5 million restricted activity days and nearly a half-million lost school days for children aged 5–17.[3] This should not be surprising because the 1986–87 data reveal that more than half of the children aged 5–17 have dental caries in permanent teeth.

The epidemiology of dental caries in the primary dentition of children has received less attention and investigation than that of permanent teeth. Comparing the 1986–87[1] and 1979–80[2] national surveys, the decline in dmfs for children aged 5–9 was only 26% versus a 36% decline in the DMFS of 12-year-olds. Based on these data it can be concluded that the prevalence of caries in the primary dentition was not only greater than that in the permanent dentition, but its decline from 1979–80 to 1986–87 was less dramatic.

The epidemiology of primary tooth caries needs more exploration. Underscoring this point further is the fact that the actual distribution of primary tooth caries within the US population is unknown. The National Preventive Dentistry Demonstration Project (NPDDP) found that 60% of the dental caries occurred in 20% of the children.[4] This distributional index is quoted often; however it should be noted that these caries distributional data apply to permanent teeth only, not primary teeth. A high-risk group of children with primary tooth caries has not been identified nor characterized; that was a major aim of our study. This information is essential to determine whether preventive methods should be applied universally to all children or targeted to those with elevated risk. In a time of limited oral health care resources, these questions have important health policy implications.

Data on primary tooth caries

The limited information available on primary tooth caries comes from studies in primary dentitions of narrowly defined population groups in developing or industrialized countries.[5] Data from these studies have not been collected by standardized methods, making comparisons very difficult. Moreover, most of these studies have been cross-sectional, so estimates of caries incidence are impossible to infer.

A few studies have attempted to estimate changes in primary tooth caries by comparing caries prevalence in successive cross-sectional studies. A decrease in prevalence of primary tooth caries has been reported from cross-sectional studies of national samples of US children from 1963 to 1987 (Table 1).

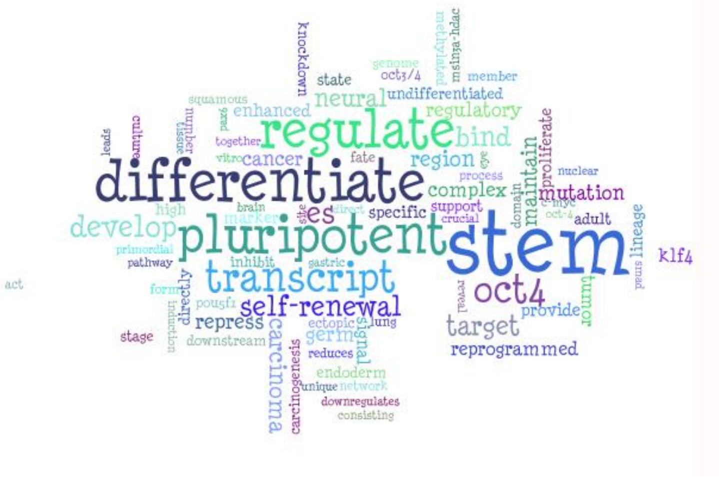
Outside the United States, a decrease in primary tooth caries has not been demonstrated clearly. Truin, et al. studied children in The Hague, The Netherlands,

Table 1. Prevalence of dental caries in primary teeth of children in US population studies

Date	*Age*	*deft*	*dfs*
1963–65[2]	6–11	3.0	n.r.
1971–74[2]	6–11	2.7	n.r.
1979–80[2]	5–9	2.6	5.3
1986–87[1]	5–9	1.9	3.9

n.r. = not reported.

200 Pediatric Dentistry: May/June 1994 – Volume 16, Number 3



Why summarizing?

- Quickly access key document information
  - disregarding redundant or marginally relevant information

- Guarantee content accessibility in case of limited bandwidth or reduced visualization quality

Text summarization

- Problem
  - Generation of a summary of a collection of textual documents
- Issues
  - Collection size
    - Single document vs multiple documents
  - Document characteristics and structure
    - E.g., semi-structured vs. unstructured
  - Language
  - Collection updates
  - Summary goal

Single-document summarization

- Problem
  - Generation a summary of a single document
- Examples
  - Abstracts of a scientific paper
  - Crib of a teaching document

Single-document summarization

- Advantages
    - Reduced computational complexity
    - Limited scalability issues
        - As long as the complexity of the analyzed document is fairly low
- Disadvantages
    - Need for exploration of the document structure
        - Homogeneous portions of text
    - Higher sensitivity to noise

Multi-document summarization

- Problem
  - Generation of a summary of a collection of documents
- Examples
  - Summary of the news articles published on different online newspapers and related to the same topic
  - Summary of all the online reviews of the same product

Multi-document summarization

- Advantages
  - Lower sensitivity to noise
  - Lower impact of the document structure

- Disadvantages
  - Higher computational complexity
  - Need for clustering homogenous documents
  - Presence of imbalances in the document distribution may bias the results
  - Scalability issues
    - Memory and time

Multi-lingual summarization

- Problem
  - Summarize collections of documents potentially written in different languages
  - Documents within the same collection are **all written in the same language**

- Challenges
  - Need for solutions portable to different languages
  - No need for combining models related to different languages

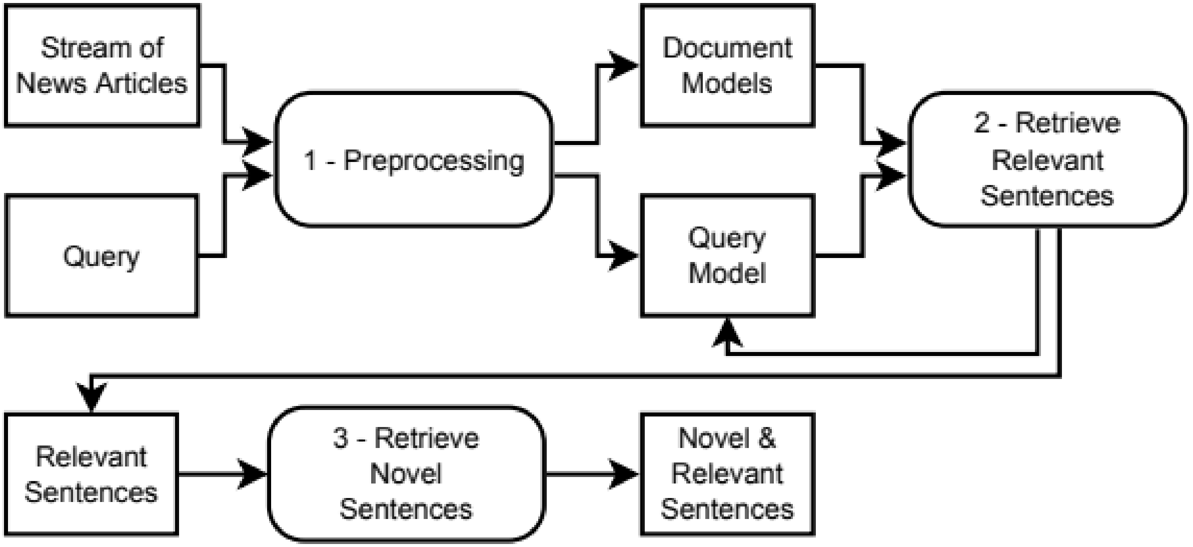
Cross-lingual summarization

- Problem
  - Summarize collections of documents potentially written in different languages
  - Documents within the same collection **can be written in different languages**

- Challenges
  - Need for solutions portable to different languages
  - Need for combining models related to different languages

Temporal summarization

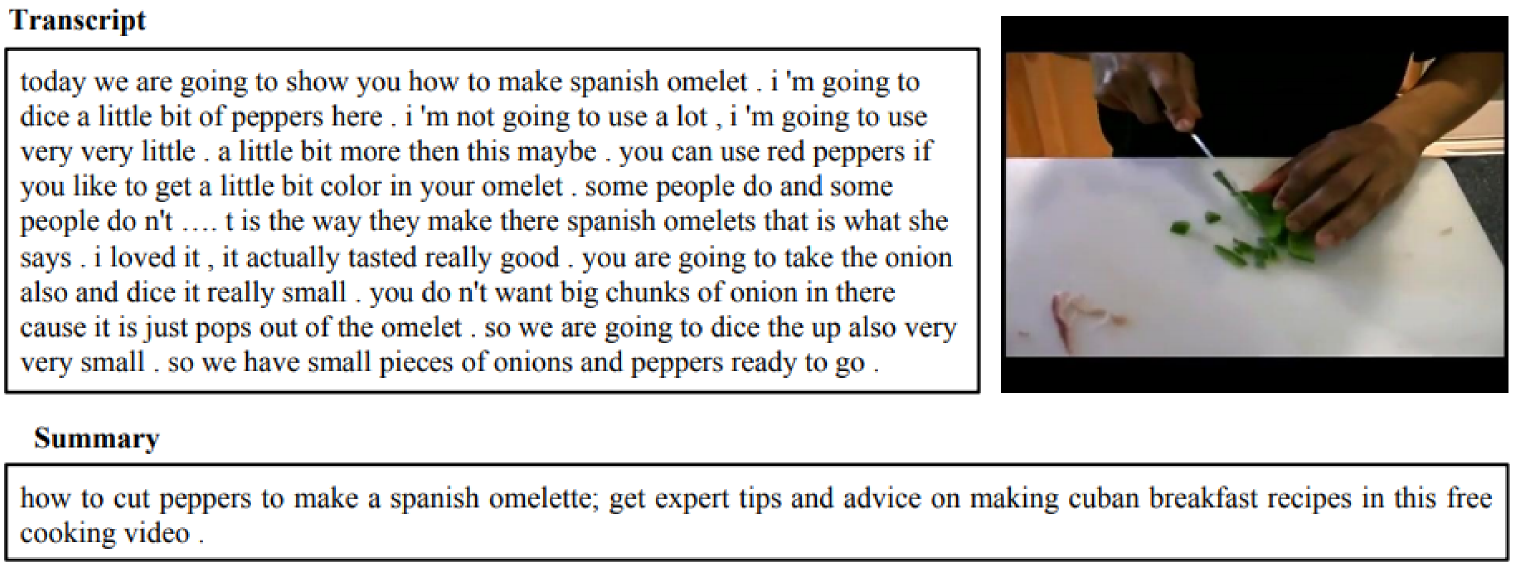
- Problem
  - Summarize a timestamped sequence of textual documents
    - E.g., a news stream
  - Identify salient concepts with minimal latency and redundancy



T. Schubotz and R. Krestel, "Online Temporal Summarization of News Events," 2015 IEEE/WIC/ACM WI-IAT, Singapore, 2015, pp. 409-412.doi: 10.1109/WI-IAT.2015.159

Multimodal summarization

- Problem statement
  - Generate textual summaries of data generated from multimodal sources
  - Adapt textual summaries to specific contexts, users, or data platform
    - E.g., Use metadata extracted from videos



Transcript

today we are going to show you how to make spanish omelet . i 'm going to dice a little bit of peppers here . i 'm not going to use a lot , i 'm going to use very very little . a little bit more then this maybe . you can use red peppers if you like to get a little bit color in your omelet . some people do and some people do n't .... t is the way they make there spanish omelets that is what she says . i loved it , it actually tasted really good . you are going to take the onion also and dice it really small . you do n't want big chunks of onion in there cause it is just pops out of the omelet . so we are going to dice the up also very very small . so we have small pieces of onions and peppers ready to go .

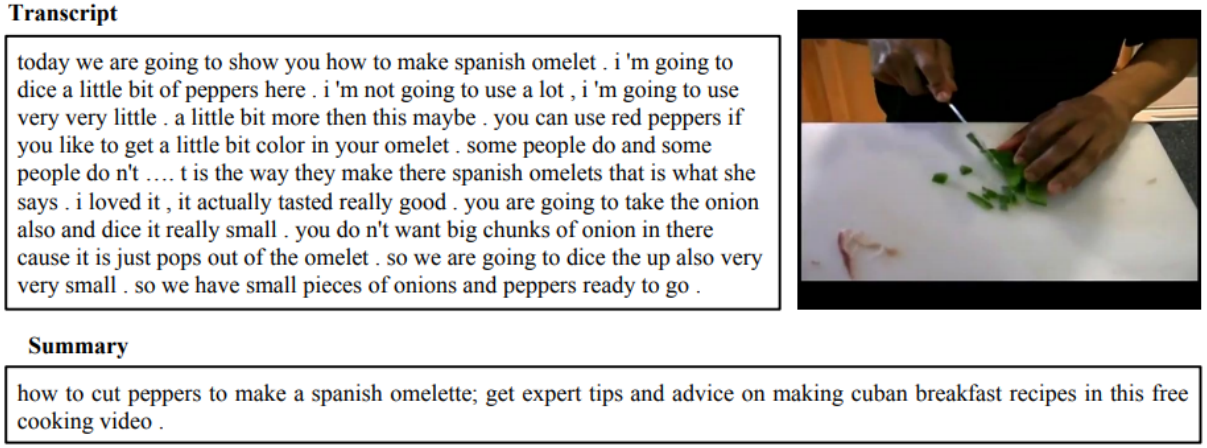
Summary

how to cut peppers to make a spanish omelette; get expert tips and advice on making cuban breakfast recipes in this free cooking video .

Multimodal Abstractive Summarization for How2 Videos.
Shruti Palaskar, Jindrinch Libovicky, Spandana Gella, Florian Metze. https://arxiv.org/abs/1906.07901

Multimodal summarization

- Multimodality can be either at the input or at the output layers
- When the modality changes from input to output: **Cross-modal summarization**



Transcript

today we are going to show you how to make spanish omelet . i 'm going to dice a little bit of peppers here . i 'm not going to use a lot , i 'm going to use very very little . a little bit more then this maybe . you can use red peppers if you like to get a little bit color in your omelet . some people do and some people do n't …. t is the way they make there spanish omelets that is what she says . i loved it , it actually tasted really good . you are going to take the onion also and dice it really small . you do n't want big chunks of onion in there cause it is just pops out of the omelet . so we are going to dice the up also very very small . so we have small pieces of onions and peppers ready to go .

Summary

how to cut peppers to make a spanish omelette; get expert tips and advice on making cuban breakfast recipes in this free cooking video .

Multimodal Abstractive Summarization for How2 Videos.
Shruti Palaskar, Jindrinch Libovicky, Spandana Gella, Florian Metze. https://arxiv.org/abs/1906.07901

Generic vs. query-relevant summarization

- Generic document summarization
    - The summary should reflect the major content of the documents without any additional information
    - Model generation is solely based on the analyzed documents

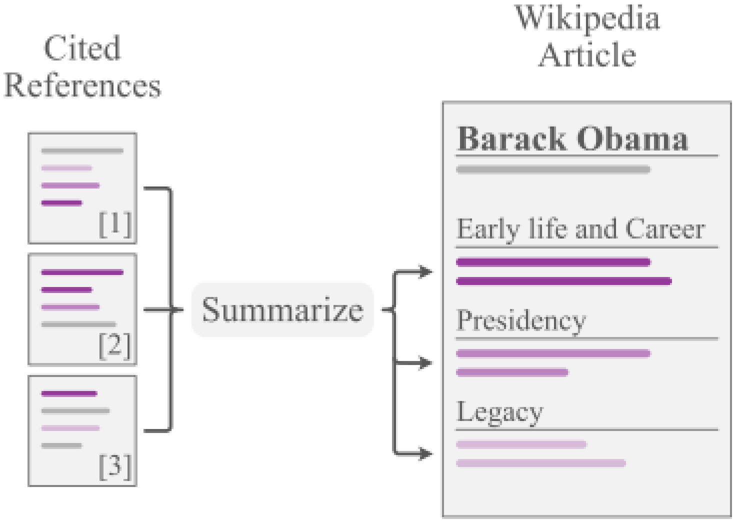
- Query-relevant document summarization
    - The summary focuses on the information expressed in the given queries
        - the summaries must be biased to the given queries
    - Model generation should take query content into account

Query-relevant summarization vs. IR

- Summaries have further constraints beyond pertinence
  - E.g., Minimal redundancy, minimum latency, etc.
- Summaries can be abstractive
- The ordering of the summary content is not always relevant
  - E.g., Generate a summary consisting of 665 bytes

Aspect-based summarization

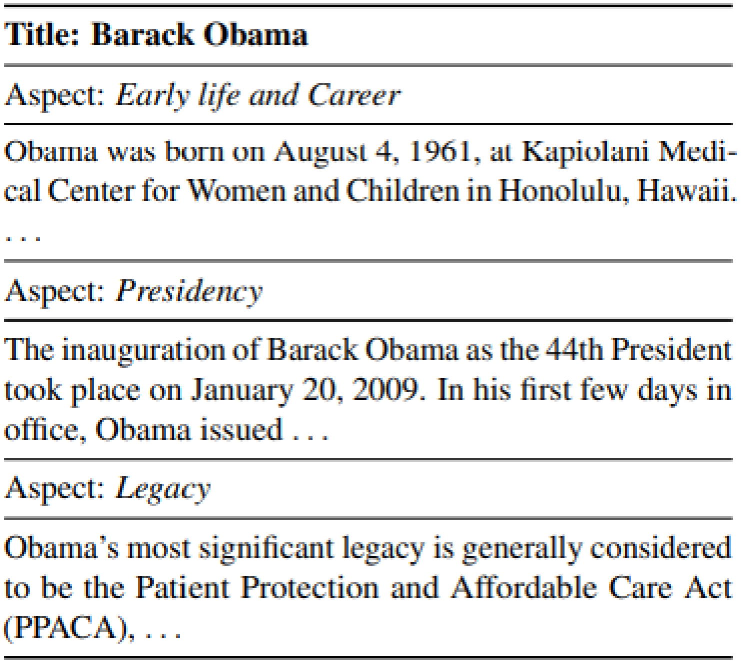
- Generic document summarization
  - Allow readers to fulfill focused information needs more easily and quickly
- Can be either domain-specific or open-domain
  - Whether they are applicable to various domains or not



WikiAsp: A Dataset for Multi-domain Aspect-based Summarization. Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, Graham Neubig. ACL 2020

Aspect-based summarization

- Aspects can be derived from either section titles or further document annotations
  - E.g., Wikipedia sections



Title: Barack Obama
Aspect: *Early life and Career*
Obama was born on August 4, 1961, at Kapiolani Medical Center for Women and Children in Honolulu, Hawaii. . . .
Aspect: *Presidency*
The inauguration of Barack Obama as the 44th President took place on January 20, 2009. In his first few days in office, Obama issued . . .
Aspect: *Legacy*
Obama's most significant legacy is generally considered to be the Patient Protection and Affordable Care Act (PPACA), . . .

WikiAsp: A Dataset for Multi-domain Aspect-based Summarization. Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, Graham Neubig. ACL 2020

Example of aspect-based summarization

- Generate Wikipedia pages automatically by using cited references as document source for summarization
  - E.g., books, websites, etc.
- Let $\{R_1, R_2, \ldots, R_M\}$ be a collection of M cited references for an article $S = \{s_1, s_2, \ldots, s_N\}$ of N sections
- Problem: learn a model $f : R \rightarrow S$
  1. identify and gather information from cited references
  2. generate a section-by-section summary, where each section contains the appropriate type of information: section title and one or more paragraphs

WikiAsp: A Dataset for Multi-domain Aspect-based Summarization. Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, Graham Neubig. ACL 2020

Incremental summarization

- Problem
  - Generate a summary from the input document collection
  - As soon as the collection changes the summary content should be updated
    - Incremental models are needed to avoid re-calculating the model from scratch
- Challenges
  - Most data mining techniques are non-incremental
  - The relevance of the old document content should be updated according to the new content

Evaluation methods

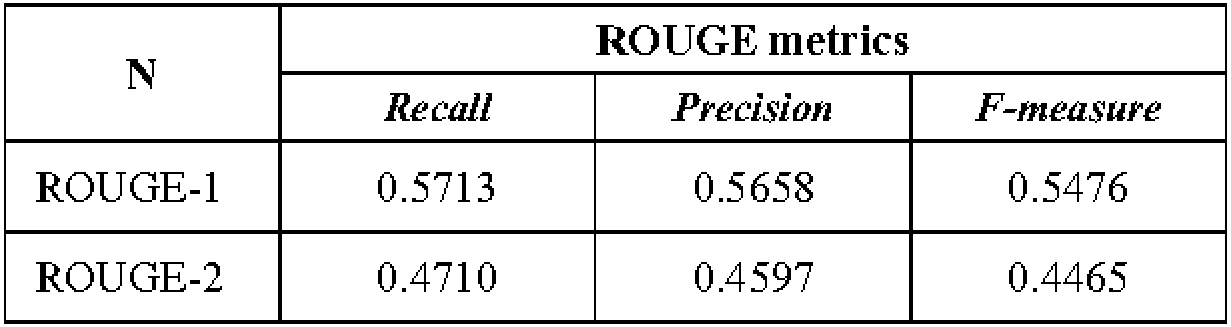
- Evaluate the quality of the output summaries is crucial for assessing the effectiveness of automated summarization systems
- Quantitative evaluation
  - Similarity with the ground truth
- Qualitative evaluation
  - Readability, pertinence, smoothness, etc.

Evaluation methods

- Intrinsic evaluators
  - Assess the quality of a summary against a reference summary or based on the feedbacks provided by a domain expert who manually evaluated the output summaries
- Extrinsic evaluators
  - Measure the effectiveness of a summarization method in accomplishing a given task

The Rouge score

- Recall-Oriented Understudy for Gisting Evaluation
- Intrinsic evaluation metric
  - Based on syntax
- It allows analysts to compare an automatically generated summary with a set of reference summaries
  - Typically human-generated
- It measures the unit overlaps between the text of the reference summary, i.e., the ground truth, and text contained in the automatically generated summary

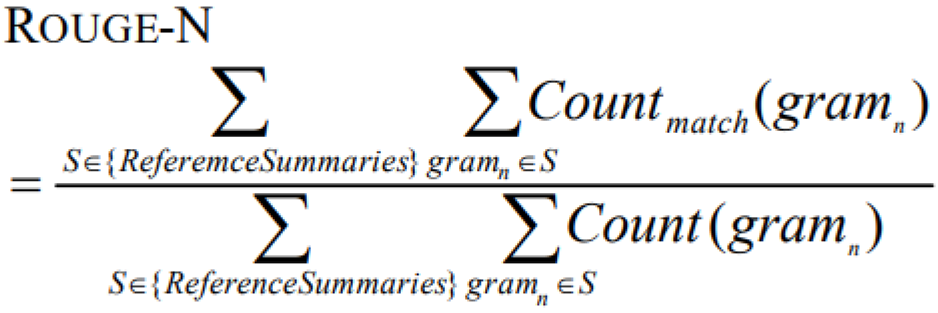


N	**ROUGE metrics**		
	Recall	***Precision***	***F-measure***
ROUGE-1	0.5713	0.5658	0.5476
ROUGE-2	0.4710	0.4597	0.4465

Chin-Yew Lin, Workshop on Text Summarization Branches Out, Barcelona, Spain, July 25 - 26, 2004

Rouge-N

- N-gram co-occurrences between reference and candidate summaries
- High order ROUGE-N with n-gram length greater than 1 estimates the fluency of summaries
- Rouge-N recall:

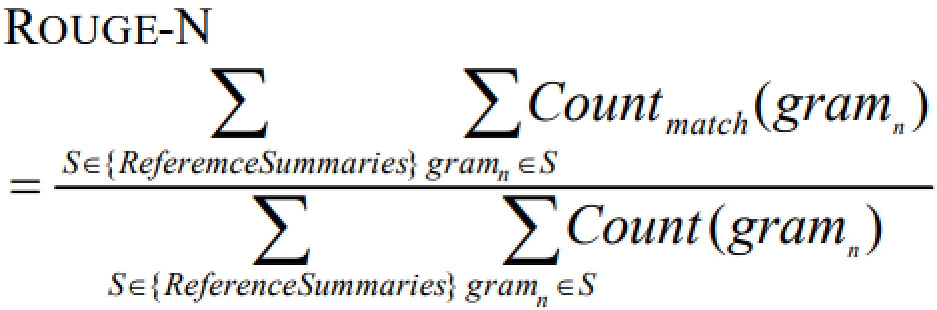


$$\text{ROUGE-N} = \frac{\sum\limits_{S \in \{ReferemceSummaries\}} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S \in \{ReferenceSummaries\}} \sum\limits_{gram_n \in S} Count(gram_n)}$$

Chin-Yew Lin, Workshop on Text Summarization Branches Out, Barcelona, Spain, July 25 - 26, 2004

Rouge-N measures

- Standard Precision, Recall, and F1 score measures
- Rouge-N recall:



ROUGE-N

$$= \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

- Considering Rouge-N Recall is advisable while evaluating fixed-size summaries
  - Recommendation: use Rouge-N F1 score otherwise!

Chin-Yew Lin, Workshop on Text Summarization Branches Out, Barcelona, Spain, July 25 - 26, 2004

Rouge-N

- Rouge evaluation ignores semantic similarity
  - Example
    - reference text: *police killed the gunman*
    - Summary 1: police kill the gunman
    - Summary 2: the gunman kill the police
  - The two summaries above have the same N-gram co-occurrence!

Chin-Yew Lin, Workshop on Text Summarization Branches Out, Barcelona, Spain, July 25 - 26, 2004

Rouge-L

- Longest Common Subsequence
  - Given two sequences X and Y, the longest common subsequence of X and Y is a common subsequence with maximum length
    - Intuition: the longer the match, the more similar the summaries
- Used for comparing text translations as well
- Example
  - reference text: *police killed the gunman*
  - Summary 1: police kill the gunman -> LCS overlap 3/4
  - Summary 2: the gunman kill the police -> LCS overlap 2/4
  - Summary 1 is better than Summary 2 (3/4 > 2/4)
- The two summaries above have the same N-gram co-occurrence!

Chin-Yew Lin, Workshop on Text Summarization Branches Out, Barcelona, Spain, July 25 - 26, 2004

Rouge-W

- Weighted Longest Common Subsequence
- Favors strings with consecutive matchings
- Can be computed efficiently using dynamic programming
- Example
  - Reference: [A B C D E F G]
  - S1: [A B C D H I K]
  - S2: [A H B K C I D]
  - Rouge-L(S1) = Rouge-L(S2)
  - Rouge-W(S1) > Rouge-W(S2)

Chin-Yew Lin, Workshop on Text Summarization Branches Out, Barcelona, Spain, July 25 - 26, 2004

Rouge-S

- Skip-Bigram
- Any pair of words in their sentence order, allowing for arbitrary gaps
- Intuition
  - Consider long distance dependency
  - Allow gaps in matches as in LCS but count all in-sequence pairs rather than longest subsequences only

Chin-Yew Lin, Workshop on Text Summarization Branches Out, Barcelona, Spain, July 25 - 26, 2004

Rouge-S example

- Example
  - reference text: *police killed the gunman*
  - Summary 1: police kill the gunman
  - Summary 2: the gunman kill the police
  - Summary 3: the gunman police killed
- Results
  - Rouge-N: Summary 4 > Summary 2 = Summary 3
  - Rouge-L: Summary 2 > Summary 3 = Summary 4
  - Rouge-S: Summary 2 > Summary 4 > Summary 3
    - Summary 2: 3/6 (“police the”, “police gunman”, “the gunman”)
    - Summary 3: 1/6 (“the gunman”)
    - Summary 4: 2/6 (“the gunman”, “police killed”)

Chin-Yew Lin, Workshop on Text Summarization Branches Out, Barcelona, Spain, July 25 - 26, 2004

Rouge-SU

- Skip-Bigram plus unigram-based co-occurrence statistics
- Similar to Rouge-S
- It counts **both** unigrams and bigrams

Chin-Yew Lin, Workshop on Text Summarization Branches Out, Barcelona, Spain, July 25 - 26, 2004

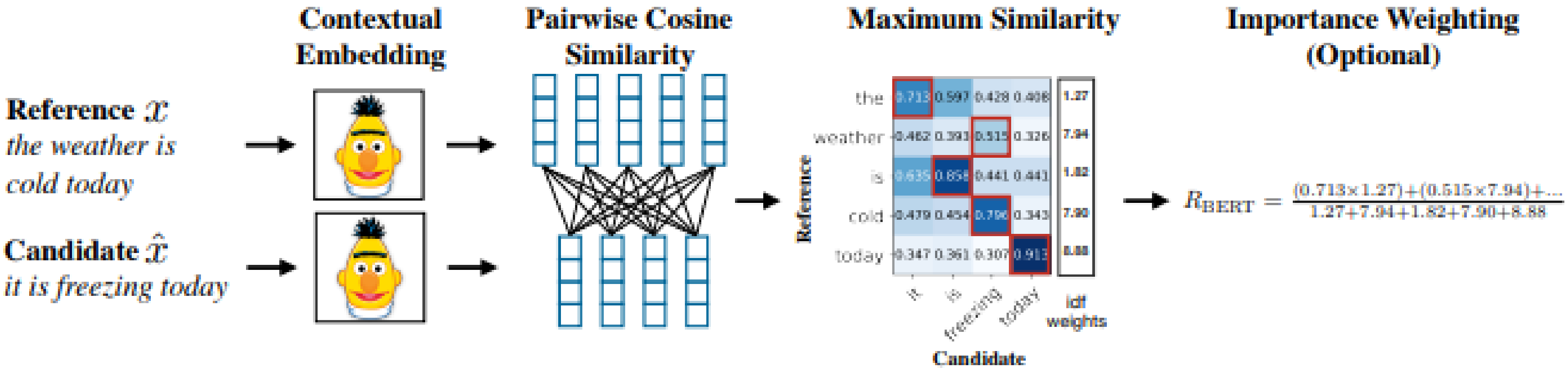
JRouge

- Java-based Rouge extension designed for non-European languages
- Improvements
  - Uses unicode regular expressions to match characters and numerals of any Unicode text
  - Uses the Stanford NLP toolkit to split the sentences
  - Provides consistent results in different Operating Systems
  - Calculates scores more precisely than the original ROUGE since it does not truncate and round the results to the fifth digit after the decimal point. Instead, JRouge uses the whole extent of the Double primitive type of Java to provide precise scores

Chin-Yew Lin, Workshop on Text Summarization Branches Out, Barcelona, Spain, July 25 - 26, 2004

BERT score

- Automatic evaluation metric for text generation
- Correlates human judgments using BERT
- Suitable for both abstractive and extractive summarization



Reference \ Candidate	it	is	freezing	today	idf weights
the	0.713	0.597	0.428	0.408	1.27
weather	0.462	0.393	0.515	0.326	7.94
is	0.635	0.858	0.441	0.441	1.82
cold	0.479	0.454	0.796	0.343	7.90
today	0.347	0.361	0.307	0.913	8.88

$$R_{\text{BERT}} = \frac{(0.713\times1.27)+(0.515\times7.94)+\ldots}{1.27+7.94+1.82+7.90+8.88}$$

BERTScore: Evaluating Text Generation with BERT. Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi. ICLR 2019.

BERT score

- Precision, Recall, and F1 Score metrics

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \ , \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \ , \quad F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}$$

where the similarity between the reference and candidate tokens is computed using the cosine similarity $\frac{\mathbf{x}_i^\top \hat{\mathbf{x}}_j}{\|\mathbf{x}_i\|\|\hat{\mathbf{x}}_j\|}$

BERTScore: Evaluating Text Generation with BERT. Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi. ICLR 2019.

Mean Reciprocal Rank

- Extrinsic evaluation metric, established in Information Retrieval
- Input:
  - a query, a ranked list of retrieved object, and the expected ranked list of objects
- Output
  - inverse of the rank of each of the retrieved objects

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{Q} \frac{Relevance\_Label\_Value}{\mathrm{rank}_i}$$

Mean Reciprocal Rank

- Can be tailored to extractive text summarization
- Example
  - Object: sentence.
  - Summary: ranked list of sentences.
  - Golden summary: expected ranked list of sentences.

Benchmark datasets (small)

- **Document Understanding Conference 2001**
  - single- and multi-document summarization task
  - 60 collections of documents which have been partitioned into training and test sets
- **Document Understanding Conference 2004**
  - Multi-document summarization
  - variant of the DUC 2001 dataset
- **Document Understanding Conference 2007**
  - Variant of the DUC’04 collection
  - update summarization task
- **MultiLing Pilot 2011**
  - Multi-lingual summarization
  - A set of WikiNews documents translated in Arabic, Czech, English, French, Greek, Hebrew and Hindi by native speakers
  - Each article belongs to one of the 10 topics under consideration and for each topic at least 3 reference summaries are provided
- **Text REtrieval Conference 2013**
  - Temporal summarization task
  - It consists of a collection of news articles, which have been chronologically ordered to simulate a stream of news articles ranging over a set of events
  - Tasks: filtering and summarization or summarization only

Benchmark datasets (large)

- **CNN/DailyMail dataset**
  - Multi-document news summarization
- **Big Patent dataset**
  - Single-document patent summarization
- **TLS-Covid19**
  - Sars-Cov-2 news timeline summarization
- **ScisummNet**
  - Scientific paper summarization based on cited text spans analysis

Additional reading on ROUGE

- ROUGE: A Package for Automatic Evaluation of Summaries. Ming-Wei Chang Kenton Lee Kristina Toutanova. Proceedings of the ACL 2004.
- Download and read the paper: https://aclanthology.org/W04-1013.pdf

Acknowledgements and copyright license

- Copyright licence
  - Attribution + Noncommercial + NoDerivatives
- Acknowledgements
  - I would like to thank Dr. Moreno La Quatra, who collaborated to the writing and revision of the teaching content
- Affiliation
  - The author and his staff are currently members of the Database and Data Mining Group at Dipartimento di Automatica e Informatica (Politecnico di Torino) and of the SmartData interdepartmental centre
    - https://dbdmg.polito.it
    - https://smartdata.polito.it

Thank you!