AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 771: Computer Processing of Natural Language
Lecture 22: Summary Generation

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

12TH NOVEMBER, 2024

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 23

- Opening Segment
  - Announcements
- Main Lecture

Main Section
- Summary generation
- Methods
  - Extractive - traceable to original content
  - Abstractive – non traceable to original content
  - Compressive – remove content but not information
- Applications

- Concluding Segment
  - About Next Lecture – Lecture 24

Recap of Lecture 22

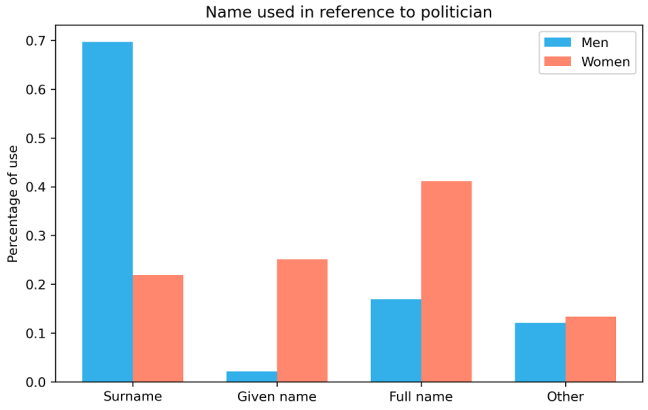
- Sentiment Analysis
- Methods
  - Lexicon-based Methods
  - Learning-based Methods
- Usability Considerations - Ethical Issues
- Business Application Case Study

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0274317

"*Discussion board comments were analyzed using statistical algorithms.*

The researchers analyzed a total of 10 million English-language comments in conversations about male and female politicians from 312 different countries.

About half of Reddit users are from the United States, *followed by UK and Canadian users with roughly 7.5% each. 90% of the politicians mentioned in the study are from the United States.*"



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0274317

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Quantifying gender biases towards politicians on Reddit

Sara Marjanovic, Karolina Stańczak, Isabelle Augenstein

Published: October 26, 2022 • https://doi.org/10.1371/journal.pone.0274317

Article | Authors | Metrics | Comments | Media Coverage | Peer Review

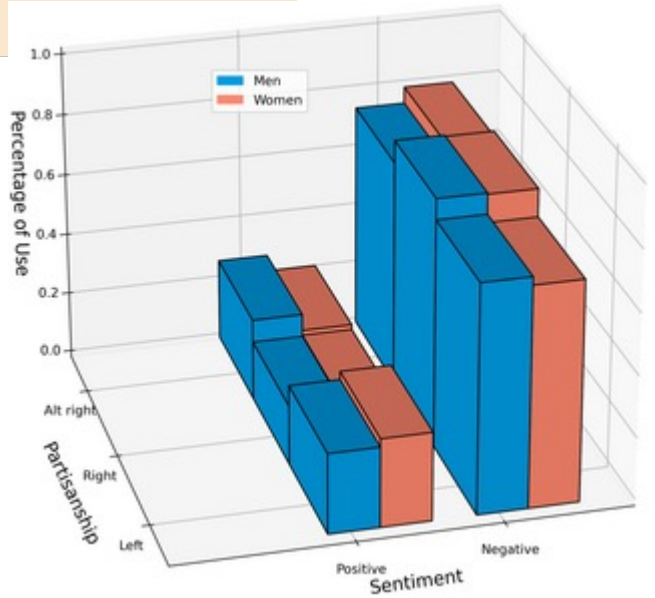
Abstract

Despite attempts to increase gender parity in politics, global efforts have struggled to ensure equal female representation. This is likely tied to implicit gender biases against women in authority. In this work, we present a comprehensive study of gender biases that appear in online political discussion. To this end, we collect 10 million comments on Reddit in conversations *about* male and female politicians, which enables an exhaustive study of automatic gender bias detection. We address not only misogynistic language, but also other manifestations of bias, like benevolent sexism in the form of seemingly positive sentiment and dominance attributed to female politicians, or differences in descriptor attribution. Finally, we conduct a multi-faceted study of gender bias towards politicians investigating both linguistic and extra-linguistic cues. We assess 5 different types of gender bias, evaluating coverage, combinatorial, nominal, sentimental and lexical biases extant in social media language and discourse. Overall, we find that, contrary to previous research, coverage and sentiment biases suggest equal public interest in female politicians. Rather than overt hostile or benevolent sexism, the results of the nominal and lexical analyses suggest this interest is not as professional or respectful as that expressed about male politicians. Female politicians are often named by their first names and are described in relation to their body, clothing, or family; this is a treatment that is not similarly extended to men. On the now banned far-right subreddits, this disparity is greatest, though differences in gender biases still appear in the right and left-leaning subreddits. We release the curated dataset to the public for future studies.

Analysis of Bias

Sentimental Biases

When people discuss male and female politicians, do they express equal sentiment and power levels in the words chosen?



https://journals.plos.org/plosone/article?id=10.1371/journal.pone.027

"Though gender differences are significant in both left-leaning and alt-right subreddits, the assosciation strength is negligible, like the differences measured via the sentiment lexicon."

•**Coverage biases –** the amount of interest devoted.

•**Combinatorial biases** – are female politicians given a different role than male ones – are they mentioned, for example, in the context of other women, or as the one woman among many men?

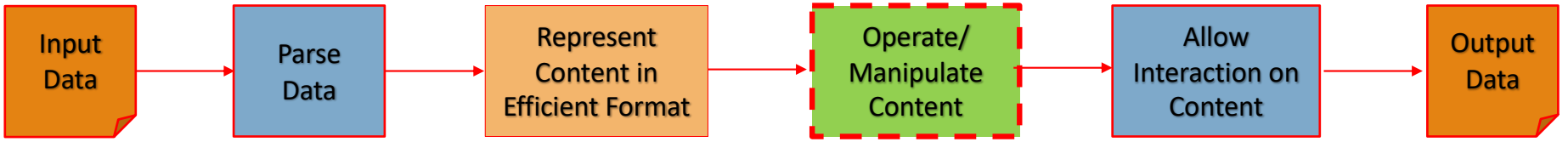
•**Nominal biases** – how a politician is addressed (madam, minister, Mette, Hillary etc.) and what types of words are used to describe them?

•**Sentimental biases** – are equal sentiments and levels of power expressed in word choice regarding male and female politicians? – e.g., the quantity of positive words vs. negative as well as neutral words (such as 'goddess/dumb/republican).

•**Lexical biases** – what categories of words do people use to describe male and female politicians? For example, categories such as family, body and dress or profession.

Main Lecture

Text Summarization



Motivation

- We have a variety of data around
  - Volume: Single document, Multiple documents
  - Variety: Text, numbers/ tables, figures, …
  - Languages
- Can one get a quick insight about the data ? Meaning ?
  - Choice 1: **Important** ideas in the input
  - Choice 2: **Representative** ideas in the input
  - Choice 3: **Example** ideas in the input
    - Random selection is sufficient?

Example Text

1. https://www.nytimes.com/article/2020-nobel-prize-winners.html
2. 2020 Nobel Prize Winners: Full List
3. Nobel Prize season begins every October as committees in Sweden and Norway name laureates in a variety of prizes in the sciences, literature and economics, as well as peace work. The announcements started last week with the awarding of the prize in Physiology or Medicine. They wrapped up on Monday, when the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel was announced.
4. The Nobel Prizes most years are presented to recipients in Stockholm and Oslo in December. Because of the coronavirus pandemic, the committees are changing their approaches. Some of the events in Stockholm will be canceled in favor of a digital ceremony for the Nobelists, and medals and diplomas are to be distributed to the recipients? embassies and handed over in their home countries. Recipients may be invited to the award ceremony for 2021, if possible.
5. The Oslo ceremony for the peace prize will be smaller than in most years, with a limited audience.
6. The Nobel committee also announced another change last month: Each prize will rise to 10 million Swedish krona, 1 million more than in the previous year. That?s a hike in the prize value of about $112,000 in current exchange rates.
7. The 2020 Winners
8. Physiology or Medicine
9. Drs. Harvey J. Alter, Michael Houghton and Charles M. Rice on Monday received the prize for their discovery of the hepatitis C virus. The Nobel committee said the three scientists had ?made possible blood tests and new medicines that have saved millions of lives.?
10. Physics
11. Roger Penrose, Reinhard Genzel and Andrea Ghez are the recipients of the Nobel Prize in Physics for 2020.Credit...Pool photo by Fredrik Sandberg
12. Roger Penrose, Reinhard Genzel and Andrea Ghez received the prize on Tuesday for their discoveries that have improved understanding of the universe, including work on black holes.
13. Chemistry
14. The Nobel Prize in Chemistry was jointly awarded on Wednesday to Emmanuelle Charpentier and Jennifer A. Doudna for their work on the development of Crispr-Cas9, a method for genome editing.
15. Literature
16. The Nobel Prize in Literature was awarded on Thursday to Louise Glack, one of America?s most celebrated poets, ?for her unmistakable poetic voice that with austere beauty makes individual existence universal.?
17. Peace Prize
18. The Nobel Peace Prize was awarded on Friday to the World Food Program for its efforts to combat a surge in global hunger amid the coronavirus pandemic, which has swept around the world with devastating impact.
19. Economic Science
20. Paul R. Milgrom and Robert B. Wilson were awarded the Nobel in economic science on Monday for improvements to auction theory and inventions of new auction formats.

A Few Ways to Generate Insights

- Word tag cloud (textual data; based on frequency)
  - Important
- Topic analysis (textual data, based on learned patterns)
  - Important
- **Summary generation** – formal study
  - Representative
  - Important
  - Example

Example Text

Summary Generated

1. 2020 Nobel Prize Winners: Full List Nobel Prize season begins every October as committees in Sweden and Norway name laureates in a variety of prizes in the sciences, literature and economics, as well as peace work. ?
2. Peace Prize The Nobel Peace Prize was awarded on Friday to the World Food Program for its efforts to combat a surge in global hunger amid the coronavirus pandemic, which has swept around the world with devastating impact.
3. Literature The Nobel Prize in Literature was awarded on Thursday to Louise Glack, one of America?s most celebrated poets, ?for her unmistakable poetic voice that with austere beauty makes individual existence universal.

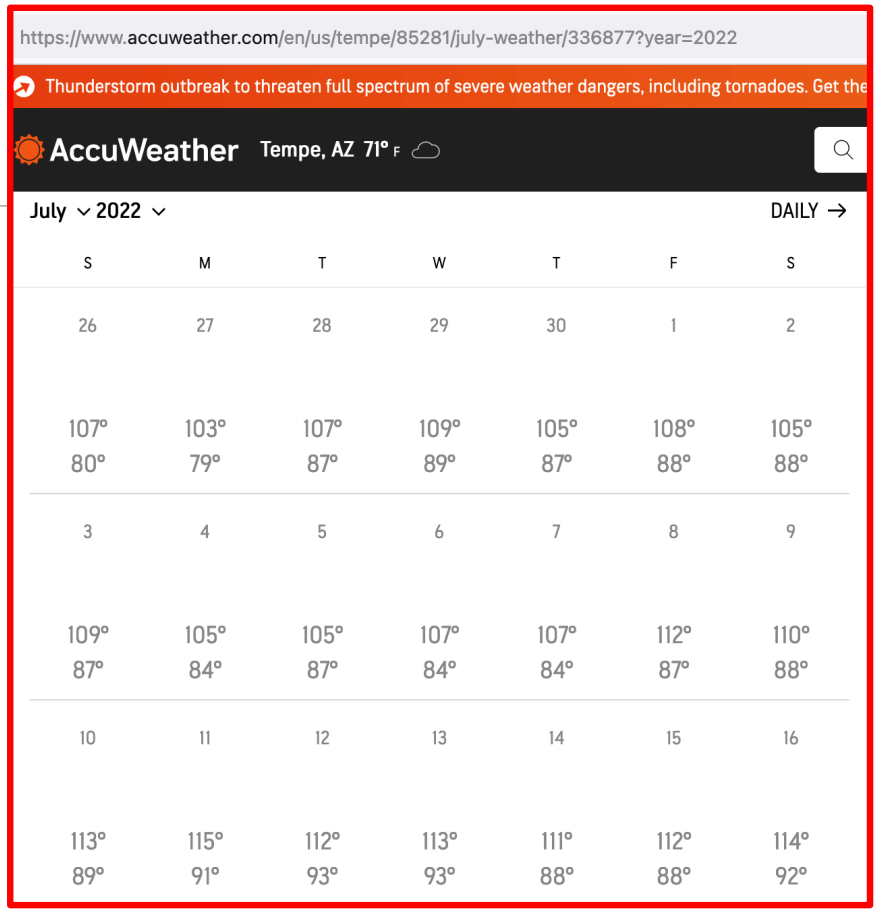
1. https://www.nytimes.com/article/2020-nobel-prize-winners.html
2. 2020 Nobel Prize Winners: Full List
3. Nobel Prize season begins every October as committees in Sweden and Norway name laureates in a variety of prizes in the sciences, literature and economics, as well as peace work. The announcements started last week with the awarding of the prize in Physiology or Medicine. They wrapped up on Monday, when the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel was announced.
4. The Nobel Prizes most years are presented to recipients in Stockholm and Oslo in December. Because of the coronavirus pandemic, the committees are changing their approaches. Some of the events in Stockholm will be canceled in favor of a digital ceremony for the Nobelists, and medals and diplomas are to be distributed to the recipients? embassies and handed over in their home countries. Recipients may be invited to the award ceremony for 2021, if possible.
5. The Oslo ceremony for the peace prize will be smaller than in most years, with a limited audience.
6. The Nobel committee also announced another change last month: Each prize will rise to 10 million Swedish krona, 1 million more than in the previous year. That?s a hike in the prize value of about $112,000 in current exchange rates.
7. The 2020 Winners
8. Physiology or Medicine
9. Drs. Harvey J. Alter, Michael Houghton and Charles M. Rice on Monday received the prize for their discovery of the hepatitis C virus. The Nobel committee said the three scientists had ?made possible blood tests and new medicines that have saved millions of lives.?
10. Physics
11. Roger Penrose, Reinhard Genzel and Andrea Ghez are the recipients of the Nobel Prize in Physics for 2020.Credit...Pool photo by Fredrik Sandberg
12. Roger Penrose, Reinhard Genzel and Andrea Ghez received the prize on Tuesday for their discoveries that have improved understanding of the universe, including work on black holes.
13. Chemistry
14. The Nobel Prize in Chemistry was jointly awarded on Wednesday to Emmanuelle Charpentier and Jennifer A. Doudna for their work on the development of Crispr-Cas9, a method for genome editing.
15. Literature
16. The Nobel Prize in Literature was awarded on Thursday to Louise Glack, one of America?s most celebrated poets, ?for her unmistakable poetic voice that with austere beauty makes individual existence universal.?
17. Peace Prize
18. The Nobel Peace Prize was awarded on Friday to the World Food Program for its efforts to combat a surge in global hunger amid the coronavirus pandemic, which has swept around the world with devastating impact.
19. Economic Science
20. Paul R. Milgrom and Robert B. Wilson were awarded the Nobel in economic science on Monday for improvements to auction theory and inventions of new auction formats.

Summarizing Weather of a Place

Week 1 [103-107; 79-88]

Week 2 [105-112; 84-88]

Week 3 [111-115; 88-92]



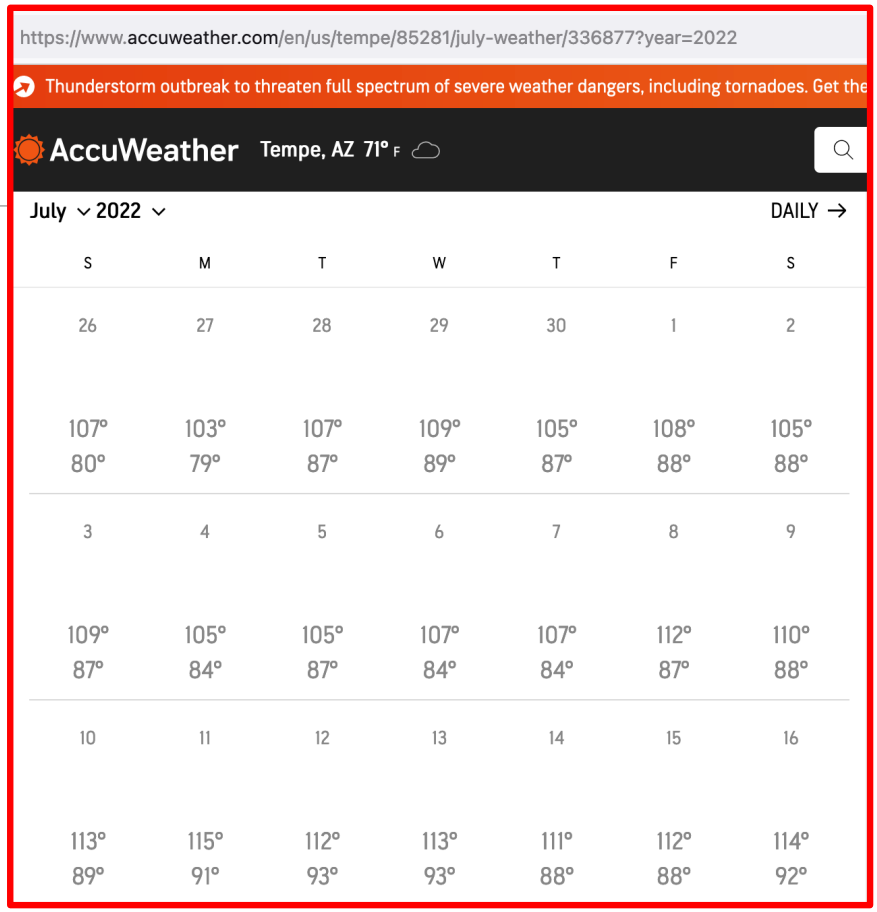
https://www.accuweather.com/en/us/tempe/85281/july-weather/336877?year=2022

Summarizing Weather – Appropriate?

Week 1 [103-107; 79-88]:
The minimum temp varied more than max

Week 2 [105-112; 84-88]:
The max temp varied more than temp

Week 3 [111-115; 88-93]:
Both temps varied about the same



https://www.accuweather.com/en/us/tempe/85281/july-weather/336877?year=2022

Summarization Type and Methods

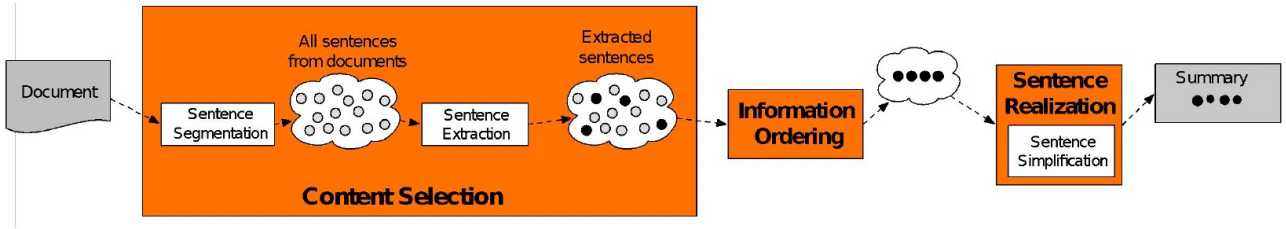
- **Extractive**: extract summary **with** content in the input
  - Should be able to trace output back to the input
  - Interpretability is easy to show

- **Abstractive**: create summary **from ideas** in the content in the input
  - but using at least some different words
  - Useful in creative task – digital journalism

- **Compressive**: remove redundant from input
  - shorter sentence by removing redundant information
  - Preserve grammar and important content in the input

Generic Summarization Approach

1. content selection: choose sentences to extract from the document
2. information ordering: choose an order to place them in the summary
3. sentence realization: clean up the sentences



Credit: Chan Young Park's slides

Evaluation of Text Summarization

- **Intrinsic** – How good is the output ?
- Considerations
  - Quality of output - readability
  - Coherence of information – out of context sentences ?
  - Compression of information

Comparison against a reference summary

- Is this the only summary possible?
- A scale of what summary should contain
- Constraints on summary characteristics. E.g., number of words

Where is the refence summary?

- Experts do not agree on summaries
  - Variance between experts
  - Not even their when separated 8-weeks apart nearly 50% times
- But experts often agree on important sentences that make up into a summary

Summarization Evaluation: An Overview, Inderjeet Mani, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/sum-mani.pdf 2019

Evaluation of Text Summarization

- **Intrinsic** – How good is the output ?
- Considerations
  - Quality of output - readability
  - Coherence of information – out of context sentences ?
  - Compression of information

Automatic Evaluation

- Against a set of criteria
- Criteria can be task independent or task-dependent
- Can also be about sentences to include and their rank
- Can also be about topics in content and in summary

Metrics

- ROUGE (Recall Oriented Understudy for Gisting Evaluation) score
- METEOR (Metric for Evaluation of Translation with Explicit Ordering) score
- BLEU (bilingual evaluation understudy) score
- SUMMAC summary automatic scoring

Summarization Evaluation: An Overview, Inderjeet Mani, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/sum-mani.pdf 2019

Evaluation of Text Summarization

- **Extrinsic** – What is the impact of the output ?
- Overall idea
  - Describe a task that a person wants to do with a summary
  - Given the generated summary, how well can a person do that job
- Comments
  - Extrinsic useful for business sponsor but not directly to technical developer

Summarization Evaluation: An Overview, Inderjeet Mani, http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings2/sum-mani.pdf 2019

Coding Illustration - Extractive

https://github.com/biplav-s/course-nl/tree/master/l20-textsumm

Coding Illustration – LLM Based

https://github.com/biplav-s/course-nl-f22/blob/main/sample-code/l21-24-llm-tasks/Summarization-withTransformers.ipynb

Evaluation in Example Text

Summary Generated

1. 2020 Nobel Prize Winners: Full List Nobel Prize season begins every October as committees in Sweden and Norway name laureates in a variety of prizes in the sciences, literature and economics, as well as peace work. ?
2. Peace Prize The Nobel Peace Prize was awarded on Friday to the World Food Program for its efforts to combat a surge in global hunger amid the coronavirus pandemic, which has swept around the world with devastating impact.
3. Literature The Nobel Prize in Literature was awarded on Thursday to Louise Glack, one of America?s most celebrated poets, ?for her unmistakable poetic voice that with austere beauty makes individual existence universal.

1. https://www.nytimes.com/article/2020-nobel-prize-winners.html
2. 2020 Nobel Prize Winners: Full List
3. Nobel Prize season begins every October as committees in Sweden and Norway name laureates in a variety of prizes in the sciences, literature and economics, as well as peace work. The announcements started last week with the awarding of the prize in Physiology or Medicine. They wrapped up on Monday, when the Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel was announced.
4. The Nobel Prizes most years are presented to recipients in Stockholm and Oslo in December. Because of the coronavirus pandemic, the committees are changing their approaches. Some of the events in Stockholm will be canceled in favor of a digital ceremony for the Nobelists, and medals and diplomas are to be distributed to the recipients? embassies and handed over in their home countries. Recipients may be invited to the award ceremony for 2021, if possible.
5. The Oslo ceremony for the peace prize will be smaller than in most years, with a limited audience.
6. The Nobel committee also announced another change last month: Each prize will rise to 10 million Swedish krona, 1 million more than in the previous year. That?s a hike in the prize value of about $112,000 in current exchange rates.
7. The 2020 Winners
8. Physiology or Medicine
9. Drs. Harvey J. Alter, Michael Houghton and Charles M. Rice on Monday received the prize for their discovery of the hepatitis C virus. The Nobel committee said the three scientists had ?made possible blood tests and new medicines that have saved millions of lives.?
10. Physics
11. Roger Penrose, Reinhard Genzel and Andrea Ghez are the recipients of the Nobel Prize in Physics for 2020.Credit...Pool photo by Fredrik Sandberg
12. Roger Penrose, Reinhard Genzel and Andrea Ghez received the prize on Tuesday for their discoveries that have improved understanding of the universe, including work on black holes.
13. Chemistry
14. The Nobel Prize in Chemistry was jointly awarded on Wednesday to Emmanuelle Charpentier and Jennifer A. Doudna for their work on the development of Crispr-Cas9, a method for genome editing.
15. Literature
16. The Nobel Prize in Literature was awarded on Thursday to Louise Glack, one of America?s most celebrated poets, ?for her unmistakable poetic voice that with austere beauty makes individual existence universal.?
17. Peace Prize
18. The Nobel Peace Prize was awarded on Friday to the World Food Program for its efforts to combat a surge in global hunger amid the coronavirus pandemic, which has swept around the world with devastating impact.
19. Economic Science
20. Paul R. Milgrom and Robert B. Wilson were awarded the Nobel in economic science on Monday for improvements to auction theory and inventions of new auction formats.

Learning for Extractive Summarization

- **Neural Extractive Text Summarization with Syntactic Compression**
  - https://arxiv.org/abs/1902.00863
- Extracting candidates and learning to score importance in single documents
- Specifically,
  - choose sentences from the document
  - identify possible compressions based on constituency parses
  - scores compressions with a neural model to produce the final summary

Detailed Review of Summarization Methods

Chan Young Park's slides: http://demo.clab.cs.cmu.edu/algo4nlp19/slides/summarization.pdf

Coding Illustration - Abstractive

- **Intuition**
  - Summary as a ‘machine translation’ task, where translation is to same language but newer words
- Sample code: https://github.com/biplav-s/course-nl/tree/master/l20-textsumm

Discussion: State of Art and Metrics

http://nlpprogress.com/english/summarization.html

Compressive Summarization

- Summary as a content deletion task preserving crucial information in the input
- Metrics
  - F1 score – for measuring content quality
  - Compression factor – for measuring compression efficiency

- Sanity check – baseline – method
  - Random work chooser/ remover !

Reference:
http://nlpprogress.com/english/summarization.html

Lecture 23: Concluding Comments

- We looked at Summary generation
- Methods
  - Extractive - traceable to original content
  - Abstractive – non traceable to original content
  - Compressive – remove content but not information
- Applications

Concluding Segment

Course Project

Discussion: Course Project

Theme: Analyze quality of official information available for elections in 2024 [in a state]

- Take information available from
  - Official site: State Election Commissions
  - Respected non-profits: League of Women Voters
- Analyze information
  - State-level: Analyze quality of questions, answers, answers-to-questions
  - Comparatively: above along all states (being done by students)
- Benchmark and report
  - Compare analysis with LLM
  - Prepare report

- Process and analyze using NLP
  - Extract entities
  - Assess quality – metrics
    - Content – *Englishness*
    - Content – *Domain* -- election
  - … other NLP tasks
  - Analyze and communicate overall

Major dates for project check

- Sep 10: written – project outline
- Oct 8: in class
- Oct 31: in class // LLM
- Dec 5: in class // Comparative

Obtaining Election Data

Here are a few things to do:
A) **Official data** backed by laws: state election commission
a) Find the state's election commission
b) Find the Q/As they provide. They may be as FAQs or on different web pages.
c) Collect the Q/A programmatically

B) **Secondary data** sources: non-profit
a) Find Q/As from Vote 411 which is supported by the non-profit: LWV.

For reference, for SC,
A) Official - https://scvotes.gov/voters/voter-faq/
B) Secondary - https://www.vote411.org/south-carolina

For extraction, one or more approaches:
- Manually annotating
- BeautifulSoup,
- Tika
- or other open source libraries.

Election Q/A for Your State

- Format in .json; name file as "**xy**_qa.json", where xy is the two-character US state acronym
- Fixed attributes in .json
  - state: xy
  - num_questions: **a**, where a is the number of questions
  - num_answers: **b**, where b is the number of answers
  - contributor: student name
- questions: List of Q/As with attributes for each it:
  - q // question
  - a // answer
  - s // source url from where the information is taken
  - t // time when the information is taken – UTC format
- Store it in your github repo; put in sub-dir like "project/data"
- Instructor will keep it in common place inside course github repo and share.

Election Q/As for Multiple States

- Instructor will keep it common place inside course github repo and share.
- You will be able to access Q/As of all states from common location
  - To compare data across all states

Discussion

- How will you use a LLMs for election data analysis ?
- When and Why? (conversely, not)

<Student Name>

CSCE 771 Fall 2024: Milestone#1

1. State Selected:
2. Election data sites:
   - Official site (e.g.,State Election Commission) url
   - Secondary site (e.g., League of Women Voters) url
1. Report how data collected and Q/A statistics

1. Take on NLP methods you will use and why for Q/A analysis
   1. State-level (right)
   2. Comparatively: above along states being done by peers

Initial analysis of questions (Q)

*

Initial analysis of answers (A)

*

Initial analysis of an answer (a_i) for a question (q_i)

*

Project Report Guidelines

- Use template of ACM Computing Surveys – Latex or Word - https://www.acm.org/publications/authors/submissions
- Consider your report as a paper. Sections to have will be similar
  - Abstract: 1-line each on what, how, result // Optional
  - Introduction: motivation for the work // Optional
  - **Problem** // Clearly state input and output
  - **Related Work** // What are closely related work?
  - **Approach** // How does your solution/ code work?
  - **Evaluation** // How is the result better than a baseline? What better could have been done ?
  - **Discussion** // About results, what more could be done, anything else interesting
  - **Conclusion** // Optional
  - **References**

About Next Lecture – Lecture 24

Lecture 24 Outline

- Conversation Agents
  - Rule based methods
  - (Deep) learning based methods
- Applications
- Ethical Issues

18	Oct 22 (Tu)	Entity extraction, linking
19	Oct 24 (Th)	Events extraction, spatio-temporal analysis
20	Oct 29 (Tu)	Topic Analysis
21	Oct 31 (Th)	PROJ REVIEW
	Nov 5 (Tu)	
22	Nov 7 (Th)	NLP Task: Sentiment
23	Nov 12 (Tu)	NLP Task: Summarization
24	Nov 14 (Th)	Conversation Agents
25	Nov 19 (Tu)	Ethical Concerns with NLP, Trusted AI and Societal Impact
26	Nov 21 (Th)	Working with LLMs for NLP Tasks - programming, Quiz